aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–22

DOI: 10.1177/1745691618792261
www.psychologicalscience.org/PPS
SAGE

Viewing Adaptive Social Choice Through the Lens of Associative Learning

Oriel FeldmanHall[1] and Joseph E. Dunsmoor[2]
[1]Department of Cognitive, Linguistic & Psychological Sciences, Brown University, and
[2]Department of Psychiatry, University of Texas at Austin

Abstract
Because humans live in a dynamic and evolving social world, modeling the factors that guide social behavior has remained a challenge for psychology. In contrast, much progress has been made on understanding some of the more basic elements of human behavior, such as associative learning and memory, which has been successfully modeled in other species. Here we argue that applying an associative learning approach to social behavior can offer valuable insights into the human moral experience. We propose that the basic principles of associative learning—conserved across a range of species—can, in many situations, help to explain seemingly complex human behaviors, including altruistic, cooperative, and selfish acts. We describe examples from the social decision-making literature using Pavlovian learning phenomena (e.g., extinction, cue competition, stimulus generalization) to detail how a history of positive or negative social outcomes influences cognitive and affective mechanisms that shape moral choice. Examining how we might understand social behaviors and their likely reliance on domain-general mechanisms can help to generate testable hypotheses to further understand how social value is learned, represented, and expressed behaviorally.

Keywords
learning, associative, motivation, goals, reward, social cognition

In our everyday social lives, we continually make decisions about how to engage with others. Many of these decisions involve instances in which self-benefit can be enhanced at the cost of another's well-being (Bartels, Bauman, Cushman, Pizarro, & McGraw, 2015; Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009; Turiel, 1983). This moral tension touches on any number of real-life decisions, from trivial predicaments such as exaggerating the truth to more consequential dilemmas such as cheating on a spouse or embezzling from a company. The field of moral decision making continues to make inroads into identifying a myriad of factors that shape these choices, illustrating that the moral decision space is dynamic (Van Bavel, FeldmanHall, & Mende-Siedlecki, 2015) and contextually bound (Akitsuki & Decety, 2009; Cikara, Farnsworth, Harris, & Fiske, 2010; Cushman, Young, & Greene, 2009; Decety, Michalska, & Kinzler, 2012; FeldmanHall, Mobbs, et al., 2012; Forbes & Grafman, 2010; Greene et al., 2009; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005; Van Bavel et al., 2015). For example, small shifts in the environment (Aquino & Reed, 2002; Tversky & Kahneman, 1974, 1981; Valdesolo & DeSteno, 2007) or even one's degree of emotional engagement can swiftly alter decisions to engage or refrain from harming another for self-gain (Cushman, Gray, Gaffey, & Mendes, 2011; Decety et al., 2012; FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; Greene et al., 2009; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Hutcherson & Gross, 2011; Mendez & Shapira, 2009; Teper, Inzlicht, & Page-Gould, 2011; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005).

Telling a lie and embezzling from a company are moral quandaries that span degrees of severity and significance. Yet both moral quandaries have a common property—increasing one's monetary, emotional, or physical benefit at the expense of harming another.

Corresponding Author:
Oriel FeldmanHall, Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI 02906
E-mail: oriel.feldmanhall@brown.edu

Facing a decision to harm another for self-gain likely draws on a multitude of qualitatively different signals about how to act: from direct experience (King-Casas et al., 2005; Murty, FeldmanHall, Hunter, Phelps, & Davachi, 2016) to vicariously learned norms that harming another is immoral (Blair, 1995; Buckholtz, 2015; Gino & Galinsky, 2012; Kouchaki, 2011; P. A. Miller, Eisenberg, Fabes, & Shell, 1996). But how do we arrive at these moral actions?

How humans and other animals learn that stimuli signal meaningful outcomes has been widely studied for the past century using models of Pavlovian conditioning (Domjan, 2005; Pavlov, 1927). These paradigms continue to play a central role in the neuroscience of learning, memory, and emotion (J. LeDoux, 2003), as well as statistical and machine learning (Courville, Daw, & Touretzky, 2005; Sutton & Barto, 1998), neuroeconomics (Ruff & Fehr, 2014), and clinical research (Blair, 2013). Indeed, much contemporary research on decision making incorporates a conditioning framework to examine how organisms learn and represent stimuli in their environment to guide value-based behaviors (Clark, Hollon, & Phillips, 2012; Dayan & Berridge, 2014).

Extending animal models of learning to understand human behavior has proven highly beneficial in conceptualizing and understanding a range of complex and mostly nonsocial real-world behaviors (Le Pelley, Mitchell, Beesley, George, & Wills, 2016). For example, this approach has fruitfully shed light onto the mechanisms that govern the processes of causal learning, evaluative conditioning, and fear conditioning—to name a few. The ability to extrapolate known learning effects to new types of situations is informative because it elucidates the environmental determinants (i.e., the regularities in the environment such as stimulus-stimulus relations) that shape discrete behavioral patterns (De Houwer, Barnes-Holmes, & Moors, 2013; De Houwer, Gawronski, & Barnes-Holmes, 2013). In other words, learning experiences can have systematic and predictable effects on shaping choices. This is perhaps best exemplified by conditioning-based models of psychopathology in which learning phenomena described by Pavlovian conditioning—including acquisition, extinction, and stimulus generalization—help to explain the etiology and maintenance of psychopathologies ranging from anxiety and stress-related disorders to addiction and psychopathy (Blair, 2013; Brewin, 2001; Foa & Kozak, 1986; J. E. LeDoux, 2000). Because Pavlovian conditioning is built on a history of rigorous behavioral and neurobiological experimentation and makes clear predictions for which factors affect learning and behavior (Wasserman & Miller, 1997), it provides tractable models for clinical research (Mineka & Zinbarg, 2006).

As with clinical research, the field of moral research is fraught with issues of how to define, study, and fractionate a conceptually nebulous domain. Much of the moral research to date has focused on judgments (e.g., "Is it appropriate to kill one to save five?") or decisions that typically occur as isolated choices (e.g., "Do I, in this one instance, harm another for monetary gain?"). These types of moral judgments and decisions can likely be explained to some degree by an individual's history of observable and unobservable (i.e., latent) reinforcement during similar previous experiences. Yet an individual's past learning history is rarely accounted for in laboratory-based moral research, and thus the degree to which learning and memory processes affect moral behavior has remained largely unspecified. The past few years, however, have seen a push toward incorporating an individual's learning history in understanding social and moral cognition, often from the perspective of behavioral economics or reinforcement learning.

Here, we consider whether the learning procedures and processes of Pavlovian conditioning offer valuable frameworks for understanding dynamic human moral behavior. In the basic Pavlovian-conditioning procedure, a neutral cue (often something simple such as a light or a tone) is paired with a naturally salient stimulus (often something aversive such as a shock or appetitive such as food). If the conditioned stimulus (CS; e.g., the light) is a reliable predictor for the unconditioned stimulus (US; e.g., the shock), the CS alone can elicit a conditioned response (CR; e.g., freezing). Note that a variety of Pavlovian-conditioning paradigms can be used to probe underlying psychological process by which the subject learns an association between the CS and US. Distinguishing between Pavlovian learning as a procedure and a process has been a matter of considerable theoretical and empirical interest in psychology for nearly a century. We propose that Pavlovian learning paradigms can be used to illustrate how a diverse set of moral behaviors are learned and expressed, including altruistic, cooperative, punitive, or trustworthy behaviors.

We also discuss different accounts for how mental processes underlying associative learning might guide behavior during complex social experiences. Such accounts can help to explain how a specific history of pairing social phenomena with positive or negative outcomes can come to influence—and ultimately bias—complex moral behaviors. By building on extensive knowledge from two mostly disconnected fields, Pavlovian learning and social decision making, we aim to operationalize a framework for future empirical research on moral learning that includes the discrete cognitive and affective mechanisms that systematically drive moral action. We further posit that learning to assign moral value to a social stimulus is largely—but in some cases not solely—governed by domain-general processes (Griffiths & Tenenbaum, 2011). That is, although

Pavlovian learning principles are likely to subserve many instances in which individuals learn about moral value, given the salient nature of moral phenomena, there may be times, we hypothesize, in which the acquisition of moral value does not follow traditional associative learning principles.

A Theoretical Foundation

Traditional learning theories describe two predominant routes by which stimuli acquire value and control behavior: instrumental (or operant) and Pavlovian (or classical) conditioning (Table 1). Whereas Pavlovian conditioning explains how a neutral stimulus comes to elicit seemingly automatic behaviors through direct or indirect pairings with a salient stimulus (Pavlov, 1927), instrumental conditioning describes how certain overt behaviors are strengthened or weakened through effective reinforcement or punishment of those behaviors. Crucially, Pavlovian and instrumental responses are not entirely separable, and these systems commonly interact to guide behavior. For instance, in Pavlovian-instrumental transfer, a CS (a tone paired with food) can enhance an animal's response toward the same reward (e.g., pressing a lever to accrue even more food). In this case, the association between the tone and food influences a goal-directed behavior that was learned independently from Pavlovian conditioning. Although researchers outside the intellectual tradition of associative learning occasionally still characterize Pavlovian conditioning as a purely reflexive, inflexible, and automatic process (see Rescorla, 1988), contemporary views of this phenomenon involve cognitively mediated information processing and value learning systems. Thus, associative learning invokes cognitive expectancies and mental representations in the formation of a conditioned behavior.

It is noteworthy that learning about social value can be exploited in different ways during a decision-making task. Theoretical work that integrates reinforcement learning (Daw & Doya, 2006; Daw & Frank, 2009; Daw & Shohamy, 2008; Dayan & Daw, 2008; Niv, 2009) with social decision making has suggested that failures in learning from reinforcement contingencies can explain (in part) deficits in prosocial behavior (Blair, 2013; Blair, Jones, Clark, & Smith, 1997; Budhani & Blair, 2005; Budhani, Richell, & Blair, 2006; Finger et al., 2011; D. G. V. Mitchell et al., 2006; Moll et al., 2005; White et al., 2013). More recently, a reinforcement framework has been explicitly applied to the domain of morality (Buckholtz, 2015; Christopoulos, Liu, & Hong, 2017; Cushman, 2013; Gesiarz & Crockett, 2015). These models of social behavior have established several important theoretical predictions for how moral learning may unfold, positing that the state or context of the decision space is critical in determining how value representations are generated (see Cushman, 2013 for a detailed account). To date, moral learning research has been largely buoyed by decision-making accounts of moral behavior that involve repeated social exchanges with the same individual (i.e., instrumental learning; see Table 1). This work has demonstrated increases in prosocial (or antisocial) responses after explicit positive (or negative) reinforcement when repeatedly engaging with a trustworthy, supportive, or generous person (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Hackel, Doll, & Amodio, 2015; Jones et al., 2011; King-Casas et al., 2005; Klucharev, Hytonen, Rijpkema, Smidts, & Fernandez, 2009; Rilling et al., 2002). Reinforcement models thus provide a simple and elegant mechanistic account for how histories of past decisions influence future social choice.

