Research Article

Psychological Science
1–11

DOI: 10.1177/0956797617706394
www.psychologicalscience.org/PS

Associative Learning of Social Value in Dynamic Groups

Oriel FeldmanHall[1], Joseph E. Dunsmoor[2], Marijn C. W. Kroes[2], Sandra Lackovic[2], and Elizabeth A. Phelps[2,3,4]

[1]Department of Cognitive, Linguistic & Psychological Sciences, Brown University; [2]Department of Psychology, New York University; [3]Center for Neural Science, New York University; and [4]Nathan Kline Institute, Orangeburg, New York

Abstract

Although humans live in societies that regularly demand engaging with multiple people simultaneously, little is known about social learning in group settings. In two experiments, we combined a Pavlovian learning framework with dyadic economic games to test whether blocking mechanisms support value-based social learning in the gain (altruistic dictators) and loss (greedy robbers) domains. Subjects first learned about an altruistic dictator, who subsequently made altruistic splits collectively with a partner. Results revealed that because the presence of the dictator already predicted the outcome, subjects did not learn to associate value with the partner. This social blocking effect was not observed in the loss domain: A kind robber's partner, who could steal all the subjects' money but stole little, acquired highly positive value—which biased subjects' subsequent behavior. These findings reveal how Pavlovian mechanisms support efficient social learning, while also demonstrating that violations of social expectations can attenuate how readily these mechanisms are recruited.

Keywords

social learning, Pavlovian blocking, social value, classical conditioning, open data

Received 10/25/16; Revision accepted 4/2/17

Successfully navigating through the complex and large social world requires constant assessments of whether social interactions produce rewarding outcomes. Individuals must routinely learn whether a person can be trusted, is dependable, or should be cooperated with—oftentimes while engaging with multiple people at once. Research within the nonsocial domain illustrates that both humans and nonhuman animals are highly adept at learning from reward and punishment contingencies, and regularly exhibit value-based decision making (Glimcher, 2009; O'Doherty, 2004; Rangel, Camerer, & Montague, 2008; Schultz, Dayan, & Montague, 1997). This work has resulted in a well-characterized account of how associative-learning mechanisms underpin value-based choice (Pavlov, 1927; Rescorla & Wagner, 1972; Sutton & Barto, 1998). However, much less is known about how value-based learning occurs within the social domain and in group settings (Ruff & Fehr, 2014)—despite evidence that optimal social decision making is fundamentally dependent on the actions of other individuals (Rilling, King-Casas, & Sanfey, 2008). In the experiments reported here, we asked the critical questions of how humans learn social value about others in dynamic group environments, and whether these value representations influence subsequent social choice.

Imagine encountering an individual who repeatedly demonstrates that she is trustworthy. The knowledge gained through simply associating this individual with trustworthiness promotes continued trust over ensuing encounters (McKelvey & Palfrey, 1992). Classic associative-learning accounts (Vurbic & Bouton, 2014) can be used to explain how such direct and repetitive experiences influence behavior (Klucharev, Hytönen, Rijpkema,

Corresponding Author:
Oriel FeldmanHall, Department of Cognitive, Linguistic & Psychological Sciences, Brown University, 190 Thayer St., Providence, RI 02912
E-mail: oriel.feldmanhall@brown.edu

Smidts, & Fernández, 2009), including choosing to trust someone who has proven to be highly trustworthy (King-Casas et al., 2005). Continuous reinforcement helps explain how social behaviors can be learned over time.

However, in today's large and ever-changing social world, the conditions under which one gleans information are rarely limited to isolated, repeated interactions with the same individual. Rather, one often learns about individuals in the company of others, which requires simultaneous social evaluations. One is also likely to initially meet an individual alone and later see that individual among friends, which requires that evaluations be updated if outcomes change depending on the context. For example, if you initially encounter a kind individual and later reencounter the same individual alongside a stranger who also exhibits kindness, have you learned that the stranger is kind? By leveraging an associative-learning framework to examine these questions, we probed whether people bind social value to other individuals in a group setting, update these values when conditions change, and apply learned value associations across social domains to make adaptive choices.

Following in the tradition of human causal-judgment research (Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003), we asked subjects to play a series of dyadic dictator games in which they could learn whether dictators were characteristically altruistic or selfish—which can be considered a form of associative conditioning. Subjects engaged with many dictators, some of whom first made offers alone and later made offers alongside a new dictator. In an ensuing trust game, subjects decided how much of their own money to entrust to each dictator, as well as novel, never-before-seen strangers.

By giving subjects the opportunity to entrust their own money to each dictator, we were able to test whether (a) subjects learned to associate social value to specific dictators, even one who never made offers by himself (i.e., who made offers only as another dictator's partner), and (b) whether value acquired in one social domain—altruism—influenced subsequent behavior in another domain—trust. This framework allowed us to test multiple competing hypotheses for how social learning occurs in dynamic groups. If a dictator's behavior when making decisions alone is consistent with his later behavior when making decisions with a partner (e.g., the dictator always offers altruistic splits to the subject), it is possible that the dictator's partner will acquire the same value as the dictator, because both are yoked to the same positive outcome. This social category-learning account (Kahneman & Miller, 1986; Kashima, Woolcock, & Kashima, 2000) posits that the dictator's partner will obtain value because he is associated with a dictator previously known to be altruistic and because the two dictators continue to be altruistic together. In this case, the value learned about the first dictator is transferred to his partner because the two activate similar exemplar representations (Smith & Zarate, 1992).

