Dissociating cognitive and affective uncertainty using a General Linear Classifier

Jordan Richard Schoenherr [*1,2], and Robert Thomson[1]
[1]*United States Military Academy, Army Cyber Institute / Behavioural Sciences and Leadership*
[2]*Carleton University, Department of Psychology / Institute for Data Science*
[*]<Jordan.Schoenherr@Carleton.ca>

Abstract

A number of categorization models have been proposed that consider classification performance and uncertainty in terms of prototypes, category boundaries, and exemplars. Like other models of categorization, category boundary models (e.g., GLCs) only consider cognitive uncertainty while failing to consider affective uncertainty. Using a modified GLC, measures of affective uncertainty were obtained by combining exemplar-based information (e.g., response frequency) and categorical information (e.g., categorization accuracy) in varying proportions (0/100, 25,75, 50/50, 75/25, and 100/0). We provide evidence that categorical and exemplar-based representations likely inform affective uncertainty in simple categorization tasks.

Introduction

Uncertainty is a fundamental concept in information theory (Shannon, 1948) explicitly or implicitly informing many early models of cognition (Lachman, Lachman, & Butterfield, 1979). Uncertainty is typically understood in terms of confusability between multiple stimuli due to an absence of information or as the result of stimulus similarity. This reflects the cognitive dimension. Contemporary models of decision-making have considered affect in terms of an antecedent to information processing (e.g., Slovic et al., 2002). More recently, studies have attempted to dissociate cognitive and affective uncertainty (Burleigh & Schoenherr, 2014; Schoenherr & Burleigh, under review). In the present study, we examine whether a simple model of categorization based on General Recognition Theory (GRT; Ashby & Townsend, 1986) can be adapted to understand differences between cognitive and affective uncertainty rather than postulating separate response mechanisms. We examine three different computational methods for modelling affective uncertainty that incorporate categorical and exemplar-based information and contrast them against models that use a single source information (i.e., categorical or exemplar knowledge). We then consider, and reject, a Bayesian model of this phenomenon.

Cognitive and Affective Uncertainty in a Categorization Task

Categorization Models

A number of similarity-based categorization models have been proposed that assume that novel stimuli are classified based on their similarity to summary representations, instances, and category boundaries (for a review, see Pothos & Wills, 2011). Collectively, these models assume that participant will be uncertainty to the extent that a given exemplar shares features with an exemplar associated with a contrasting category. While recent categorization models have considered cognitive uncertainty (e.g., Paul et al., 2011), affective uncertainty has yet to be considered.

Uncanny Phenomena

In the context of engineering, Mori (1970) suggested that the extent to which an object (e.g., robot, mask, doll, prosthetic) was associated with negative affect ('eeriness') was a nonlinear function of the extent to which it shared features with humans, i.e., humanlikeness. The result is an "uncanny valley" (hereafter, referred to as the uncanny valley hypothesis, or UCV), or minima in a function describing positive affect along a continuum.

With considerable qualification, the UCV has received some support from a number of empirical studies (e.g., Cheetham, Suter, & Jäncke, 2011; MacDorman & Ishiguro, 2006). Studies that present participants with stimuli from human and nonhuman categories have found that stimuli containing characteristics from both categories are associate with greater decisional uncertainty which has been interpreted as affective uncertainty (e.g., Cheetham et al., 2011). However, these results must be qualified. In a study by Burleigh, Schoenherr, and Lacroix (2013), stimuli that deviated in terms of humanlikeness (e.g., facial morphology, the number of polygons) did not demonstrate an uncanny valley (Experiment 1). However, when human and nonhuman features were blended in novel stimuli, negative affect was observed (Experiment 2). These results can be taken as suggesting that the uncanny valley is not dependent on humanlikeness, but requires two contrasting category and exemplars that share features of each.

Evidence also suggests that the UCV can be understood in terms of categorization processes more generally. Burleigh and Schoenherr (2014; Schoenherr & Burleigh, under review) suggest that negative affect might be the result of uncertainty that arises from comparing novel stimuli that share features from two contrasting categories. The absence of familiarity with the novel exemplars relative to two or more well-known categories leads to negative affect. Thus, presentation frequency should be a key determinant of affect uncertainty.

Burleigh and Schoenherr (2014) obtained evidence to support this account using a unidimensional categorization task. In a training phase, participants were provided with exemplars selected from two nonhuman categories and were provided with feedback. In a transfer phase, participants categorized old and new stimuli (novel extrapolation items) and provided eeriness and typicality ratings. Burleigh and Schoenherr (Burliegh & Schoenherr, 2014; Schoenherr & Burleigh, under review) observed that cognitive uncertainty (errors) and affective uncertainty (eeriness ratings) are highest around the category boundary. Crucially,

they additionally observed that affective uncertainty (but not cognitive uncertainty) was also high for stimuli located far away from the category boundary that had not been presented during training (extrapolation items). They suggested that this finding can be accounted in terms of frequency-based effects in studies of preference (Bonanno & Stillings, 1986) and the mere exposure effect (e.g., Borenstein, 1989). Consequently, while categorization performance was determined by a categorical representation (i.e., a category boundary), ratings of negative affect were additionally affected by exemplar presentation frequency during the learning phase.

