NeuroImage 167 (2018) 121–129

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A common neural network differentially mediates direct and social fear learning

Björn Lindström [a,b,*,1], Jan Haaker [a,c,1], Andreas Olsson [a]

[a] Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Sweden
[b] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Zurich, Switzerland
[c] Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Germany

ARTICLE INFO

Keywords:
Fear conditioning
Social learning
fMRI
DCM
Amygdala
Associability

ABSTRACT

Across species, fears often spread between individuals through social learning. Yet, little is known about the neural and computational mechanisms underlying social learning. Addressing this question, we compared social and direct (Pavlovian) fear learning showing that they showed indistinguishable behavioral effects, and involved the same cross-modal (self/other) aversive learning network, centered on the amygdala, the anterior insula (AI), and the anterior cingulate cortex (ACC). Crucially, the information flow within this network differed between social and direct fear learning. Dynamic causal modeling combined with reinforcement learning modeling revealed that the amygdala and AI provided input to this network during direct and social learning, respectively. Furthermore, the AI gated learning signals based on surprise (associability), which were conveyed to the ACC, in both learning modalities. Our findings provide insights into the mechanisms underlying social fear learning, with implications for understanding common psychological dysfunctions, such as phobias and other anxiety disorders.

1. Introduction

Humans, and many other animals, can acquire fears through observing conspecifics being subjected to aversive events (Bandura and McClelland, 1977; Debiec et al., 2017; Olsson and Phelps, 2007; Rachman, 1977). This capacity for social learning without direct exposure to aversive consequences is an adaptation that allows the organism to avoid the costs of individual learning, such as predation and poisoning (Laland, 2004; Lindström et al., 2016). Social learning might, however, not always be adaptive: a large proportion of human fears and phobias is acquired through social means, speaking to the generality of this learning mechanism (Askew and Field, 2008; Rachman, 1977).

Despite the cross-species importance of social fear learning, its neural and computational underpinnings are poorly understood, which stands in contrast to our quickly advancing knowledge about fear learning through direct, Pavlovian, conditioning (henceforth *direct fear learning*). Animal studies of direct fear learning have delineated the amygdala as critical for the acquisition, storage and expression of conditioned fear (LeDoux, 2012). The amygdala is thought to be the site where information about the stimulus predicting danger (the conditioned stimulus, CS) becomes associated with information about the aversive event (the unconditioned stimulus, US). Similarly, research has demonstrated that the amygdala is critical for direct fear learning also in our species (Delgado et al., 2006; LaBar et al., 1998; Phelps and LeDoux, 2005).

Recent studies in rodents (Debiec and Sullivan, 2014; Jeon et al., 2010; Knapska et al., 2006), and fMRI studies in humans (Meffert et al., 2015; Olsson et al., 2007), suggest that the amygdala also plays a central role in acquiring fears through social fear learning. Together with a range of overt behavioral similarities (Olsson and Phelps, 2007), this shared neural circuitry suggests that partially overlapping mechanisms are involved in direct and social fear learning. However, because the available studies have not directly contrasted direct and social forms of fear learning within the same participant, this conclusion has been tentative. This situation mirrors the ongoing debate within the wider field of learning and decision making, where many conclusions about the neural and computational overlap between individual and social experiences are constrained by comparison between studies or reverse inference (Ruff and Fehr, 2014).

In the present study, we addressed the dearth of knowledge about the neural underpinnings of social fear learning by investigating shared and unique features of hemodynamic responses (fMRI) during direct and social fear learning within the same participant. Such a direct comparison

* Corresponding author. Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Sweden.
E-mail address: bjorn.r.lindstrom@gmail.com (B. Lindström).
[1] Equal contribution.

https://doi.org/10.1016/j.neuroimage.2017.11.039
Received 21 March 2017; Received in revised form 27 October 2017; Accepted 18 November 2017
Available online 21 November 2017

within participants is crucial for analyzing the similarities and differences between direct and social fear learning. Moreover, we combined formal learning theory and dynamic causal modeling (DCM) to better understand how the computational mechanisms involved in direct and social fear learning converge and diverge.

The computational mechanisms underlying direct fear conditioning can be well characterized by classical formal learning theory, such as the Rescorla-Wagner (R-W) (Rescorla and Wagner, 1972a) and Pearce-Hall (P-H) models (Pearce and Bouton, 2001) that describe how the CS and US signals are combined algebraically to associate cues with aversive events. The R-W model encompasses the idea of error-driven learning, where the mismatch (i.e., prediction errors) between delivered reinforcements and the organism's predictions of reinforcement leads to updated associations. One basic assumption in the R-W model is that the rate of learning is constant (Rescorla and Wagner, 1972a). In contrast, the P-H model explicitly describe how more surprising outcomes change the rate of learning (termed *associability*) (Pelley, 2004). Associability increases in proportion to the absolute prediction error on the last interaction with a stimulus, allowing the agent to adapt to changing environments, which by definition, leads to larger prediction errors, and thereby higher associability. These two accounts have recently been combined in a hybrid model, which is closely related to optimal (Bayesian) statistical inference (Roesch et al., 2012). The computations posited by the hybrid model have been linked to neural signals within the amygdala and insular cortex in humans (Boll et al., 2013; Li et al., 2011), as well as the amygdala in rodents (Roesch et al., 2012). These structures implement a surprise-based associability signal, which gates learning from prediction errors when the environment is changing. Whether the amygdala and insular cortex subserve similar roles in social fear learning is presently unknown.