There are, however, other avenues in addition to reinforcement learning that describe how social value is learned. For example, learning can occur in the absence of observable reinforcement (e.g., latent learning), with multiple competing sources of information (e.g., cue competition), or without explicit action or goal-directed responses. It can also be acquired rapidly (e.g., one-shot learning). Therefore, the question we pose is not how do we represent the value of a moral action itself (which can be characterized by reinforcement learning accounts among other theoretical models), but rather how do we initially learn to assign (or withhold assigning) value to a social stimulus to later guide a moral action? This question is critically important because humans do not always have the opportunity to accumulate feedback about whether a specific individual is morally trustworthy, cooperative, or fair. We often come to develop positive and negative associations with people without experiencing direct feedback. Given that Pavlovian conditioning has been a valuable framework for yielding insights into many complex learning processes (De Houwer, Thomas, & Baeyens, 2001; Dunsmoor & Murphy, 2015; Gluck & Bower, 1988; Le Pelley, Oakeshott, & McLaren, 2005; Shanks, 2010) that map onto many of our everyday social and moral exchanges, how then might we incorporate such a framework to augment the study of moral learning?

A Case for Pavlovian Social Learning and Moral Choice

Pavlovian responses account for a remarkable amount of behavior across species—from elemental learning systems in a sea slug (Kandel & Schwartz, 1982) to complex judgment and decision making in humans (Clark et al., 2012). Indeed, flexible cognitive processes

Table 1. Understanding Moral Action From a Pavlovian Learning Perspective

Conditioning phenomena	Description	Moral example	Behavioral prediction
Direct experiences			
Basic classical conditioning effects			
Acquisition	An association between a neutral cue (CS) and a salient or meaningful outcome (US). CS-US association imbues CS with value and CS elicits a conditioned behavior.	Decider encounters receiver (CS) at a negative event (US), and the receiver acquires negative emotional value.	Given receiver's negative emotional value, decider more likely to exhibit antisocial behaviors toward receiver: *selfishly* keeps money and applies high-intensity shocks.
Extinction	After acquisition, subsequent encounters with CS in absence of US reduces associative value and conditioned behavior.	Decider encounters receiver at a negative event, but subsequent encounters with receiver are uneventful.	Receiver's initial negative emotional value diminishes and decider acts *prototypically*: keeps some money and applies medium-intensity shocks.
Latent inhibition	Before acquisition, numerous experiences with CS in the absence of US impedes the ability to later form CS-US association.	Decider repeatedly encounters receiver in neutral situations. Later, receiver is encountered at a negative event.	Decider does not attribute negative emotional value to the receiver and acts *prototypically*: keeps some money and applies medium-intensity shocks.
Cue competition			
Blocking	Ability to form a CS-US association is impaired if the CS is combined with another cue that is already associated with the US.	Decider encounters receiver at a negative event, but Mary, who the receiver already associates with negative events, is also at the event.	Receiver does not acquire negative emotional value. Decider treats receiver *prototypically*: keeps some money and applies medium-intensity shocks.
Unblocking	Ability to form a CS-US association is not impaired in the presence of another cue that was previously associated with a weaker or stronger US.	Decider initially encounters Mary at a negative event. Decider then encounters Mary and the receiver at a much worse event.	Receiver acquires negative emotional value and decider behaves antisocially: *selfishly* keeps money and applies high-intensity shocks.
Retrospective revaluation			
Backward blocking	The reverse of blocking: two cues are associated with the US. One cue is then presented alone with the US. The other cue retrospectively loses its associative value.	Decider encounters receiver and Mary at a negative event. Later, the decider encounters only Mary at another negative event.	Receiver loses negative emotional value. Decider treats receiver *prototypically*: keeps some money and applies medium-intensity shocks.
Release from overshadowing	Two cues are associated with a US. One cue is then presented without a US. Associative value is solely attributed to the other cue. Release from overshadowing likely occurs if the two CSs are each initially associated with the US but only weakly associated with one another.	Decider encounters receiver and Mary at a negative event. Later, the decider encounters only Mary at a fun event.	Decider retrospectively attributes negative emotional value to the receiver. Decider behaves antisocially: *selfishly* keeps money and applies high-intensity shocks.
Secondary extinction	Similar procedure as release from overshadowing, but presentations of one cue without the US acts as a form of (secondary) extinction that transfers to the other cue. Both cues lose associative value. Secondary extinction is more likely to occur if the two CSs are highly similar or strongly associated with one another.	Decider strongly associates the receiver and Mary together and encounters them both at a negative event. Later, the decider encounters only Mary at a fun event. The strong association between the receiver and Mary allows extinction to generalize from Mary to the receiver.	Receiver's initial negative emotional value diminishes and decider acts *prototypically*: keeps some money and applies medium-intensity shocks.

(*continued*)

Table 1. (Continued)

Conditioning phenomena	Description	Moral example	Behavioral prediction
Indirect experiences			
Vicarious learning	An indirect form of acquisition that involves learning a CS-US association by observing others react to the CS.	Decider observes others reacting negatively to the receiver.	Decider behaves antisocially: *selfishly* keeps money and applies high-intensity shocks.
Sensory preconditioning	Two cues are associated with one another in the absence of the US. Later, one cue is paired with the US. Associative value transfers to the other cue.	Decider routinely encounters Mary and receiver together in nonemotional situations. Later, the decider encounters only Mary at a negative event.	Receiver acquires negative emotional value. Decider treats receiver antisocially: *selfishly* keeps money and applies high-intensity shocks.
Stimulus generalization	After acquisition, other stimuli that are similar to the CS evoke a learned response as well. Magnitude of response tends to be positively related to the amount of similarity to original CS.	Decider encounters someone who looks very similar to the receiver during a negative event.	Decider treats receiver antisocially: *selfishly* keeps money and applies high-intensity shocks.

Note: CS = conditioned stimulus; US = unconditioned stimulus.

in humans, such as value-based decision making, may have evolved out of basic reflexive learning processes exemplified by Pavlovian-conditioning phenomena in simple organisms. Many Pavlovian-conditioning paradigms are cognitively mediated (e.g., involving stimulus-outcome representations or higher order associations; Rescorla, 1988) and therefore constitute part of a normative learning system responsible for flexible value-based decision making (Dayan & Berridge, 2014; Doll, Simon, & Daw, 2012; Huys et al., 2011). Viewed through this lens, a Pavlovian learning system provides a useful description for how humans represent and learn social value that later contributes to flexible moral action.

Because Pavlovian conditioning is well placed in the framework of experimental psychology, it provides several validated learning paradigms, testable hypotheses, and straightforward predictions to explore how learning experiences determine how stimuli acquire value to affect a variety of social behaviors. Moreover, taking a functional view of conditioning—that is, conditioning maps regularities in the environment and the specific pairings of stimuli onto discrete changes in behavior—allows there to be theoretical freedom from any assumptions about the kinds of mental processes that subserve the relationship between CS-US pairings and downstream behavior (De Houwer, 2018). Such an account is critical to the study of moral behavior for three reasons.

First, although the field of decision neuroscience has identified cognitive-emotional mechanisms underlying value-based learning, it is less clear whether moral behavior also relies on similar domain-general learning processes (Ruff & Fehr, 2014). Applying a Pavlovian learning account to moral behavior provides a structure for examining a handful of critical questions: What produces learning? What are the products of learning? How do humans represent their social world?

Second, a challenge in moral research is the sheer diversity of social situations researchers can choose to explore. As it stands, the moral field describes a range of behaviors from utilitarian (i.e., maximizing utility or happiness for all)-deontological (i.e., rule-based ethics) to prosocial-antisocial. Seemingly small paradigm modifications—for example, pulling a lever to cause someone's death versus pushing someone to their death (Greene et al., 2001)—can dramatically shift behavior. Accordingly, just as Pavlovian learning has been used to describe a range of human and animal behavior (e.g., phobias, drug seeking, formation of preferences and attitudes), it might also provide a set of mechanisms by which the acquisition of values systematically shapes an array of moral behaviors.

Finally, classical-conditioning paradigms can make straightforward predictions for how emotion contributes to the representation of social value and its ability to systematically bias moral choice. Together, such an account affords a relatively constrained psychological model to study how social value is acquired, which may help to elucidate the rules that determine how relationships between social stimuli are strengthened or weakened to guide adaptive moral behavior.

We begin by describing an example of a moral dilemma[1] in which helping or harming another is juxtaposed against self-gain and then propose a theoretical framework for whether various moral and immoral actions can be explained on the basis of prior Pavlovian learning experiences. This includes illustrating how a variety of experimental paradigms (drawn from the

basic principles of Pavlovian conditioning) can motivate divergent moral actions depending on how a stimulus is paired with a salient outcome. Further, we question how changes in the environment and emotional contingencies can alter how other people acquire value, both via direct and indirect experience, and how this process can influence the formation of social attitudes and moral action (Cikara, Bruneau, Van Bavel, & Saxe, 2014; Cikara & Van Bavel, 2014; Parish & Fleetwood, 1975; Van Bavel & Cunningham, 2010; Xiao & Van Bavel, 2012).

Imagine this: You are studying in a café looking out of the window, contemplating how you are going to pay back your student loans on your busboy salary. While brainstorming ways to make more money, you spot an old woman outside the café open her wallet to reveal hundreds of bills—by your estimate at least $10,000. A thought occurs to you: You could dash out, push the woman over, and snatch her wallet. Essentially, you could steal her money, substantially relieving your college debt. Nobody is around, and your chance of getting away with robbing the old woman is high. What do you do?

Let us suspend disbelief at this embellished moral dilemma for a moment and take it as a proxy for the archetypal tensions that arise in many moral situations: the tension between harming another and self-benefit. The competing pressures captured in this class of dilemma underpins many a moral quandary—robbery, infidelity, and harming others for money, power, or fame. This tension has been captured in the pain-versus-gain (PvG) laboratory task (FeldmanHall, Mobbs, et al., 2012), in which the subject—the decider—is given the opportunity to make ample money (up to $300) by applying a series of painful, but not physically damaging, electric shocks to another subject—the receiver (Fig. 1a). The decider has three options: He or she can choose to keep all the money, thereby applying many painful, high-intensity shocks to the receiver; keep some of the money, thereby applying a few medium-intensity shocks to the receiver; or give up the money entirely, thereby preventing any shocks from reaching the receiver. The altruistic, prosocial decision is to forgo the money entirely and administer no harm. And yet the enticement of making $300 is strong: Multiple experiments from our lab (FeldmanHall et al., 2015; FeldmanHall, Dalgleish, & Mobbs, 2013; FeldmanHall, Dalgleish, et al., 2012; FeldmanHall, Mobbs, et al., 2012) illustrate that approximately 80% of subjects keep more than half of the money, thereby subjugating the receiver to repeated, slightly painful electric shocks (typical behavior illustrated in Fig. 1b). This paradigm has proven useful in examining moral and altruistic behavior, and other labs have subsequently adopted variants of this task to fruitfully investigate factors that influence moral choice (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017).