Present Study

The results of Burleigh and Schoenherr (2014) suggest that uncanny valley phenomena need not be the result of special kinds of knowledge or processes related to human categories, but might best be understood in terms of general categorization processes. If true, this means that computational models of categorization should be able to account for the patterns of results associated with the UCV. Following the suggestions of Burleigh and Schoenherr (2014), we consider an account of the UCV that uses a category boundary to classify stimuli from two contrasting categories. Given the widespread use of GRT (Ashby & Townsend, 1986) and category boundary models (e.g., Ashby & Gott, 1988), we used a previous implementation of this model (Alfonso-Reese, 2006) to simulate categorization performance using a General Linear Classifier (GLC). GRT assumes that stimuli are represented in multidimensional space and that categorization occurs by means of the adoption of a decision-boundary. Following the presentation of a stimulus, a GLC adjusts the location of the category boundary in multidimensional space until an optimal decision boundary is identified. The relative location of an old or new stimulus to the decision boundary will determine the category of the stimulus.

Beyond GRT, we additionally assume that participants retain knowledge of both a category boundary and specific exemplars (e.g., RULEX; Nosofsky, Palmeri, & McKinley, 1994). Exemplar-based information was modelled using the response frequency of the GLC, i.e., the *perceived* frequency of stimulus presentation. By pooling categorization accuracy and model stimulus response frequency, we examine possible measures of affective uncertainty.

Model and Results

The model was based on the GRT Toolbox (Alfonso-Reese, 2006) developed in MATLAB. We used the unidimensional variant of the General Linear Classifier (GLC) model (lindiscrim1dvals.m) to accommodate the single dimension used by Burleigh and Schoenherr (2014) and used a low-level of noise (noise = 1) due to the high level of performance of participants in that study. The category boundary was assigned to the mid-point of the stimulus distribution (i.e., Stimulus 8).

Stimulus sets were created in MATLAB to replicate those used by Burleigh and Schoenherr (2014). We examined 2 separate stimulus sets corresponding to two different kinds of training conditions. Corresponding to the equal frequency (EF) condition, the GLC was provided with 5 stimuli from each category (i.e., Stimuli 3-7 and 9-13) that were randomly sampled without replacement. Each of the 10 stimuli were presented 4 times in a block of trials.

In the unequal frequency (UF) condition, the same number of training stimuli (40) were provided to participants. However, the presentation frequency of each of the 4 stimuli (i.e., Stimuli 3-6 and 8-13) increased as a function of distance away from the category. Each successive position away from the stimuli resulted in a doubling of presentation frequency, such that the stimuli located close to the category boundary (Stimuli 6 and 8) were presented 2 times whereas as the most distant stimuli (Stimuli 3 and 13) were presented 8 times. In this way, the GLC becomes sensitized to extreme values along the stimulus continuum in the UF condition but receives unbiased training in the EF condition. The model was provided with 10 training blocks. Table 1 contains the distributions for the training blocks in the EF and UF conditions.

The test session proceeded in an identical manner for the EF and UF conditions. All 15 stimuli were presented to the model. In each of the two training blocks, the model received only 2 presentations of each stimuli.

Response Coding

Correct responses were determined by the assignment of a given exemplar to a region along a perceptual continuum, i.e., a category. Stimulus 8 was used as a mid-point and therefore it was neither correct nor incorrect. Stimuli 1-7 were assigned to Category 1 whereas stimuli 9-15 were assigned to Category 2. Responses were coded as correct if the appropriate category label was assigned to a stimulus.

Model Fit

Figure 1a provides the model output for the training phase wherein only a subset of stimuli were provided. Like Burleigh and Schoenherr, response accuracy is at the lowest at the category boundary, reflecting high decisional uncertainty. As Figure 1b demonstrates, participants were highly accuracy even with the two novel stimuli from each category. To this end both the GLC and human data suggest that a linear category boundary was used to categorize stimuli. A correlational analysis demonstrates the similarity between the human and model categorization data with a strong positive correlation, $r(14) = .980$, $p < .001$.