An alternative hypothesis, drawn from cognitive psychology, proposes that because the monetary split continues to be altruistic, the dictator's partner provides no new information, and thus subjects will not learn the social value associated with the partner. This phenomenon is known as blocking (Kamin, 1969; Rescorla & Wagner, 1972)—a basic, albeit fundamental, Pavlovian learning mechanism that is robustly observed across species within nonsocial environments. According to this associative-learning account, the dictator will block his partner from acquiring social value, and subjects should consequently trust the dictator's partner as they would trust a stranger, despite having direct experience of the partner's altruistic behavior.

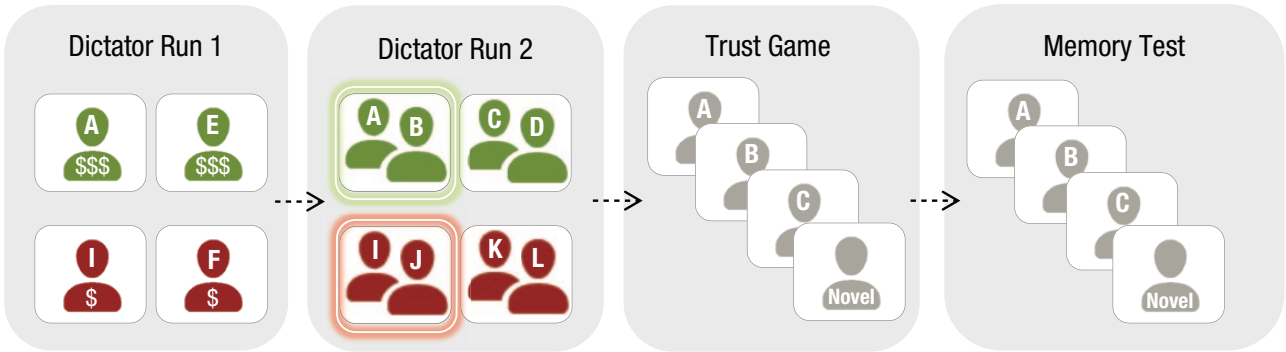
Although blocking has been used to explain efficient error-driven learning (Gluck & Bower, 1988), it is not always the case that people ignore additional cues when learning about novel stimuli (Dickinson, Shanks, & Evenden, 1984). For example, blocking is not observed in some category-learning tasks, in which people learn more about a stimulus than is necessary to perform a classification (Bott, Hoffman, & Murphy, 2007). In fact, prior experiences with a stimulus or its outcome can modify or even reverse the effectiveness of blocking (Dickinson, Hall, & Mackintosh, 1976; Le Pelley, 2004). This modulation of blocking may also extend to social learning, such that specific learning dynamics critically determine how much—or little—one learns in the social domain.

Experiment 1

Method

Task procedures. To test these competing accounts of social learning, we employed a series of social economic games to examine learning within the gain domain. Inspired by classical-conditioning paradigms involving compound pair cues and their associated outcomes (Lovibond et al., 2003), we asked our subjects in Experiment 1 to complete four tasks in the following order (Fig. 1): (a) Dictator Run 1, in which all subjects learned about an altruistic dictator (stimulus A) and a selfish dictator (stimulus I); (b) Dictator Run 2, in which subjects were again exposed to these dictators, but this time in compound with partners (stimulus A was paired with test stimulus B, and stimulus I was paired with test stimulus J); (c) a trust game, which tested whether the social value of the initial dictators' partners was learned; and, finally,

a



b

Dictator Run 1			Dictator Run 2		Trust Game	Memory Test
Player (each presented 5x)	Mean Offer	Change	Player (each presented 5x)	Mean Offer	Player (each presented 4x)	Player
Altruistic Dictator (Stimulus A)	$4.04	→ Add Novel B	Altruistic Dictator Pair (Stimulus A + B)	$4.04	A	A
Altruistic Dictator (Stimulus E)	$4.04	→ Same	Altruistic Dictator (Stimulus E)	$4.04	B	B
Selfish Dictator Pair (Stimulus G + H)	$0.18	→ Same	Selfish Dictator Pair (Stimulus G + H)	$0.18	C	C
Selfish Dictator (Stimulus F)	$0.18	→ Same	Selfish Dictator (Stimulus F)	$0.18	D	D
Selfish Dictator (Stimulus I)	$0.18	→ Add Novel J	Selfish Dictator Pair (Stimulus I + J)	$0.18	E	E
			Altruistic Dictator Pair (Stimulus C + D)	$4.04	F	F
			Selfish Dictator Pair (Stimulus K + L)	$0.18	G	G
					H	H
					I	I
					J	J
					K	K
					L	L
					M (novel)	M
					N (novel)	N
						Foils

Fig. 1. Task structure of Experiment 1. As illustrated with examples in (a), all subjects played two runs of dictator games before reencountering the same dictators in a trust game. In Dictator Run 1, subjects received various altruistic splits from good dictators (i.e., stimuli A and E) and selfish splits from bad dictators (e.g., stimuli I and F). After each split, subjects rated how they felt. In Dictator Run 2, subjects again received altruistic and selfish splits from dictators, only this time some of the dictators who had made decisions by themselves in Run 1 made decisions together with partners. For example, dictator I, who made selfish splits by himself in Run 1, made selfish splits collectively with dictator J in Run 2. Subjects then played a trust game, in which they decided how much of $10 to entrust in a series of individuals: the dictators from Runs 1 and 2, as well as novel, never-before-seen people who had no history with the subjects. Finally, subjects completed a surprise memory test, in which they were asked to report how much money each dictator had offered them in the dictator games. Panel (b) lists all stimuli presented in each of the four phases of the experiment.