The participant in the PvG task traditionally plays the role of the decider and is assumed to have no knowledge of, or prior experience with, the receiver. In such a preparation, moral choice is not biased by direct prior experience with the receiver; thus, the moral decision could be explained by emotional cues that signal critical information, such as how much pain the receiver is in (FeldmanHall et al., 2015) or what the receiver's gender is (FeldmanHall et al., 2016). It is important to note that although the original version of the task was not designed to manipulate how the decider feels per se about the receiver (e.g., whether the receiver has positive or negative emotional value), here we assume having positive feelings about the receiver will influence how readily prosocial choices are endorsed over antisocial ones. The assumption that increasing positive (or negative) emotional value can bias prosocial (or antisocial) choice is borne out of classic work that views behavior as largely governed by likes and dislikes (Allport, 1935; De Houwer et al., 2001).[2] Individuals associated with positive affect are found to be more likeable than those associated with negative affect (Byrne & Clore, 1970; Jones et al., 2011; Lott & Lott, 1974), which influences behavior, including enhancing empathy (Batson et al., 1997; Tangney, Stuewig, & Mashek, 2007), reducing intergroup bias (Gaertner et al., 1999; Pettigrew, 1997), and promoting prosocial actions (Aknin, Van de Vondervoort, & Hamlin, 2018).

Moral Learning From Direct Experience

Acquisition of learned value through direct conditioning

In its most basic form, associative learning helps determine animals' behavior in the presence of cues associated with positive or negative outcomes. In this way, a simple classical-conditioning account can explain how the receiver in a PvG task could acquire value through a past emotional experience that would later guide moral-based choices (Table 1). This framework assumes that there may be nothing inherently moral about the initial emotional experiences. Yet these nonmoral—but emotionally charged interactions—can have profound downstream effects on whether an individual later chooses to take a moral or immoral action. Imagine the decider has encountered the receiver (construed here as the CS) during an unpleasant emotional experience (the US)—for example, a social situation in which the decider was made to feel awful (e.g., overhearing

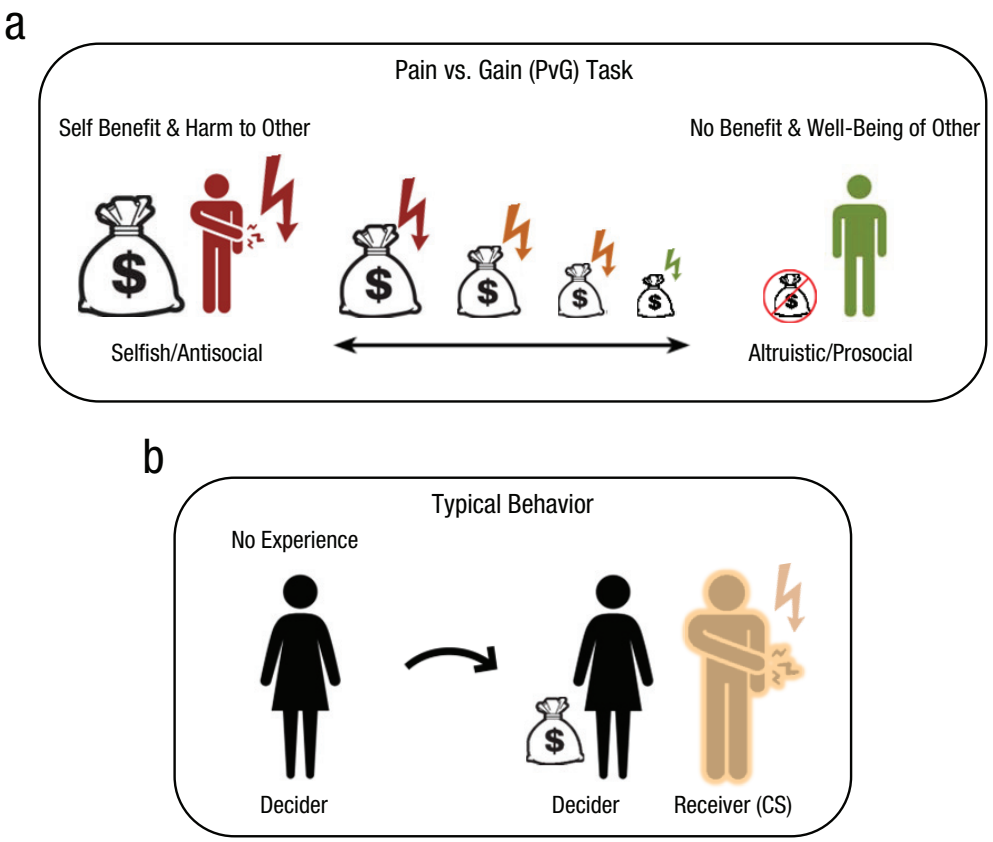

Fig. 1. Schematic of the choices available in the pain-versus-gain (PvG) task. Deciders can (a) make an immoral/antisocial choice, deciding to maximize self-benefit by keeping the money and applying painful shocks to the receiver (another participant, denoted in red), or behave altruistically, giving up all the money and preserving the physical welfare of the receiver (denoted in green). Deciders can also choose to attenuate the pain and thus keep only some of the money—behavior that falls into a moral middle ground. When (b) the decider has no previous relevant experience, typical behavior in the PvG task is to keep some of the money and apply some shocks.

extremely offensive conversations at a cocktail party). Through this CS-US pairing, encountering the receiver again in a similar context signals to the decider to expect another lousy time (Fig. 2a). Through this pairing, the receiver is assigned negative value, and this CS-US association can affect subsequent interactions with the receiver in an array of different contexts. Thus, if the decider encounters the receiver in the PvG task, a decision to administer pain for money may be biased such that the decider is now more likely to enhance her own monetary benefit at the expense of the receiver (i.e., choosing to keep the money and apply high-intensity shocks; Fig. 2b).

Note that the choice in this scenario is not a tit-for-tat expression, per se. As an analogy, although the tone does not cause an electrical shock in a fear conditioning experiment, the CS-US association establishes the tone as a signal of imminent danger and can guide decisions to terminate or escape from the CS (J. LeDoux & Daw, 2018). In a similar way, although the receiver in our scenario may not have directly contributed to the negative event at the party (she did not make the extremely offensive remarks herself), the association serves to trigger a negative emotional response in the decider that biases the moral choice to selfishly make money at the expense of the receiver's welfare. This is akin to *evaluative conditioning* (De Houwer et al., 2001), in which neutral stimuli (e.g., a novel object) acquire emotional value merely by pairings with a displeasing event (e.g., a noxious odor), leading to a dislike of that object and subsequent decisions to disengage with it in the future (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; C. J. Mitchell, Anderson, & Lovibond, 2003). Note that causal factors, such as the decider blaming the receiver for making offensive remarks at the party, are likely to influence decisions to be altruistic or selfish in ways that are more parsimoniously explained by instrumental conditioning processes,

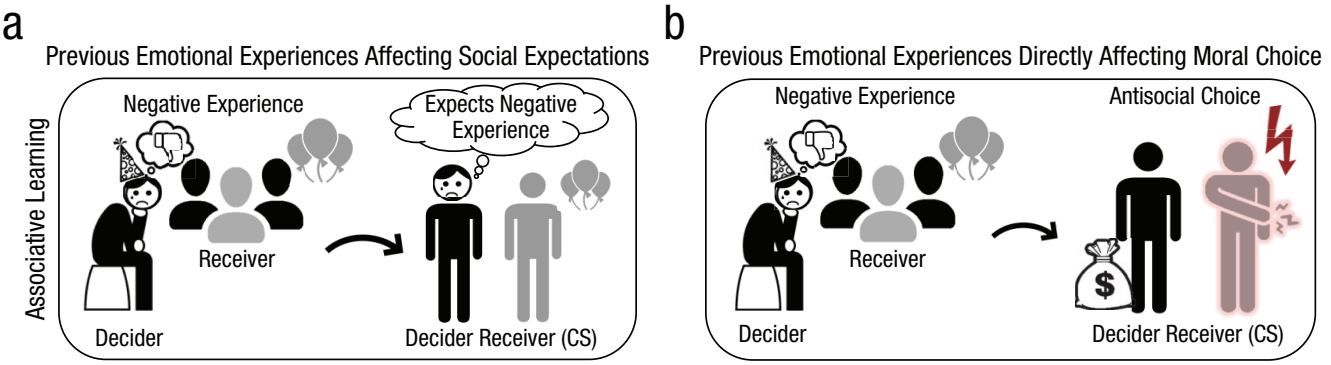

Fig. 2. Moral learning from direct experience. In the most straightforward form of Pavlovian learning, (a) the receiver (conditioned stimulus, CS) directly acquires value through either a negative or positive experience (unconditioned stimulus). Future encounters with the receiver at another party generates the expectancy for another bad time and might motivate a decision to leave. A previous negative experience with the receiver in a social setting influences the decider's moral decision in a separate context, in this case (b) a decision to antisocially apply shocks and keep money in the PvG task.

including retributive actions (Henrich et al., 2005; Mellers, Haselhuhn, Tetlock, Silva, & Isen, 2010; Peysakhovich & Rand, 2016).

Extinction and latent inhibition

A classical-conditioning framework provides a straightforward and intuitive explanation for how direct previous experience with the receiver influences the decider's moral choice in the PvG task. Simply put, the receiver evokes a negative emotional reaction in the decider that in turn contributes toward the decider's antisocial moral action. However, a CS-US pairing is not always necessary or sufficient for the acquisition of a Pavlovian association: The circumstances that produce conditioning are sensitive to the frequency in which two events are paired, that is, the base rate of the US occurrence (Rescorla, 1968), the informational relationship on which stimuli differ (Rescorla & Wagner, 1972), and the ability to relate a cue to its outcome (Garcia, 1966). In fact, research on basic associative learning processes has revealed that depending on the information the CS provides about the US (Table 1), classical conditioning can involve far more complex learning processes that can help explain a wide variety of behaviors beyond those depicted in Figure 2.

Let us imagine that there are two different deciders in the PvG task, Sarah and Abigail, each separately tasked with making a choice to antisocially accept money for shocking the receiver or altruistically foregoing the money and preserving the receiver's welfare. By changing the frequency or temporal pairing of cues and events, we can understand how these two different deciders can come to make very different moral decisions (Fig. 3). For instance, Sarah and Abigail encounter the receiver for the first time at the awful party described previously; Sarah later sees the receiver at a few more parties that are not so bad, whereas Abigail never encounters the receiver again until the PvG task. During the PvG experiment, Abigail selfishly decides to keep the money and delivers several high-intensity shocks to the receiver, but Sarah delivers only a handful of medium-intensity shocks for some money (the typical response; Fig. 1b). Abigail's selfish behavior is explained by a simple Pavlovian association from her one negative interaction with the receiver. However, Sarah's choice was influenced by subsequent neutral experiences—the "not-so-bad parties" with the receiver—that effectively reduced the negative emotional value that had previously been assigned to the receiver. This process, known as *extinction* (Table 1), illustrates how repeatedly encountering the receiver within a more neutral environment can attenuate the original negative association and cause a prosocial moral response in the PvG task.

Although experimental extinction in classical conditioning is ubiquitous, attitudes and preferences formed during evaluative conditioning appear to be fairly resistant to extinction (De Houwer et al., 2001). Such findings suggest that behaviors that rely on valence and preferences toward the CS (e.g., affective priming) are more difficult to affect through extinction than behaviors that rely on expectancies for the US (Vansteenwegen, Francken, Vervliet, De Clercq, & Eelen, 2006). If associative learning in the moral domain shares aspects of evaluative conditioning, then extinction may not be a simple matter of future experiences with a disliked individual (i.e., the receiver in our example) in the absence of a negative outcome (Gawronski et al., 2018).

