Negative Evidence and Inductive Reasoning in Generalization of Associative Learning

Jessica C. Lee

Peter. F. Lovibond

Brett. K. Hayes

Danielle J. Navarro

University of New South Wales

Corresponding author: Jessica Lee (jessica.lee@unsw.edu.au)

Author Note:
We would like to thank Alexis Porter for assistance with data collection. This study was funded by an Australian Research Council Discovery Grant (DP160101907) awarded to Peter Lovibond and Brett Hayes.
The data and ideas have been presented at the 2018 International Psychonomics Meeting, at the 2017 Australian Learning Group Midyear Meeting, and at the 2016 Australian Learning Group Christmas Workshop. The data and ideas have not previously been published elsewhere.

Abstract

When generalizing properties from known to novel instances, both positive evidence (instances known to possess a property) and negative evidence (instances known not to possess a property) must be integrated. The current study compared generalization based on positive evidence alone against a mixture of positive evidence and perceptually dissimilar negative evidence in an interdimensional discrimination procedure. In two experiments, we compared generalization following training with a single positive stimulus (that predicted shock) against groups where an additional negative stimulus (that did not predict shock) was presented in a causal judgement (Experiment 1) and a fear conditioning (Experiment 2) procedure. In contrast to animal conditioning studies, we found that adding a “distant” negative stimulus resulted in an overall increase in generalization to stimuli varying on the dimension of the positive stimulus, consistent with the inductive reasoning literature. We show that this key qualitative result can be simulated by a Bayesian model that incorporates helpful sampling assumptions. Our results suggest that similar processes underlie generalization in inductive reasoning and associative learning tasks.

Keywords: generalization, negative evidence, inductive reasoning, interdimensional discrimination, Bayesian model

Negative Evidence and Inductive Reasoning in Generalization of Associative Learning

Generalization is ubiquitous in everyday life. Being able to use previous experiences to guide behavior and judgements in novel situations is critical since stimuli that we encounter are always novel in some way. For example, traffic lights that signal when to stop and go differ in appearance between countries, and a person we have just met might behave differently the next time we see them in a different context. Often, we need to make judgements about whether a given property will apply from one instance to another (e.g., whether a novel green light signals “go”), given what we know about other familiar instances (i.e., other green lights). In this sense, generalization can be conceptualized as an inductive inference that is made on the basis of available evidence. Experiences with stimuli and their consequences or properties can be thought of as evidence, and can consist of both positive evidence (stimuli that cause an outcome or possess a property) or negative evidence (stimuli that do not cause an outcome or do not possess a property).

Our goal in this paper is to explore how we generalize when confronted with positive and negative evidence that are *dissimilar*. We will explore this question using two different learning procedures (causal judgement and fear conditioning), and from two different psychological perspectives – associative learning and inductive reasoning. Despite differences in experimental procedures, nomenclature, and explanatory constructs in these two literatures, we will argue that there are important parallels between the way generalization is conceptualized in each domain (see also Dunsmoor & Murphy, 2015).

Bridging these two research traditions through the common theme of generalization has theoretical value. Firstly, although the role of reasoning and rules is often acknowledged in associative generalization (see McLaren, Forrest, McLaren, Aitken, & Mackintosh, 2014; Mitchell, De Houwer, & Lovibond, 2009), there have been few attempts to specify the processes and computations that underlie such reasoning (but see Soto, Gershman, & Niv, 2014; Tenenbaum & Griffiths, 2001). In this respect, associative learning would benefit from utilizing theories and insights from the inductive reasoning literature. Secondly, much of the research on generalization in associative learning has been carried out with non-human animals. Differences in behaviour between humans and animals are interesting from a comparative perspective, and have implications for specifying how associative and reasoning processes might interact in dual-process models of learning (see McLaren et al., 2014).

There are also important clinical implications from investigating the processes underlying generalization. Over-generalization has been implicated as a contributing factor to the maintenance of anxiety disorders such as Post-Traumatic Stress Disorder (PTSD; Morey et al., 2015), panic disorder (Lissek et al., 2010) and Generalized Anxiety Disorder (GAD; Lissek et al., 2014; see Lissek ,2012, for a review). Notably, over-generalization of fear often involves conceptual relations between stimuli. For example, patients with PTSD typically experience symptoms in response not only to stimuli that are physically similar to the circumstances of their original trauma, but also to stimuli that are conceptually related (e.g., that share themes of abandonment or blame; Ehlers & Clark, 2000). Studying the conditions under which we generalize adaptively is likely to enhance our understanding of the processes that occur when generalization becomes maladaptive. In the next section, we provide a brief review of the role of negative evidence in inductive reasoning, and highlight the similarities between property induction in reasoning and stimulus generalization in associative learning.

Negative Evidence in Inductive Reasoning

A typical inductive reasoning argument involves presenting participants with premises of an argument, and asking participants to either judge the likelihood of the conclusion being true, or provide an overall rating of argument strength. In a given argument (such as the following examples, adapted from Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015), participants are asked to assume that all the premises above the line are true:

Premise: Mozart’s music causes alpha waves in the brain. (1)

Conclusion: Nirvana’s music causes alpha waves in the brain.

In the above example, we might expect a moderate amount of generalization from the positive evidence to the conclusion category, since Mozart’s and Nirvana’s music are part of the same superordinate category “music”, but belong to different subcategories within music and thus are perceptually distinct. When there is a single positive premise, a large amount of empirical research (see Hayes & Heit, 2018; Heit, 2000, for reviews) has established that similarity between premise and conclusion categories is a key component in explaining generalization in the absence of any other information. Not surprisingly then, similarity features heavily in computational models of induction (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Shepard, 1987; Sloman, 1993), as well as early theories of stimulus generalization (Estes, 1950; Shepard, 1957).

While there has been much research on how people generalize from positive evidence in inductive reasoning, relatively little attention has been devoted to the effect of negative evidence (but see Heussen, Voorspoels, Verheyen, Storms, &

Hampton, 2011; Kalish & Lawson, 2007; Voorspoels et al., 2015). Exploring the role of negative evidence is paramount since models of inductive reasoning should be able to explain how negative evidence is integrated with positive evidence to produce generalization. Consider the following argument:

Premise: Mozart's music causes alpha waves in the brain. (2)
Premise: Metallica's music does NOT cause alpha waves in the brain.

Conclusion: Nirvana's music causes alpha waves in the brain.

The above argument should seem to be weaker than argument 1 with the addition of *close negative* evidence from a different subcategory as the positive evidence. Intuitively, Nirvana's music is more similar to Metallica's music than Mozart's, and thus participants should show a *decrease* in generalization from the positive evidence to the conclusion category, relative to when only the positive evidence was present. Now consider the argument:

Premise: Mozart's music causes alpha waves in the brain. (3)
Premise: The sound of falling rocks does NOT cause alpha waves in the brain.

Conclusion: Nirvana's music causes alpha waves in the brain.

Note that now the negative evidence comes from the broader category of "sounds", and thus can be considered *distant negative* evidence. In contrast to arguments involving close negative evidence, arguments incorporating distant negative evidence have been found to *increase* generalization from the positive evidence to the conclusion category (Heussen et al., 2011; Voorspoels et al., 2015), relative to arguments containing the single premise of positive evidence. Thus, it appears that the direction of change in generalization (increasing or decreasing) following exposure to negative evidence depends on the *similarity* between the positive and negative premise.

The ability of distant negative evidence to increase generalization is particularly interesting because it challenges many existing models of inductive reasoning (e.g., Blok, Medin, & Osherson, 2007; Osherson, Stern, Wilkie, Stob, & Smith, 1990; Sloman, 1993). These models predict the principle of *monotonicity*, which refers to the tendency for generalization to the conclusion category to increase with the amount of positive evidence, and decrease with the amount of negative evidence in an argument (Heit, 2000; Osherson et al., 1991). The ability of distant negative evidence to increase generalization is a violation of this principle. Such violations however, can be explained by Bayesian models of induction (Heit, 1998; Kemp & Tenenbaum, 2009; Tenenbaum & Griffiths, 2001). This class of models conceptualize inductive reasoning as a form of belief updating. Participants are assumed to hold prior beliefs in hypotheses about the extension of the property in question, and update those hypotheses upon encountering positive and negative evidence (e.g., the premises in arguments like those shown earlier). Notably, many Bayesian models also fail to explain the non-monotonicity effect observed in Voorspoels et al. (2015). Critically, only models that incorporate a particular *sampling assumption* (i.e., beliefs about how the evidence was selected) are able to account for the ability of distant negative evidence to increase generalization.

Intuitively, considering Argument 3, if participants assume that the negative evidence was sampled *randomly*, then they might assume that the sound of falling rocks is so different from Mozart’s music that it offers little informational value in evaluating whether Nirvana’s music has the same property. In contrast, if participants assume that the negative evidence was chosen to be *helpful*, they might interpret it as highlighting the boundary of the property (i.e., music causes alpha waves). This would increase generalization to the conclusion category (Nirvana’s music) since it is also a type of music, causing a violation of monotonicity. This Bayesian approach has been very successful at providing formal accounts of generalization data in inductive reasoning tasks, and as we will demonstrate, can also be applied to associative learning tasks.

Interdimensional Discrimination in Associative Learning

Another reason that the effect of distant negative evidence is interesting is because it draws an intriguing parallel with generalization as assessed in the animal conditioning literature. Conditioning paradigms are typically conducted by pairing a conditioned stimulus (CS+) with an outcome (unconditioned stimulus, US) to establish a conditioned response (CR) to the CS+, or by rewarding an instrumental response (R) made in the presence of a stimulus (S+). Unlike a typical inductive reasoning task, but similar to some concept learning experiments (e.g., Lee & Livesey, 2017; Navarro, Dry, & Lee, 2012; Posner & Keele, 1968), generalization is assessed by varying a feature or dimension of the CS+ and assessing conditioned responding to each test stimulus in the absence of the outcome, producing a generalization *gradient* along the test dimension. A key difference between the learning and inductive reasoning literatures is that the bulk of generalization research in associative learning has used non-human animals such as pigeons. Hence, theories of discrimination and generalization have been formulated specifically to explain how different training procedures impact the shape of generalization gradients in animals (see Honig & Urcuioli, 1981; Gilbert & Sutherland, 1969; and Mackintosh, 1974, for reviews). Although much of the associative learning research on generalization has involved conditioning with pleasant or appetitive outcomes, there is good evidence that similar associative principles operate in the generalization of responses to an aversive US such as electric shock (e.g., Dunsmoor & LaBar, 2013) or unpleasant odors (e.g., Li, Howard, Parrish, & Gottfried, 2008).

