2018-09-05

Social learning of threat and safety

Andreas Olsson[1], Philip Pärnamets[1,2], Eric C. Nook[3], Björn Lindström[1,4]

[1]Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Solna, Sweden

[2]Department of Psychology, New York University, New York, NY, USA

[3]Department of Psychology, Harvard University, Cambridge, MA, USA

[4]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8001 Zürich

Corresponding author:

Andreas Olsson

Division of Psychology,

Department of Clinical Neuroscience,

Karolinska Institutet,

Solna, Sweden

Phone: +46-8-52482459

Email: andreas.olsson@ki.se

Word count: 6080 (inl. Abstract, references, and fig captions), 2 figures

Acknowledgments: We thank Tove Hensler for comments on an earlier draft and assistance with the manuscript. This research was supported by an Independent Starting Grant (284366; Emotional Learning in Social Interaction) from the European Research Council, and the Knut and Alice Wallenberg Foundation (KAW 2014.0237) to A. Olsson.

Abstract

In rapidly changing environments, humans and other animals often glean information about the value of objects and behaviors through social learning. For example, in humans, observing others' behaviors and their consequences, enables the transmission of a wide range of value-based information about what stimuli should be avoided and approached. We survey important developments in our understanding of the behavioral, computational and neural aspects of social learning of threat and safety. In particular, we discuss the study of social learning trough observation, which has enabled comparisons across species. This research shows that observational threat and safety learning draw on mechanisms partially shared with direct (Pavlovian) threat conditioning and extinction learning. Importantly, however, the outcome of social learning is distinguished from asocial forms of learning by its dependence on the learner's processing of social information. Here, we highlight the role of empathic processes during observational learning. We conclude by underscoring the importance of studying social learning across species using behavioral, computational and neural measures.

Keywords: Vicarious learning, observational fear, empathy, amygdala, fMRI

To cope with challenges and opportunities in their environment, animals need to learn the value of stimuli and contexts. In our hyper-social species, such information is rapidly propagated between individuals. For example, in the wake of a natural disaster, you might have received alerts and warnings from friends on social media, hopefully followed by 'safe' status postings. Research has shown that such emotionally charged information, as well as information about value; whom to trust and what food to prefer, spreads at a great speed between individuals in social networks (Christakis & Fowler, 2009). Regardless if these networks are real or virtual, the effects of vicariously experienced information on the individual can be dramatic. Indeed, social learning between peers and across generations shapes the individual's behavior across all walks of life, and has been ascribed a key role in the development of clinical disorders (Rachman, 1972), as well as the evolution of culture (Boyd & Richerson, 2009). Importantly, because social learning can be studied across species in well controlled experiments, it constitutes a unique link between brain, behavior and the society. Yet, most of what is known about the neural and computational properties of learning derives from the study of learning in a social vacuum (Debiec & Olsson, 2017). The chief reason for this lack of progress in understanding social learning is the immense complexity added when several individuals are dynamically interconnected.

In this chapter, we will survey important developments in cognitive and affective neuroscience that examines social learning of threat and safety. We begin by introducing core behavioral, computational and neural mechanisms of learning, followed by a discussion of a social cognitive process with particular relevance to social learning; empathy. Then, similarities and differences between social and non-social forms of learning are highlighted. We conclude by underscoring the importance of studying social learning across species and levels of analyses.

Learning through direct experiences

Neural correlates of threat learning

Pavlovian conditioning is commonly used in the laboratory to study the neural systems and cellular mechanisms underlying emotional learning and memory. In a Pavlovian threat conditioning paradigm, a neutral stimulus (conditioned stimulus, CS) is paired with a directly experienced naturally aversive event (unconditioned stimulus, US) endowing the CS with an ability to trigger a conditioned (threat) response (CR). This paradigm enables the study of the formation and maintenance of directly learned aversions. Direct threat conditioning critically involves the amygdala, a subcortical structure in the temporal lobe consisting of several interconnected parts with relevance to conditioning, in particular the basolateral part (BL, consisting of lateral, LA, basal, BA, and basomedial, BM nuclei) and central (CeA) nucleus (Pitkänen, Savander, & LeDoux, 1997). Sensory information from the midbrain, thalamus and cortex converges in the LA where CS-US associations are formed during learning (Rogan, Stäubli, & LeDoux, 1997). The LA and BA neurons also receive information from other brain structures involved in different aspects of threat learning, among them the hippocampus and prefrontal cortex, which provides contextual information (Maren, Aharonov, & Fanselow, 1997; Morgan, Romanski, & LeDoux, 1993). The LA projects to the CeA, which propagates information downstream to brain regions controlling behavioral, autonomic and somatic defensive responses, such as freezing and the release of stress hormones (LeDoux, Iwata, Cicchetti, & Reis, 1988).