For example, it is widely acknowledged that Pavlovian extinction is not a process of erasure or forgetting the original CS-US association, as evidenced by numerous

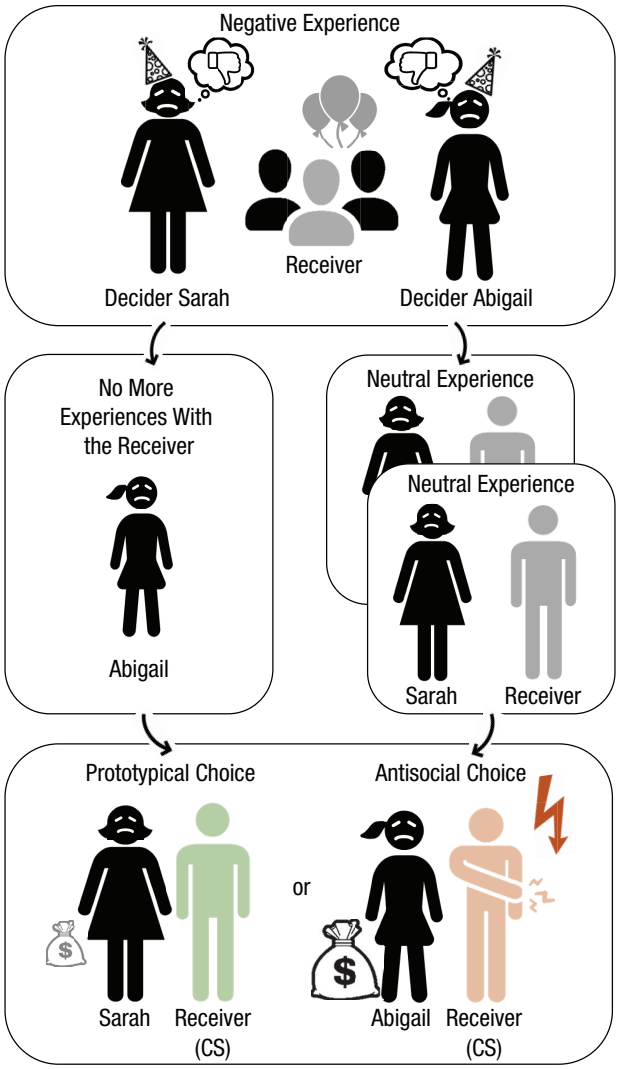

Fig. 3. Conditioning and timing. Conditioning is highly sensitive to the frequency and temporal nature by which two events are paired. Changing the frequency or timing of the association can bias subsequent moral choice, illustrated here by two different deciders who have slightly different histories with the receiver yet make very different moral choices. Whereas Abigail's one negative experience with the receiver biased her decision to behave antisocially in the PvG task, Sarah was able to extinguish the negative pairing with the receiver through repeated neutral encounters, a phenomenon known as extinction.

studies on the return of conditioned behavior (Bouton, 2002). Even if our decider, Sarah, has a number of subsequent affectively neutral experiences with the disliked receiver, her prosocial attitude may simply revert back to the original negative association after some time (referred to as spontaneous recovery), especially if she encounters the receiver in a separate context away from where the neutral experiences occurred (referred to as renewal) or if she recently had a similar negative experience around a different set of individuals (referred to as reinstatement). To prevent such postextinction return of conditioned behavior, researchers on Pavlovian conditioning have implemented several behavioral and pharmacological strategies to strengthen the extinction process (e.g., prolonged extinction under multiple contexts), including persistently altering the original CS-US association (see Dunsmoor, Niv, Daw, & Phelps, 2015). If moral attitudes and behavior recruit a similar domain-general system, then the same techniques should also yield a similar pattern of findings. One specific, testable prediction is whether persistently presenting an individual in a neutral or positive emotional framework can alter a preexisting negative moral attitude toward that individual. Indeed, we are aware of no work that has tested the durability or strength of an association between an emotional experience and the moral value assigned to an individual. However, given the salient nature that moral phenomena seem to play in daily life (Gantman & Van Bavel, 2014), it may be even more difficult to erase a preexisting CS-US association within the moral domain—an intriguing possibility that merits further empirical research.

Reversing the temporal order detailed previously, such that Sarah engages with the receiver in other less emotionally charged events *before* the dreadful party, could also result in the receiver failing to elicit an emotional response in Sarah following the awful party. This effect is referred to as *latent inhibition* (Table 1) and occurs when prior experiences with a cue in the absence of a US reduces the potential for that cue to later form an association with the US (Lubow, 1973). In these models, prior experience with a stimulus in the absence of any meaningful outcome reduces the amount of attention paid toward that stimulus, referred to as the CS's associability (Pearce & Hall, 1980). This loss in associability reduces the capacity for the stimulus to form a strong association with the US. Latent inhibition has been used to describe how early nonfearful encounters with dentists, animals, or heights provide protection from developing phobias given a later traumatic experience with these situations (Mineka & Zinbarg, 2006).

Within the moral domain, having encounters with an individual who is trustworthy, cooperative, or merely neutral—construed here as preexposure effects—prevents negative associations from developing even if the individual is later encountered in a highly negative environment (Delgado, Frank, & Phelps, 2005; Fouragnan et al., 2013). For example, Delgado et al. (2005) presented subjects with moral information describing whether another individual, a future partner for a trust game, was a praiseworthy, neutral, or suspect moral character. Once subjects were exposed to information about their partner's moral nature, they ignored the partner's actual behavior in the

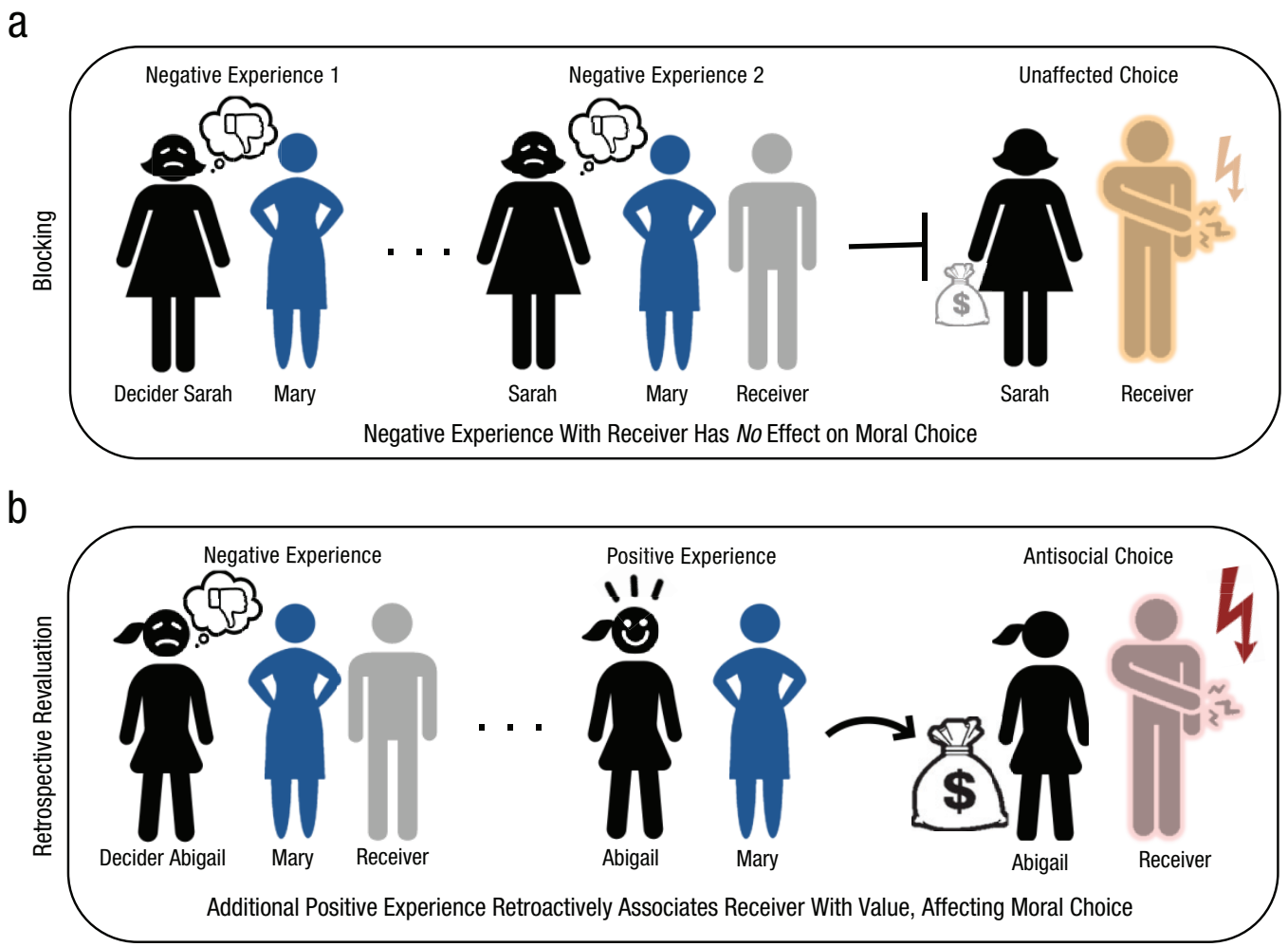

Fig. 4. Moral learning from direct experience. In a scenario in which blocking occurs, (a) Sarah has an initial negative experience with Mary. In a subsequent negative experience, Sarah encounters Mary and the receiver together. Because Mary was already associated with a negative emotional event (Negative Experience 1), the receiver is blocked from acquiring value. Thus, behavior in the pain-versus-gain (PvG) task should be unaffected, and moral choice should reflect behavior typically observed in this task (i.e., keeping some money and applying some shocks). In a scenario in which retrospective revaluation occurs, (b) Mary and the receiver are both present for the first negative experience. A subsequent experience with Mary in a positive context results in the receiver acquiring additional negative emotional value. Thus, Abigail's moral decision may be biased toward an antisocial choice in the PvG task (i.e., keeping the money and applying the shocks to the receiver).

subsequent game. Knowing that a partner was a moral beacon led to a failure in attending to and updating information of a partner behaving in untrustworthy ways. This inability to decipher which partners were trustworthy resulted in subjects investing their money in untrustworthy partners (e.g., a loss of money). In this case, preexposure to an individual's moral character buffered against subsequent learning from the individual's actual behavior when engaging with that person in real time.

Cue-Competition Effects

There are also real-world situations in which prior experience is not limited to isolated encounters with the same individual. For example, learning may occur when engaging with multiple people in large group settings that would require simultaneous evaluations of multiple individuals.

Blocking

To consider such a situation, imagine a slight twist to the party scenario. Sarah and Abigail encounter the receiver at the awful party, but this time in the company of another person, Mary. Sarah has met Mary under past similar negative social situations, but Abigail has never seen Mary before (Fig. 4a). Although Sarah and Abigail have both encountered the receiver only once at the awful party, when making a decision in the PvG task, Abigail has a negative emotional response to the

receiver and Sarah does not. Given Sarah and Abigail had the same experience with the receiver, why does only Abigail behave antisocially?