The conditioning procedure of most relevance to the current study is *interdimensional discrimination*. Interdimensional discrimination involves training with a “positive” stimulus which is followed by an outcome (CS+) and a “negative” stimulus which is not followed by an outcome (CS-), where the CS+ and CS- lie on different dimensions (e.g., a tone as the CS+ and white noise as the CS-). Generalization of responding is tested along the dimension of the CS+ (e.g., by varying the frequency of the tone). The typical finding is that compared to “single cue” training where the animal learns an association between a single CS+ and an outcome (e.g., food), adding a CS- from another dimension sharpens the generalization gradient (Lyons, 1969; Switalski, Lyons, & Thomas, 1966; Thomas & Wheatley, 1974). For example, Switalski et al. trained pigeons to peck in response to a keylight of a particular color (e.g., green) by rewarding the pecking response with food, and then administered a generalization test where pecking responses were recorded to different keylights varying on the color dimension. This produced a peaked gradient with the highest level of pecking recorded at the trained CS+. Subsequently, the animals were given interdimensional discrimination training where pecking to the same green keylight seen in the first training phase (the CS+) was again followed by food, but pecking on trials with a black keylight with a superimposed white vertical line was never followed by food (the CS-). They were then administered another generalization test. Compared to the generalization gradient following single cue training, the gradient was more sharply peaked following interdimensional discrimination training, with less responding at the extreme ends of the dimension.

Associative explanations of this effect rely on the notion of competition amongst stimuli for a finite amount of associative strength with the outcome, and involve selective attention and error-correction mechanisms. Sutherland and Mackintosh's (1971, see also Sutherland, 1964) stimulus analyzer theory posits that during discrimination training, animals learn to attend to dimensions of the stimuli that are predictive of outcomes through a process of error-correction. The amount of attention directed at a particular dimension (the strength of the analyzer) governs the sharpness of the generalization gradient at test (with zero attention producing flat gradients). Sutherland and Mackintosh's account therefore predicts a sharper generalization gradient if the test dimension is a (somewhat) predictive feature during discrimination training.

A related idea proposed by Wagner (1969) posits that discrimination training serves to "neutralize" learning of incidental stimuli in the animal's environment (e.g., smell of the experimental chamber, location of the keylight) that might otherwise become associated with the outcome in single cue training due to their co-occurrence with the outcome. During discrimination training these incidental stimuli are equally correlated with the outcome and no outcome, and thus become relatively non-predictive compared to the stimuli of interest to the experimenter. The critical point is that the relative predictive validity of stimuli determines which stimuli animals learn about (Wagner, Logan, Haberlandt, & Price, 1968). Less attention to irrelevant features or stimuli increases the likelihood that animals will attend to, and learn more about the critical (predictive) features of interest to the experimenter, producing a sharper generalization gradient on test when the relevant dimension is varied.

It might be apparent that interdimensional discrimination training in conditioning bears some resemblance to the presentation of distant negative evidence in an induction task. Although the paradigms differ markedly, pairing a stimulus with an outcome (CS+) can be seen as presenting positive evidence, and pairing a stimulus with no outcome (CS-) can be seen as presenting negative evidence. Discriminating between a CS+ and a CS- on different dimensions is also similar to presenting negative evidence from a broader superordinate category to the positive evidence. In both cases, information from positive and negative evidence must be integrated in order to respond appropriately to a novel instance.

Notably, if we accept that these paradigms are somewhat analogous, then the inductive reasoning and animal conditioning literatures make opposing predictions about how generalization will be affected by *distant negative evidence*. On one hand, the property induction literature shows that relative to an argument with a single positive premise, presenting additional negative evidence from a distant category *increases* generalization of the target property to a conclusion from the same category (Heussen et al., 2011) or a close subcategory (Voorspoels et al., 2015). In contrast, in the associative learning literature the presentation of a "distant" CS- (i.e. a CS- on a different dimension in interdimensional discrimination) leads to a *decrease* in generalization to stimuli that are perceptually similar to the CS+ (Lyons, 1969; Switalski, Lyons, & Thomas, 1966; Thomas & Wheatley, 1974). While the effects of interdimensional discrimination in animals is well established, it is not known whether humans will display the same behavior in a similar task.

The main aim of the current experiments was to test whether interdimensional discrimination leads to an increase or a decrease in generalization to stimuli varying on the CS+ dimension when making predictions about a hypothetical shock outcome (Experiment 1) or a physical shock outcome (Experiment 2). This will reveal whether human generalization is consistent with associative theories and empirical results in non-human animals, or is better thought of as a process of generating an inductive hypothesis based on the positive and negative evidence that has been observed. To further examine this issue, we also explored whether our experimental data could be explained by a formal Bayesian reasoning model.

In our experiments, we compared generalization gradients between groups receiving different training procedures. All groups received training with the same CS+ (the positive stimulus), which was an aqua (bluey-green) colored rectangle, and all groups received the same generalization test phase where they were presented with rectangles varying on the color (restricted to blue-green) dimension (see Figure 1). The Single Positive group received single cue training with a single positive stimulus (CS+), while the Distant Negative received interdimensional discrimination training with an additional black and white checkered rectangle that served as the CS- (the negative stimulus, see Figure 1). The choice of the distant negative stimulus was driven by the requirement that it be part of the same superordinate category (rectangles), but outside the same perceptual subcategory as the CS+ and test stimuli (colored rectangles).

In Experiment 1, we also included a Close Negative group to verify that any group differences found were due to the distant negative stimulus and not the addition of *any* negative stimulus per se. This group can be seen as analogous to the addition of “close negative evidence” in the inductive reasoning literature (see Heussen et al., 2011; Voorspoels et al., 2015)[1]. Experiment 1 used a causal judgement scenario where participants made judgements about an outcome (shock) occurring to a hypothetical character, whereas Experiment 2 used a physical shock in a fear conditioning paradigm.

Experiment 1

Experiment 1 examined the effect of presenting a distant negative stimulus on generalization from a positive stimulus in a causal judgement paradigm. The task was to predict the occurrence of shock delivered to a hypothetical "Mr. X". Training stimuli were presented to participants as symbols on a shock machine. Participants were told that these symbols would be predictive of shock and were instructed to use them to make their predictions. Three groups of participants (Single Positive, Close Negative, and Distant Negative) received 12 presentations with the same positive stimulus, which was an aqua-colored rectangle. The positive stimulus was followed by the shock outcome on 100% of trials[2]. The Single Positive group received training with only the positive stimulus, while the other groups also received additional (intermixed) training trials with a negative stimulus, which was never followed by the outcome. For the Close Negative group, the negative stimulus was a slightly bluer or slightly greener rectangle, with the relational difference counterbalanced. In the Distant Negative group, the negative stimulus was a black and white checkered rectangle (see Figure 1).

Method

Participants

One hundred and forty (99 female, *M* age = 19.9, *SD* age = 3.65) University of New South Wales students participated in exchange for partial course credit. Participants were randomly allocated to the Single Positive group (*n* = 47), Close Negative group (*n* = 47), or Distant Negative group (*n* = 46). Recruitment continued until there were at least 40 participants in each group after exclusions.

Apparatus

The experiment was programmed using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and run using Matlab on standard PC computers connected to a 23-inch monitor. Participants made responses using a standard PC keyboard and mouse in individual cubicles.

Stimuli

Figure 1 shows the stimuli used in the training and test phases of Experiment 1. There were 11 stimuli (S1-S11) in total that varied along the hue dimension, with the stimulus at the midpoint (stimulus 6, S6) representing the CS+ for all three groups. The minimum hue value (HSV values) was .396 and the maximum was .583, with saturation and brightness held constant at 1 and .75 respectively. The size and shape of the rectangles were also kept constant at 180 pixels in height and 360 pixels in width for all stimuli presented during training. The direction of the dimension (green to blue or blue to green) was counterbalanced between participants such that S1 or S11 could be the greenest (or bluest) stimulus. The CS+ (S6) was an aqua (HSV values .489, 1, .75) rectangle. The CS- in the Close Negative group consisted of the stimulus at S4 on the dimension that was either slightly greener or slightly bluer, depending on counterbalancing. The CS- in the Distant Negative group was a black and white checkered rectangle.

Procedure

The procedure was approved by the University of New South Wales Human Research Advisory Panel. The experiment was composed of a training phase, test phase, and questionnaire. Participants were told a cover story where their task was to help a hypothetical “Mr. X” predict when shocks (the outcome) were going to occur based on a predictive symbol (the stimulus) presented on the shock machine.

On each training trial, participants were presented with a stimulus inside a black 600x600 pixel box along with the words “The following symbol appears on the machine.” After 1s, the question “How likely do you think it is that a shock will be delivered to Mr. X?” appeared along with a visual analogue rating scale. The scale ranged from “Certain NO shock” to “Certain shock” and the endpoints were marked with ticks. All ratings were converted to range between 0 and 100 for analysis. Participants used the mouse to click on any point on the scale, and could change their rating. After participants were happy with their rating, they pressed spacebar to continue to the next trial. Each stimulus was presented 12 times which meant that there were 12 training trials in the Single Positive group, and 24 trials in the Close Negative and Distant Negative groups. Trials were randomized with the constraint that the same stimulus could not appear more than twice in a row for the Close Negative and Distant Negative groups. The CS+ was always followed by shock and the CS-, if present, was never followed by shock.

After the training phase, participants were told that they would no longer receive feedback, but that they should continue making predictions about shock based on the symbols presented. The test phase consisted of presentation of the 11 test stimuli varying the color dimension in randomized order and holding the shape constant at the CS+ value, followed by presentation of 11 test stimuli varying the shape dimension in randomized order, holding the color constant at the CS+ value.

