Towards a new understanding of fear generalization and its neural origin

Selim Onat
Departments of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
Martinistr. 52, D-20246 Hamburg, Germany
Email address: sonat@uke.de

Abstract

Forming generalizations from previous experiences is a complex skill, which requires a delicate coordination between several basic cognitive abilities. In menacing situations, this ability is called "fear generalization". It allows humans to predict harmful events and is necessary for survival. Impairments of this ability may lead to overgeneralizations – a phenomenon we know from anxiety disorders. By and large, fear generalization has been studied with one type of experimental paradigm. Stimuli forming a carefully controlled perceptual similarity gradient have been the basis to quantify behavioral and neuronal "fear generalization profiles". This paradigm has provided fruitful insights into how learnt fear generalizes to perceptually similar events. Yet, a number of findings suggest that fear generalization is more adaptive than predicted by a mechanism which is solely based on perceptual similarity. In this opinion article, I aim to bring new perspectives onto fear generalization as a complex, adaptive process. I will investigate the following major hypotheses: (1) Fear generalization can be understood as the optimal result of a Bayesian inference problem. (2) In real-world conditions, fear generalization builds on conceptual knowledge rather than perceptual similarity alone. (3) Brain structures involved in fear generalization can be causally linked to modulate fear responses adaptively. To test these hypotheses, I propose use of tools including fMRI, EEG as well as intracranial electrical stimulation and LFP recordings in presurgical epilepsy patients. With the combination of these tools, the expected findings have the potential to revolutionize our understanding of fear generalization and anxiety disorders.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27311v1 | CC BY 4.0 Open Access | rec: 31 Oct 2018, publ: 31 Oct 2018

Introduction

One way of dealing with the ungraspable complexity of the environment consists of making generalizations[1,2]. Previously learnt regularities of the environment can be useful when applied to novel situations. For example, a novel nutriment can be categorized as inedible based on past experiences with truly harmful ones. This competence called fear generalization (FG) is a remarkably high-level cognitive ability that builds upon more basic skills such as object recognition and categorization, statistical learning, perceptual learning, memory, affective processing and conceptual learning. FG provides an important opportunity to study how basic cognitive abilities, which are typically studied in isolation, function collectively to generate adaptive behavior in a complex world. Notably, dissonance between these abilities manifests as maladaptive behavior and may result in mental health disorders[3–8], such as specific phobia. These are characterized by an overgeneralization of previous harmful encounters, leading to the perception of truly safe situations as harmful. Therefore, understanding the neuronal and computational mechanisms of FG is crucial both for basic, as well as clinical neuroscience.

The study of human FG has benefited enormously from well-established experimental paradigms dating back to Pavlov[9–14]. The rationale behind these paradigms consists of characterizing how learning generalizes to other events based on their perceptual similarity with a harmful item. During conditioning, humans learn the characteristics of truly harmful (CS+) and safe (CS–) events. The harmful quality of the CS+ is established by pairing it with an aversive outcome (UCS; e.g. mild electric shock on the hand) using well-established conditioning paradigms[15], where learning can be objectively monitored. Empirically, FG is characterized by measuring fear-related responses to other stimuli organized to form a continuous similarity gradient (Fig. 1). Typically, responses decay with decreasing similarity to the CS+ resulting in graded fear tuning profiles[1]. The strength of this paradigm consists of parametric characterization of behavioral and neuronal fear tuning profiles based on their peak positions and widths[16] (Fig. 1). Hence, it provides a powerful paradigm to investigate neuronal mechanisms responsible for enacting adaptive and selective fear responses.

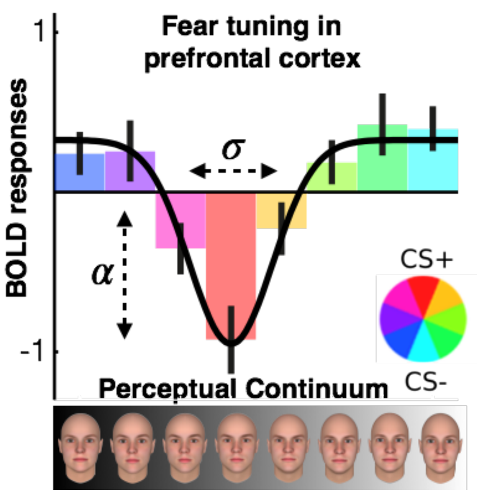

Figure 1 Faces form a circular similarity gradient (color: distance from the CS+, see color wheel). Example BOLD responses form a fear tuning profile, which can be parametrically characterized.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27311v1 | CC BY 4.0 Open Access | rec: 31 Oct 2018, publ: 31 Oct 2018

A Roadmap with Five Milestones to achieve Progress in Fear Generalization Research