Direct threat conditioning in patients and healthy human participants using functional magnetic resonance imaging (fMRI) supports that the basic neural and computational mechanisms for acquiring and expressing learned threat are conserved across species (LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Phelps & LeDoux, 2005). Like in other animals, the human amygdala is interconnected with cortical regions, including the hippocampus and

ventromedial prefrontal cortex necessary for encoding and retrieving contextual information, and regulating conditioned threat responses. A major role of the hippocampal-PFC circuitry is thus to disambiguate the meaning of cues in the given context (Maren, Phan, & Liberzon, 2013), be it social or non-social. Other cortical regions linked to the amygdala, among them the anterior insula (AI) the ACC have been implicated in the aversive experiences of receiving, anticipating and controlling painful and otherwise aversive experiences (Craig, 2009; Shackman et al., 2011). For example, a recent meta-analysis (Fullana et al., 2016) consistently implicated the ACC and AI during human threat learning, supporting the role of these regions in homeostatic autonomic and behavioral regulation. Interestingly, recent research has suggested some of these serves separate and crucial functions during social learning (Apps, Rushworth, & Chang, 2016).

Formal theories of learning

Several formal theories and computational models have been proposed to account for direct or Pavlovian and instrumental learning in conditioning experiments. Prominent examples are the Rescorla-Wagner (RW; Rescorla & Wagner, 1972) , temporal difference (TD; Sutton & Barto, 1998) and Pearce-Hall (PH; Pearce & Hall, 1980) models. All models have in common that they posit that the organism learns by first predicting outcomes based on the available cues and then comparing the actual outcome with the predicted. This discrepancy, the prediction error, is used to update associations between cues and their predicted outcomes. In the RW model, and the closely related TD model that generalizes RW across multiple timesteps, the change in predicted outcome is given by the prediction error multiplied by a constant learning rate that determined by the salience of cue and the reinforcer. These models are sometimes called US driven learning models, implying that the amount of learning on a trial will be dependent on the strength of the US received. By contrast, the Pearce-Hall model is a CS driven model, meaning that how much is learned from

a given cue is determined from recent experiences with that cue. In the PH model, an associability term governs how much is learned from a given reinforcer. Associability evolves as a function of the absolute, or unsigned, prediction error on the previous trial, capturing the animal's surprise at the previous outcome. In the cognitive neurosciences, researchers have recently begun combining ideas from both RW and PH models, known as the hybrid model (Le Pelley, 2004), which has had success in explaining both direct and social threat learning.

Neural correlates of formal theories

Phasic responses of dopaminergic neurons in the midbrain have been found to be consistent with predictions from RW/TD models; that is excitatory firing for positive prediction errors and reduced firing for negative prediction errors (Schultz, Dayan, & Montague, 1997). These associations have most frequently been reported for the substantia nigra (SN), ventral tegmental area (VTA), and the ventral striatum (a major projection site for the VTA) (VS). While neural correlates of prediction errors often are investigated in the context of reward learning tasks, prediction errors are also found in aversive learning tasks featuring punishing reinforcers, such as air puffs or shocks (Matsumoto & Hikosaka, 2009). Such aversive prediction errors have been reported in a range of structures, including the amygdala (McHugh et al., 2014), the striatum (Delgado, Li, Schiller, & Phelps, 2008), and the PAG (Roy et al., 2014). Furthermore, fMRI studies have found correlates of both unsigned prediction errors and the associability term from the hybrid learning model in the amygdala (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013; Li, Schiller, Schoenbaum, Phelps, & Daw, 2011). Other work has emphasized the interaction between amygdala and the dorsal anterior cingulate cortex (dACC), suggesting a critical role for dACC in interpreting the sign of errors and highlighting the complex interplay between various brain systems for learning (Klavir, Genud-Gabai, & Paz, 2013). In sum, there is considerable evidence that the mammalian brain implements and uses the quantities posited by formal learning theory both

for rewarding and threatening stimuli. Characterizing how these quantities are processed in social learning in an important topic in current research.