According to learning theory, a prior negative experience with Mary interfered with Sarah's ability to assign negative emotional value to the receiver. In classical conditioning, if the US (a bad time) is already predicted by a CS (Mary), then there is no error in learning when the US occurs in the presence of that CS and an additional cue (the receiver). In Sarah's experience, the receiver does not contribute any new information to having a bad time. That is, because Mary is already associated with a bad time, the negative associations are solely attributed to Mary, preventing the receiver from receiving associative strength. This phenomenon is referred to as *forward blocking* (Kamin, 1969).

The seminal discovery of blocking revolutionized learning theory in the mid-20th century by showing that learning was not the result of a mere co-occurrence of CSs and USs (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981). Rather, learning relies on the "surprise" of receiving the US—prediction errors (a mismatch between one's expectations and reality)—and the quality of the CS in predicting the US. Blocking is well established in nearly all classical-conditioning preparations with appetitive or aversive outcomes (Miller & Witnauer, 2016) and is frequently observed in human causal learning experiments that examine nonsocial decision making (Dickinson, Shanks, & Evenden, 1984) and recently even within the social domain (Seid-Fatemi & Tobler, 2015). Applying this logic to situations in which an individual is learning the social value of others—for instance, in an economic game structured to examine altruism—it is possible that if a specific individual is already known to be altruistic, this positively predicted outcome will interfere with learning new information about other individuals present at the time of the learning episode.

Recent research from our own lab illustrates that when engaging with multiple people at once, a blocking mechanism prevents "redundant" individuals from acquiring social value (FeldmanHall, Dunsmoor, Kroes, Lackovic, & Phelps, 2017). Following in the tradition of human causal judgment research (Lovibond, 2003), in our experimental structure, participants play a series of dyadic dictator games in which they can learn through interactions as a respondent whether dictators are characteristically altruistic or selfish. Once this initial learning episode has occurred (which can be considered a form of associative learning with a direct history of reinforcement), in a second round, the same dictators again deliver the same altruistic or selfish monetary splits, only this time the dictator is making this decision collectively with a partner whom the participant has not encountered before. This round is akin to a blocking procedure in which the outcome is already predicted by a stimulus that is now presented in combination with a novel stimulus. In our study, the test for associative blocking involved a trust game in which participants were asked how much of their money they would separately entrust to the dictator and to their partner. This allowed us to test whether participants learned to associate any social value to individuals who were experienced only in combination with a dictator.

Results showed that participants entrusted much of their money to altruistic dictators and little to selfish dictators who were encountered alone in the first round—as was to be expected given that participants had a direct history of reinforcement. However, there was no difference in how much money participants entrusted to the dictators' partners who were paired with either altruistic or selfish outcomes, despite the partners being present when participants received these outcomes in the second round of the dictator game. In other words, there was a failure to associate the dictators' partner with either positive or negative social value, which was revealed through the subsequent decision to treat these partners the same as a complete stranger for whom there was no history of reinforcement. Evidence of this blocking effect is consistent with the idea that there is no prediction error in learning when the same outcome occurs in the presence of both the dictator and the dictator's partner. Such a putatively domain-general blocking phenomenon illustrates that when interacting with multiple people at once, people do not associate value with individuals who seem to offer no new information, suggesting that within the social domain, a Pavlovian blocking mechanism can straightforwardly explain adaptive social decision making.

Unblocking

Blocking effects are predicated on the outcome being of equivalent magnitude when cues are presented together as they are when presented alone. If, however, the magnitude of the outcome changes—for example, the party with the receiver actually is much worse than the first party without the receiver—then there is an opportunity for the receiver to acquire associative strength. This is referred to as *unblocking* (Table 1) and is explained by classic associative learning models that propose that prediction errors stem from the difference between the expected and actual outcome (e.g., Rescorla & Wagner, 1972). Increasing the intensity of the US (an awful time) generates a prediction error that allows the blocked cue (the receiver) to acquire associative strength and evoke a CR when presented alone. Critically, unblocking by an increase in US intensity is

considered a rational inference by the subject (Dickinson et al., 1984). That is, if Cue A alone predicts a strong outcome (Mary predicts a bad time) but Cues A and B together predict an even stronger outcome (Mary and the receiver together predict a terrible time), then Cue B (the receiver) contributes to the negative predictive value.

An interesting and counterintuitive form of unblocking can occur when Mary predicts a terrible time (strong outcome) and Mary and the receiver together predict a somewhat bad time (weak outcome). According to conventional associative learning models (Rescorla & Wagner, 1972), decreasing the strength of the US should cause unblocking: The decrease in US should confer inhibitory properties onto the novel CS because the presence of the CS is now associated with a weaker outcome. In other words, the receiver would be regarded as predicting a party that was not as bad as expected (maybe their presence helped to keep the party from being as bad as it would be otherwise). However, it has also been observed in some conditioning experiments that unblocking through decreasing the strength of the US can paradoxically increase excitatory learning (Dickinson, Hall, & Mackintosh, 1976). That is, a CS that would normally be blocked from acquired associative strength will form a stronger association with the US if the strength of the US is lowered (Dickinson & Mackintosh, 1979; Holland, 1988). Thus, a downshift in US intensity (the initially terrible but now merely bad time) paradoxically increases the chance that the receiver will be associated with bad parties. Although this effect is not predicted by the Rescorla-Wagner model, it is predicted by learning theories that provide an associability mechanism to determine how much is learned about the receiver (Pearce & Hall, 1980). In these models, the surprise of the outcome on compound trials (Mary + receiver)—but not the intensity of the outcome per se—is the important factor. The mere surprise that the event is not as bad as expected allows the receiver to acquire a negative association.

This effect raises an intriguing and counterintuitive possibility for the two deciders in our moral example. Both Sarah and Abigail encounter Mary at the first awful party. Abigail goes on to have an equally awful time at a second party with both Mary and the receiver; because the first and second party were equally awful, nothing is learned about the receiver. In contrast, Sarah is surprised to find that the second party is only somewhat bad, and she is not having quite the awful time she was expecting given the presence of Mary. According to an associability mechanism of unblocking, Sarah's surprise allows her to learn about the receiver (i.e., the receiver is associated with bad—but not necessarily awful—parties). Thus, despite the fact that Abigail had a much worse time around the receiver, it is only Sarah who develops a negative emotional response toward the receiver. Remarkably, according to this framework, Abigail's awful time at the party protected the receiver from acquiring any negative emotional value.

Indeed, in our own work, we have observed evidence of a similar unblocking mechanism during socially dyadic exchanges (FeldmanHall et al., 2017). For example, in one experiment, participants play a series of robber games in which they can learn through interactions whether robbers are characteristically kind and steal little or are greedy and steal a lot. Similar to the dictator experiment described previously, some robbers first steal alone and then subsequently steal with a partner. Unlike in the gain domain (dictator games), a blocking mechanism was not systematically observed in the loss domain. Rather, when losses loomed, people were able to learn about and entrust much of their own money with seemingly "redundant kind robbers," partners who refrained from stealing large amounts of money (i.e., unblocking), but they failed to associate value—and thus entrusted money—to greedy robbers' partners (i.e., blocking).

These asymmetric effects suggest that the expectations and normative behavior of social loss generate a different set of predictions than those generated by social gain. It is possible that unlike receiving monetary windfalls through altruistic acts, individuals expect to have money stolen from them (if the option to steal is present), and a violation of this expectancy allows an additional stimulus to acquire associative value. In this context, the surprising outcome is that stealing still does not ensue even with the addition of another robber (i.e., because the potential loss is not as bad as expected, learning occurs). Accordingly, whereas a Pavlovian blocking mechanism captures social decision making in the gain domain, in the loss domain, this mechanism fails to fully account for the underlying learning processes, suggesting that moral intent can alter how prediction errors are processed, which in turn shifts whether moral value is acquired.

Retrospective revaluation

In accordance with contemporary learning theory, further variations in temporal cue pairings can give rise to a different set of attitudes or behaviors toward the receiver (R. R. Miller & Witnauer, 2016). For example, *backward blocking* is the inverse of a forward blocking design in which the subject first learns that Stimuli A and B together predict the US before later learning that A alone predicts the US. This learning effect is a form of retrospective revaluation (Table 1) and reveals an

underlying complexity to associative learning by illustrating that the representation of a stimulus is retroactively updated even when the stimulus is not presented (Dickinson & Burke, 1996; R. R. Miller & Matute, 1996; Shanks, 2010). In our moral scenario, Mary and the receiver are initially paired during a negative emotional experience. The decider later negatively experiences Mary alone and recounts that the negative experience is associated with Mary and not the receiver. Effectively, the first negative experience is retrospectively attributed solely to Mary, and the receiver loses negative emotional value. This would result in the receiver being treated more altruistically.

Release from overshadowing

Human causal learning experiments that manipulate the likelihood of a particular outcome illustrate how both excitatory and inhibitory learning effects can occur. Imagine that both of our deciders equally associate Mary and the receiver with a negative experience. Abigail meets Mary again, but this time she has a wonderful time (Fig. 4b). Sarah never meets Mary again. In the PvG task, Abigail exhibits more antisocial behavior toward the receiver than Sarah does. What explains this difference in moral behavior?

One possibility is that Abigail's subsequent positive experience with Mary alone updated the association between Mary, the receiver, and the initial bad experience in such a way that the receiver now acquires even more negative emotional value. This phenomenon is known as *release from overshadowing* (Matzel, Schachtman, & Miller, 1985). In a conditioning framework, experience with Stimulus A (Mary) alone following a compound presentation of A and B (Mary + receiver) with a US produces a prediction error that diminishes the associative value of A (i.e., extinction). This experience can generate a revaluation of the previously experienced outcome on compound A and B trials such that, in retrospect, the presence of A in fact reduced the intensity of the US. In other words, if Mary is associated with having a good time, and Mary and the receiver together predict a bad time, then the receiver alone should predict an even worse time. In this framework, the presence of Mary prevented the first party from being worse than it would have been had she not been there, and the receiver retroactively receives an increase in negative emotional value.

Explicit versus implicit associations

Conditioning phenomena that involve multiple cues, such as blocking and release from overshadowing, are fundamental to contemporary associative learning theory accounts of behavior. These signature classical-conditioning phenomena show that the amount that an animal can learn is determined by an array of parameters beyond the co-occurrence of a cue and an outcome. Although much of the research on cue competition has been in the domain of simple animal behavior, there is interest in whether these effects occur in more complex reasoning processes characteristic of human judgment and decision making (Shanks, 2010). For instance, propositional-cognitive accounts of human reasoning propose that learning phenomena are in many cases determined by a cognitive evaluation of the relationship between stimuli and outcome, and classical-conditioning phenomena are not purely automatic or nonconscious as traditionally assumed (De Houwer, 2009; C. J. Mitchell, De Houwer, & Lovibond, 2009). However, because propositional models do not explain basic learning phenomena in animals that lack the ability to express learning in propositional terms or how CRs can be acquired when conscious awareness is dramatically diminished (e.g., masked or subliminal presentations of CSs; Knight, Waters, & Bandettini, 2009), they remain controversial to the field of Pavlovian conditioning and associative learning in general (see Baeyens, Vansteenwegen, & Hermans, 2009; C. J. Mitchell et al., 2009).