Accounts of FG greatly differ on how they attempt to conceptually explain graded fear tuning profiles. According to different perceptual models, graded responses are a mere reflection of perceptual similarity to the behaviorally relevant stimulus[17,18]. However, for FG to be an *adaptive process* in a complex world, it must be flexible and be regulated independently from perceptual factors. Several findings suggest that this is indeed the case; (1) Fear tuning profiles do not always peak on the objectively most harmful stimulus[19]. Such peak-shift are well-documented[20,21] and indicate that FG is prone to subjective biases. (2) Patients with anxiety disorders typically show wider fear-tuning than healthy controls[3,4,6–8,22,23], even though they have been presented with the same perceptual stimulus material. (3) It has been shown that participants readily generalize to semantically related objects that are part of the same category but which do not necessarily bear close physical resemblance (e.g. a hammer and a saw)[24]. *Despite these observations, there are up to date no theoretical frameworks to understand how flexibility and adaptivity emerge during FG.* In this work, I aim to go beyond the current conception of FG as a sole result of physical similarity and I will deliver insights on how we can conceptualize FG as a process that can be adaptively tailored to ensure survival in a complex world. To realize this conceptual shift, I propose the following milestones to be incorporated in a research agenda:

Milestone (1): Descriptive to Normative Transition

Two factors presumably contribute to fear tuning profiles. (1) Previous work provided evidence that uncertainty about the occurrence of harmful events is an important factor for FG[16]. However, the precise role of uncertainty in tailoring FG strategies in an adaptive manner is still largely unknown. (2) There is evidence that prior belief about the harmfulness of different stimuli along the generalization gradient plays an important role that might lead to biases in fear tuning profiles[19]. *The term "bias" implicitly suggests an error in detection performance. However, this behavior can be compatible with an agent behaving optimally while integrating different sources of knowledge with the aim of disambiguating the source of harmful event in uncertain conditions.* Normative models can characterize optimality of behavior. I propose therefore use of a modern theoretical framework of predictive coding[25–27], where Bayesian inference[28–30] takes a central role, as a powerful tool to advance our understanding of FG as an optimal, adaptive phenomenon.

Milestone (2): Role of categorization during fear generalization

Categorization constitutes a fundamental mechanism for organizing and transferring knowledge[31]. It can therefore support FG[24]. Yet, our knowledge on its contribution to FG is scarce[32]. The role of categorical knowledge can be investigated in two forms.
(1) Categorical knowledge can be used to transfer knowledge in abstract ways between events independently from perceptual similarity[31,33]. Hence, *the emergence of categorical knowledge predicts a qualitative change in fear tuning[34]. This transition can be investigated in humans with representational similarity analysis[35] of multivariate activity patterns recorded in EEG and fMRI.* This can inform us where, when and how abstract aversive representations form during FG.

(2) Categorical knowledge can also lead to ambiguity, especially when a given item simultaneously belongs to multiple categories. Understanding how ambiguity is resolved has clinical relevance[36–38]. Using the correct category for FG among many competing ones can only be achieved by collecting statistical regularities about the occurrence of harmful events. It is important to understand whether this ambiguity is resolved faster in healthy humans in comparison to groups suffering from anxiety disorders.

Milestone (3): Establishing causation between neural activity and fear generalization

Understanding neuronal underpinnings of FG requires ultimately establishing causality between neuronal activity and fear tuning profiles. This is especially true for FG paradigms, as many brain regions while they exhibit fear-tuned profiles, may not necessarily be of great importance for FG. Since most of the research about neuronal mechanisms of FG is of correlational nature, studies establishing causality can provide crucial insights. *In particular, parametrically organized stimulus gradients offer the possibility to quantify effects of causal interventions by biasing FG profiles.* Knowing which brain structures shape FG is not only of high interest for the progress of field but also of utmost importance for research into the etiology of anxiety disorders and the development of clinical applications.

Milestone (4): Space and time of neuronal dynamics

Neuronal activity unfolds both in time and space. In the past, investigations of FG has benefited enormously from fMRI[4,6,16,18,22], but much less so from EEG (but see [39]). Consequently, our understanding of fast temporal dynamics of FG during a single trial is scarce. For a thorough characterization of rapid neuronal mechanisms of FG, methods such as EEG and iEEG must be used. Representational similarity analysis[35,40] (RSA) of multivariate activity patterns is an appropriate method to investigate FG both with EEG[41] and fMRI[42]. This is because FG relies ultimately on a subjective metric that evaluates the relatedness of different stimuli to CS+. Hence, similarity of activity patterns between conditions can capture subjective strategies used during FG. In combination with EEG, RSA is a powerful method[42] to investigate temporal dynamics of FG.