(d) a surprise memory test in which episodic memory for all the dictators, including the partners, was explicitly probed.

Given that the players subjects engaged with were actually computer algorithms yoked to predetermined reinforcement rates (e.g., highly altruistic or selfish), we used a deception manipulation intended to create a realistic dyadic social environment. Subjects were photographed in front of a white wall and told that their picture, along with their responses to "how much of $10 would you split with a future player?" would be fed forward to the next subject. A similar question was employed for the trust game. This procedure was explained as the most efficient way to use subjects' responses without needing multiple people to come in for an experimental session at once. Subjects were told that in the event that their decision as the dictator or second player in the trust game was used in subsequent experimental sessions, they would be mailed a check based on the outcome of that specific decision. Extensive posttask debriefing revealed that subjects believed the social manipulation (see the Supplemental Material available online for further details).

Dictator game. In both runs of the dictator game, subjects played the role of receiver, receiving portions of a $10 endowment from another player or pair of other players (see the Supplemental Material for additional details). In the first run (initial conditioning), subjects learned through repeated interactions about a dictator (stimulus A) who consistently made altruistic splits (approximately $4; subjects were explicitly told that splits above $5 would not occur) and a dictator (stimulus I) who consistently made selfish splits (splits of approximately $0.18). In this run, these dictators made splits alone. Once a split was made, subjects rated how they felt on a 5-point visual analogue scale.

In the second run of the dictator game, subjects reencountered the same dictators (a within-subjects design), only this time each of these dictators was paired with a never-before-seen partner to form a compound cue. For instance, the altruistic dictator (stimulus A)—who made altruistic splits by himself in Run 1—also made altruistic splits when deciding with a partner (stimulus B) how to split the money in Run 2 (subjects were told that the two members of a dictator pair jointly and equally contributed to deciding how to split the money). Accordingly, the altruistic dictator's partner was paired with a positive outcome, and thus had the potential to acquire positive social value.

Critically, subjects played with various other dictators and dictator pairs in both runs. These other players either served as distractors, so that learning was nontrivial (i.e., not too easy), or provided comparisons for evaluating the magnitude of how well social value was learned (Lovibond et al., 2003). For instance, two distractor dictators (stimuli E and F) always made splits alone and were present in both Runs 1 and 2, and two dictator pairs (stimuli C + D and K + L), who were presented only in Run 2, served as a test for whether learning could occur from just the second run (results for this test are presented in the Supplemental Material; see Fig. 1b for a full description of all dictators in both runs). Effectively, this task structure enabled us to temporally manipulate information regarding an individual's social value in order to examine how value is learned in complex and dynamic group environments.

Trust game. In the subsequent test phase, subjects played a trust game. On each trial, they were endowed with $10 and could choose how much to entrust in a second player knowing that whatever amount was transferred would be multiplied by 4 and that the second player would decide whether to transfer back half of the increased sum or keep all the money himself (this feedback was not revealed until after the game, when one trial is randomly selected to be paid out). Subjects were given multiple opportunities to entrust their money with each dictator (dictators were always presented separately during the trust game). In addition, to get a baseline measurement of willingness to trust, we presented strangers (stimuli M and N) with no prior positive or negative associations from Runs 1 and 2 as second players in the trust game. In this way, we were able to examine whether previous learning from Runs 1 and 2 biased how subjects treated each individual in the trust game (i.e., whether dictators who exhibited altruistic behavior were trusted with greater sums of money than selfish dictators, or even strangers who had no previous associations).

Memory test. Finally, to test whether subjects explicitly remembered the social value associated with each dictator (including all partners), we assessed episodic memory for the splits made by each dictator in a surprise memory test. Subjects reported the amounts shared with them on a visual analogue scale from $1.00 to $5.00, in $0.01 increments. Subjects' recognition memory for all individuals presented in the previous tasks was also assessed (for additional details about the memory test, see the Supplemental Material).

Subjects. Our sample size was based on extant research using the same dictator game (Murty, FeldmanHall, Hunter, Phelps, & Davachi, 2016), as well as on classic research on blocking in humans (Beckers, De Houwer, Pineno, & Miller, 2005). We recruited 45 subjects from New York University and the surrounding New York City community (27 females; mean age = 21.5 years, *SD* = 2.9).

Subjects were paid an initial $15 and received additional compensation based on the result of one randomly selected trial from the dictator game and one randomly selected trial from the trust game (up to $25 in additional compensation). Informed consent was obtained from each subject in a manner approved by New York University's Committee on Activities Involving Human Subjects.

Results

Behavioral results. Results from Experiment 1 favored the associative-learning account: The altruistic dictator's partner failed to acquire positive social value in Run 2, presumably because the altruistic dictator previously predicted the same positive outcome when presented alone in Run 1. A repeated measures analysis of variance (ANOVA) of money entrusted in the altruistic dictator (stimulus A), the altruistic dictator's partner (stimulus B), the altruistic dictator pair from Run 2 (stimulus C + D; amounts averaged across these two players), and the novel individuals (stimuli M and N; amounts averaged across these two players) was significant, $F(3, 132) = 4.28$, $p = .006$, $\eta_p^2 = .113$. Subjects trusted the altruistic dictator with the most money ($M = \$5.06$, $SD = 2.7$), and with significantly more money than they entrusted to his partner ($M = \$4.06$, $SD = 2.7$), paired-samples $t(44) = 2.40$, $p = .023$ (Fig. 2). The altruistic dictator's partner was entrusted with the same amount of money as a stranger ($M = \$3.95$, $SD = 2.6$), paired-samples $t(44) = -0.44$, $p > .250$. Control tests probing whether the failure to associate the altruistic dictator's partner with positive social value was attributable to a general failure in learning revealed no such effect (see Manipulation Learning Check in the Supplemental Material).