Following completion of the test phase participants progressed to a written questionnaire where they were asked to report any rules or hypotheses they had derived concerning the features of the stimuli that predicted shock or the relationship between the stimuli and shock. The questionnaire consisted of both free-report and forced-choice questions probing these rules, as well as a forced-choice question asking participants how they thought the stimuli were selected[3].

Results and Discussion

Exclusion Criteria

Participants were excluded from the analyses if they indicated they were colorblind (*n* = 3). Participants were also excluded if their average causal judgement for the CS+ in the final block of training (i.e. last four trials) was < 80 or if their average causal judgement for the CS- in the final block of training was > 20. After applying this criterion 130 participants remained (44 in the Single Positive group, 42 in the Close Negative group, and 44 in the Distant Negative group).

Data Analysis

Since our aim was to investigate the effect of adding negative evidence, we treated the generalization gradient in the Single Positive group as a baseline, and compared the Close Negative and Distant Negative groups to the Single Positive group separately. For the test stimuli, we looked at overall linear and quadratic trends in responses to stimuli at different points on the blue-green dimension, and tested contrasts if there appeared to be differential group differences on the left- and right-hand side of the CS+. To control the family-wise error rate at .05, we used Holm-Bonferroni corrections to the critical alpha value (this did not change any of the reported results). Note that we will not present the results of the shape generalization test, since this dimension is irrelevant for the discrimination and the color dimension was of primary interest. In general, we found slightly peaked gradients in all groups for the shape dimension (see Supplementary Materials for the shape generalization gradients). Note that these peaked gradients are in contrast to associative theory, which would predict flat generalization gradients for stimulus dimensions that are not predictive of the outcome. Data for the reported experiments in this paper and in Supplementary Materials are available at osf.io/htk6j.

Training

The training data are shown in Figure 2. All three groups showed evidence of learning, with participants increasing their causal judgements to the CS+ over training presentations, as reflected in a significant linear trend for the presentation factor, *F*(1,127) = 310.7, $p < .001$, $\eta_p^2 = .710$. Relative to the Single Positive group, the Close Negative group showed a significantly steeper learning curve for the CS+, *F*(1,84) = 5.05, $p = .027$, $\eta_p^2 = .057$, while the Distant Negative group did not, $F < 1$. The pattern of results regarding the linear trend in CS+ ratings reflects the flattening of the gradient after trial 2 for both the Single Positive and Distant Negative groups, but not in the Close Negative group. There were no group differences in accuracy on the final CS+ trial, *F*(2,127) = 1.14, $p = .322$, $\eta^2 = .018$.

Considering just the Close Negative and Distant Negative groups in Figure 2, there appears to be differential responding to the CS+ and CS-. This observation was confirmed by a significant difference in overall causal judgements to the CS+ and CS-, *F*(1,84) = 5603.7, $p < .001$, $\eta_p^2 = .985$, and this difference diverged over training trials, *F*(1,84) = 315.1, $p < .001$, $\eta_p^2 = .790$. This effect further interacted with group, *F*(1,84) = 10.6, $p = .002$, $\eta_p^2 = .112$. Thus, there was evidence that participants learned to differentiate between positive and negative stimuli, but there was faster learning in the Distant Negative group. This is not surprising due to the positive and negative stimuli in the Close Negative group being more perceptually similar and therefore confusable.

Generalization Test

The results for the generalization test varying the color dimension are shown in Figure 3. The figure shows that the groups differed in the shape of the generalization gradient and in their absolute levels of causal ratings. The comparison of most interest was between the Single Positive and Distant Negative groups. We analyzed these data in a 2x(11) ANOVA with group as the between-subjects factor and stimulus (S1-S11) as the within-subjects factor, and conducted linear and quadratic trend analyses. Overall, there were significantly higher causal ratings for the test stimuli in the Distant Negative group than in the Single Positive group, $F(1,86) = 16.6$, $p < .001$, $\eta_p^2 = .162$. Thus, we can conclude that adding a distant negative stimulus increased generalization to perceptually similar stimuli along the color dimension. This result is consistent with the effect of adding distant negative evidence to a single positive premise in an inductive reasoning argument, which results in increased generalization to the conclusion category (Heussen et al., 2011; Kalish & Lawson, 2007; Voorspoels et al., 2015). For these two groups, there was an overall quadratic trend across the whole dimension, $F(1,86) = 110.8$, $p < .001$, $\eta_p^2 = .563$, that was significantly weaker in the Distant Negative group, $F(1,86) = 17.1$, $p < .001$, $\eta_p^2 = .166$. There was no significant overall linear trend, $F < 1$, and no interaction with group, $F < 1$. There was also no group difference for causal ratings given to the CS+, $F < 1$, meaning that the overall group difference could be interpreted as a difference in generalization and not in learning about the stimuli presented. Thus, not only did the distant negative stimulus increase generalization over the whole dimension, the gradient itself was also flatter.

Comparing the Close Negative group to the Single Positive group, there was a marginally non-significant effect of group, $F(1,84) = 3.92$, $p = .051$, $\eta_p^2 = .045$. This effect is probably due to the lower ratings seen in the Close Negative group on the left-hand side of the figure, but higher ratings on the right-hand side of the figure. These observations were confirmed statistically. The gradient on the left-hand side in the Close Negative group was significantly lower, $F(1,84) = 37.2$, $p < .001$, $\eta_p^2 = .307$, and steeper, $F(1,84) = 13.8$, $p < .001$, $\eta_p^2 = .141$, than the gradient in the Single Positive group. On the right-hand side, the Close Negative group showed higher causal ratings than the Single Positive group, $F(1,84) = 9.86$, $p = .002$, $\eta_p^2 = .105$, but there was no interaction with the linear trend, $F(1,84) = 2.76$, $p = .101$, $\eta_p^2 = .032$. There was an overall linear, $F(1,84) = 61.4$, $p < .001$, $\eta_p^2 = .422$, and quadratic trend across the dimension, $F(1,84) = 146.9$, $p < .001$, $\eta_p^2 = .636$. The linear trend interacted with group, $F(1,84) = 48.8$, $p < .001$, $\eta_p^2 = .367$, due to the negative stimulus in the Close Negative group steepening the gradient between the CS+ and CS-. The quadratic trend did not interact with group, $F < 1$. There was no significant difference between the Close Negative and Single Positive group in their ratings of the CS+, $F(1,84) = 1.51$, $p = .223$, $\eta^2 = .018$, again suggesting that any obtained group differences were not due to differences in learning about the CS+.

In summary, it is clear that “distant negative” evidence, implemented in the form of a visual stimulus on another dimension, increased generalization to stimuli within the same category (solid-colored rectangles). Although the paradigm was more similar to interdimensional discrimination training used in conditioning paradigms (e.g., Switalski & Lyons, 1966; Thomas & Wheatley, 1974), we observed a generalization pattern that was opposite to that seen in non-human animals. Rather than interdimensional discrimination sharpening the generalization gradient relative to non-discriminative single cue training, we found that in humans, interdimensional discrimination flattened the generalization gradient and increased generalization over the whole dimension. Overall, this is more consistent with the generalization patterns found in studies of property induction involving combinations of dissimilar positive and negative evidence (e.g., Voorspoels et al., 2015).

Adding a “close negative” stimulus in contrast, produced asymmetrical results on the gradient, with less generalization on the side containing the CS-, but more generalization on the opposite side. On the left-hand side of the gradient (the side containing the negative stimulus), the effect is analogous to the effect of adding close negative evidence in Argument 3 in decreasing generalization from the positive evidence to the conclusion category. In Argument 3, Metallica’s music (the negative evidence) is more similar to Nirvana’s music than Mozart’s music (the positive evidence), and therefore generalization from Mozart to Nirvana decreases relative to an argument with just the positive evidence. Likewise, stimuli that are perceptually similar to the CS- produce low causal ratings, but our results also show that stimuli that are more similar to the CS+ than the CS- show a similar elevation in generalization on the right-hand side to the Distant Negative group. In inductive reasoning terms, the close negative stimulus produced a monotonicity effect for one set of generalization stimuli, but non-monotonicity for another set of generalization stimuli. Whether the generalization stimuli are more similar to the positive or negative stimulus seems to determine the direction of this effect. Note however, that this result also constitutes an example of the “area shift” phenomenon commonly observed in animals and humans after intradimensional discrimination training (e.g., see Cheng & Spetch, 2002; Doll & Thomas, 1967), and is consistent with associative theories that posit an interaction of excitatory and inhibitory processes (Blough, 1975; Ghirlanda & Enquist, 1998; McLaren & Mackintosh, 2002; Spence, 1937).

A potential reason for the discrepancy between our results and those of animal conditioning studies is that the paradigm used in this experiment (causal judgement) differs from a traditional conditioning procedure in a number of important ways. For instance, there is no biologically significant outcome in a causal judgment paradigm, and our cover story meant that participants were making predictions about an outcome occurring to a hypothetical character, Mr. X. It might be that causal judgement scenarios encourage more deliberative thinking, and therefore encourage the use of reasoning processes in comparison to conditioning procedures. Despite the fact that many important conditioning effects replicate in human causal learning tasks, and that causal judgement paradigms are commonly assumed to tap into associative processes (Shanks, 2007), we considered it important to test whether the effect of distant negative evidence on generalization replicated in a conditioning paradigm with a biologically significant outcome. Thus, Experiment 2 was conducted using a fear conditioning procedure.

Experiment 2

The aim of Experiment 2 was to test the effect of adding a distant negative stimulus on generalization of fear conditioning. Surprisingly, there is a stark absence of human studies that have examined the effect of interdimensional discrimination on generalization in a conditioning paradigm (but see Struyf, Iberico, & Vervliet, 2014). Where possible, we minimized changes from Experiment 1. However, due to the inclusion of a physical shock, a few procedural details were changed to prevent habituation to shock and to account for the fact that the shock could not occur during the test phase. We omitted the Close Negative group from this experiment, since our primary aim was to explore the effect of distant negative evidence on generalization, and because the observed generalization gradients in this group in Experiment 1 were consistent with both associative learning and inductive reasoning approaches.

