Fear Generalization of Emotional Stimuli Can Be Explained By a Bayesian Inference Model

Lukas Neugebauer (l.neugebauer@uke.de)
Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf
Martinistr. 52, 20246 Hamburg, Germany

Christian Büchel (buechel@uke.de)
Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf
Martinistr. 52, 20246 Hamburg, Germany

Abstract:

The amount of generalization that organisms show from learned associations to new stimuli can often be explained by perceptual dissimilarity, i.e. distance in psychological space. However, this doesn't seem to be true using emotionally relevant stimuli like fearful faces. We propose that this can be understood in a Bayesian framework in which the organism infers a mapping of psychological space onto outcome probabilities by integrating prior assumptions with new information. This approach allows for the incorporation of domain specific prior knowledge. We employed face stimuli that differ on one fear relevant (emotional expression) and one fear irrelevant (identity) dimension in combination with Pavlovian conditioning to investigate generalization at several time points. We can show that generalization is skewed towards the fear relevant pole in the beginning, but gravitates towards the actually reinforced stimulus over time. Our Bayesian model that comprises a prior belief state about the structure of the predictive relationship between the psychological space and an aversive outcome can reproduce the experimental data.

Keywords: Generalization, Fear conditioning, Bayesian inference, Cognitive modeling

Generalization describes the cognitive capacity of organisms to translate knowledge from known situations to new ones. This ability is considered an evolutionary advantage because it allows for a more efficient interaction with the environment.

Generative models of generalization

It is generally accepted that similarity plays a decisive role in generalization along perceptual dimensions. In conditioning paradigms one finds stronger generalization from learned associations to stimuli that are more similar to the CS+ (Onat & Büchel, 2015). Most theoretical approaches conceptualize similarity as distances in a psychological space in which more similar stimuli are closer to each other. Shepard (1987) introduced the idea of a *consequential region*, which is a part of psychological space that predicts a consequence. While a seminal contribution to models of generalization, Shepard's approach is only applicable to learning from one single consequential observation. Tenenbaum & Griffiths (2001) extended this idea in a rational analysis of generalization and proposed a Bayesian model in which the prior knowledge is captured in a probability distribution over all potential consequential regions. Other work has used and refined this approach but some issues remain. E.g. that the model can't be applied to probabilistic outcome structures since consequential regions are deterministically linking the psychological space to an outcome. Additionally, the neural representation of a space of discrete hypotheses seems biologically implausible. We propose an alternative formulation that deals with these issues.

Our Modeling Approach

Associative Map As an alternative to consequential regions, we propose the idea of associative maps which map the psychological space onto outcome probabilities for any given consequence. Associative maps capture the idea that different regions in psychological space can lead to the same outcome with different probabilities. E.g. a very moldy bread will lead to gastrointestinal issues with a higher probability than an only slightly moldy piece of bread. An associative map in N dimensions for an outcome is characterized by the following parameters:

1. μ - 1xN coordinate vector of the center of the associative map

2. **θ** – 1xN vector that defines weights for exponential decay of outcome probability in every dimension
3. **φ**– the probability of the outcome at the midpoint

When constraining every dimension in psychological space to the range of [0,1] we can parameterize the prior probability distributions p(H) on the parameters of the model like this:

$$\mu_i \sim Beta(a_{m,i}. b_{m,i})$$
$$\theta_i \sim Gamma(s_i, p_i)$$
$$\varphi \sim Beta(a_s, b_s)$$

This formulation allows for a flexible incorporation of prior knowledge. E.g. we can capture the impact of fear relevant dimensions that lead to skewed generalization gradients like emotional expression (Dunsmoor, Mitroff, & LaBar, 2009). The prior distributions on μ comprise domain knowledge whereas the priors on θ cover general assumptions about the strength of probability decay in any dimension. E.g. after experiencing a threatening situation involving a lion, it is adaptive to generalize to mountain lions, but not to domestic cats.

Bayesian inference Every point in parameter space defines an associative map. The inferred outcome probability of generalization depends on the weighted distance d from the inferred midpoint μ. For a stimulus at the position σ, the weighted distance in N dimensions is given by

$$d(\sigma) = \sqrt{\sum_{i=1}^{N} \left(\theta_i * (\mu_i - \sigma_i)\right)^2}.$$

The outcome probability p for this stimulus is then given by

$$p(\sigma) = \exp(-d(\sigma)) * \varphi.$$

The likelihood of an observation consisting of a stimulus and an outcome given the prior distributions is

$$p(\sigma, out \mid H) = (p(\sigma))^{out} * (1 - p(\sigma))^{1-out}$$

where *out* is 1 for consequential and 0 for non-consequential observations. Consequently, the likelihood of a set of observations consisting of a sequence of N stimuli Σ and respective outcomes O is defined as

$$p(\Sigma, O \mid H) = \prod_{i=1}^{n} p(\Sigma_i, O_i \mid H)$$

The posterior probability of model parameters is then proportional to the product of the prior probability and the likelihood according to Bayes' rule:

$$p(H \mid \Sigma, O) \propto p(\Sigma, O \mid H) * p(H)$$

Methods

Stimulus space

To test the model's ability to distinguish purely perceptual and informed generalization we constructed a stimulus space that consists of one neutral and one fear relevant dimension (emotional expression). For this we designed computer generated faces on a 5 x 5 grid by creating different morphing steps between two identities and then adding different levels of a happy or an angry emotional expression. Figure 1 shows the angry version. In both cases, the stimulus in the center (indicated by black circle) is the CS+, i.e. it is probabilistically reinforced with an electric shock.