A similar pattern was observed for the selfish dictators, who kept most of the money. A repeated measures ANOVA of the money entrusted in the selfish dictator (stimulus I), the selfish dictator's partner (stimulus J), the selfish dictator pair from Run 2 (stimulus K + L; amounts averaged across these two players), and the novel individuals (stimuli M and N) was significant, $F(3, 132) = 3.10$, $p = .029$, $\eta_p^2 = .07$. The selfish dictator, who by himself made selfish splits in Run 1 and then collectively made selfish splits with his partner in Run 2, was entrusted with the least amount of money ($M = \$3.42$, $SD = 2.8$), and with significantly less than was entrusted to his partner ($M = \$3.93$, $SD = 2.6$), paired-samples $t(44) = -2.22$, $p = .031$ (Fig. 2). Rather, the selfish dictator's partner was entrusted with effectively the same amount as a stranger ($M = \$3.95$, $SD = 2.6$), paired-samples $t(44) = -0.13$, $p > .250$.

At first blush it may appear surprising that subjects did not learn the social value of either the altruistic dictator's partner or the selfish dictator's partner, despite

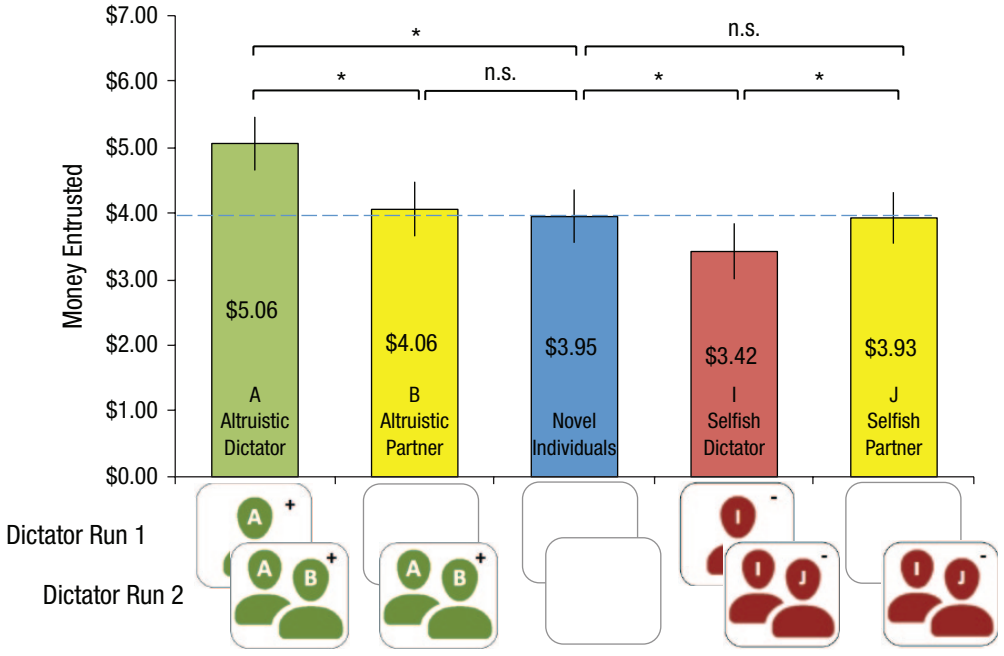


Fig. 2. Money entrusted to different types of dictators and novel players in the trust game in Experiment 1. The dashed blue line highlights the amount of money that was entrusted to novel players. The diagrams below the data bars indicate the runs in which the stimuli were presented during the dictator games; "+" denotes altruistic dictators, and "–" denotes selfish dictators. Error bars represent ±1 *SEM*, and asterisks indicate significant differences in pairwise comparisons ($p < .05$).

having directly experienced positive and negative outcomes in their presence. However, according to classic Pavlovian learning theory (Rescorla & Wagner, 1972), when a stimulus such as an altruistic dictator is already associated with positive outcomes, later encountering both the altruistic dictator and his partner together results in the partner acquiring no social value—because the positive outcome is already fully predicted by the presence of the initial altruistic dictator. In our experiment, this phenomenon resulted in the previously learned association with a dictator interfering with (i.e., blocking) subjects' ability to form an association with the dictator's partner (Kamin, 1969). Such a finding illustrates that social value learning relies on a difference between the expected outcome and the actual outcome (i.e., prediction error). That is, if a single stimulus already predicts a positive or negative outcome, an added stimulus could be considered redundant, and thus fail to acquire associative value.