Based on a previously suggested neural model of social fear learning (Olsson and Phelps, 2007), we hypothesized that social fear learning would involve many of the same brain regions as direct fear learning, representing a cross-modal (self/other) core "aversive learning network". We predicted that the amygdala would be at the center of this aversive learning network, and that one key computational role of the amygdala would be to gate learning based on associability, or attention, signals, in both direct and social fear learning. In addition to the amygdala, this aversive learning network should involve regions linked to the aversive value of pain, both self-experienced and empathic, such as the anterior insula (AI) and the anterior cingulate cortex (ACC), because the US response driving fear learning rests on such representations. Furthermore, because social fear learning by definition must involve distinctly social processes (simply because another individual is the recipient of the US), we predicted that regions commonly involved in social cognition and theory of mind, such as the dorsomedial prefrontal cortex (DMPFC), the superior temporal sulcus (STS), and the **temporo-parietal** (TPJ), would be uniquely involved in social fear learning (Olsson and Phelps, 2007). This involvement might either be additive, so that the "aversive learning" network underlying social fear learning would involve additional nodes relative to direct fear learning, or modulatory, so that the same network nodes would receive differential input depending on the modality (self/other).

2. Methods

2.1. Participants

Twenty-eight healthy adults (15 female. Mean age = 22.8, SD = 3.33), right-handed participants who were free from self-reported life-time psychiatric or neurological disease and medication were recruited. All participants provided written informed consent and were paid 350 SEK (approximately 38 USD) for their participation. Prior to analysis, one participant was removed as this person aborted the experiment during the Direct phase, leaving 27 participants in the analyzed sample. All procedures were approved by the regional ethics board at Karolinska Insitutet, Stockholm, Sweden.

2.2. Stimuli and experimental timing

For the direct learning phase, two differently colored squares (violet and orange) served as CSs (see Fig. 1). The US following the CS+ consisted of a 100-ms DC-pulse electric stimulation applied to the participant's right wrist. For the social learning phase we created movies (1920 × 1080 mm, 25 FPS, in .avi format) using Adobe Premiere Pro CS5.5 that showed the male demonstrator sitting in front of a computer screen watching two differently colored squares (yellow and blue), serving as observational CSs. During the social US, which as in the direct phase followed the CS+, the demonstrator reacted to the shocks by slightly twitching the arm and blinking (resulting from an electric stimulation of the shock electrode that was visibly attached to the demonstrator's right wrist). The demonstrator acted calmly while watching the presentations of the CS-. Each movie had a 6 s duration. During the inter-trial-interval (ITI) a fixation cross was presented to the participants in the same manner in both phases (i.e., without a video, see Fig. 1). The experiment was controlled by Presentation® Software (NeuroBehavioral Systems, Albany California, USA).

2.3. Experimental procedure

Before the experimental task, participants were attached to SCR and shock electrodes and underwent a standard work-up procedure in order to adjust the level of the shock to be experienced as "uncomfortable but not painful" (mean = 35.4 mV, SD = 10.0 mV).

The experimental task consisted of two learning phases (direct and social learning, counterbalanced in order across participants), using a fear discrimination and reversal paradigm, with delay conditioning and partial reinforcement. In the direct learning phase, each CS was presented 24 times in total out of which 12 presentations (50%) of the CS+ co-terminated with an electrical stimulation as the US, whereas the presentation of the CS- was never paired with a shock. The US's followed a pseudo-random (sets of 8 CS+ trials were randomly paired with 4 USs in order to approximately match the number of US within different parts of the experiment) assignment. The contingencies of the CS+ and the CS- were reversed after 24 trials. The participants were in no way instructed about the reversal. The social learning phase consisted of the same number of CS-trials and social US presentation as the direct phase (24 presentation of each observational CS, with 12 trials were the observational CS+ co-terminated with a social US [the shock delivered to the demonstrator]), had the same pseudo-random US assignment, and included a reversal of the contingencies after 24 trials. As for the direct phase, participants were in no way instructed about the reversal.

The participants were uninformed of the purpose and content of the experiment, the two learning phases, and the contingencies of the CSs or of a reversal of contingencies. Before the direct learning phase, subjects were given the following instructions: "*During the next stage you will watch two colored squares. You might receive an electrical stimulation*" Prior to the social learning stage, subjects were instructed with the following: "*During the next stage you will watch another person that will be watching two colored squares. That person might receive an electrical stimulation*".

2.4. Subjective ratings

As the primary measure of learning, participants were asked to indicate (yes/no) by button presses at each CS onset if they were expecting an electrical stimulation (during the direct fear learning phase) or if they were expecting an electrical stimulation for the demonstrator (during the social fear learning phase). The participants were informed that their US expectancy ratings did not affect the likelihood that either they (direct phase) or the demonstrator (social phase) would receive electric shocks. We in addition collected skin conductance responses (SCR) as a secondary measure. However, SCR analysis was prohibited by