Learning in a Social World.

Social learning of threat

Traditionally, the study of learning and its underlying processes has focused on the organism in a social vacuum. In the natural ecology of many species, however, much of the learning of new skills and the value of objects and contexts takes place in social situations, where other individuals serve as intentional or unintentional demonstrators. Social learning, here broadly defined as learning from, or in interaction with, others is often adaptive, as it minimizes exposure to threats, and gives access to others' innovations. Work in theoretical biology shows, however, that social learning is not always adaptive, because information gleaned from others might be error prone. So called social learning strategies, SLS, are described to specify when and from whom the individual should learn from to optimize its behavior in different situations (Boyd & Richerson, 2009; Kendal et al., 2018). It remains unclear, however, if SLS are distinctly social or emerge from basic asocial learning mechanisms. Moreover, the neural bases of SLS are so far unexplored.

An well studied example of social learning studied across species, including rodents (Jeon et al., 2010; Kavaliers, Choleris, & Colwell, 2001; Knapska et al., 2006) and monkeys (Mineka & Cook, 1993), is observational threat learning. In humans, early work on observational, or "vicarious", threat learning, used experimental confederates serving as demonstrators, who responded with distress to CS (Berger, 1961; Hygge & Öhman, 1976). Seminal studies by Mineka and her colleagues (Mineka, Davidson, Cook, & Keir, 1984) showed that cage reared monkeys quickly acquired long lasting threat responses towards snakes after only one exposure to a conspecific's facial expressions of distress. In these

studies, the relationships in strength between the demonstrator's expressed distress, the observer's immediate response to the demonstrator's distress, and the subsequent threat expression in the observer, were comparable to the relationships between US, UR, and CR in direct threat conditioning. Taken together with similar findings in other animals, these findings supports the view that vicarious threat learning partly relies on the same learning mechanisms as classical conditioning. More recently, these findings have been confirmed and extended to human children (Askew & Field, 2007) and adults (Hooker, Germine, Knight, & D'Esposito, 2006; Lindström, Haaker, & Olsson, 2018; Olsson & Phelps, 2004) using both behavioral and neural measures. The workhorse of much recent experimental work is the video-based paradigm introduced by Olsson & Phelps (2004) and described in detail by Haaker, Golkar, Selbing, and Olsson (2017, see fig 1). A "demonstrator" is filmed making stereotypical threat and pain responses when receiving electrical shocks during exposure to CS. The procedure gives a high degree of control over what the participant is exposed to, and can readily be combined with, for example, measures of visual attention (Kleberg, Selbing, Lundqvist, Hofvander, & Olsson, 2015). This paradigm can also be modified to examine the role of observer-demonstrator match in terms of social group, and observational safety learning (Golkar et al. 2013, see below). The findings outlined above also suggest that social cognitive processes, such as empathy, are important for the social US to be effective.

---------- insert fig. 1 about here ----------

The role of empathic processes

Research on empathic processes is of special interest to the understanding of social threat learning. Empathy is a collection of psychological processes allowing people to share

and understand others' thoughts and feelings. One formulation of empathy subdivides it into three specific processes: affect sharing (resonating with another person's emotions), mentalizing (understanding the contents of another person's mind), and prosocial motivation (wanting to reduce others' suffering; Zaki, 2014). Social learning can depend on both affect sharing and mentalizing processes (Olsson et al., 2016), and prosocial behaviors can be both learned and serve as reinforcers themselves in social interaction (Lockwood, Apps, Valton, Viding, & Roiser, 2016).

The last few decades have seen an explosion of interest in charting the neural bases of empathy, with good results. One paradigm that induces affect sharing during fMRI scanning has participants observe photos of people experiencing pain or videos of people discussing emotional experiences, and has revealed that networks of brain regions thought to underlie affect sharing overlap with those thought to contribute to action, sensation, affect, and interception, such as the premotor cortex, visual cortex, inferior frontal gyrus, inferior parietal lobule, and insula (Zaki & Ochsner, 2012). By contrast, tasks that assess mentalizing involve asking participants to specifically take the perspective of an observer or answer questions about what another person believes about a situation. These studies show that mentalizing is subserved by a network of brain regions that allow people to construct mental models and "project" themselves into other times, places, and perspectives (Mitchell, Banaji, & Macrae, 2005; Saxe & Wexler, 2005). These regions include the medial PFC, temporoparietal junction (TPJ), superior temporal sulcus (STS), precuneus, and the temporal poles. It is worth noting that networks underlying affect sharing and mentalizing are separable from each other (Zaki & Ochsner, 2012). It should be clear that studies examining empathy are likely to include several aspects of social learning.