Milestone (5): Clinical relevance

Ultimately research in FG must elucidate why anxiety disorders are associated with wide fear tuning profiles. Therefore, it is important that experiments to be systematically conducted also in groups with anxiety disorders. Milestones 1 and 2 on the roadmap can potentially provide key insights on anxiety disorders: Milestone (1) brings a normative approach that will give a computational account of wider fear tuning profiles observed in anxiety disorders. Milestone (2) probes patients for a compromised strategy of updating their internal hypothesis about the source of threat.

Three Empirical Directions

I propose here 3 experimental approaches to proceed on the above roadmap.

Direction(1): Fear Generalization as Harm Prediction: A Bayesian Integration Framework

This experiment addresses milestones 1 ("Normative Framework") and 5 ("Clinical Relevance") of the roadmap. The goal of this package is to cast FG as cognitive ability used to predict future harm. To do so, I propose *conceiving FG as a Bayesian inference problem for recovering the cause of threat in an uncertain environment.* Here, fear tuning reflects the degree of belief about the potential harm of different stimuli. In order to reduce uncertainty about the source of threat, humans integrate different sources of available knowledge:
(1) Learnt threat likelihood and
(2) prior beliefs.
Likelihood reflects the probability of different stimuli to predict objectively harmful outcomes; therefore, it reflects the conditioning regime imposed by the experimenter. Prior beliefs reflect one's previous opinion (i.e. before conditioning) on different stimuli to be harmful. The integration of these sources results in the observed fear tuning, which reflects the posterior, the integrated high-level FG.

1.1 A new Bayesian framework for fear generalization

As a first step, I propose to manipulate explicitly uncertainty levels associated with the prediction of harmful events. This framework predicts that humans will rely more heavily on their prior beliefs when the sensory evidence for the prediction of harmful events is less reliable. This requires controlling uncertainty and using stimulus material where humans can use prior knowledge.

Controlling uncertainty: To introduce uncertainty as an experimental factor one needs to control UCS administration with a probability distribution along the generalization gradient during the conditioning phase. Depending on the width of the probability distribution (min: only one item along the gradient predicts harmful outcome; max: all items equally predict harmful outcome), participants will be provided with more or less reliable sensory information about the source of threat. This allows us to parametrically control uncertainty associated with the delivery of harmful outcomes and provides an elegant extension of the classical conditioning paradigms.

Introducing prior knowledge: Prior beliefs consist of the accumulated experience reflecting regularities acquired across longer time-scales in real life. Therefore, to bring prior knowledge into the game we must use stimuli that have ecological validity. Faces are an excellent choice for this objective, as they are a rich source of information in social situations. Even though prior beliefs are not directly accessible, we can observe their influence by using social priors that people commonly associate with faces. I propose using gender[43], emotional expression[19], pupil size[44], ethnicity[45] and gaze direction[46]. Independent evidence from fear learning literature indicates how these features modulate fear learning (e.g. males perceived more dangerous than females[43]). Therefore, the perspective proposed here will bring together these disparate observations about social priors in an encompassing theoretical framework. These features can be (1) manipulated parametrically, and (2) used as facial elements without interfering with the identity of a face that was previously learnt to predict UCS. By parametrically introducing prior biases along the generalization gradient, it will be possible to test the predictions of the Bayesian framework at different uncertainty levels. The availability of good software support to generate faces makes this a feasible goal.

Predictions: The optimal Bayesian integration makes two clear predictions on empirical fear tuning profiles depending on how threat-likelihood and prior knowledge are aligned with each other, and the uncertainty levels in the threat likelihood.

(i) With increasing uncertainty in threat-likelihood, FG will rely more strongly on prior beliefs. Hence, stronger deviations away from the objectively most harmful stimulus (i.e. where the likelihood peaks) with increasing uncertainty are expected (Fig. 2).

(ii) At a given uncertainty-level, fear tuning must become increasingly sharper with increasing alignment of prior beliefs and threat-likelihood (i.e. smaller separation between the peaks of likelihood and prior). This is because the evidence from threat-likelihood and prior beliefs will add up to produce highly selective fear tuning. Learning with fear-relevant stimuli[47] can be taken as an illustration of this effect. When objects that are also dangerous in real life are used as stimuli stronger and more persistent learning can be established. The predictions will be statistically tested based on the parameters (e.g peak position, tuning width) describing the empirical fear tuning profiles.