Despite propositional theories' limitations for explaining certain aspects of basic conditioning, these models are likely useful in the context of explaining the acquisition of social behaviors and in particular how inferential processes contribute to the expression of moral choice. One area advanced by the propositional accounts is how inferential processes contribute to associative learning, particularly in humans who routinely extract higher order regularities from even simple learning experiences. Indeed, propositions about stimulus relations are implicated in both evaluative conditioning (Corneille & Stahl, 2018) and causal learning (Beckers, De Houwer, Pineno, & Miller, 2005).

When applied to moral learning, social psychology research illustrates examples supporting both associative and propositional accounts. For instance, within the laboratory, there are indications that social attitudes and preferences are learned implicitly within a Pavlovian-conditioning framework (Parish & Fleetwood, 1975; Sherif, 1969). Research on stereotyping (Banaji & Hardin, 1996) and intergroup biases (Cikara & Van Bavel, 2014) further reveals that these negative attitudes and preferences can bias decision making during reinforcement learning paradigms (Lindstrom et al., 2014) and experimental economic tasks (Kubota, Li, Bar-David, Banaji, & Phelps, 2013). In these cases, negative value acquired because of race or group membership does not

necessarily require declarative knowledge of the CS-US association (C. J. Mitchell et al., 2003).

Yet world history is rife with real-world examples of individuals consciously and knowingly attributing negative value to a person or group on the basis of their race or religion, which would accord with the idea that inferential processes such as propositional learning help shape moral choices (Waldmann & Holyoak, 1992). Within our framework, such a propositional account would be akin to instructing (De Houwer, 2009; Lagnado, Waldmann, Hagmayer, & Sloman, 2007) Sarah that the receiver should be treated like an animal (e.g., calling the receiver "vermin"). Likewise, if Sarah vicariously observed other individuals treating the receiver in a dehumanizing manner, she would subsequently treat the receiver poorly by keeping the money and administering high-intensity shocks. In either case, it seems that depending on the learning framework, acquiring social biases can occur implicitly and automatically through repeated CS-US pairings or explicitly through cognitive evaluations of the relationship between stimuli and outcomes.

Moral Learning From Indirect or No Experience

Of course, there are also many real-world situations in which prior experience is limited or nonexistent and there is little direct knowledge to guide choice. In the case of indirect experience, research demonstrates that humans can vicariously learn the value of stimuli from social observation (Olsson, Nearing, & Phelps, 2007; Olsson & Phelps, 2004, 2007) and instruction (Behrens, Hunt, Woolrich, & Rushworth, 2008; Mobbs et al., 2015). Although these cases illustrate an ability to learn from other individuals, other phenomena in Pavlovian conditioning make valuable predictions for how an individual behaves in the absence of any social observation or direct history of experiencing a CS in combination with a US.

Sensory preconditioning and acquired equivalence

Imagine, for example, that the decider has met the receiver and Mary together a number of times but in the absence of any meaningful emotional context. The decider begins to associate the receiver and Mary with each other. If Mary alone later acquires negative emotional value, the receiver would also gain negative emotional value through an effect referred to as *sensory preconditioning* (Table 1; Brogden, 1939; Dunsmoor, White, & LaBar, 2011; Walther, 2002). The receiver-Mary association established before conditioning allows Mary's acquired emotional value to transfer to the receiver.

In a related scenario, imagine the receiver and Mary are not associated with one another and have never been encountered together but are both associated with a common trait or attribute—for example, they are big fans of jazz music. Through this common association, the receiver and Mary are considered more alike. Any new positive information learned about Mary (e.g., Mary loves doing charity work) may therefore transfer to the receiver. This transfer effect is referred to as *acquired equivalence* (Hall, Mitchell, Graham, & Lavis, 2003; Hayes, 2001; Honey & Hall, 1989; Shohamy & Wagner, 2008; Sidman, 2009). Later, when the decider encounters the receiver in the PvG task, the positive value of Mary generalizes to the receiver, which may bias the choice in favor of giving up money and preventing shocks (Fig. 5)—who would shock a jazz enthusiast? One open question is whether the common trait

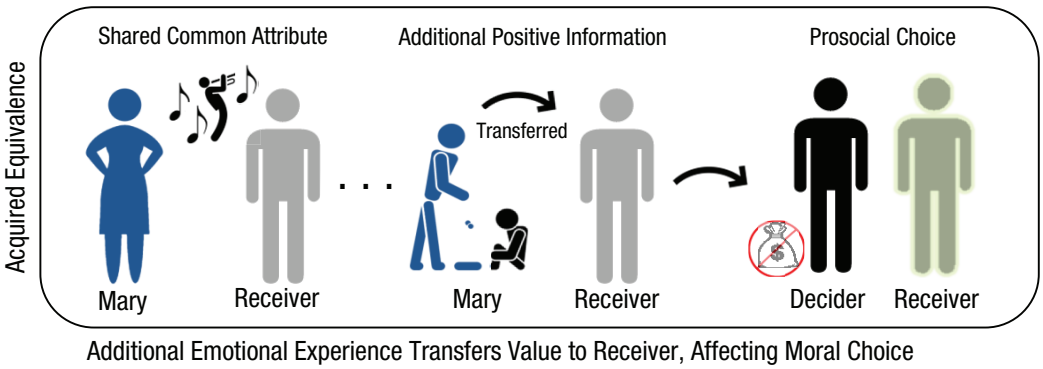

Fig. 5. Moral learning from indirect experience. In this scenario depicting how acquired equivalence occurs, two stimuli (Mary and the receiver) share a common trait of enjoying jazz. Subsequent knowledge that Mary is good (i.e., gives to charity) results in that positive value being transferred to the receiver. Considering that the receiver has been paired (albeit indirectly) with positive emotional value, the decider may give up the money and preserve the receiver's welfare.

or attribute must be qualitatively moral in nature for information to generalize. It is possible that only moral information (and not, e.g., information about hobbies) will be strong enough to establish an associative link between two individuals capable of biasing downstream behavior. Regardless, transfer effects such as sensory preconditioning and acquired equivalence demonstrate that learning is not always determined by direct reinforcement and can instead rely on preexisting associative networks established over the course of ongoing latent learning. When an emotional event occurs, learning spreads to a number of associated stimuli not present at the time of reinforcement. As such, transfer effects are likely some of the most ubiquitous conditioning phenomena in nature.

Stimulus generalization

Let us take one final example. Our decider, Abigail, has known Mary since college. She knows Mary to be a selfish and egotistical person who works in a morally questionable position at a major Wall Street firm. Our other decider, Sarah, also knows Mary but considers her a close friend who is generous and considerate and who donates a considerable portion of her paycheck to global charities that build schools in poverty-stricken countries. Later, in the laboratory in which they are participating in the PvG task, Abigail and Sarah separately notice that the receiver bears a remarkable resemblance to Mary. Learning theory predicts that, on the basis of this strong physical resemblance, the receiver will be treated as if she *were* Mary, an effect referred to as *stimulus generalization* (Table 1). Thus, Abigail and Sarah should behave antisocially and prosocially, respectively, toward the receiver.

Generalization of associative learning has been documented across species (Dunsmoor & Paz, 2015) and occurs in both the perceptual (things that look alike) and nonperceptual (things that share a conceptual similarity) domains. A long-standing question in associative learning was whether stimulus generalization was merely a failure to discriminate between different stimuli (Lashley & Wade, 1946). However, the commonly accepted view of stimulus generalization is that it is an active cognitive process in which behavior is expressed despite the capacity to detect perceptual differences from what was previously learned (Guttman & Kalish, 1956; Shepard, 1987). This would suggest that the deciders in our example do not simply confuse the receiver for another person of whom they have prior social knowledge.

Active generalization is adaptive because stimuli rarely occur in the same form from one encounter to the next and might differ considerably in perceptual form. And yet these different stimuli might portend the same consequence and should therefore be treated similarly. Thus, the ability to generalize learning across stimuli and situations is essential and helps to ensure survival in an ever-changing environment by applying prior experience to novel situations as appropriate. Within the social domain, similar information processing models have been proposed that suggest that information about close others can transfer from one individual to another (Andersen & Baum, 1994).

Note that similarity-based stimulus generalization tends to be graded such that the strongest response (i.e., intense like or dislike) is most frequently elicited by the original CS, and behavioral generalization diminishes as similarity to the CS diminishes. These gradients of behavioral generalization have been demonstrated in the domain of fear conditioning (Dunsmoor & Paz, 2015; Paz, 2014). Recent work from our lab extends stimulus generalization research from Pavlovian fear conditioning to complex dyadic social situations, revealing that decisions to trust a stranger—in the absence of direct knowledge about the stranger's reputation—relies not only on the ability to generalize from past experiences but also on the degree of similarity to these past experiences (FeldmanHall et al., 2018).

In this task, subjects played an iterative trust game with three partners who exhibited highly trustworthy, somewhat trustworthy, or highly untrustworthy behavior. After learning who can be trusted (a form of associative conditioning), subjects selected new partners for a second trust game. Unbeknownst to the subject, each potential new partner was morphed with one of the three players from the original trust game. We observed that subjects strongly preferred to play with strangers who implicitly resembled the original player that they had previously learned was trustworthy and avoided playing with strangers who resembled the untrustworthy individual. These decisions to either trust or distrust strangers formed asymmetrical generalization gradients that converged toward baseline as perceptual similarity to the original player diminished. That is, individuals were even more distrusted if they minimally resembled someone previously associated with untrustworthy and aversive outcomes, exhibiting a better-safe-than-sorry approach (Schechtman, Laufer, & Paz, 2010).

This suggests that in social situations a domain-general learning mechanism that draws on prior experience can reduce the ambiguity of a stranger's social value, ultimately facilitating potentially adaptive decisions to trust (or withhold trust from) unfamiliar others. That people seem to rely on an efficient, albeit basic, learning heuristic that facilitates adaptive engagement accords with the idea that a similarity-based generalization mechanism can be highly adaptive because it enables many stimuli—in this case, unfamiliar

individuals—to acquire value from minimal learning. Even without any direct experience of untrustworthiness, individuals implicitly deemed as untrustworthy are systematically avoided.

Caveats and Future Directions

To probe the cognitive and affective mechanisms underlying moral learning, we explored how various conditioning processes might bias altruistic behavior in the PvG task, which pits self-benefit against harm to another. This type of moral behavior (and paradigm) was chosen for conceptual simplicity and contiguity; however, the Pavlovian perspective outlined here could be applied to a host of moral behaviors, including prosocial decisions to trust, cooperate with, or punish another for wrongdoing and antisocial actions to cheat, steal, or murder. In each of these situations, an associative learning framework would delineate similar processes by which social stimuli acquire emotional value and influence subsequent moral action. In addition, gleaning information about an individual in a group setting—regardless of whether the social dynamic is about trust, cooperation, or altruism—parallels classical-conditioning experiments that probe learning when multiple cues are present and competing. Accordingly, a domain-general Pavlovian model of moral learning is especially powerful because it is likely that the same suite of learning mechanisms applies to a diverse set of social and moral situations.