Memory results. To understand whether this behavioral blocking effect was due to a failure in explicitly associating the altruistic dictator's partner with altruism and the selfish dictator's partner with selfishness, we examined how accurately subjects remembered each of the dictator's offers from Runs 1 and 2. Accuracy was computed by taking the absolute difference between the remembered split (from the surprise memory test) and the average of the actual splits. A repeated measures ANOVA revealed that subjects more accurately remembered splits from the altruistic dictator and the altruistic dictator pair than those from the altruistic dictator's partner, $F(2, 50) = 7.12$, $p = .002$, $\eta_p^2 = .22$; post hoc tests comparing accuracy for other players with accuracy for the altruistic dictator's partner were all significant, ps $< .01$. Note that this was true despite the fact that *all* altruistic dictators gave the same average split. A similar pattern was observed for the selfish dictators; subjects more accurately remembered the splits from the selfish dictator and the selfish dictator pair than those from the selfish dictator's partner, though the test failed to reach significance, $F(2, 56) = 1.98$, $p = .148$. Thus, it seems that the behavioral blocking effect observed for the dictators' partners during the trust game may be attributed to blocked episodic memory of the dictators' partners' offers.

Evidence of a Pavlovian blocking mechanism—from both a behavioral and a memory perspective—makes a case for efficient learning in social contexts (Seid-Fatemi & Tobler, 2015): People who appear to add no critical information fail to become associated with social value. Indeed, it is well established that blocking occurs when no new information about reward probability is elicited from a learning episode. This indicates that the psychological process of surprise is a critical feature of learning. Although typical behavior in a dictator game is to give on average 30% of the monetary pie, prior research has revealed that altruistic behavior is quite variable, and there are many demonstrations of dictators behaving selfishly and keeping large portions of the pie (Engel, 2011). This variable behavior—which spans altruistic benevolence to selfish enhancement—is observed across tasks and cultures, contradicting traditional economic models that tout the dominance of rational behavior (e.g., to share nothing and keep all the money). Given that a dictator can keep all the money without negative consequences, the psychological explanation for offering a split of any size in the dictator game hinges on the notion that societies have norms that govern wealth sharing: People routinely make some effort to trade off their material benefit to comply with social norms of sharing with others (Bolton, Katok, & Zwick, 1998; Dreber, Ellingsen, Johannesson, & Rand, 2013). This suggests that both selfish (small) and altruistic (large) splits in the dictator game operate within a framework of socially expected behavior. Accordingly, as long as a dictator's behavior is normatively expected, blocking effects for a partner associated with that dictator are facilitated, because the environment leaves little room for surprise.

Experiment 2

A traditional Pavlovian account, however, is agnostic to changes in stimulus valence, and consequently may fail to capture all aspects of social learning. For example, considering the various ways in which framing has profound effects on choice (Kahneman & Tversky, 1984; Tversky & Kahneman, 1981), there may be certain contexts (Kahneman & Tversky, 1979) that give rise to surprise, which should in turn give rise to unblocking effects (Dickinson et al., 1976). Within the social domain, if learning is thought to be sensitive to prior beliefs and inferences about the world, behavior that deviates from socially normative expectations may enable learning. For example, one of the oldest proverbs in moral philosophy states that if given the opportunity to steal without consequence, a person will invariably steal (Plato, trans. 1950), and there is growing evidence that this is true across psychological domains (Greenberg, 1993; Greene & Paxton, 2009; Mazar, Amir, & Ariely, 2008). Violations of this social expectation (e.g., robbers who steal little) would be surprising, and unblocking should occur for individuals who behave in such unexpected ways. In Experiment 2, we tested the possibility that blocking is sensitive to the framing of social contexts, theorizing that prior expectancies—beliefs and inferences about future social events—will influence learning such that seemingly redundant individuals will acquire social value.

Method

Accordingly, Experiment 2 mirrored the structure of the first experiment with two key differences. First, prior to playing any games, subjects completed a short math task in which they could earn up to $5. Second, instead of playing a series of dictator games, subjects played a series of robbery games. These games were structurally identical to the dictator games except that robbers could steal up to $5 from subjects (rather than give up to $5 to subjects). Robbers were the same as the dictators in Experiment 1, except that in the robbery game, kind robbers "stole" very little (approximately $0.18), whereas greedy robbers stole most of the subjects' money (approximately $4). The kind robber (stimulus A), who was initially presented alone in Run 1, continued to steal tiny amounts when presented with a partner (stimulus B) in Run 2. In contrast, the greedy robber (stimulus I) single-handedly stole most of subjects' money in Run 1, and continued to steal most of subjects' money when paired with a partner (stimulus J) in Run 2. Subjects played a trust game with each of the robbers, as well as with strangers, before completing a surprise memory test. Thus, Experiment 2—which probed whether blocking of social value also occurs in the loss domain, where there are different social expectations—was structurally and monetarily matched to Experiment 1.

For this experiment, we again recruited 45 subjects from New York University and the surrounding New York City community (22 females; mean age = 22.1 years, *SD* = 3.1). As in Experiment 1, informed consent was obtained from each subject in a manner approved by New York University's Committee on Activities Involving Human Subjects.

Results

Behavioral results. With the possibility of being robbed of their earned money, subjects no longer failed to associate social value to the kind robber's partner. A repeated measures ANOVA of the money entrusted to the kind robber (stimulus A), the kind robber's partner (stimulus B), the kind pair (stimulus C + D; amount averaged across players), and the novel individuals (stimuli M and N; amount averaged across players) revealed a significant effect of robber type, $F(3, 132) = 9.62$, $p < .001$, $\eta_p^2 = .26$. The kind robber's partner was trusted with significantly more money (M = \$5.11, SD = 2.8) than a stranger (M = \$4.15, SD = 2.6), paired-samples $t(44) = 3.02$, $p = .004$, albeit still less than what was entrusted to the kind robber (M = \$5.94, SD = 2.8; Fig. 3), paired-samples $t(44) = 2.87$, $p = .006$. Interestingly, this unblocking effect was observed only for kind robbers, who stole very small amounts, and not for greedy robbers, who stole most of the subjects' money. A repeated measures ANOVA of the