Method

Participants

Seventy-one first-year Psychology students (*M* age = 19.3, *SD* = 2.1) at the University of New South Wales participated for partial course credit. Participants were randomly allocated to either the Single Positive group or the Distant Negative group. Recruitment continued until there were 30 participants in each group after exclusions.

Apparatus

As in Experiment 1, the experiment was programmed and run in Matlab using the Psychophysics Toolbox (Brainard, 1999; Pelli, 1999). Shocks were delivered via stainless steel electrodes attached to the medial and distal segments of participants' index finger of the non-dominant hand. Skin conductance was recorded via electrodes attached to the distal segments of participants' middle and ring fingers of the same hand. AD instruments hardware and LabChart software were used to record skin conductance throughout the experiment. The electrodes were secured with Leukoplast tape and isotonic gel was applied to the skin conductance electrodes unless the initial reading was very high.

Participants made their expectancy ratings using a semicircular dial with a rotary pointer, which ranged from 0% (labelled "CERTAIN NO SHOCK" at approximately 0 degrees) to 100% (labelled "CERTAIN SHOCK" at approximately 180 degrees). 50% (labelled "equal chance of shock or no shock" was placed at the 90 degree mark on the dial, and there were tick marks placed at intervals of 10 along with the corresponding number. Just under 0 degrees was labelled “Off”. The dial was clamped to the table in front of participants’ dominant hand.

Procedure

The procedure was approved by the University of New South Wales Human Research Ethics Committee. Prior to beginning the experiment, participants underwent a shock work-up procedure to allow them to select an appropriate level of shock. It was emphasized that the participant’s chosen level of shock should feel “definitely uncomfortable, but not painful” and should be sufficient to make them anxious throughout the experiment. Participants completed the experiment in a dimly lit cubicle with the door closed while the experimenter monitored the experiment from outside. The instructions emphasized that participants were to use the symbols presented to them to learn to predict when shock was going to occur.

Training. The training phase consisted of four presentations of each stimulus, such that the maximum number of shocks received by either group was four. The number of training trials was reduced from Experiment 1 due to the need to minimize habituation to the shock. On each trial, after a 10s baseline period, a stimulus was presented at the center of the screen for 10s. During this presentation time, participants were asked to make a rating of how much they expected a shock to follow the stimulus on screen using the expectancy dial. After the stimulus presentation, there was a 2s feedback period where the stimulus disappeared and a shock of 0.5s duration was either delivered (if the stimulus was a CS+) or not delivered (if the stimulus was a CS-) at the beginning of this period. Presentation of shock was always accompanied by visual feedback consisting of the word “SHOCK!” as well as a picture of a lightning bolt. This meant that feedback was presented for 2s, or a blank screen would be presented for 2s if there was no shock on that trial. Following the feedback period, the message “Please turn the expectancy dial back to the “Off” position” was presented for 2s. The ITI was variable and ranged between 15-25s.

Expectancy Ratings Test. After the training phase, the program paused and the experimenter entered the cubicle and told the participant that due to ethical reasons, the number of shocks given to participants had to be limited, and thus the shock electrodes had to be disconnected for the next phase. This procedure was administered in order to equate the testing conditions to Experiment 1, where generalization was assessed in the absence of feedback about the shock outcome. Participants were to make hypothetical expectancy ratings of shock, imagining that the shock electrodes were still connected and that shock was still possible. They were told that there would be no feedback about whether shock was delivered or not. After checking for understanding of the instructions, the experimenter left the cubicle and resumed the experiment from outside. The expectancy ratings test consisted of the 11 test stimuli varying the hue dimension presented in randomized order.

Skin Conductance Test. Prior to the skin conductance test, the program paused and the experimenter returned to the cubicle and explained that in this phase, the shock electrodes would be reconnected, and that therefore shocks were now possible. Participants were told to continue making expectancy ratings as before. The skin conductance test consisted of three trials only: the CS+ (S6) and the two extreme test stimuli (S1 and S11), presented in randomized order, and there were no shocks presented. The skin conductance data did not produce any significant group differences in the test phase. Therefore they will not be presented here (see Supplementary Materials for the skin conductance data).

Results and Discussion

Exclusion Criteria

Participants who indicated that they were colorblind were excluded from analyses (*n* = 2). The same exclusion criterion in Experiment 1 was used, except that average ratings for the CS+ and CS- were calculated for the final two trials in each group. All participants had to give a causal rating > 80 for the CS+, and also give a causal rating < 20 for the CS- if they were in the Distant Negative group. Seven participants from the Distant Negative group, and four participants from the Single Positive group failed this criterion. After exclusions, 30 participants remained in both Single Positive and Distant Negative groups.

Training

The results from the training phase are shown in Figure 4. Expectancy ratings to the CS+ increased over the training trials, as reflected in the significant linear trend contrast, $F(1,58) = 495.1$, $p < .001$, $\eta_p^2 = .895$, and this trend did not differ between groups, $F(1,58) = 1.13$, $p = .292$, $\eta_p^2 = .019$. There was no difference between the groups for accuracy to the CS+ on the final trial, $F(1,58) = 1.87$, $p = .177$, $\eta^2 = .031$. In the Distant Negative group, participants gave significantly higher ratings to the CS+ than the CS-, $F(1,29) = 1451.8$, $p < .001$, $\eta_p^2 = .980$, and this difference diverged over training trials, $F(1,29) = 862.4$, $p < .001$, $\eta_p^2 = .967$. Thus, despite the reduced number of training trials relative to Experiment 1, both groups showed clear evidence of learning about the stimuli.

Generalization Test

The generalization gradients are shown in Figure 5. Despite the change in procedures from Experiment 1, it can be seen that the overall increase in ratings in the Distant Negative group has replicated. This observation was confirmed statistically, with a significant main effect of group, $F(1,58) = 6.81$, $p = .012$, $\eta_p^2 = .105$, indicating an increase in generalization in the Distant Negative group relative to the Single

Positive group. There was a significant quadratic trend across generalization stimuli, $F(1,58) = 72.4$, $p < .001$, $\eta_p^2 = .555$, but no linear trend, $F < 1$. Neither of these trends interacted with group, *F*s < 1, and there was no group difference in ratings given to the CS+, $F < 1$. Unlike Experiment 1, there was no significant group interaction with the quadratic trend. Thus, there was no evidence that the shape of the generalization gradient differed between groups.

Comparison of Figures 3 and 5 suggests that this difference between the results in the two studies primarily reflects differences in expectancies to test stimuli with the most extreme values on the color dimension (i.e. those that differed maximally from the CS+) in the distant negative conditions. In Experiment 1, expectances for the shock symbol remained relatively high in this condition (around 70 on the causal rating scale). In Experiment 2, the analogous expectancies for actual shock were lower, falling close to the mid-range of the expectancy scale. Given the procedural differences between the studies, it is difficult to be certain about the reason for this discrepancy. To speculate, this may have been due to the use of the expectancy dial in this experiment as opposed to the visual analogue scale displayed on the screen in Experiment 1. In particular, "50% chance of shock" is labelled on the dial but not on the visual analogue scale, and may have served as an anchor to which participants were drawn when rating stimuli at the extreme ends of the dimension (which they would have the highest amount of uncertainty about). Nevertheless, the overall increase in expectancy ratings can readily be interpreted as a "categorical" increase in generalization across the dimension, and is consistent with what was observed in Experiment 1.

Overall, we replicated the major results of Experiment 1, confirming that the effect of distant negative evidence in increasing generalization is not restricted to causal judgement paradigms with hypothetical outcomes. The use of an actual shock in a fear conditioning paradigm more closely approximates the conditioning studies performed in animals, and yet participants still showed generalization consistent with empirical results in inductive reasoning tasks. One limitation was that we obtained a null result on the skin conductance measure. It is possible to interpret this finding as demonstrating a dissociation between conditioned responses and conscious expectancy (e.g., see McAndrew, Jones, McLaren, & McLaren, 2012), where conditioned responding is attributed to an automatic associative mechanism and expectancy measures to higher-order reasoning mechanisms. However, since we obtained a null effect, a more likely explanation is that our skin conductance measure was not sufficiently sensitive. Skin conductance is highly variable between participants when there is a single test trial (Vervoort, Vervliet, Bennett, & Baeyens, 2014), and extinguishes quickly at test once the shock is no longer presented.

Nevertheless, we now have evidence across two learning procedures of an effect on outcome expectancy akin to the non-monotonicity effect reported by Voorspoels et al. (2015) in a property induction task. Although our results suggest that inductive reasoning processes are operating during associative learning, the question remains as to exactly how participants reason about negative evidence. Intuitively, experience with the CS+ might allow the learner to formulate a number of hypotheses about the categories of stimuli that cause shock. For example, it could be that just aqua rectangles cause shock, all colored rectangles cause shock, or all rectangles or all shapes cause shock. Participants in the Single Positive group have no further evidence and thus distribute their belief approximately equally amongst these hypotheses. Participants in the Distant Negative group who also learn that a checkered rectangle does not cause shock, can rule out inconsistent hypotheses (i.e. that all shapes or all rectangles cause shock), and redistribute their belief amongst the credible remaining hypotheses. However, participants seem to favor the largest or broadest non-falsified hypothesis (Voorspoels et al., 2015). In other words, after learning that a checkered rectangle does not lead to shock, they are more likely to believe that all colored rectangles lead to shock, rather than, that only the specific aqua-colored rectangle leads to shock. Interestingly, this bias towards the broader hypothesis is justified if participants are treating the negative stimulus as a "near-miss" exemplar given to them by a helpful teacher (Kalish & Lawson, 2007); in other words, if participants have a "helpful" sampling assumption. The role of the negative stimulus then, can be seen as directing attention to (what participants assume to be) the relevant dimension for generalization. This interpretation of the results is broadly consistent with the relevance framework of induction (Medin, Coley, Storms, & Hayes, 2003), which posits that negative evidence highlights the relevant category (or in this case, stimulus dimension) for induction. In the next section, we offer a more detailed discussion for why the effect occurs, and outline and test a formal Bayesian model.