(iii)

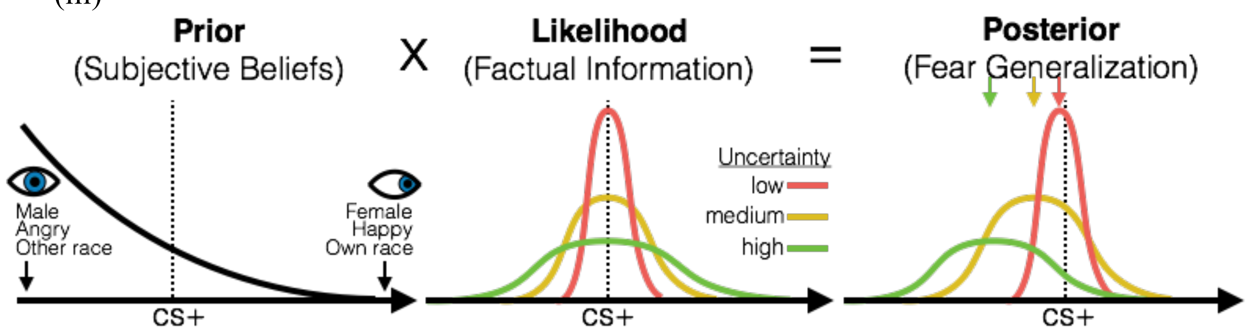

Figure 2 Optimal Bayesian integration for FG predicts larger deviations in fear tuning with increasing uncertainty levels in the likelihood.

Experiment: During the conditioning phase, faces along the generalization gradient will be probabilistically paired with UCSs based on a Gaussian distribution (i.e. likelihood). While the width parameter of this Gaussian controls uncertainty, its peak position determines which face predicts best UCS, hence the alignment between prior and likelihood. Following learning, during the test phase, same faces will be presented, but additionally they will exhibit facial features of social priors. The predictions of the Bayesian framework can be tested using fear-tuning profiles derived from autonomic nervous system activity in the form of skin-conductance responses and complemented with explicit ratings of UCS likelihood.

1.2 Computational characterization of neuronal fear tuning profiles

Previous work[16] showed that fear tuning in prefrontal cortex is significantly wider than in insula, which is characterized by a sharp fear tuning. The Bayesian framework allows us to understand these descriptive observations in computational terms. Wide prefrontal fear tuning is compatible with a fear representation that results from the integration of different sources of information. On the other hand, sharp insular fear tuning can reflect stimulus-UCS contingencies before the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27311v1 | CC BY 4.0 Open Access | rec: 31 Oct 2018, publ: 31 Oct 2018

integration of prior knowledge. Hence, it is compatible with a representation based on threat-likelihood.

This experiment and the encompassing theoretical framework has the potential to advance our understanding of how the brain forms aversive representations and whether these can be understood in terms of optimal aversive representations.

1.3. Clinical study

The Bayesian framework can also provide a basis to advance our understanding of anxiety disorders. For example, this view can be used to investigate a widespread anxiety disorder. Patients with specific phobias (ICD10 - F40.2) exhibit intense fear responses that are triggered by very specific situations[47,48] and exhibit overgeneralization behavior during FG[7,49]. This offers an optimal scenario to study predictions of the Bayesian framework. *This type of behavior can be obtained by an inability to form optimal threat representations, or alternatively by the presence of over-precise prior beliefs.* By comparing the width of the threat-likelihood and the recovered priors with healthy individuals, we will be able to identify which of these effects cause intense fear responses in phobia patients.

Direction (2): Role of Categorization during Fear Generalization

This work package addresses milestones 2 ("*Categorical Knowledge*"), 4 ("*Neurodynamics*") and 5 ("*Clinical Relevance*") of the road map. The first part is concerned with the emergence of categorical knowledge during FG and the associated changes in neuronal representations and dynamics[34]. In the last part, I propose using hierarchically organized categories of commonly-known objects[50] to induce ambiguity and investigate FG strategies in anxiety patients and healthy controls[36–38].

2.1 Two-stage model of fear generalization

FG has been mainly investigated with stimuli organized along a similarity gradient. FG based on perceptual similarity could simply constitute one specific form. This type of similarity-based generalization can be seen as the predecessor of category-based generalization, which instead requires an abstraction from superficial perceptual aspects. However, as long as the organism has not yet experienced enough aversive events, it will not be possible to extract features that can abstractly describe these events. *At this initial stage, grouping different stimuli based on their perceptual similarity could be the best available strategy.* However, gradual learning leads to the emergence of harmful and safe categories. *I predict that this "Aha!" moment will be marked by a transition from fuzzy generalization profiles across perceptually similar stimuli to a binary yes-or-no type generalization profile reflecting their category membership.*

To capture this transition, it is necessary to establish a novel experimental paradigm where the CS+ and CS– will characterize two probabilistic category structures[31,51,52] defined across two facial features (e.g. gender and age). Across interleaved conditioning and test phases, participants can be given the possibility to extract the underlying category structure[53]. Importantly, it is crucial to pit the category membership of faces against perceptual similarities to investigate their independent contributions over the course of the learning. To this end, all stimuli will be characterized both (1)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27311v1 | CC BY 4.0 Open Access | rec: 31 Oct 2018, publ: 31 Oct 2018

by their similarity to previous harmful faces, and (2) by their category membership. By modeling fear-related responses (i.e. SCR, explicit ratings) with these two predictors I will quantify the contribution of perceptual and categorical factors. I predict that with the emergence of categorical knowledge the contribution of perceptual factors will diminish. This will therefore establish an important link between two cognitive abilities that were so far studied separately.