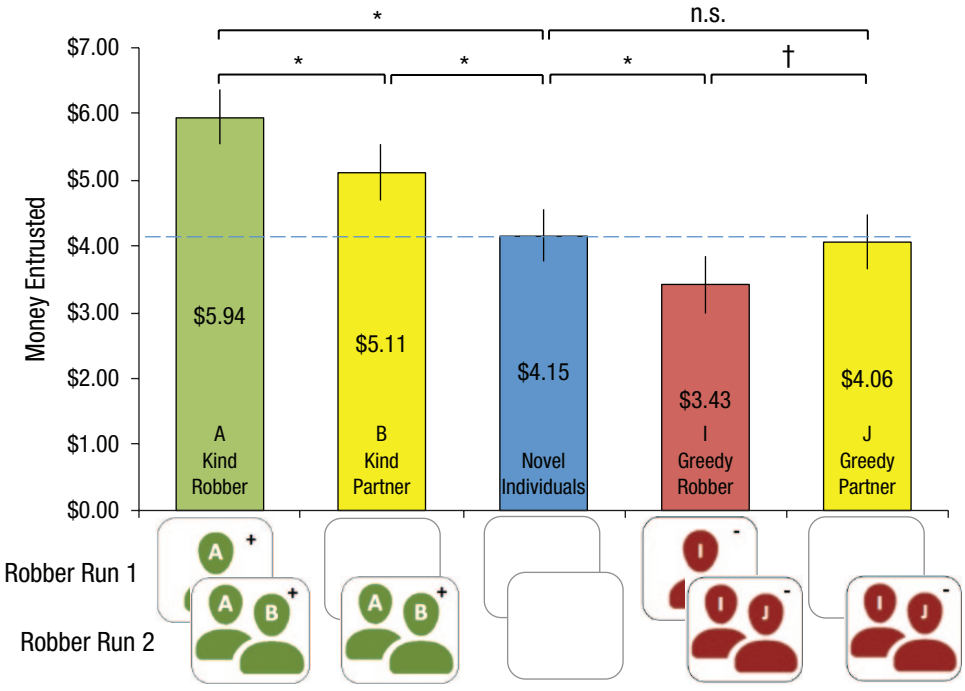


Fig. 3. Money entrusted to different types of robbers and novel players in the trust game in Experiment 2. The dashed blue line highlights the amount of money that was entrusted to novel players. The diagrams below the data bars indicate the runs in which the stimuli were presented during the dictator game; "+" denotes kind robbers, and "–" denotes greedy robbers. Error bars represent ±1 *SEM*; asterisks indicate significant differences in pairwise comparisons (at least at $p < .05$), and the dagger indicates a marginally significant difference ($p < .067$).

money entrusted to the greedy robber (stimulus I), the greedy robber's partner (stimulus J), the greedy pair (stimulus K + L; amount averaged across players), and the novel individuals (stimuli M and N) also revealed a significant effect of robber type, $F(3, 132) = 2.70$, $p = .050$, $\eta_p^2 = .07$. However, the greedy robber's partner was entrusted with the same amount of money ($M = \$4.06$, $SD = 2.9$) as a stranger ($M = \$4.15$, $SD = 2.6$), paired-samples $t(44) = -0.38$, $p = .702$, and, as indicated by a marginally significant t test, with more than was entrusted to the greedy robber ($M = \$3.43$, $SD = 2.9$; Fig. 3), paired-samples $t(44) = 1.88$, $p = .067$. Thus, the findings were similar to those of Experiment 1, as the greedy robber blocked his partner from acquiring social value.

Memory results. When we probed subjects' episodic memory for how much money each robber stole, we found asymmetrical blocking effects that mirrored those observed in the behavioral data. Subjects were equally accurate at remembering how much money the kind robber, his partner, and the kind pair stole, $F(2, 68) = 1.39$, $p > .250$, but were less accurate in remembering how much the greedy robber's partner stole, compared with how much other similarly greedy robbers stole, $F(2, 76) = 3.44$, $p = .037$, $\eta_p^2 = .083$.

Social expectations across the loss and gain domains. To understand why subjects were able to associate the kind robber's partner with social value but failed to associate the greedy robber's partner (or any of the dictators' partners in Experiment 1) with value, we examined subjects' reported subjective feelings after interacting with each dictator and robber. We theorized that if there were differences between how subjects felt about being robbed and how they felt about being given money, these differences might elucidate specific—and possibly divergent—expectations linked to social contexts involving gains and losses. Critically, in the two experiments, subjects retained the same monetary payout (e.g., because subjects were first endowed with $5 in the robbery game, there was approximately a $4 payout after a small amount of money was stolen, and this mirrored the altruistic split offered during the dictator game).

To assess whether social gain and loss were differentially experienced, we took subjects' ratings on the 5-point analogue scale (5 = *very happy*, 1 = *very unhappy*) and subtracted 3 (the midpoint, a neutral rating), such that positive feelings were indicated by a positive difference score and negative feelings by a negative difference score. We then compared the difference scores across experiments, specifically examining how subjects reported feeling about the altruistic dictators versus the kind robbers (stimuli A and E) and how they reported feeling about the selfish dictators versus the greedy robbers (stimuli I and F; all matched in their associated value). If the amount of money paid out was the central feature of the task, subjects' subjective feelings should have been similar when they received large amounts of money and when they had fairly little money stolen, as they ended up with the same amount of money regardless of whether they interacted with altruistic dictators or kind robbers (the same reasoning applies to the comparison of subjective feelings after interacting with selfish dictators and greedy robbers).