The role of empathizing in observational threat learning was supported by studies showing that individuals high in psychopathic traits do not readily acquire conditioned

responses compared to normal controls (Aniskiewicz, 1979). Other work has demonstrated that observers higher in trait empathy acquire greater conditioned responses (Kleberg et al., 2015), and that by instructing observers to make empathic appraisals of demonstrators also enhances their responses (Olsson et al., 2016). A recent extension of the observational threat learning paradigm using live, naïve participant pairs tested if observational threat learning could be modulated by empathic experience sharing. The degree of synchrony between observers' and demonstrators' electrodermal activity, which indexes autonomic nervous system activity, during learning was found to predict the strength of the observers' later conditioned responses (Pärnamets, Espinosa, & Olsson, 2018). This aligns with work on rodent models which similarly indicates that experiential sharing of emotional states might underpin observational learning in those species (Meyza, Bartal, Monfils, Panksepp, & Knapska, 2017). It is an intriguing possibility that the functionality of empathic experiences in many cases might be the learning opportunity that they offer to the animal.

Neural mechanisms involved in observational threat learning

A critical question in the study of social threat and safety learning is to what extent social learning relies on the same neural mechanisms as non-social forms of learning. As in Pavlovian learning the amygdala has been heavily implicated in observational threat learning (Jeon et al., 2010; Knapska et al., 2006; Olsson, Nearing, & Phelps, 2007). In humans, overlapping amygdala activity was found in participants watching a video of the demonstrator receiving shocks paired with CS, and when later encountering the same CS without receiving any shocks, indicating that a specific CS-US association was formed and expressed in the same regions of the amygdala (Olsson et al., 2007). The same study also reported activations in the anterior insula (AI) and ACC in the CS+ > CS- contrast during the test stage, and this activity during observation predicted the strength of the CR (electrodermal activity) during the test stage, consistent with the roles these areas play in empathic processing.

The joint role of the ACC and amygdala for observational threat learning has been directly investigated in studies in rodents. For example, Jeon et al. (2010) showed that during observational learning, theta band synchronization increased between the ACC and basolateral amygdala (BLA), indicating a close interaction between these regions during learning. Selectively deactivating either region impaired observational learning, showing that both regions play causal role in the formation of threat memories during social learning. These findings have been extended and refined by Allsop et al. (2018) using optogenetic techniques to selectively inhibit cells projecting from ACC to BLA (ACC->BLA). The results showed that the ACC, more specifically, its input to the BLA is critical for learning about the aversive value of a cue predicting aversive treatment of a demonstrator. These findings suggest that the homologous circuitry in the primate ACC might play a similar role. In support of this, studies tracing the white tract fibers of the primate brain (Vogt & Paxinos, 2014) show that the gurys of the ACC (ACCg), is uniquely connected with the neural circuitry implicated in mentalizing and simulation of others' actions; the medial PFC, TPJ and the action system.

A recent fMRI study directly investigated the contributions of three of the core brain regions discussed so far - the amygdala, AI and ACC - to both direct and observational threat learning by contrasting the two types of learning within subjects (Lindström et al., 2018). The behavioral expectancy ratings data from both the direct and observational learning conditions were best described by the hybrid model, which both provided the first evidence that this model applies to observational learning and suggest overlap in the mechanisms underlying the two types of learning. Furthermore, overlapping activity in both the amygdala, the AI, and the ACC in the two types of learning indicates commonalities in the underlying neural systems. The associability term from both direct and observational learning were found in the right AI, in line with earlier findings from direct learning (Li et al., 2011). The researchers also

investigated the flow of information between the amygdala, AI and ACC in response to the UCS using dynamic causal modeling (DCM). The DCM analysis indicated that the US signal likely entered the network through the amygdala for direct learning, and through the AI for observational learning, consistent with the role of the AI in empathic processes.