2.2 fMRI on category learning during fear generalization

Bringing this paradigm to fMRI, it will be possible identify neuronal mechanisms responsible for the emergence of categorical knowledge during FG. *This will allow to address the extent to which perceptual and categorical FG shares common neuronal mechanisms.* As category-based FG relies on the use of more abstract knowledge, it is possible that it depends on different neuronal mechanisms than perceptual FG. This echoes an important dichotomy in the categorization research regarding abstract vs. similarity based categorical learning[54]. RSA[40] is powerful and sensitive method that can contribute to the elucidation of neuronal mechanisms. As it evaluates between-condition similarity of multi-voxel activity patterns, it predicts different similarity geometries depending on whether FG proceeds with categorical[24,55] or perceptual factors. It is therefore the appropriate tool for the identification of neuronal mechanisms of FG when multiple factors are available.

2.3 EEG on category learning during fear generalization fear generalization

Temporal dynamics of neuronal activity during FG within a single trial are largely unknown. Using the same experimental paradigm in an EEG setting, one can investigate fast neuronal dynamics (e.g. time-frequency analysis) of FG and characterize temporal unfolding of perceptual and categorical factors. Recently, it has been shown that RSA can be applied on EEG source space[41], and provides one sound way to test the contributions of perceptual and categorical factors during FG. Therefore, bringing this paradigm to EEG-lab can provide synergistic insights to fMRI. In particular, use of RSA it will be possible to merge insights gathered from EEG and fMRI modalities[42].

2.4 Overgeneralization across hierarchically organized categories

Overgeneralization in individuals with anxiety disorders has been observed in FG paradigms using perceptual gradients[3,6,8,23]. There is evidence that this finding can be accounted, at least in part, by perceptual confusion[22]. Moreover, perceptual performance is influence by learning[56,57]. *It is therefore crucial to test the finding of overgeneralization in situations that do not require fine perceptual discrimination.*

In hierarchically organized taxonomies, a single item simultaneously belongs to multiple categories (e.g. sub-ordinate, basic-level). Therefore, during a FG experiment with such stimuli, it should be impossible to unambiguously assign a CS+ to a given hierarchical level. Hence, with such stimuli one can measure overgeneralization independent of perceptual factors. In this experiment we will test FG performance of patients with specific phobias. Overgeneralization, if true, predicts that patients will consistently tune in on higher levels in the hierarchy in comparison to healthy individuals for the prediction of harmful events.

Direction (3): Establishing Causality between FG and Neuronal fear tuning

This experiment addresses the milestones (3) ("*Establishing causation*") and (4) ("*Neurodynamics*") of the roadmap. To achive this, I propose physiological experiments to be conducted on presurgical epilepsy patients with implanted deep electrodes as they offer the unique possibility to directly investigate activity of neuronal populations during complex cognitive tasks[58,59]. These recordings will provide a detailed picture of population dynamics during FG in the form of local-field potentials at precisely known neuronal sites. Most importantly, through stimulation of neuronal activity with subthreshold electrical currents[60,61], it is possible to establish causality between neuronal activity and FG by directly observing changes in fear tuning profiles.

Importantly, FG paradigms relying on perceptual similarity are extremely well-suited for introducing biases in fear tuning profiles during learning with subthreshold electrical stimulation. Furthermore, it is possible to investigate the dynamic emergence of fear tuning by pairing the CS+ face with UCS at unpredictable moments (Fig. 3A). This results in the emergence of fear tuning that dynamically grows along the course of the experiment (Fig. 3A, shock symbols).

3.1 Dynamic emergence of fear generalization profiles in affective brain structures

Fear tuning in human brain has been almost exclusively shown with fMRI[4,16,18,19]. Given the sluggish nature of BOLD responses, our knowledge on neuronal signatures responsible for encoding fear responses is largely unknown. Local field potentials capture population dynamics at high temporal resolution, and thus provide a great source of information. The parametric nature of the FG experiment makes it possible to identify fear tuning across different frequency channels. This will provide important insights for understanding neuronal mechanisms of FG.