A Bayesian Account of Reasoning and Associative Learning

Voorspoels et al. (2015) outlined a Bayesian computational model of how negative evidence is used in inductive reasoning. Bayesian approaches to inductive reasoning (Heit, 1998; Kemp & Tenenbaum, 2009; Tenenbaum & Griffiths, 2001) are probabilistic models that assume participants have prior beliefs in hypotheses and update those beliefs when presented with new information, using Bayes rule to do so. Importantly, this belief updating is governed by a likelihood function that describes a subjective assessment of the probability of encountering this information. Because of this, these models are sensitive to the *sampling assumptions* held by participants –

beliefs a learner has about how the information was selected. Sensitivity to sampling has been implicated in a variety of reasoning phenomena, including premise non-monotonicity effects (Ransom et al., 2016), the role of negative evidence (Voorspoels et al., 2015) and the value of evidential diversity (Hayes, Navarro, Stephens, Ransom & Dilevski, submitted), all of which have proved amenable to Bayesian modelling.

Given these successes, it is natural to ask if we can account for our data using a Bayesian model of associative learning. On the one hand, the structure of the task mirrors the reasoning problems from Voorspoels et al. (2015), suggesting that a Bayesian inductive generalization model is appropriate. On the other hand, our learning problem requires people to learn novel CS-US associations across trials and a plausible model should produce appropriate stimulus generalization gradients for the relevant stimulus dimension (color). Bayesian models for reinforcement learning (Courville Daw, & Touretzky, 2006; Gershman, 2015; Gershman & Niv, 2012) have also been very successful, but have a different structure to their reasoning counterparts. Accordingly, it is not obvious how these approaches might be integrated.

In this section we describe an initial attempt to adapt Bayesian accounts of inductive reasoning to an associative learning task (a detailed description is presented in Appendix A). We assume the learning goal is to infer an *association map*, $A(x)$, a function that assigns an associative strength between a stimulus x and the outcome o for every possible stimulus (Figure 6a). The learner places a prior $P(A)$ over possible association maps that imposes a smoothness constraint, such that stimuli near each other in psychological space are assumed to have similar association strengths, but still allows for some degree of “patchiness”, where stimuli can have idiosyncratic associations (Figure 6b). This approach is similar to the Bayesian associative learning model of Soto et al. (2014), where the learner can infer a number of “latent causes”, which correspond to a particular consequential region (i.e. stimuli which cause the outcome) in the stimulus space (i.e. association map). The smoothness bias produces generalization of learned associations: when trained on a specific CS+, the learner increases the association strength for that item, but smoothness ensures that association strengths for all nearby items in the stimulus space tend to rise with it. This formalism provides a Bayesian version of the “covering maps” used in classic connectionist models of learning and generalization (Kruschke, 1992), though in our formalism the maps need not be smooth, borrowing from classic models in image segmentation (Geman & Geman 1984, Mumford & Shah 1985). Formally, the posterior probability of an association map A is given by:

$$P(A\ |x, o)\ \propto P(x, o|A)P(A)$$

where the likelihood function $P(x, o|A)$ describes the probability of the observations on the current trial if A is the correct map.

The critical property of the Bayesian model lies in the fact that the likelihood factorizes into two terms, $P(x, o|A) = P(o|x, A)\ P(x|A)$, the first of which is a prediction accuracy term and the second of which describes the sampling assumption. The prediction accuracy term $P(o|x, A)$ describes the probability of the *outcome* *o* when presented with stimulus x, if association map A is correct: maps that make better predictions are assigned higher posterior probability. Bayesian framing notwithstanding, this term plays a very similar role to the prediction error term in standard reinforcement learning models (e.g., Sutton & Barto, 1990).

The key component of the model is the sampling assumption, $P(x|A)$, which describes the learner’s belief about the probability they would have been presented with *stimulus* x if association map A were correct. Under a *weak sampling* assumption the learner assumes the stimulus items were chosen randomly (Figure 6c), and so $P(x|A)$ is a constant value and has no effect on learning. As shown on the right of Figure 6e, the Bayesian model produces exactly the same generalization behaviour in the Single Positive group as in the Distant Negative group, and produces a slight decrease in generalization for the Close Negative group.

Alternatively, if human learners have beliefs about the way in which other humans (including experimenters) choose to communicate concepts, they may adopt a *helpful sampling* assumption (Figure 6d). Under these assumptions, the stimuli are assumed to be chosen by the experimenter in order to teach a particular concept. Previous Bayesian accounts of helpful sampling (e.g., Shafto, Goodman, & Griffiths, 2014; Goodman & Frank, 2016; Voorspoels et al., 2015) have sought to derive the rules for helpful evidence selection from general communicative principles (Grice, 1975), but for the current purposes we assume that a helpful teacher prefers to (a) select training items that vary on a relevant (i.e., diagnostic) stimulus dimension, and (b) chooses the relevant features of a stimulus to avoid ambiguous items (i.e., prefers association strengths near 0 or 1). As shown on the left of Figure 6e, this version of the model produces generalization behavior that is very similar to human performance in the task. While the model itself is somewhat complex, the explanation for its behavior is simple. In the distant CS- condition, the model assumes that texture (i.e., solid color vs. checkeredness) is the relevant basis for induction, and hence that the specific hue of the CS+ is not as important, leading to wider generalization.

Although the qualitative predictions of the model under helpful sampling assumptions match the higher overall gradient observed in the Distant Negative group, there were more subtle model predictions that were not observed. For example, for the helpful sampling condition, the model predicts a small difference between the

Distant Negative and Close Negative groups on the right side of the gradient, but this effect does not appear to be present in the data. Further studies are needed to test the source of this discrepancy, as well as to formally test quantitative model predictions against the data. Nevertheless, it is clear that a Bayesian model incorporating helpful sampling can account for the key empirical finding of a higher gradient in the Distant Negative group.

General Discussion

Across two experiments using a causal judgement paradigm (Experiment 1) and a fear conditioning paradigm (Experiment 2), we found that discrimination training between a positive stimulus (CS+, an aqua rectangle which predicted shock) and a dissimilar "distant" negative stimulus (CS-, a checkered rectangle which did not predict shock) increased generalization across the color (blue-green) dimension, relative to a group who received training with only the single positive stimulus. We proposed that the interdimensional discrimination procedure, whereby participants are required to discriminate between a CS+ and a CS- on different dimensions, was somewhat analogous to presenting "distant" negative evidence from a broader superordinate category to the positive evidence in an inductive reasoning argument (cf. Voorspoels et al., 2015).

Notably, associative learning and inductive reasoning accounts make divergent predictions about the effects of adding negative evidence. Our results were broadly in line with non-monotonicity effects found in previous inductive reasoning studies (Heussen et al., 2011; Kalish & Lawson, 2007; Voorspoels et al., 2015), and inconsistent with the empirical data observed in animal conditioning studies (Lyons, 1969; Switalski, Lyons, & Thomas, 1966; Thomas & Wheatley, 1974). Moreover, similarly to Voorspoels et al. (2015), a novel Bayesian model of associative learning successfully simulated our results under helpful sampling assumptions.

Implications of the Model

In one respect the success of the Bayesian model might be considered unsurprising, as no individual component of the model is novel and already has some degree of support in the empirical literature. On the other hand, it seems remarkable – to us, at least – that a model pieced together from such disparate parts could work as well as it does. The core associative learning mechanism is similar in spirit to other probabilistic models of human reinforcement learning (Courville et al., 2006; Gershman, 2015; Gershman & Niv, 2012), but the stimulus generalization component is more closely tied to concept learning models (Kruschke, 1992; Shepard, 1987; Tenenbaum & Griffiths, 2001), and the higher order reasoning component is built from Bayesian models of pragmatic inference (Goodman & Frank, 2016; Shafto, Goodman, & Griffiths, 2014; Voorspoels et al., 2015).

Nevertheless, despite its eclectic origins, the model behaves in a very sensible way. When a simpler, “weak sampling” model is used to drive learning, our Bayesian model behaves in a manner that is at least somewhat reminiscent of other associative learning models. Human learning appears to be somewhat richer: the fact that helpful sampling best simulated our results suggests that the experimental context in an associative learning task might be interpreted pedagogically by the participant, in which the participant has expectations for what kind of evidence the experimenter should present to them if they expect them to learn (Shafto, Goodman, & Griffiths, 2014). As mentioned above, a helpful sampling assumption justifies why participants might see the negative evidence as marking the category boundary of the stimuli which cause shock, with the intuition being that a helpful teacher/experimenter would have chosen different and appropriate stimuli to mark different boundaries.

That said, it is an open question how a learner might solve the "communicative intention" problem. Some models of this process (e.g., Shafto et al., 2014; Voorspoels et al., 2015) assume a "fully recursive" theory of mind where teacher and learner behaviour mutually constrain one another; whereas other models rely on limited recursion (e.g., Goodman & Frank 2016; Ransom, Voorspoels, Perfors, & Navarro, 2017) and others avoid recursive inference entirely (e.g., Tenenbaum & Griffiths 2001; Hayes et al., submitted). These approaches would be difficult to distinguish within the experimental paradigm adopted here, but this would be a worthwhile line of future investigation. It is also important to note that a full exploration of the model parameter space is currently lacking and although outside the scope of the current study, would be a worthwhile endeavour to test the generality of the model.

Associative Learning and Reasoning

Our results suggest that similar processes underlie property generalization in inductive reasoning, and stimulus generalization in associative learning. Despite differences in the stimuli (written statements about categories vs. visual stimuli), presentation of evidence (premises in an argument vs. trial-by-trial pairings), presentation of the generalization test (as the conclusion of an argument vs. in a separate test phase), and the nature of the property to be generalized (hypothetical property vs. causal relationship to shock), the processes assumed to operate in inductive reasoning tasks also appear to operate in associative learning tasks. This outcome is surprising because our paradigm was drawn directly from the associative learning literature, and the relevant theories (Sutherland & Mackintosh, 1971;

Wagner, 1969) and empirical data (Lyons, 1969; Switalski, Lyons, & Thomas, 1966; Thomas & Wheatley, 1974) make a very clear prediction in the opposite direction. Interestingly, the only other study we are aware of comparing generalization following interdimensional discrimination with single cue training in humans (Baron, 1973) also found a somewhat broader and higher gradient, although an interpretation of this result was not given. The task itself was also quite different, requiring identification of a target stimulus rather than outcome predictions. It thus appears that the effect of interdimensional discrimination training on generalization can differ between animals and humans.