---------- insert fig. 2 about here ----------

Like the study by Lindström et al. (2018), other work has used formal theories to better understand the contributions of different neural regions to observational threat learning primarily by investigating the role of prediction errors. Meffert, Brislin, White and Blair (2015) conducted a study where participants learned about objects serving as CS through their pairings with observing happy or angry facial expressions (US) directed towards the CS. Prediction errors were calculated over the repeated exposure trials and found to correlate with amygdala activity for both happy and angry emotional expressions, suggesting amygdala involvement in learning about specifically social US. The role of prediction errors in amygdala in direct learning is well characterized and involve NMDA receptors in the lateral amygdala (Johansen, Cain, Ostroff, & LeDoux, 2011). Prediction errors are also downregulated by involvement of opioidergic circuits in the periaqueductal gray (PAG; McNally & Cole, 2006), a region projecting to the amygdala and involved in regulating freezing and other defensive behaviors as well as in analgesia. In an observational threat learning study on humans (Haaker, Yi, Petrovic, & Olsson, 2017), Naltrexone, an opioid antagonist, was administered prior to learning. Compared to placebo controls, Naltrexone treated participants exhibited stronger CRs (electrodermal activity) and stronger activation to the US in the amygdala and in the PAG. When comparing Naltrexone participants to placebo

controls, an increased functional connectivity was displayed between the PAG and the STS, a region associated with the integrative processing of social stimuli and mentalizing.

Observational safety learning

Equally important to learning what is potentially dangerous is to learn when something that was previously dangerous no longer poses a threat. This form of safety learning has traditionally been studied through extinction protocols where the participant is repeatedly, and directly, exposed to the CS in the absence of the US (Bouton, 2002). Extinction training has become the standard experimental protocol to understand both the etiology and the treatment of dysfunctional fear and anxieties (Craske, Hermans, & Vervliet, 2018). A growing literature has shown that safety learning through direct extinction involves the ventromedial PFC and its interaction with the amygdala in both rodents (Milad & Quirk, 2002) and humans (Phelps et al. 2004; see Dunsmoor et al. 2015 for a review). A major goal for the study of social safety learning is to understand whether social safety learning involves a change of the CS-US associations (the fear memory) or the strengthening of the inhibitory safety memories formed during extinction.

Observing a demonstrator approach the target of a phobia in a calm and controlled manner has been shown to reduce anxiety and increase approach behavior towards that target (Bandura, Grusec, & Menlove, 1967). Using a modified version of the video-based threat learning paradigm as described above, research has demonstrated that undergoing observational safety learning was more effective in preventing the recovery of directly conditioned threat responses (during a subsequent reinstatement test) as compared to direct extinction (Golkar et al., 2013). A first on observational safety learning using fMRI (Golkar, Haaker, Selbing, & Olsson, 2016) replicated these findings, and found that vmPFC activity

increased during safety learning to the reinforced, but not extinguished, CS, and that this activity was coupled with an increase in connectivity to the amygdala. The the vmPFC activity was interpreted as tracking the relative cue value. More work in needed to fully understand its role in observational safety learning.

Social instrumental learning

Learning is not only passive, but crucially also involves actively intervening in the environment to learn how actions can bring about rewarding or punishing consequences – instrumental learning (Balleine & Dickinson, 1998). There has been considerable work on how stimulus-action-outcome contingencies are learned and the computational properties of the underlying neural systems (Dolan & Dayan, 2013; Ruff & Fehr, 2014). However, less is known about the computational and neural mechanisms involved when learning from others.

In one experiment, participants made choices between options that were probabilistically rewarded or punished. Participants made choices without and with social information derived from viewing a demonstrator make choices, as well as seeing the outcome of the observer's choice (Burke, Tobler, Baddeley, & Schultz, 2010). Increased social information monotonically increased the quality of participants' choices. When social information was restricted to the demonstrators' actions, observational action prediction errors (the difference between the observed and predicted action) were expressed in the dorsolateral prefrontal cortex (dlPFC) activity, thought to reflect increased uncertainty in selection given the choice of the demonstrator. When social information included both the actions of, and outcomes for, the demonstrator, observational prediction instead correlated with vmPFC activity and inversely with ventral striatal activity, indicating the full integration of these quantities into the brain's valuation circuits. The behavioral findings from this experiment were replicated and refined in a study using the same conditions but additionally manipulating the skill of the demonstrator (Selbing, Lindström, & Olsson, 2014). Participants performed

better when observing both skilled and unskilled demonstrators relative to when learning on their own. The demonstrator's skill level modulated an imitation rate parameter in an RL model, which determined how much the demonstrator's choices affected the participant. Together, these studies show that participants readily and adaptively use observational information from others' choices and this process can be well described using formal learning theories.