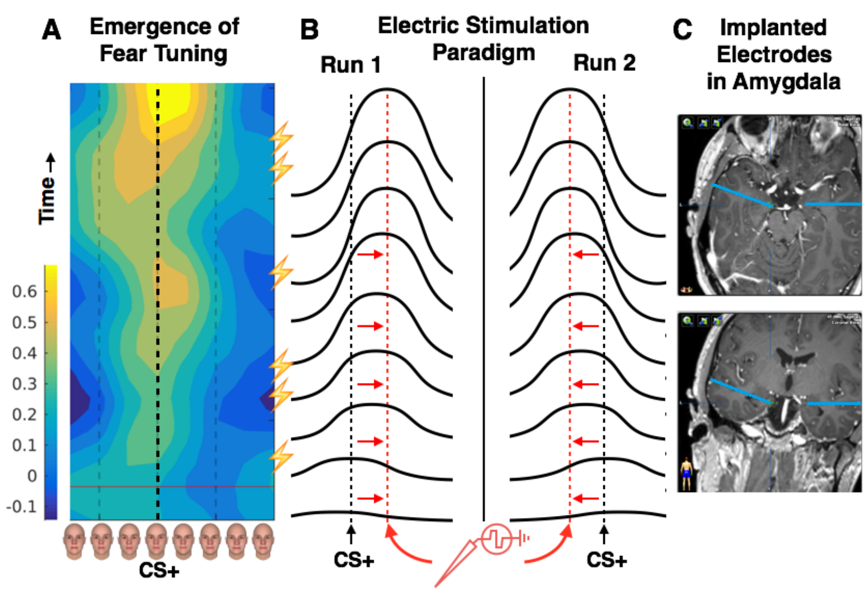

Figure 3 Emergence of FG across time and faces in Insula. (color: BOLD amplitude, shock symbols: UCS delivery with CS+ presentation). FG is shown only for trials where no UCS is administrated. B. Causal intervention with electric stimulation biases fea

3.2 Establishing causality

Understanding how neuronal activity causally contributes to the regulation of behavioral fear tuning profiles is crucial for an understanding of neuronal mechanisms implicated in anxiety disorders[62]. Causal intervention through electrical stimulation is a powerful method that can be

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27311v1 | CC BY 4.0 Open Access | rec: 31 Oct 2018, publ: 31 Oct 2018

used to investigate neuronal sites that can potentially bias fear tuning profiles[61]. The parametric nature of fear tuning profiles provides an objective and quantitative method to investigate these biases at the behavioral level.

To achieve this objective, I will aim to introduce biases on fear tuning profiles via subthreshold electrical stimulation. I will use the loudness of white noise auditory bursts as UCSs with presurgical patients. Therefore along the generalization gradient faces will be paired with increasing loudness levels. The CS+ face will be paired with the loudest UCS. I will aim to increment the aversive quality of faces closely neighboring the CS+ face with electrical stimulation in a reversible manner across two different runs (Fig. 3B). For stimulation, we will use electrode contacts that are functionally related to FG, which will be characterized previously. Using this methodology I will investigate the causal contribution of different neuronal sites to the production of fear tuning profiles.

References

1. Shepard RN. Toward a universal law of generalization for psychological science. *Science*. 1987;237(4820):1317-1323.

2. Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. *Behav Brain Sci*. 2001;24(4):629–640. doi:10.1017/S0140525X01000061.

3. Lissek S, Rabin S, Heller RE, et al. Overgeneralization of conditioned fear as a pathogenic marker of panic disorder. *Am J Psychiatry*. 2009;167(1):47-55. doi:10.1176/appi.ajp.2009.09030410.

4. Greenberg T, Carlson JM, Cha J, Hajcak G, Mujica-Parodi LR. Ventromedial Prefrontal Cortex Reactivity Is Altered in Generalized Anxiety Disorder During Fear Generalization. *Depress Anxiety*. 2013;30(3):242-250. doi:10.1002/da.22016.

5. Lissek S, Kaczkurkin AN, Rabin S, Geraci M, Pine DS, Grillon C. Generalized Anxiety Disorder Is Associated With Overgeneralization of Classically Conditioned Fear. *Biol Psychiatry*. 2014;75(11):909-915. doi:10.1016/j.biopsych.2013.07.025.

6. Cha J, Greenberg T, Carlson JM, DeDora DJ, Hajcak G, Mujica-Parodi LR. Circuit-Wide Structural and Functional Measures Predict Ventromedial Prefrontal Cortex Fear Generalization: Implications for Generalized Anxiety Disorder. *J Neurosci*. 2014;34(11):4043-4053. doi:10.1523/JNEUROSCI.3372-13.2014.

7. Dymond S, Dunsmoor JE, Vervliet B, Roche B, Hermans D. Fear Generalization in Humans: Systematic Review and Implications for Anxiety Disorder Research. *Behav Ther*. http://www.sciencedirect.com/science/article/pii/S0005789414001348. Accessed March 2, 2015.

8. Jasnow AM, Lynch JF, Gilman TL, Riccio DC. Perspectives on fear generalization and its implications for emotional disorders. *J Neurosci Res*. 2017;95(3):821-835. doi:10.1002/jnr.23837.