It is useful to consider why we might obtain different patterns of generalization in humans, given that associative theories of stimulus generalization are often assumed to be universal, applying to a range of stimuli and experimental procedures (Sutherland & Mackintosh, 1971; Wagner, 1969). It may be that despite attempting to approximate the conditioning procedures used in animals, the experimental context simply does not facilitate the operation of selective attention and associative mechanisms in humans. The associative mechanisms proposed to explain generalization following conditioning are couched in terms of cue competition, with the animal assumed to attend to, and learn about, the most predictive features out of the stimulus dimensions (Sutherland & Mackintosh, 1971) and incidental cues in the animal’s environment (Wagner, 1969). In our experiments, although there were incidental stimuli present in the experimental environment (e.g., other objects in the testing room, the dim lighting etc.), it seems unlikely that the participant would consider them as being related to the outcome in any way. Therefore, interdimensional discrimination fails to neutralize learning about these stimuli because participants do not learn about these incidental stimuli even in single cue training. It is also possible that selective attention mechanisms do not operate unless the stimuli are complex (e.g., containing a large number of features) and require effort or time to process (c.f. Sutherland and Mackintosh, 1971). Our stimuli were quite simple and contained few dimensions. Participants saw a very small number of stimuli (1 or 2) and could easily attend to all dimensions of the stimuli at once and encode them in a simple verbal description (e.g., “aqua rectangle”). The experimental situation is thus quite different from that of animals undergoing single cue or discrimination training where a large number of stimuli potentially compete for associative strength.

Our results highlight general issues with conducting associative learning research in humans using “abstract” or simple perceptual stimuli. A noteworthy aspect of our results is that unlike the stimuli used in inductive reasoning tasks that often belong to categories and have clear hierarchical taxonomy (e.g., animals, music), our stimuli were simple perceptual stimuli. The use of such stimuli is pervasive throughout the associative learning literature, and the implicit assumption is that these abstract stimuli minimize the influence of semantic or categorical knowledge such that generalization can be assessed on purely perceptual grounds. However, the overall increase in causal judgements and shock expectancy ratings found in our Distant Negative groups can be interpreted as a “categorical” increase in belief that colored rectangles lead to shock. This interpretation implies that participants perceive even very simple perceptual stimuli (colored rectangles) as belonging to a hierarchy with larger categories of stimuli (e.g., all rectangles, all shapes etc.).

One implication is that although the stimuli typically used in generalization studies are designed to be free of semantic knowledge, in reality, participants may consider them as part of broader categories and this can influence their generalization. In our task, it is possible that presenting a checkered rectangle highlights the categorical difference between patterns and colored stimuli, and this is responsible for the overall increase in generalization across the color dimension. Presenting a CS- that is also a colored rectangle may not lead participants to think of the category “colored rectangles”, since there is no contrast category to compare it against. Note also that our use of simple stimuli and continuous reinforcement may explain why in Experiment 1, we did not obtain the peak shift effect in the Close Negative group, despite this discrimination procedure being quite similar to previous studies that have found peak shift (e.g., Cheng & Spetch, 2002; Struyf et al, 2014; but see Lee, Hayes, & Lovibond, in press, for an alternative explanation).

Conclusion

In conclusion, we have demonstrated that discrimination training between a positive (CS+) and a distant negative (CS-) stimulus on another dimension is analogous to the presentation of distant negative evidence in inductive reasoning arguments in increasing generalization to similar stimuli. This result is inconsistent with the literature on interdimensional discrimination training in animal generalization studies, and suggests that the associative mechanisms formulated to explain animal generalization do not always operate in human associative learning tasks. Rather, the results suggest that participants are actively reasoning about hypotheses in a similar way to when they are evaluating inductive arguments. We successfully simulated the empirical results using an adapted Bayesian model (Tenenbaum & Griffiths, 2001) incorporating a helpful sampling assumption.

Like a growing number of other researchers (e.g., De Houwer, Beckers & Vandorpe, 2005; Dunsmoor and Murphy, 2015), we believe that associative and cognitive approaches to learning share some important commonalities and that insights from one field can be used to inform the other. Although our tasks and stimuli were inspired by associative approaches, we found that the data were more consistent with theoretical accounts developed in the domain of inductive reasoning. Integrating cognitive and associative approaches can produce a more nuanced understanding of the mechanisms by which people generalize their learning.

Footnotes

1. Note that this condition is the same as the intradimensional discrimination procedure that generates the peak shift phenomenon (Hanson, 1959; see Purtle, 1973, for a review).
2. See supplementary materials (Appendix B) for an additional experiment with partial (75%) reinforcement that showed a similar pattern of results.
3. In this experiment, as well as in Experiment 2, dividing participants according to self-reported rules did not change the pattern of group differences and thus we will not report the results from the questionnaire here (see Appendix E and F in supplementary materials).

References

Baron, A. (1973). Postdiscrimination gradients of human subjects on a tone continuum. *Journal of Experimental Psychology, 101,* 337–342.

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007). Induction as conditional probability judgment. *Memory & Cognition, 35*, 1353-1364.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 104,* 3-21.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433-436.

Cheng, K. & Spetch, M. L. (2002). Spatial generalization and peak shift in humans. *Learning and Motivation, 33*, 358–389.

Cross, G. R. & Jain, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 25-39.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294–300.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. *Tutorials in Quantitative Methods for Psychology, 1*, 42-45.

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior, 33*, 239-249.

Doll, T. J. & Thomas, D. R. (1967). Effects of discrimination training on stimulus generalization for human subjects. *Journal of Experimental Psychology*, *75,* 508–512.

Dunsmoor, J. E., & LaBar, K. S. (2013). Effects of discrimination training on fear generalization gradients and perceptual classification in humans, *Behavioral Neuroscience, 127,* 350-356.

Dunsmoor, J. E. & Murphy, G. L. (2015). Categories, concepts, and conditioning: How humans generalize fear. *Trends in Cognitive Sciences, 19*, 73–77.

Ehlers, A., & Clark, D. (2000). A cognitive model of Posttraumatic Stress Disorder. *Behavior Research & Therapy*, *38,* 319-345.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57,* 94–107.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence, 6*(6), 721.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLOS Computational Biology, 11,* e1004567.

Gershman, S. J. & Niv, Y (2012). Exploring a latent cause model of classical conditioning. *Learning & Behavior, 40*, 255-268.

Goodman, N. D. & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*, 818-829.

Ghirlanda, S. & Enquist, M. (1998). Artificial neural networks as models of stimulus control. *Animal Behaviour, 56*, 1383–1389.

Gilbert, R. M. & Sutherland, N. S. (1969). *Animal discrimination learning*. London, England: Academic Press.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. L. (ed.) *Syntax and semantics: Speech acts*, 3, 41–58. New York: Academic Press.

Hanson, H. M. (1959). Discrimination training effect on stimulus generalization gradient for spectrum stimuli. *Science, 125*(3253), 888–889.

Hayes, B. K., Navarro, D. J., Stephens, R., Ransom, K. & Dilevski, N. (submitted). The diversity effect in inductive reasoning depends on strong sampling. *Manuscrupt submitted for publication*.

Hayes, B. K. & Heit, E. (2018). Inductive Reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science, 9*, e1459

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaskford & N. Chater (Eds.), *Rational models of cognition* (pp. 248-274). Oxford, England: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569–592.

Heussen, D., Voorspoels, W., Verheyen, S., Storms, G., & Hampton, J. A. (2011). Raising argument strength using negative evidence: A constraint on models of induction. *Memory & Cognition, 39*, 1496-1507.

Honig, W. K. & Urcuioli, P. J. (1981). The legacy of Guttman and Kalish (1956): Twenty-five years of research on stimulus generalization. *Journal of the Experimental Analysis of Behavior, 36*, 405–445.

Kalish, C. W. & Lawson, C. A. (2007). Negative evidence and inductive generalisation. *Thinking & Reasoning, 13*, 394–425.

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22.

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (forthcoming). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Lee, J. C. & Livesey, E. J. (2017). The effect of encoding conditions on learning in the prototype distortion task. *Learning & Behavior*, *45*, 164–183.

Li, S. Z. (2009). *Markov random field modeling in image analysis*. Springer Science & Business Media.

Li, W., Howard, J. D., Parrish, T. B., & Gottfried, J. A. (2008). Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science, 319(5871),* 1842–1845.

Lissek, S. (2012). Toward an account of clinical anxiety predicated on basic, neutrally mapped mechanisms of Pavlovian fear-learning: The case for conditioned overgeneralization. *Depression and Anxiety, 29*, 257-263.

Lissek, S., Kaczkurkin, A. N., Rabin, S., Geraci, M., Pine, D. S., & Grillon, C. (2014). Generalized Anxiety Disorder is associated with overgeneralization of classical conditioned fear. *Biological Psychiatry, 75*, 909-915.

Lissek, S., Rabin, S., Heller, R. E., Lukenbaugh, D., Geraci, M., Pine, D. S., & Grillon, C. (2010). Overgeneralization of conditioned fear as a pathogenic marker of Panic Disorder. *The American Journal of Psychiatry, 167*, 47-55.

Lyons, J. (1969). Stimulus generalization along the dimension of S+ as a function of discrimination learning with and without error. *Journal of Experimental Psychology*, *81*, 95–100.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London, England: Academic Press.

McAndrew, A., Jones, F. W., McLaren, R. P., & McLaren, I. P. L. (2012). Dissociating expectancy of shock and changes in skin conductance: An investigation of the Perruchet Effect using an electrodermal paradigm. Journal of Experimental Psychology: *Animal Behavior Processes, 38*, 203-208.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*, 177–200.

McLaren, I. P. L., Forrest, C. L. D., McLaren, R. P., Aitken, M. R. F., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account. *Neurobiology of Learning and Memory, 108*, 185-195.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*, 517–532.