Our understanding of social learning has developed dramatically over recent years thanks to both theoretical and empirical advancements, including the use of experimental models comparable across species. For example, research on observational threat and safety learning has shown that these learning procedures draw on computational and neural mechanisms partially shared with direct (Pavlovian) threat conditioning and extinction learning, respectively. Importantly, however, social learning is distinguished from direct forms of learning by its dependence on social cognition, including empathic processes. The experimental study of social learning offers a unique opportunity to bridge knowledge about computational and neural mechanisms with the study of behaviors that support social phenomena at a societal scale.

REFERENCES

Allsop, S. A., Wichmann, R., Mills, F., Burgos-Robles, A., Chang, C. J., Felix-Ortiz, A. C., … Tye, K. M. (2018). Corticoamygdala transfer of socially derived information gates observational learning. *Cell*, *173*(6), 1329–1342. http://doi.org/10.1016/j.cell.2018.04.004

Aniskiewicz, A. S. (1979). Autonomic components of vicarious conditioning and psychopathy. *Journal of Clinical Psychology*, *35*(1), 60–67. http://doi.org/10.1002/1097-4679(197901)35:1<60::AID-JCLP2270350106>3.0.CO;2-R

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron*, *90*(4), 692–707. http://doi.org/10.1016/J.NEURON.2016.04.018

Askew, C., & Field, A. P. (2007). Vicarious learning and the development of fears in childhood. *Behaviour Research and Therapy*. http://doi.org/10.1016/j.brat.2007.06.008

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407–419. http://doi.org/https://doi.org/10.1016/S0028-3908(98)00033-1

Bandura, A., Grusec, J. E., & Menlove, F. L. (1967). Vicarious extinction of avoidance behavior. *Journal of Personality and Social Psychology*, *5*(1), 16–23. http://doi.org/10.1037/h0024182

Berger, S. M. (1961). Incidental learning through vicarious reinforcement. *Psychological Reports*, *9*(3), 477–491. http://doi.org/10.2466/pr0.1961.9.3.477

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans.

European Journal of Neuroscience, *37*(5), 758–767. http://doi.org/10.1111/ejn.12094

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry*, *52*(10), 976–86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12437938

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), 3281 LP-3288. Retrieved from http://rstb.royalsocietypublishing.org/content/364/1533/3281.abstract

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431–6. http://doi.org/10.1073/pnas.1003111107

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York, NY: Little, Brown.

Craig, A. D. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70. http://doi.org/10.1038/nrn2555

Craske, M. G., Hermans, D., & Vervliet, B. (2018). State-of-the-art and future directions for extinction as a translational model for fear and anxiety. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1742). Retrieved from http://rstb.royalsocietypublishing.org/content/373/1742/20170025.abstract

Debiec, J., & Olsson, A. (2017). Social fear learning: From animal models to human function. *Trends in Cognitive Sciences*, *21*(7), 546–555. http://doi.org/10.1016/j.tics.2017.04.010

Delgado, M. R., Li, J., Schiller, D., & Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal*

Society B: Biological Sciences, *363*(1511), 3787 LP-3800. Retrieved from http://rstb.royalsocietypublishing.org/content/363/1511/3787.abstract

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. http://doi.org/10.1016/J.NEURON.2013.09.007

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron*, *88*(1), 47–63. http://doi.org/10.1016/j.neuron.2015.09.028

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*. http://doi.org/10.1038/mp.2015.88

Golkar, A., Haaker, J., Selbing, I., & Olsson, A. (2016). Neural signals of vicarious extinction learning. *Social Cognitive and Affective Neuroscience*, *11*(10), 1541–1549. http://doi.org/10.1093/scan/nsw068

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

Haaker, J., Golkar, A., Selbing, I., & Olsson, A. (2017). Assessment of social transmission of threats in humans using observational fear conditioning. *Nature Protocols*, *12*, 1378. Retrieved from http://dx.doi.org/10.1038/nprot.2017.027

Haaker, J., Yi, J., Petrovic, P., & Olsson, A. (2017). Endogenous opioids regulate social threat learning in humans, *8*, 15495. Retrieved from http://dx.doi.org/10.1038/ncomms15495

Hooker, C. I., Germine, L. T., Knight, R. T., & D'Esposito, M. (2006). Amygdala response to facial expressions reflects emotional learning. *Journal of Neuroscience*, *26*(35), 8915–