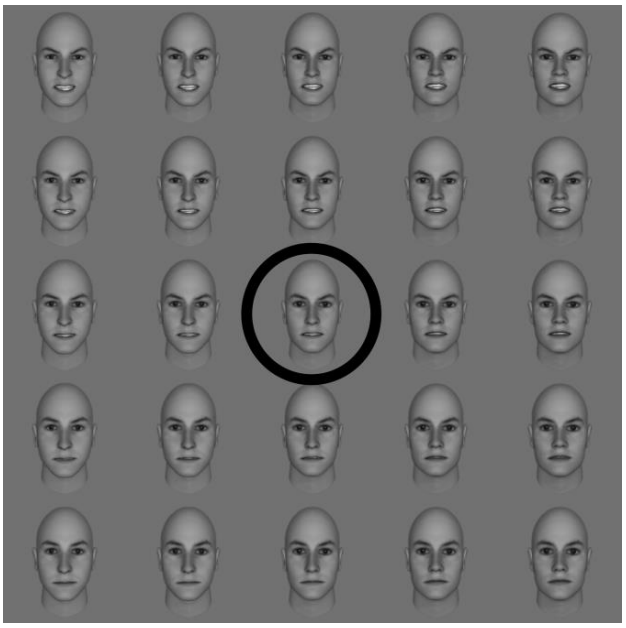

Figure 1: Stimulus space, angry version. CS+ indicated by black circle.

Conditioning paradigm

In this study, we use a conditioning paradigm that consists of 20 microblocks that each consist of all 25 stimuli. The CS+ is shown twice and reinforced once per microblock with a small electric shock. Microblocks are arranged in mesoblocks of five microblocks each. We collect shock expectation ratings before conditioning and after every mesoblock, i.e. five in total.

Results

Model predictions

In line with Dunsmoor et al. (2009) we assume that prior knowledge on the fear relevance of dimensions has a decisive impact. E.g. angry faces are considered more likely to predict a negative outcome then neutral ones while happy faces might be considered a safety signal a priori. To capture this in the model, we assume an informed prior on the midpoint parameter for the respective dimension. This results in a generalization gradient that is heavily skewed in the emotional dimension. Given these assumptions, the observations become more and more unlikely with increased number of trials. Thus, we expect this gradient to become increasingly specific and to gravitate towards the CS+ over time. The visualized predictions for these assumptions about both stimulus sets can be found in figure 2.

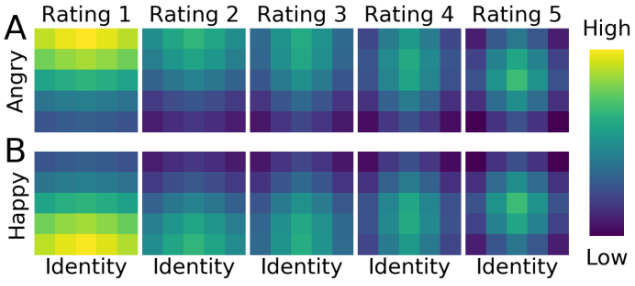

Figure 2: Model predictions for shock expectation ratings using the (A) angry and (B) happy face stimuli.

Ratings

The averaged shock expectancy ratings over subjects for the different ratings can be seen in figure 3.

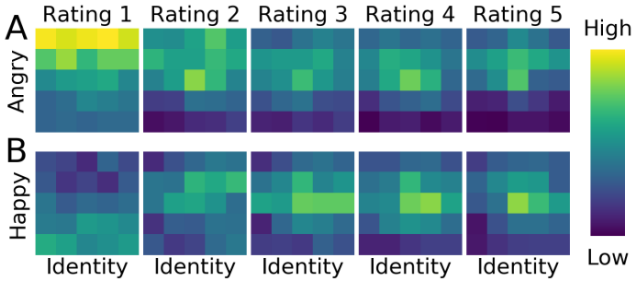

Figure 3: Averaged shock expectation ratings at different points in time for the (A) angry and (B) happy face stimuli.

Angry face stimuli As expected, subjects report higher shock expectation for angrier looking faces without an impact of identity before conditioning. With increasing amounts of information this prior assumption is overwritten due to the increasingly smaller likelihood of the observations under these assumptions.

Happy face stimuli Subjects initially report lower shock expectation for happy than neutral faces. From the second rating onwards, the generalization gradient is more or less centered on the CS+ and becomes more specific over time.

Discussion

The results from the angry face condition are well in line with our expectation. The shock expectation ratings are compatible with the idea of Bayesian integration of the assumption that angry faces are more likely to be predictive of an aversive outcome with counterfactual evidence from the conditioning. Before observing data the gradient relies entirely on prior knowledge. Since the observations become increasingly unlikely with more trials, the posterior distribution in parameter space shifts towards an associative map that is centered on the CS+.

Results from the happy condition show a different picture. Again, in the beginning there is a gradient in the expected direction. However, since it's not quite as strong, the prior on the midpoint parameter is quickly overpowered by the observations' likelihood. The generalization gradient becomes increasingly steep as the observations are more likely under this assumption.

In summary, the model clearly captures the characteristics of shock expectancy ratings in the angry condition and is in line with the results from the happy condition.

References

Dunsmoor, J. E., Mitroff, S. R., & LaBar, K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, *16*(7), 460–469. https://doi.org/10.1101/lm.1431609

Onat, S., & Büchel, C. (2015). The neuronal basis of fear generalization in humans. *Nature Neuroscience*, *18*(12), 1811–1818. https://doi.org/10.1038/nn.4166

Shepard, R. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*, 1317–1323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(04). https://doi.org/10.1017/S0140525X01000061