We should note, however, that although Pavlovian principles can be applied to many moral actions, empirical work over the past century has revealed that some learning effects are more robust than others. For example, much of the animal literature demonstrates that secondary extinction processes predominate over release from overshadowing processes (Holland, 1999), whereas the opposite result seems to predominate for humans (Lovibond, 2003). Indeed, a crucial distinction between a reinforcement learning framework and a Pavlovian framework is the description for how learning occurs in the absence of overt reinforcement. From a Pavlovian perspective, effects such as sensory preconditioning, latent inhibition, stimulus generalization, or retrospective revaluation can guide future moral decision making in ways not strictly accounted for by traditional reinforcement learning accounts.

Furthermore, although we argue that it is likely that many social situations rely on domain general Pavlovian learning processes, it is also possible that given the salient nature of social phenomena, there may be cases in which such a framework might fail to account for social learning. Indeed, evidence that, depending on the situation, certain social expectations and immoral intentions can dictate whether a Pavlovian learning mechanism is recruited (suggests that other learning mechanisms may also play a role in the representation of social value (Courville, Daw, & Touretzky, 2006). This may be because asymmetrical expectations of socially normative behavior (Chakroff, Russell, Piazza, & Young, 2017) can influence whether a prediction error will arise. Future work aimed at discovering which learning mechanisms predominate given a set of contextual constraints and which mechanisms are highly unstable within the moral learning domain will help to characterize the core learning mechanisms that support the representation of social value.

Finally, the proposal that learning effects such as blocking and retrospective revaluation can affect moral choice does not necessitate a call to dual-process theories per se (Evans, 2008). As outlined here, depending on the demands of the situation, preferences for other people that arise out of stimulus pairings could either be implicit and reflexive or explicit and reflective. It is possible that the circumstances under which social value is learned determines the relative contribution of automatic versus deliberative strategies or a combination thereof. As learning models continue to be investigated and fine-tuned within the nonsocial domain, so too will our understanding of the underlying moral learning processes.

Conclusion

Associative learning is one of the most well examined and documented areas in psychological science and has been used to describe a host of behaviors observed across species. These models can be extended to highly complex social decision making, providing an ideal test bed for understanding moral learning. By operationalizing a "your-pain-for-my-gain" task to examine the theoretical implications of moral choice relying on Pavlovian learning principles, we describe how changes in an individual's moral behavior can be traced to the pairing of value and social stimuli. We discussed the various conditions under which previously learned emotional value can be acquired, transferred, or blocked from one individual to the next in a dyadic social situation—which in turn may have subsequent effects on moral choice. This account can help to explain the contexts that facilitate learning, the process by which emotional value is incorporated into the acquisition of social value, and the resulting antisocial or prosocial behavior. By bridging the literature on classical conditioning and morality, our hope was to help identify and characterize possible core learning mechanisms that support the representation of social value, as well as generate a number of specific, testable

predictions that provide clear avenues for future research.

Action Editor

Timothy McNamara served as action editor and June Gruber served as interim editor-in-chief for this article.

Acknowledgments

We thank Gregory L. Murphy for helpful feedback, commentary, and spirited philosophical debates.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Note that these principles could be applied to many thought experiments, including Foot's famous Trolley dilemma (Foot, 1978), or any number of economic games used to investigate social decision making (e.g., the dictator, trust, and ultimatum games, etc.). To test these principles using dyadic economic games, a researcher would have to structure a task to measure how learning about moral outcomes (altruism, trust, etc.) occurs in the presence of multiple players, which would enable, for example, the examination of cue-competition effects.
2. This relationship between positive affect and prosocial choice can be parsimoniously explained by emotion processes that broadly and flexibly bias moral behavior depending on the conceptual content of the situation (Cameron et al., 2015).

References

Akitsuki, Y., & Decety, J. (2009). Social context and perceived agency affects empathy for pain: An event-related fMRI investigation. *NeuroImage*, *47*, 722–734. doi:10.1016/j.neuroimage.2009.04.091

Aknin, L. B., Van de Vondervoort, J. W., & Hamlin, J. K. (2018). Positive feelings reward and promote prosocial behavior. *Current Opinion in Psychology*, *20*, 55–59. doi:10.1016/j.copsyc.2017.08.017

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (Vol. 2, pp. 798–844). Worcester, MA: Clark University Press.

Andersen, S. M., & Baum, A. (1994). Transference in interpersonal relations: Inferences and affect based on significant-other representations. *Journal of Personality*, *62*, 459–497. doi:10.1111/J.1467-6494.1994.Tb00306.X

Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440.

Baeyens, F., Vansteenwegen, D., & Hermans, D. (2009). Associative learning requires associations, not propositions. *Behavioral & Brain Sciences*, *32*, 198–199.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*, 136–141. doi:10.1111/J.1467-9280.1996.Tb00346.X

Bartels, D. M., Bauman, C., Cushman, F., Pizarro, D., & McGraw, P. (2015). Moral judgment and decision-making. In G. K. G. Wu (Ed.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 479–518). Chichester, England: Wiley.

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., . . . Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, *72*, 105–118.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron*, *64*, 756–770. doi:10.1016/j.neuron.2009.11.017

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 238–249. doi:10.1037/0278-7393.31.2.238

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*, 245–249.

Blair, R. J. (1995). A cognitive developmental approach to mortality: Investigating the psychopath. *Cognition*, *57*, 1–29.

Blair, R. J. (2013). The neurobiology of psychopathic traits in youths. *Nature Reviews Neuroscience*, *14*, 786–799. doi:10.1038/nrn3577

Blair, R. J., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology*, *34*, 192–198.

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*, 1558–1571. doi:10.1016/j.neuron.2013.10.024

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry*, *52*, 976–986.

Brewin, C. R. (2001). A cognitive neuroscience account of posttraumatic stress disorder and its treatment. *Behaviour Research and Therapy*, *39*, 373–393. doi:10.1016/S0005-7967(00)00087-5

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*, 323–332.

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129.

Budhani, S., & Blair, R. J. R. (2005). Response reversal and children with psychopathic tendencies: Success is a function of salience of contingency change. *Journal of Child Psychology and Psychiatry*, *46*, 972–981. doi:10.1111/j.1469-7610.2004.00398.x

Budhani, S., Richell, R. A., & Blair, R. J. R. (2006). Impaired reversal but intact acquisition: Probabilistic response reversal deficits in adult individuals with psychopathy. *Journal of Abnormal Psychology*, *115*, 552–558. doi:10.1037/0021-843x.115.3.552

Byrne, D., & Clore, G. L. (1970). A reinforcement model of evaluative processes. *Personality: An International Journal*, *1*, 103–128.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, *19*, 371–394.

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209.

Christopoulos, G. I., Liu, X. X., & Hong, Y. Y. (2017). Toward an understanding of dynamic moral decision making: Model-free and model-based learning. *Journal of Business Ethics*, *144*, 699-715.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125. doi:10.1016/j.jesp.2014.06.007

Cikara, M., Farnsworth, R. A., Harris, L. T., & Fiske, S. T. (2010). On the wrong side of the trolley track: Neural correlates of relative social valuation. *Social Cognitive and Affective Neuroscience*, *5*, 404–413. doi:10.1093/scan/nsq011

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science*, *9*, 245–274. doi:10.1177/1745691614527464

Clark, J. J., Hollon, N. G., & Phillips, P. E. M. (2012). Pavlovian valuation systems in learning and decision making. *Current Opinion in Neurobiology*, *22*, 1054–1061. doi:10.1016/j.conb.2012.06.004

Corneille, O., & Stahl, C. (2018). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*. Advance online publication. doi:10.1177/1088868318763261

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. *Advances in Neural Information Processing Systems*, *17*, 313–320.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300. doi:10.1016/j.tics.2006.05.004

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, USA*, *111*, 17320–17325. doi:10.1073/pnas.1408988111

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*, 879–885. doi:10.1038/nn.4557

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*, 273–292. doi:10.1177/1088868313495594

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2011). Simulating murder: The aversion to harmful action. *Emotion*, *12*, 2–7. doi:10.1037/a0025071

Cushman, F., Young, L., & Greene, J. (2009). Multi-system moral psychology. In G. H. J. Doris, S. Nichols, J. Prinz, W. Sinnott-Armstrong, & S. Stich (Eds.), *The moral psychology handbook* (pp. 47–71). Oxford, England: Oxford University Press.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*, 199–204. doi:10.1016/j.conb.2006.03.006

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, *113*, 259–261. doi:10.1016/j.cognition.2009.09.005

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*, 593–620. doi:10.1521/Soco.2008.26.5.593

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 473–492. doi:10.3758/s13415-014-0277-8

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

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*, 209–220. doi:10.1093/cercor/bhr111

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*, 1–20. doi:10.3758/LB.37.1.1

De Houwer, J. (2018). A functional-cognitive perspective on the relation between conditioning and placebo research. *International Review of Neurobiology*, *138*, 95–111.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review*, *20*, 631–642. doi:10.3758/s13423-013-0386-3

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*, 252–287. doi:10.1080/10463283.2014.892320

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*, 853–869. doi:10.1037/0033-2909.127.6.853

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

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *49*, 60–80.

Dickinson, A., Hall, G., & Mackintosh, N. J. (1976). Surprise and attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*, 313–322. doi:10.1037//0097-7403.2.4.313

Dickinson, A., & Mackintosh, N. J. (1979). Reinforcer specificity in the enhancement of conditioning by post-trial surprise. *Journal of Experimental Psychology: Animal*

Behavior Processes, *5*, 162–177. doi:10.1037/0097-7403.5.2.162

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The Role of selective attribution. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *36*, 29–50.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*, 1075–1081. doi:10.1016/j.conb.2012.08.003

Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annual Review of Psychology*, *56*, 179–206. doi:10.1146/annurev.psych.55.090902.141409

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: How humans generalize fear. *Trends in Cognitive Sciences*, *19*, 73–77. doi:10.1016/j.tics.2014.12.003

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron*, *88*, 47–63. doi:10.1016/j.neuron.2015.09.028

Dunsmoor, J. E., & Paz, R. (2015). Fear generalization and anxiety: Behavioral and neural mechanisms. *Biological Psychiatry*, *78*, 336–343. doi:10.1016/j.biopsych.2015.04.010

Dunsmoor, J. E., White, A. J., & LaBar, K. S. (2011). Conceptual similarity promotes generalization of higher order fear learning. *Learning & Memory*, *18*, 156–160. doi:10.1101/lm.2016411

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. doi:10.1146/annurev.psych.59.103006.093629

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage*, *105*, 347–356. doi:10.1016/J.Neuroimage.2014.10.043

FeldmanHall, O., Dalgleish, T., & Mobbs, D. (2013). Alexithymia decreases altruism in real social decisions. *Cortex*, *49*, 899–904.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive and Affective Neuroscience*, *7*, 743–751.

FeldmanHall, O., Dalgleish, T., Evans, D., Navrady, L., Tedeschi, E., & Mobbs, D. (2016). Moral chivalry: Gender and harm sensitivity predict costly altruism. *Social Psychological and Personality Science*, *7*, 542–551.