Results revealed that the kind robbers, who stole only very small amounts of money, engendered significantly higher positive ratings compared with the altruistic dictators, who gave subjects a lot of money, $F(1, 89) = 4.20$, $p = .043$ (Fig. 4). In contrast, subjects felt similarly negative about the dictators who selfishly did not share the money and the robbers who greedily stole most of subjects' money, $F(1, 89) = 1.88$, $p = .174$. Thus, despite the fact that monetary outcomes in Experiments 1 and 2 were matched, the context in which the money was earned and lost (robbery game), or simply gained (dictator game) significantly influenced subjective experiences. In other words, subjects' feelings about the outcomes reflected the differential behavioral and memory blocking effects observed in Experiments 1 and 2. These results intimate that there may be specific beliefs about and expectations for normative social behavior that differ between the loss and the gain domains.

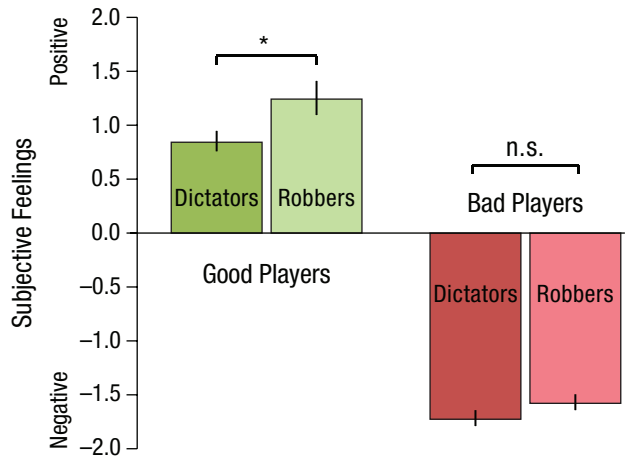


Fig. 4. Subjects' reported feelings for the dictators and robbers across Experiments 1 and 2. The graph compares subjective feelings for the altruistic dictators, kind robbers, selfish dictators, and greedy robbers. Subjective feelings were calculated as the difference between subjects' ratings and the midpoint rating (e.g., neutral) of the scale. Error bars represent ±1 *SEM*; the asterisk indicates a significant difference ($p < .05$).

Discussion

Daily life is filled with encounters that can adaptively guide choice, and yet little is understood about how humans learn about the value of others in group settings. We leveraged an associative-learning framework to investigate whether mechanisms typically observed within the nonsocial domain concisely describe the complex ways in which social stimuli acquire value to bias social choice. Results revealed that within the gain domain, a Pavlovian blocking mechanism explains social decision making; however, in the loss domain, this mechanism fails to fully account for the underlying learning processes.

We found a blocking effect 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. This blocking phenomenon illustrates that when interacting with multiple dictators at once, people do not associate value with those who seem to offer no new information. Although subjects entrusted much money to the altruistic dictators and little to the selfish dictators from Run 1, the altruistic and selfish dictators' partners were entrusted with the same amount of money—which was indiscernible from the amount entrusted to a never-encountered stranger. This blocking mechanism was systematically observed within the gain domain; however, in the loss domain, an asymmetric blocking effect was observed: People learned about and entrusted their money to seemingly redundant kind robbers' partners, who refrained from stealing (i.e., unblocking), but failed to associate value to greedy robbers' partners (i.e., blocking).

We observed a similar asymmetric blocking effect in episodic memory, such that subjects exhibited intact episodic memory for all dictators' and robbers' offers, but poor memory for the partners' offers. Although episodic memory is rarely assessed in decision-making research (Murty et al., 2016), this is, to our knowledge, the first evidence that blocked episodic memory may bias subsequent social choice. In the one instance in which behavioral unblocking occurred—for kind robbers' partners—a corresponding unblocking effect was observed for episodic memory; this unblocking effect supports the idea that episodic memory could be critical for learning the social value of other people.

An important feature of these findings is that they illustrate that social learning appears to recruit basic Pavlovian mechanisms observed across species (Pavlov, 1927). The revolutionary discovery of blocking revealed that learning is not the result of mere co-occurrence of conditioned stimuli (in this case, the dictators) and unconditioned stimuli (in this case, the social outcome of receiving money; Mackintosh, 1975; Pearce & Hall, 1980). Rather, learning relies on an outcome being "surprising"—for example, when a prediction error occurs—and on the quality of the conditioned stimuli as predictors of the outcome. In our paradigm, given that the original altruistic or selfish dictator already predicted the outcome, there was no prediction error when his partner was present, and thus the partner acquired no social value. Although this reasoning is consistent with an associative-learning framework, from a social perspective, it is surprising that the partner did not hold any informative value and was treated as a stranger—given that subjects had explicit information about the partner's behavior.

A conventional Rescorla-Wagner model (Rescorla & Wagner, 1972) can easily account for the blocking effects observed in the gain domain, as there is no error to drive learning. However, this account predicts broad blocking effects regardless of context, and it would be unable to explain the asymmetric unblocking effects observed in the loss domain. Why would conventional associative-learning models perfectly explain the learning effects within the social gain domain, but fail to explain how humans experience and learn about social loss? One possibility, captured by an associability account (Pearce & Hall, 1980), is that the link between a stimulus and its outcome critically depends on the attention paid to them (Mackintosh, 1975). Because losses, compared with gains, garner more attention and are perceived as more emotionally salient (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001), individuals may be more acutely attuned to the possibility of losing hard-earned money than to the possibility of gaining money. This account, however, cannot easily explain why social value is learned for robbers who are involved in small losses but not for those who are involved in large losses.