Morey, R. A., Dunsmoor, J. E., Haswell, C. C., Brown, V. M., Vora, A., Weiner, J., Stjepanovic, D., Wagner III, H. R., Mid-Atlantic MIRECC Workgroup, & LaBar, K. S. (2015). Fear learning circuitry is biased toward generalization of fear associations in posttraumatic stress disorder. *Translational Psychiatry, 5*, e700.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64.

Mumford, D., & Shah, J. (1985). Boundary detection by minimizing functionals. In *IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 17, pp. 137-154).

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science, 36*, 187–223.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science, 15*, 251–269.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *The Pavlovian Journal of Biological Science*.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Purtle, R. B. (1973). Peak shift: A review. *Psychological Bulletin, 80*, 408–421.

Ransom, K., Voorspoels, W., Perfors, A., & Navarro, D. J. (2017). A cognitive analysis of deception without lying. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* (pp. 992-997). Cognitive Science Society.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *The Quarterly Journal of Experimental Psychology, 60*, 291–309.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

Shepard, R. N. (1987). Toward a universal law generalization for psychological science. *Science, 237(4820),* 1317–1323.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55-89.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review, 121*, 526-558.

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*, 430–444.

Struyf, D., Iberico, C., & Vervliet, B. (2014). Increasing predictive estimations without further learning: The peak shift effect. *Experimental Psychology, 61*, 134-141.

Sutherland, N. S. (1964). Visual discrimination in animals. *British Medical Bulletin, 20,* 54-59.

Sutherland, N. S. & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Sutton R. S. & Barto A. G. (1990). Time-derivative models of Pavlovian reinforcement. In. Gabriel, M. & Moore, J. (Eds). *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. (pp. 497–537). Cambridge, MA: MIT Press.

Switalski, R. W., Lyons, J., & Thomas, D. R. (1966). Effects of interdimensional training on stimulus generalization. *Journal of Experimental Psychology*, *72*, 661–666.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Thomas, D. R. & Wheatley, K. L. (1974). Effects of interdimensional training on stimulus generalization: An extension. *Journal of Experimental Psychology*, *103*, 1080–1085.

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PLoS ONE*, *9*, e96569.

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25.

Wagner, A. R. (1969). Incidental stimuli and discrimination learning. In R. M. Gilbert & N. S. Sutherland (Eds.), *Animal Discrimination Learning* (Academic Press, pp. 83–111). London.

Wagner, A. R. Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, *76*, 2, 171-180.

Figures

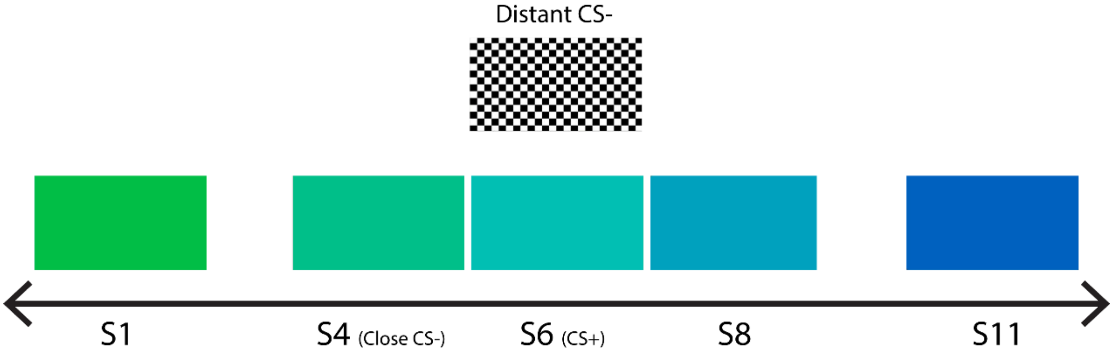

Figure 1. Stimuli used during the training and test phases. The CS+ (S6) was the same in all groups. The Close Negative group also saw a CS- (S4) that was either slightly bluer or greener, and the Distant Negative group also saw a CS- that was a checkerboard rectangle. Note that the direction of the dimension was counterbalanced such that S1 could either be the greenest or bluest stimulus, and S11 could either be the bluest or greenest stimulus.

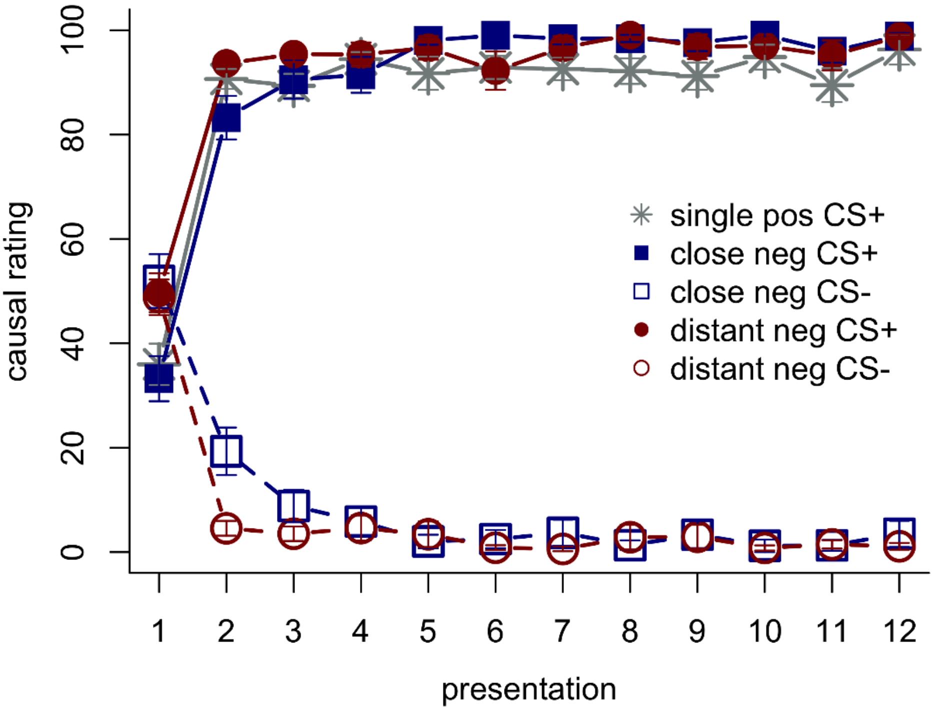

Figure 2. Mean causal ratings given to each training stimulus in Experiment 1. The CS+ was always followed by the outcome and the CS- was always followed with no outcome, and there were 12 presentations of each stimulus. Within-subject error bars calculated using the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

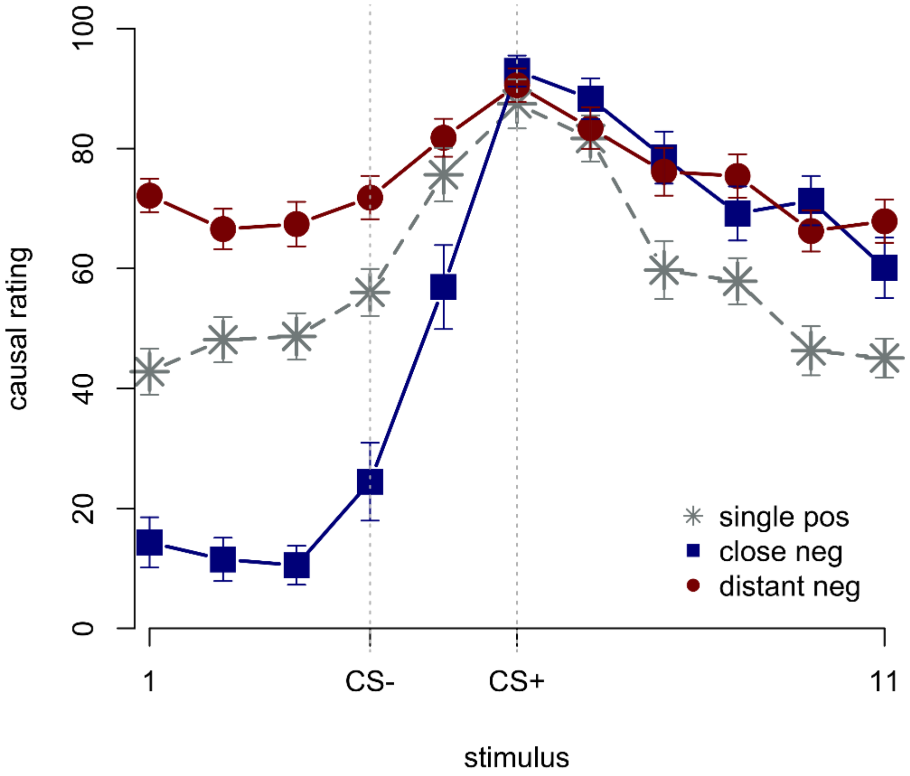

Figure 3. Generalization gradients for the color dimension in Experiment 1. CS+ (Stimulus 6) was the trained stimulus in all three groups. All other stimuli along the dimension were novel except for the CS- (Stimulus 4) in the Close Negative group. S1 and S11 were the greenest or bluest stimuli, depending on counterbalancing group. Within-subject error bars calculated using the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

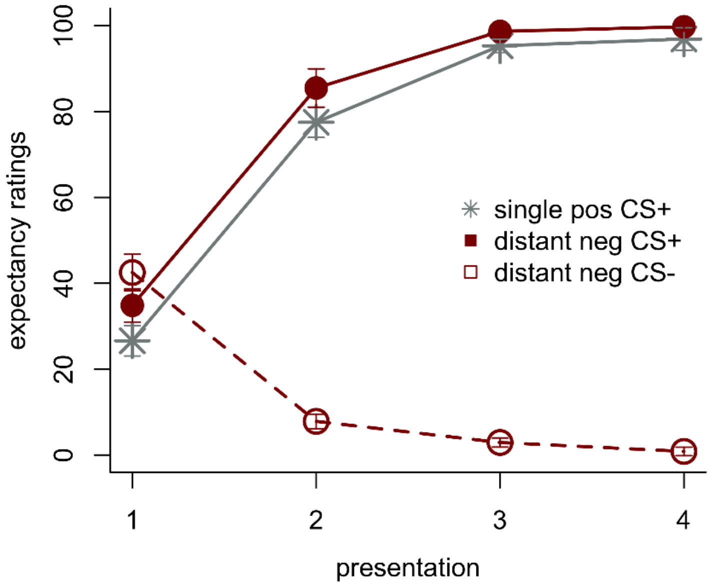

Figure 4. Mean expectancy ratings given to each training stimulus in Experiment 1. The CS+ was always followed by the outcome and the CS- was always followed with no outcome, and there were 4 presentations of each stimulus. Within-subject error bars calculated using the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