8922. http://doi.org/10.1523/JNEUROSCI.3048-05.2006

Hygge, S., & Öhman, A. (1976). Conditioning of electrodermal responses through vicarious instigation and through perceived threat to a performer. *Scandinavian Journal of Psychology*, *17*(1), 65–72. http://doi.org/10.1111/j.1467-9450.1976.tb00213.x

Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H. E., Lin, S.-Y., … Shin, H.-S. (2010). Observational fear learning involves affective pain system and Cav1.2 Ca2+ channels in ACC. *Nature Neuroscience*, *13*(4), 482–488. http://doi.org/10.1038/nn.2504

Johansen, J. P., Cain, C. K., Ostroff, L. E., & LeDoux, J. E. (2011). Molecular mechanisms of fear learning and memory. *Cell*, *147*(3), 509–524. http://doi.org/10.1016/J.CELL.2011.10.009

Kavaliers, M., Choleris, E., & Colwell, D. D. (2001). Learning from others to cope with biting flies: Social learning of fear-induced conditioned analgesia and active avoidance. *Behavioral Neuroscience*, *115*(3), 661–674. http://doi.org/10.1037/0735-7044.115.3.661

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. http://doi.org/10.1016/j.tics.2018.04.003

Klavir, O., Genud-Gabai, R., & Paz, R. (2013). Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning. *Neuron*, *80*(5), 1290–1300. http://doi.org/10.1016/J.NEURON.2013.09.035

Kleberg, J. L., Selbing, I., Lundqvist, D., Hofvander, B., & Olsson, A. (2015). Spontaneous eye movements and trait empathy predict vicarious learning of fear. *International Journal of Psychophysiology*, *98*(3), 577–583. http://doi.org/10.1016/j.ijpsycho.2015.04.001

Knapska, E., Nikolaev, E., Boguszewski, P., Walasek, G., Blaszczyk, J., Kaczmarek, L., & Werka, T. (2006). Between-subject transfer of emotional information evokes specific pattern of amygdala activation. *Proceedings of the National Academy of Sciences*, *103*(10), 3858–3862. http://doi.org/10.1073/PNAS.0511302103

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron*, *20*(5), 937–945. http://doi.org/10.1016/S0896-6273(00)80475-4

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology*, *57*(3b), 193–243. http://doi.org/10.1080/02724990344000141

LeDoux, J. E., Iwata, J., Cicchetti, P., & Reis, D. J. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *The Journal of Neuroscience*, *8*(7), 2517 LP-2529. Retrieved from http://www.jneurosci.org/content/8/7/2517.abstract

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*(10), 1250–1252. http://doi.org/10.1038/nn.2904

Lindström, B., Haaker, J., & Olsson, A. (2018). A common neural network differentially mediates direct and social fear learning. *NeuroImage*, *167*, 121–129. http://doi.org/10.1016/j.neuroimage.2017.11.039

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763–9768. http://doi.org/10.1073/pnas.1603198113

Maren, S., Aharonov, G., & Fanselow, M. S. (1997). Neurotoxic lesions of the dorsal hippocampus and Pavlovian fear conditioning in rats. *Behavioural Brain Research*, *88*(2), 261–274. http://doi.org/10.1016/S0166-4328(97)00088-0

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*, *14*(6), 417–428. http://doi.org/10.1038/nrn3492

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*, 837. Retrieved from http://dx.doi.org/10.1038/nature08028

McHugh, S. B., Barkus, C., Huber, A., Capitão, L., Lima, J., Lowry, J. P., & Bannerman, D. M. (2014). Aversive prediction error signals in the amygdala. *The Journal of Neuroscience*, *34*(27), 9024 LP-9033. Retrieved from http://www.jneurosci.org/content/34/27/9024.abstract

McNally, G. P., & Cole, S. (2006). Opioid receptors in the midbrain periaqueductal gray regulate prediction errors during Pavlovian fear conditioning. *Behavioral Neuroscience*, *120*(2), 313–323. http://doi.org/10.1037/0735-7044.120.2.313

Meffert, H., Brislin, S. J., White, S. F., & Blair, J. R. (2015). Prediction errors to emotional expressions: the roles of the amygdala in social referencing. *Social Cognitive and Affective Neuroscience*, *10*(4), 537–544. http://doi.org/10.1093/scan/nsu085