9. Anrep GV. The irradiation of conditioned reflexes. *Proc R Soc Lond Ser B Contain Pap Biol Character*. 1923:404–426.

10. Bass MJ, Hull CL. The irradiation of a tactile conditioned reflex in man. *J Comp Psychol*. 1934;17(1):47-65.

11. Hovland CI. The Generalization of Conditioned Responses: I. The Sensory Generalization of Conditioned Responses with Varying Frequencies of Tone. *J Gen Psychol*. 1937;17(1):125-148. doi:10.1080/00221309.1937.9917977.

12. Guttman N, Kalish HI. Discriminability and stimulus generalization. *J Exp Psychol*. 1956;51(1):79-88.

13. Hull CL. The problem of primary stimulus generalization. *Psychol Rev*. 1947;54(3):120-134. doi:10.1037/h0061159.

14. Lashley KS, Wade M. The Pavlovian theory of generalization. *Psychol Rev*. 1946;53(2):72-87.

15. Lonsdorf TB, Menz MM, Andreatta M, et al. Don't fear "fear conditioning": Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neurosci Biobehav Rev*. 2017;77:247-285. doi:10.1016/j.neubiorev.2017.02.026.

16. Onat S, Büchel C. The neuronal basis of fear generalization in humans. *Nat Neurosci*. 2015;advance online publication. doi:10.1038/nn.4166.

17. Lissek S. Toward an account of clinical anxiety predicated on basic, neurally mapped mechanisms of Pavlovian fear-learning: the case for conditioned overgeneralization. *Depress Anxiety*. 2012;29(4):257–263. doi:10.1002/da.21922.

18. Lissek S, Bradford DE, Alvarez RP, et al. Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Soc Cogn Affect Neurosci*. 2014;9(8):1134-1142. doi:10.1093/scan/nst096.

19. Dunsmoor JE, Prince SE, Murty VP, Kragel PA, LaBar KS. Neurobehavioral mechanisms of human fear generalization. *NeuroImage*. 2011;55(4):1878-1888. doi:10.1016/j.neuroimage.2011.01.041.

20. Purtle RB. Peak shift: A review. *Psychol Bull*. 1973;80(5):408-421.

21. Thomas DR, Mood K, Morrison S, Wiertelak E. Peak shift revisited: a test of alternative interpretations. *J Exp Psychol Anim Behav Process*. 1991;17(2):130-140.

22. Laufer O, Israeli D, Paz R. Behavioral and Neural Mechanisms of Overgeneralization in Anxiety. *Curr Biol*. 2016;26(6):713-722. doi:10.1016/j.cub.2016.01.023.

23. Dunsmoor JE, Paz R. Fear Generalization and Anxiety: Behavioral and Neural Mechanisms. *Biol Psychiatry*. 2015;78(5):336-343. doi:10.1016/j.biopsych.2015.04.010.

24. Dunsmoor JE, Kragel PA, Martin A, LaBar KS. Aversive Learning Modulates Cortical Representations of Object Categories. *Cereb Cortex*. 2014;24(11):2859-2872. doi:10.1093/cercor/bht138.

25. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2(1):79–87. doi:10.1038/4580.

26. Friston K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci*. 2010;11(2):127-138. doi:10.1038/nrn2787.

27. Vilares I, Kording K. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Ann N Y Acad Sci*. 2011;1224(1):22-39. doi:10.1111/j.1749-6632.2011.05965.x.

28. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS Neurosci*. 2004;27(12):712–719.

29. Ernst MO, Bulthoff HH. Merging the senses into a robust percept. *Trends Cogn Sci*. 2004;8(4):162–169. doi:10.1016/j.tics.2004.02.002.

30. Koerding KP, Wolpert DM. Bayesian integration in sensorimotor learning. *Nature*. 2004;427(6971):244–247. doi:10.1038/nature02169.

31. Ashby FG, Maddox WT. Human Category Learning. *Annu Rev Psychol*. 2005;56(1):149-178. doi:10.1146/annurev.psych.56.091103.070217.

32. Dunsmoor JE, Murphy GL. Categories, concepts, and conditioning: how humans generalize fear. *Trends Cogn Sci*. 2015;19(2):73-77. doi:10.1016/j.tics.2014.12.003.

33. Medin DL. Concepts and conceptual structure. *Am Psychol*. 1989;44(12):1469.

34. Ohl FW, Scheich H, Freeman WJ. Change in pattern of ongoing cortical activity with auditory category learning. *Nature*. 2001;412(6848):733-736. doi:10.1038/35089076.

35. Kriegeskorte N, Mur M, Ruff DA, et al. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*. 2008;60(6):1126-1141. doi:10.1016/j.neuron.2008.10.043.

36. Nader K, Balleine B. Ambiguity and anxiety: when a glass half full is empty. *Nat Neurosci*. 2007;10(7):807-808. doi:10.1038/nn0707-807.

37. Grupe DW, Nitschke JB. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat Rev Neurosci*. 2013;14(7):488-501. doi:10.1038/nrn3524.