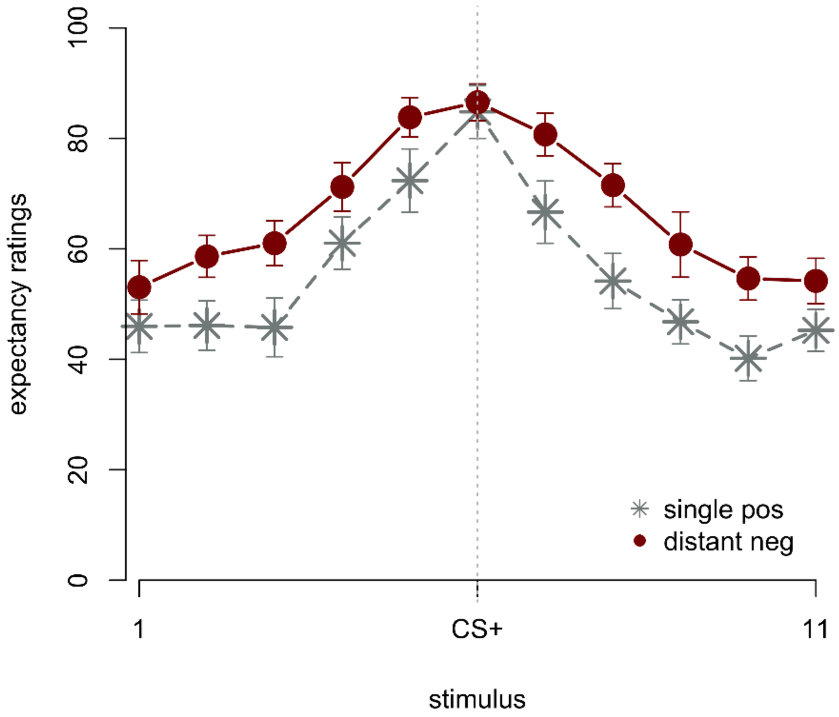

Figure 5. Generalization gradients for the color dimension in Experiment 2. CS+ (Stimulus 6) was the trained stimulus in all three groups. All other stimuli along the dimension were novel. S1 and S11 were the greenest or bluest stimuli, depending on counterbalancing group. Within-subject error bars calculated using the Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

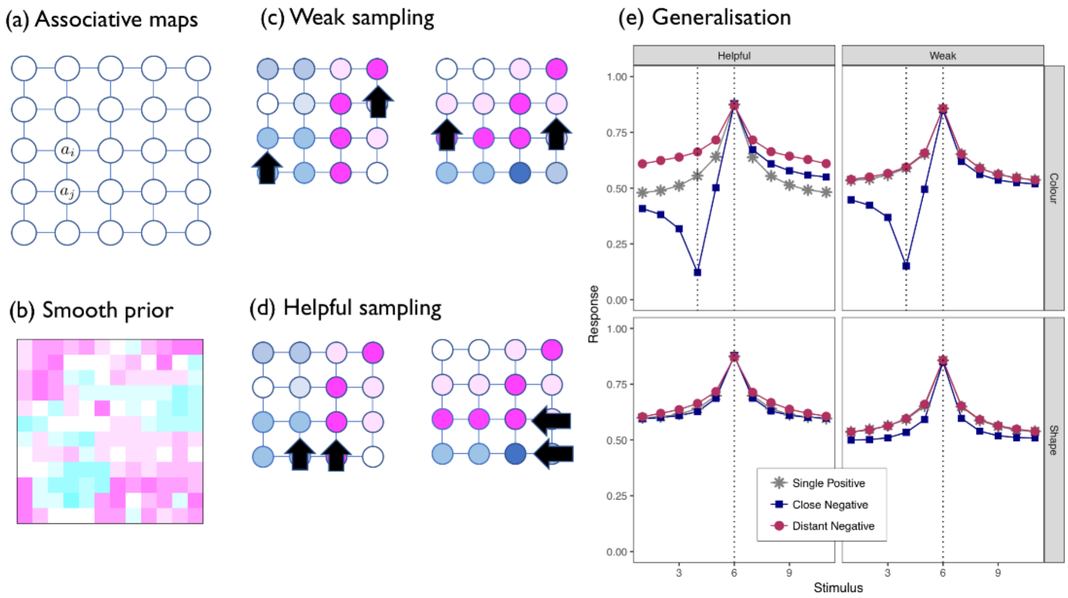

Figure 6. The Bayesian model learns an associative map (a) and has a smoothness bias (b) that ensures similar items have similar associative strengths. In a weak sampling scenario (c) the world is assumed to select stimuli randomly, independent of their true associative strength, whereas in a helpful sampling scenario (d) the experimenter is assumed to select items that produce consistent feedback and vary on relevant dimensions. Generalization gradients (e) with a helpful sampling assumption mirror the empirical data, but those with a weak sampling assumption show no effect of the distant negative evidence.

Appendix A. The Bayesian Model

In this section we describe a computational model that can account for our data as a form of Bayesian inference. Stimuli are assumed to be represented as points located in a stimulus space, and the goal is to learn an *associative map*, a function that relates the stimuli x to the probability of shock, $A(x) = P(o^{+}|x)$. Across a set of n trials the learner is presented with stimuli $X = (x_1 \dots x_n)$ and associated outcomes $O = (o_1 \dots o_n)$. Every time a new observation is made, a Bayesian learner updates the priors over associative maps $P(A)$ to a posterior $P(A\,|x, o)$, using the likelihood $P(x, o|A)$ of that observation to drive the updating:

$$P(A\,|x, o)\ \propto P(x, o|A)P(A).$$

In our task, where stimuli can vary in color, texture and shape we approximate the continuous stimulus space as a discrete lattice as illustrated in Figure 6a. The prior over the associative map takes the form of a Markov random field (e.g., Cross & Jain, 1983, Li, 2009), and assumes that similar items have similar associations: for every pair of stimuli x_i and x_j that differ on the k-th stimulus dimension at value v, the prior assumes $P\left(a_i, a_j\right) \propto (1 - |a_i - a_j|)^{\lambda_{ij}}$ where all such stimulus pairs are constrained by ϕ_{vk} , the smoothness of the associative map at that this point, $P\left(\lambda_{ij}\right) \propto \exp(-\phi_{vk}\,\lambda_{ij})$. By default, the map is assumed to be equally smooth at every location on every dimension (i.e., $\phi_{vk} = \phi$). This yields a model that depends on a single smoothing parameter ϕ that describes the strength of the learner's bias to require similar items to have similar associative properties. However, to capture the intuition that there can be quite sharp discontinuities in the map (e.g., category boundaries) the prior allows for the possibility of mutations (e.g., Kemp & Tenenbaum 2009). These are points on a stimulus dimension where the associative strength can change rapidly, captured by decreasing the smoothness by a random factor $\phi_{vk} = \phi\gamma$ at each such mutation point (i.e., where $\delta_{vk} = 1$), where γ is uniformly distributed between 0 and 1, and the *dimensional relevance* $\theta_k = P(\delta_{vk} = 1)$ describes a mutation rate that can differ for different stimulus dimensions depending on how relevant they are perceived to be. By default the learner sets a uniform prior $P(\theta) \propto 1$.

Assuming exchangeability holds, the likelihood of any CS-US pairing can be factorized into the following two terms:

$$P(x, o|A) = P(o|x, A)P(x|A)$$

each of which has a natural interpretation. The term on the left is the probability that stimulus x leads to outcome o, and is equal to the predicted probability of shock $A(x)$ or non-shock $1 - A(x)$ depending on the outome, and it is through this term that *prediction error* drives learning in this model. The term on the right is given by the learner's *sampling model*, and describes the learner's beliefs about the probability that stimulus x would have been presented if A is the correct associative map. Under a random sampling assumption (often referred to as weak sampling in the reasoning literature), there is no contingency between the associative map A and the stimulus selection itself, so $P(x|A) \propto 1$ and this term vanishes.

Inspired by Bayesian models of rational communication (e.g., Shafto, Goodman & Griffiths 2014; Goodman & Frank, 2014), we create a "*helpful sampling*" version of this model that assumes stimuli were selected by a knowledgeable teacher who follows Gricean maxims (Grice, 1975) in selecting stimuli. For example, a helpful teacher might make relevant dimensions perceptually salient (color in the Single Positive and Close Negative conditions), and/or vary those dimensions during training (color in the Close Negative condition, texture type (color vs. checkeredness) in the Distant Negative condition). Dimensions deemed relevant are presumed diagnostic, so the prior favors higher mutation rates $P(\theta) \propto \theta$ in such cases. Similarly, to be unambiguous a helpful teacher may prefer to select the relevant features proportional to the probability that they produce consistent outcomes and select irrelevant features randomly. For a CS+ this gives the selection rule $P(x|A) \propto E_{x'}[A(x')]$ where the expectation is taken across items x' that differ from x only on irrelevant features. When all dimensions are deemed communicatively irrelevant, this model reduces to the random sampling model described above. For simplicity, the helpful sampling plot in Figure 6e is produced by a model in which a single "relevant dimension" is fixed (texture type in the Distant Negative condition, color in the other conditions), but a similar result is obtained using a model that learns the relevance during training, so long as it is given a strong prior bias to assume that dimensions varied during training are relevant, and has a modest prior bias against assuming that shape is relevant (as shape was less salient). Code for the model simulations is available online at https://github.com/djnavarro/negativeassoc