Meyza, K. Z., Bartal, I. B.-A., Monfils, M. H., Panksepp, J. B., & Knapska, E. (2017). The roots of empathy: Through the lens of rodent models. *Neuroscience & Biobehavioral Reviews*, *76*, 216–234. http://doi.org/10.1016/j.neubiorev.2016.10.028

Milad, M. R., & Quirk, G. J. (2002). Neurons in medial prefrontal cortex signal memory for

fear extinction. *Nature*, *420*, 70. Retrieved from http://dx.doi.org/10.1038/nature01138

Mineka, S., & Cook, M. (1993). Mechanisms involved in the observational conditioning of fear. *Journal of Experimental Psychology: General*, *122*(1), 23–38. http://doi.org/10.1037/0096-3445.122.1.23

Mineka, S., Davidson, M., Cook, M., & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology*, *93*(4), 355–372. http://doi.org/10.1037/0021-843X.93.4.355

Mitchell, J. P., Banaji, M. R., & Macrae, N. C. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315.

Morgan, M. A., Romanski, L. M., & LeDoux, J. E. (1993). Extinction of emotional learning: Contribution of medial prefrontal cortex. *Neuroscience Letters*, *163*(1), 109–113. http://doi.org/10.1016/0304-3940(93)90241-C

Olsson, A., McMahon, K., Papenberg, G., Zaki, J., Bolger, N., & Ochsner, K. N. (2016). Vicarious Fear Learning Depends on Empathic Appraisals and Trait Empathy. *Psychological Science*, *27*(1), 25–33. http://doi.org/10.1177/0956797615604124

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*(1), 3–11. http://doi.org/10.1093/scan/nsm005

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science*, *15*(12), 822–828. http://doi.org/10.1111/j.0956-7976.2004.00762.x

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*(6), 532–552. http://doi.org/10.1037/0033-295X.87.6.532

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron, 43*(6), 897–905. http://doi.org/10.1016/J.NEURON.2004.08.042

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

Pitkänen, A., Savander, V., & LeDoux, J. E. (1997). Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. *Trends in Neurosciences, 20*(11), 517–523. http://doi.org/10.1016/S0166-2236(97)01125-9

Pärnamets, P., Espinosa, L., & Olsson, A. (2018). Physilogical synchrony between individuals predicts observational threat learning in humans. BioRxiv.

Rachman, S. (1972). Clinical applications of observational learning imitation and modeling. *Behavior Therapy, 3*(3), 379–397. http://doi.org/10.1016/S0005-7894(72)80139-4

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

Rogan, M. T., Stäubli, U. V., & LeDoux, J. E. (1997). Fear conditioning induces associative long-term potentiation in the amygdala. *Nature, 390*(6660), 604–607. http://doi.org/10.1038/37601

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, *17*, 1607. Retrieved from http://dx.doi.org/10.1038/nn.3832

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

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–1399. http://doi.org/10.1016/j.neuropsychologia.2005.02.013

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9054347

Selbing, I., Lindström, B., & Olsson, A. (2014). Demonstrator skill modulates observational aversive learning. *Cognition*, *133*(1), 128–139. http://doi.org/10.1016/j.cognition.2014.06.010

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nature Reviews Neuroscience*, *12*(3), 154–167. http://doi.org/10.1038/nrn2994

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Vogt, B. A., & Paxinos, G. (2014). Cytoarchitecture of mouse and rat cingulate cortex with human homologies. *Brain Structure & Function*, *219*(1), 185–92. http://doi.org/10.1007/s00429-012-0493-3

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–47. http://doi.org/10.1037/a0037679

Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*(5), 675–680. http://doi.org/10.1038/nn.3085

Fig.1**.** General design of the observational fear conditioning protocol depicting the observer (participant) in shaded gray, first watching the demonstrator's responses to the CS–US pairings (observational learning stage), followed by being exposed to the CS (direct test stage). Obs; Observational, CS-; Conditioned stimulus never paired with shock, CS+; Conditioned stimulus paired with shock, ITI; Intertrial interval. Adapted from Haaker, Golkar, et al., 2017.

Fig. 2. Dynamic causal modelling (DCM) of (a) direct and (b) observational threat learning (Lindström, Haaker & Olsson, 2017). The most likely input region for the US in direct and observational learning was the amygdala and AI, respectively. The dotted arrows show the most likely targets for associability gating. AI; Anterior Insula, ACC; anterior cingulate cortex, Amy; Amygdala, US; Unconditioned stimulus.