FeldmanHall, O., Dunsmoor, J. E., Kroes, M. C. W., Lackovic, S., & Phelps, E. A. (2017). Associative learning of social value in dynamic groups. *Psychological Science*, *28*, 1160–1170. doi:10.1177/0956797617706394

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, USA, 115*, E1690–E1697. doi:10.1073/pnas.1715227115

FeldmanHall, O., Mobbs, D., Evans, 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*, 434–441. doi:10.1016/j.cognition.2012.02.001

Finger, E. C., Marsh, A. A., Blair, K. S., Reid, M. E., Sims, C., Ng, P., . . . Blair, R. J. R. (2011). Disrupted reinforcement signaling in the orbitofrontal cortex and caudate in youths with conduct disorder or oppositional defiant disorder and a high level of psychopathic traits. *American Journal of Psychiatry*, *168*, 152–162. doi:10.1176/appi.ajp.2010.10010129

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin*, *99*, 20–35. doi:10.1037//0033-2909.99.1.20

Foot, P. (Ed.). (1978). The problem of abortion and the doctrine of the double effect. In *Virtues and vices and other essays in moral philosophy* (pp. 19–32). Oxford, England: Blackwell.

Forbes, C. E., & Grafman, J. (2010). The role of the human prefrontal cortex in social cognition and moral judgment. *Annual Review of Neuroscience*, *33*, 299–324. doi:10.1146/annurev-neuro-060909-153230

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*, 3602–3611. doi:10.1523/Jneurosci.3086-12.2013

Gaertner, S. L., Dovidio, J. F., Rust, M. C., Nier, J. A., Banker, B. S., Ward, C. M., . . . Houlette, M. (1999). Reducing intergroup bias: Elements of intergroup cooperation. *Journal of Personality and Social Psychology*, *76*, 388–402.

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*, 22–29. doi:10.1016/j.cognition.2014.02.007

Garcia, J. J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*, 123–124.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 57, pp. 1–52). San Diego, CA: Elsevier.

Gesiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*, Article 135. doi:10.3389/Fnbeh.2015.00135

Gino, F., & Galinsky, A. D. (2012). Vicarious dishonesty: When psychological closeness creates distance from one's moral compass. *Organizational Behavior and Human Decision Processes*, *119*, 15–26. doi:10.1016/j.obhdp.2012.03.011

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247. doi:10.1037/0096-3445.117.3.227

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364–371. doi:10.1016/j.cognition.2009.02.001

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*, 2105–2108.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*, 725–743. doi:10.1037/a0024899

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79–88.

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*, 1233–1235. doi:10.1038/nn.4080

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. *Journal of Experimental Psychology: General*, *132*, 266–276.

Hayes, S. C. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York, NY: Springer.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., . . . Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral & Brain Sciences*, *28*, 795–815.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*, 390–421. doi:10.1037/a0018916

Holland, P. C. (1988). Excitation and inhibition in unblocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*, 261–279. doi:10.1037/0097-7403.14.3.261

Holland, P. C. (1999). Overshadowing and blocking as acquisition deficits: No recovery after extinction of overshadowing or blocking cues. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *52*, 307–333. doi:10.1080/027249999393022

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*, 338–346.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*, 719–737. doi:10.1037/a0022408

Huys, Q. J. M., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLOS Computational Biology*, *7*(4), Article e1002028. doi:10.1371/journal.pcbi.1002028

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

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York, NY: Appleton-Century-Crofts.

Kandel, E. R., & Schwartz, J. H. (1982). Molecular biology of learning: Modulation of transmitter release. *Science*, *218*, 433–443.

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*, 78–83. doi:10.1126/science.1108062

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*, 140–151. doi:10.1016/j.neuron.2008.11.027

Knight, D. C., Waters, N. S., & Bandettini, P. A. (2009). Neural substrates of explicit and implicit fear memory. *NeuroImage*, *45*, 208–214. doi:10.1016/j.neuroimage.2008.11.015

Kouchaki, M. (2011). Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of Personality and Social Psychology*, *101*, 702–715. doi:10.1037/a0024552

Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological Science*, *24*, 2498–2504. doi:10.1177/0956797613496435

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, England: Oxford University Press.

Lashley, K. S., & Wade, M. (1946). The Pavlovian theory of generalization. *Psychological Review*, *53*, 72–87. doi:10.1037/h0059999

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, *23*, 727–738.

LeDoux, J., & Daw, N. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, *19*, 269–282.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*, 155–184. doi:10.1146/Annurev.Neuro.23.1.155

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*, 1111–1140. doi:10.1037/bul0000064

Le Pelley, M. E., Oakeshott, S. M., & McLaren, I. P. (2005). Blocking and unblocking in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*, 56–70.

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science*, *25*, 711–719.

Lott, A. J., & Lott, B. E. (1974). The role of reward in the formation of positive interpersonal attitudes. In T. L. Huston (Ed.), *Foundations of interpersonal attraction* (pp. 171–192). New York, NY: Academic Press.

Lovibond, P. F. (2003). Causal beliefs and conditioned responses: Retrospective revaluation induced by experience and by instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 97–106. doi:10.1037/0278-7393.29.1.97

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin*, *79*, 398–407. doi:10.1037/H0034425

Mackintosh, N. J. (1975). Theory of attention: Variations in associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298. doi:10.1037/h0076778

Matzel, L. D., Schachtman, T. R., & Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation*, *16*, 398–412. doi:10.1016/0023-9690(85)90023-2

Mellers, B. A., Haselhuhn, M. P., Tetlock, P. E., Silva, J. C., & Isen, A. M. (2010). Predicting behavior in economic games by looking through the eyes of the players. *Journal of Experimental Psychology: General*, *139*, 743–755. doi:10.1037/a0020280

Mendez, M. F., & Shapira, J. S. (2009). Altered emotional morality in frontotemporal dementia. *Cognitive Neuropsychiatry*, *14*, 165–179. doi:10.1080/13546800902924122

Miller, P. A., Eisenberg, N., Fabes, R. A., & Shell, R. (1996). Relations of moral reasoning and vicarious emotion to young children's prosocial behavior toward peers and adults. *Developmental Psychology*, *32*, 210–219. doi:10.1037//0012-1649.32.2.210

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, *125*, 370–386.

Miller, R. R., & Witnauer, J. E. (2016). Retrospective revaluation: The phenomenon and its theoretical implications. *Behavioural Processes*, *123*, 15–25. doi:10.1016/j.beproc.2015.09.001

Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders: It's not what you thought it was. *American Psychologist*, *61*, 10–26. doi:10.1037/0003-066X.61.1.10

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation*, *34*, 203–217. doi:10.1016/S0023-9690(03)00003-1

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral & Brain Sciences*, *32*, 183–198. doi:10.1017/S0140525X09000855

Mitchell, D. G. V., Fine, C., Richell, R. A., Newman, C., Lumsden, J., Blair, K. S., & Blair, R. J. R. (2006). Instrumental learning and relearning in individuals with psychopathy and in patients with lesions involving the amygdala or orbitofrontal cortex. *Neuropsychology*, *20*, 280–289. doi:10.1037/0894-4105.20.3.280

Mobbs, D., Hagan, C. C., Yu, R. J., Takahashi, H., FeldmanHall, O., Calder, A. J., & Dalgleish, T. (2015). Reflected glory and failure: The role of the medial prefrontal cortex and ventral striatum in self vs other relevance during advice-giving outcomes. *Social Cognitive and Affective Neuroscience*, *10*, 1323–1328. doi:10.1093/scan/nsv020

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*, 799–809.

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, *145*, 548–558. doi:10.1037/xge0000158

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154. doi:10.1016/j.jmp.2008.12.005

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. doi:10.1093/scan/nsm005

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science*, *15*, 822–828. doi:10.1111/J.0956-7976.2004.00762.X

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, *10*, 1095–1102.

Parish, T. S., & Fleetwood, R. S. (1975). Amount of conditioning and subsequent change in racial attitudes of children. *Perceptual and Motor Skills*, *40*, 79–86.

Pavlov, I. P. (1927). *Conditioned reflexes*. London, England: Oxford University Press.

Paz, R. (2014). Brain networks for fear extinction and generalization. *Biological Psychiatry*, *75*(9 Suppl.), 159S.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552. doi:10.1037//0033-295x.87.6.532

Pettigrew, T. F. (1997). Generalized intergroup contact effects on prejudice. *Personality and Social Psychology Bulletin*, *23*, 173–185. doi:10.1177/0146167297232006

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*, 631–647. doi:10.1287/mnsc.2015.2168

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*, 1–5.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*, 151–160. doi:10.1037/0003-066x.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*. New York, NY: Appleton-Century-Crofts.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*, 395–405. doi:10.1016/S0896-6273(02)00755-9

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*, 549–562. doi:10.1038/nrn3776

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience*, *30*, 10460–10464. doi:10.1523/JNEUROSCI.2377-10.2010

Seid-Fatemi, A., & Tobler, P. N. (2015). Efficient learning mechanisms hold in the social domain and are

implemented in the medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *10*, 735–743. doi:10.1093/scan/nsu130

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology*, *61*, 273–301. doi:10.1146/annurev.psych.093008.100519

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Sherif, M. S. C. (1969). *Social psychology* (3rd ed.). New York, NY: Harper & Row.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*, 378–389. doi:10.1016/j.neuron.2008.09.023

Sidman, M. (2009). Equivalence relations and behavior: An introductory tutorial. *The Analysis of Verbal Behavior*, *25*, 5–17.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, England: Cambridge University Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372.

Teper, R., Inzlicht, M., & Page-Gould, E. (2011). Are we more moral than we think? Exploring the role of affect in moral behavior and moral forecasting. *Psychological Science*, *22*, 553–558. doi:10.1177/0956797611402513

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*, 476–477.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*, 689–690. doi:10.1111/J.1467-9280.2007.01961.X

Van Bavel, J. J., & Cunningham, W. A. (2010). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European Review of Social Psychology*, *21*, 237–284. doi:10.1080/10463283.2010.543314

Van Bavel, J. J., FeldmanHall, O., & Mende-Siedlecki, P. (2015). The neuroscience of moral cognition: From dual processes to dynamic systems. *Current Opinion in Psychology*, *6*, 167–172.

Vansteenwegen, D., Francken, G., Vervliet, B., De Clercq, A., & Eelen, P. (2006). Resistance to extinction in evaluative conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*, 71–79.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). Hillsdale, NJ: Erlbaum.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222–236. doi:10.1037//0096-3445.121.2.222

Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. *Journal of Personality and Social Psychology*, *82*, 919–934.

Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning? *Annual Review of Psychology*, *48*, 573–607. doi:10.1146/annurev.psych.48.1.573

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*, 780–784.

White, S. F., Pope, K., Sinclair, S., Fowler, K. A., Brislin, S. J., Williams, W. C., . . . Blair, R. J. R. (2013). Disrupted expected value and prediction error signaling in youths with disruptive behavior disorders during a passive avoidance task. *American Journal of Psychiatry*, *170*, 315–323. doi:10.1176/appi.ajp.2012.12060840

Xiao, Y. J., & Van Bavel, J. J. (2012). See your friends close and your enemies closer: Social identity and identity threat shape the representation of physical distance. *Personality and Social Psychology Bulletin*, *38*, 959–972. doi:10.1177/0146167212442228