38. Lake JI, LaBar KS. Unpredictability and Uncertainty in Anxiety: A New Direction for Emotional Timing Research. *Front Integr Neurosci*. 2011;5. doi:10.3389/fnint.2011.00055.

39. McTeague LM, Gruss LF, Keil A. Aversive learning shapes neuronal orientation tuning in human visual cortex. *Nat Commun*. 2015;6:7823. doi:10.1038/ncomms8823.

40. Kriegeskorte N, Mur M, Bandettini P. Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Front Syst Neurosci*. 2008;2. doi:10.3389/neuro.06.004.2008.

41. Kietzmann TC, Gert AL, Tong F, König P. Representational Dynamics of Facial Viewpoint Encoding. *J Cogn Neurosci*. 2017;29(4):637-651. doi:10.1162/jocn_a_01070.

42. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. *Nat Neurosci*. 2014;17(3):455-462. doi:10.1038/nn.3635.

43. Navarrete CD, Olsson A, Ho AK, Mendes WB, Thomsen L, Sidanius J. Fear Extinction to an Out-Group Face: The Role of Target Gender. *Psychol Sci*. 2009;20(2):155-158. doi:10.1111/j.1467-9280.2009.02273.x.

44. Demos KE, Kelley WM, Ryan SL, Davis FC, Whalen PJ. Human Amygdala Sensitivity to the Pupil Size of Others. *Cereb Cortex*. 2008;18(12):2729-2734. doi:10.1093/cercor/bhn034.

45. Olsson A, Ebert JP, Banaji MR, Phelps EA. The role of social groups in the persistence of learned fear. *Science*. 2005;309(5735):785-787. doi:10.1126/science.1113551.

46. N'Diaye K, Sander D, Vuilleumier P. Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity. *Emotion*. 2009;9(6):798.

47. Öhman A, Mineka S. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychol Rev*. 2001;108(3):483.

48. Ipser JC, Singh L, Stein DJ. Meta-analysis of functional brain imaging in specific phobia. *Psychiatry Clin Neurosci*. 2013;67(5):311-322. doi:10.1111/pcn.12055.

49. Dymond S, Schlund MW, Roche B, Whelan R. The spread of fear: Symbolic generalization mediates graded threat-avoidance in specific phobia. *Q J Exp Psychol*. 2014;67(2):247-259. doi:10.1080/17470218.2013.800124.

50. Xu F, Tenenbaum JB. Word learning as Bayesian inference. *Psychol Rev*. 2007;114(2):245.

51. Nosofsky RM. Attention, similarity, and the identification–categorization relationship. *J Exp Psychol Gen*. 1986;115(1):39.

52. Sigala N, Logothetis NK. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*. 2002;415(6869):318-320. doi:10.1038/415318a.

53. Kietzmann TC, König P. Perceptual learning of parametric face categories leads to the integration of high-level class-based information but not to high-level pop-out. *J Vis*. 2010;10(13):20-20. doi:10.1167/10.13.20.

54. Sloman SA. The empirical case for two systems of reasoning. *Psychol Bull*. 1996;119:3–22.

55. Visser RM, Scholte HS, Beemsterboer T, Kindt M. Neural pattern similarity predicts long-term fear memory. *Nat Neurosci*. 2013;16(4):388-390. doi:10.1038/nn.3345.

56. Resnik J, Sobel N, Paz R. Auditory aversive learning increases discrimination thresholds. *Nat Neurosci*. 2011;14(6):791-796. doi:10.1038/nn.2802.

57. Holt DJ, Boeke EA, Wolthusen RPF, Nasr S, Milad MR, Tootell RBH. A parametric study of fear generalization to faces and non-face objects: relationship to discrimination thresholds. *Front Hum Neurosci*. 2014;8:624. doi:10.3389/fnhum.2014.00624.

58. Lachaux J-P, Axmacher N, Mormann F, Halgren E, Crone NE. High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Prog Neurobiol*. 2012;98(3):279-301. doi:10.1016/j.pneurobio.2012.06.008.

59. Vidal JR, Perrone-Bertolotti M, Kahane P, Lachaux J-P. Intracranial spectral amplitude dynamics of perceptual suppression in fronto-insular, occipito-temporal, and primary visual cortex. *Front Psychol*. 2015;5. doi:10.3389/fpsyg.2014.01545.

60. Cogan SF. Neural Stimulation and Recording Electrodes. *Annu Rev Biomed Eng*. 2008;10(1):275-309. doi:10.1146/annurev.bioeng.10.061807.160518.

61. Guillory SA, Bujarski KA. Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. *Soc Cogn Affect Neurosci*. 2014;9(12):1880-1889. doi:10.1093/scan/nsu002.

62. Etkin A. Neurobiology of Anxiety: From Neural Circuits to Novel Solutions? *Depress Anxiety*. 2012;29(5):355-358. doi:10.1002/da.21957.