An alternative, and perhaps more plausible, account is that inferences about and expectations of social behavior differ between the loss and gain domains, and that these inferences and expectations subsequently influence how social phenomena are attended to and experienced. If this is the case, the observed asymmetric learning within the social loss domain may be better captured by contemporary Bayesian learning accounts that describe how prior expectations (Courville, Daw, & Touretzky, 2006) govern how individuals statistically reason about the likelihood of events (Dayan & Long, 1998). If social expectations dictate that stealing typically occurs when there is opportunity, then the statistical priors of the robbery-game environment indicate a high likelihood of stealing. Violations of these statistical priors are anomalies that produce learning and enable unblocking. Applied to our paradigm, this line of reasoning suggests that the initial act of a kind robber failing to steal money in Run 1 was experienced as highly rewarding (as evidenced by subjects' subjective

reports), and a positive association with the kind robber was produced. When the kind robber was later paired with a partner, the assumption was that money would be stolen, because if people typically steal when there is opportunity to do so, the introduction of a new robber should result in a monetary loss. Because the kind robber and his partner failed to steal, the expectation was again violated, and the partner acquired positive value. In this context, the surprising outcome was that stealing did not ensue even with the addition of another robber. That blocking occurred when greedy robbers and their partners continuously stole large amounts of money can be explained with the same logic, as the expectation that people advantageously steal remained intact. A Bayesian account allows for beliefs or expectations to differ between social loss and gain, and, indeed, recent research within the moral domain has shown that moral phenomena also fall prey to asymmetrical expectations (Chakroff, Russell, Piazza, & Young, 2017).

Finding that the same basic mechanisms that govern associations between elementary sensory stimuli and appetitive or aversive outcomes also seem to govern—in part—complex social associations has important implications for understanding the building blocks of social learning. Just as associative-learning processes rely on prediction errors, so do the processes that underpin how people learn about the social value of others in dynamic groups. Thus, domain-general mechanisms are likely a fundamental feature of human learning, regardless of whether the context is social or not. Although our experiments help identify and characterize possible core learning mechanisms supporting the representation of social value, we readily recognize that there may be certain aspects of social learning that do not recruit domain-general processes. Future work aimed at disentangling the mechanisms that predominate—or those that fail to work—will help further elucidate the cognitive processes underlying complex social learning.

Action Editor

Timothy J. Pleskac served as action editor for this article.

Author Contributions

O. FeldmanHall, J. E. Dunsmoor, and M. C. W. Kroes developed and designed the experiments. O. FeldmanHall and S. Lackovic conducted the data collection. O. FeldmanHall, J. E. Dunsmoor, M. C. W. Kroes, and E. A. Phelps performed the data analysis and wrote the manuscript. J. E. Dunsmoor and M. C. W. Kroes made equal contributions to this article.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617706394

Open Practices

All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/ckwjx/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617706394. This article has received the badge for Open Data. More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges.

References

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

Bolton, G. E., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, *27*, 269–299. doi:10.1007/s001820050072

Bott, L., Hoffman, A. B., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General*, *136*, 685–699. doi:10.1037/0096-3445.136.4.685

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*, 619–639.

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.

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

Dayan, P., & Long, T. (1998). Statistical models of conditioning. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 117–123). Cambridge, MA: MIT Press.

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

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics*, *16*, 349–371. doi:10.1007/s10683-012-9341-9

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*, 583–610. doi:10.1007/s10683-011-9283-7

Glimcher, P. W. (2009). Neuroeconomics and the study of valuation. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 1085–1092). Cambridge, MA: MIT Press.

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

Greenberg, J. (1993). Stealing in the name of justice: Informational and interpersonal moderators of theft reactions to underpayment inequity. *Organizational Behavior and Human Decision Processes*, *54*, 81–103. doi:10.1006/obhd.1993.1004

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences, USA*, *106*, 12506–12511. doi:10.1073/pnas.0900152106

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136–153. doi:10.1037/0033-295x.93.2.136

Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica*, *47*, 263–291. doi:10.2307/1914185

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

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.

Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review*, *107*, 914–942. doi:10.1037//0033-295X.107.4.914

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

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *57*, 193–243. doi:10.1080/02724990344000141

Lovibond, P. F., Been, S. L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, *31*, 133–142.

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

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644. doi:10.1509/jmkr.45.6.633

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, *60*, 803–836. doi:10.2307/2951567

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

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

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. London, England: Oxford University Press/Humphrey Milford.

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

Plato. (1950). *The republic* (A. D. Lindsay, Trans.). New York, NY: Dutton.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*, 545–556. doi:10.1038/Nrn2357

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*, 159–165. doi:10.1016/j.conb.2008.06.003

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

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. doi:10.1126/science.275.5306.1593

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

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*, 3–21. doi:10.1037/0033-295x.99.1.3

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

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

Vurbic, D., & Bouton, M. E. (2014). A contemporary behavioral perspective on extinction. In F. K. McSweeney & E. S. Murphy (Eds.), *The Wiley-Blackwell handbook of operant and classical conditioning* (pp. 53–76). doi:10.1002/9781118468135