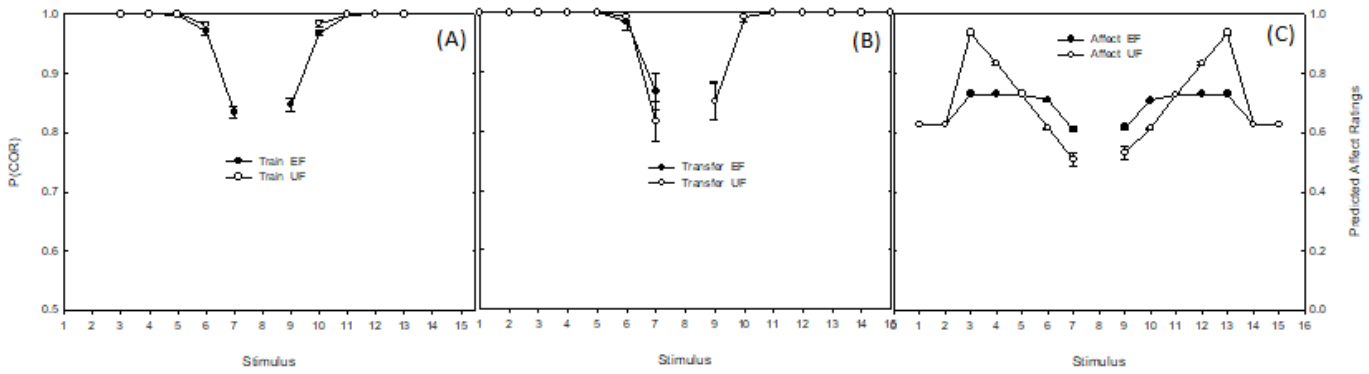

Fig 1. General Linear Classifier (GLC) predicted accuracy in the training (A) and transfer phases (B). GLC predicted negative affect in the transfer phase (C) using the 50:50 ratio for categorical and exemplar-based information.

Affective ratings were modelled using proportional combinations of two sources of information. Affective models were assessed that included only response accuracy or response frequency. Three additional possibility were also examined assuming that one source of information (categorization accuracy or training frequency) might exert a larger influence on affective ratings than another or whether they contributed equally. Table 2 provides the results.

Table 1. Correlations between models of affective responses and human data from Burleigh and Schoenherr (2014). The *p*-value is contained within parentheses.

	Categorical:Exemplar				
	100:0	75:25	50:50	25:75	0:100
Equal Freq.	-.120 (.68)	.014 (.96)	.151 (.61)	.219 (.45)	.245 (.40)
Unequal Freq.	.214 (.46)	.478 (.08)	.528 (.05)	.537 (.05)	.538 (.05)

As Table 1 demonstrates, despite the GLC providing an excellent fit to the categorization data, we did not find support that the GLC's model of response accuracy provided a reasonable fit of the affective responses in either the EF or UF conditions. This supports our claim that cognitive uncertainty (accuracy) and affective uncertainty (negative affect ratings) are differentially influenced by separate sources of information.

The results provided in Table 1 suggest that if a GLC provides a reasonable means to understand affective ratings, the basis for affective ratings might be dependent on condition. Whether equal weight was provided to both categorical and exemplar-based information (50:50) or greater weight was assigned to exemplar-based information (25:75), or exemplar-based information was solely used (0:100), the model provided a reasonable fit to the data in the UF condition. This was not the case for the EF condition. Regardless of the measure used, the GLC did not provide a reasonable model for affective responses. However, this condition appeared to reflect stimuli categorization bias. As Burleigh and Schoenherr (2014) noted, participants in the EF condition appeared to have a bias to anchor their preferences to a particular response category. The possibility of modelling such bias was not examined here but could be explored using a GLC. It is still noteworthy that, even in this condition, the best fits

followed the same trend as the UF condition: the greater the influence of exemplar-based information, the better the affective model fit.

Discussion

In the current study, we obtained evidence that a general linear classifier (GLC; Alfonso-Reese, 2006) provides a good fit for uncertainty responses. While we do not claim that categorization, or the relationship between affect and cognition, can be adequately described by a GLC, we sought a parsimonious account to contrast against more complicated accounts of this relationship (e.g., MacDorman, & Ishiguro, 2006). In the case of the present study, our model used few parameters and, with the exception that response frequency influences affective responses, no additional assumptions were required beyond those of GRT (Ashby & Townsend, 1986). In terms of the categorization results, the GLC provided an excellent fit to the data with equivalent performance to that of the human data. With minor assumptions, the pattern of affective ratings was also captured. Measures that assumed affective uncertainty and cognitive uncertainty were equivalent, provided the poorest fits for the data. In contrast, we found that measures that assume exemplar-based representations inform affective responses produced the best fits.

The current study and GLC-based model of the uncanny valley can be contrasted with a Bayesian account (Moore, 2012). This account assumes the uncanny valley is a function of two factors: (1) *perceptual tension* near category boundaries based on uncertainty between perceptual cues (modeled by a displacement function), and (2) the relative frequency of each category. Being a Bayesian account, the key difference between category boundaries is the assumption that the non-human category's probability distribution has a broader spread than the human category's distribution. This differential spread is required to match the non-monotonicity of Mori's original *affinity* axis. Crucially, Moore's account begins with Mori's function and then works backwards to describe a set of probabilistic processes which could fit this function rather than describing specific elements of the data.

Consequently, Moore's account is lacking in two respect. First, like many early Bayesians accounts, they fail to consider how prior distributions are acquired through learning. Second, like other discussions of the UCV, this account conflates cognitive and affective uncertainty responses thereby assuming only a single learning and response system. In contrast, our results suggest that multiple stimulus representations or multiple learning systems are required to understand discrepancies between affect and categorization.

The findings presented here provide support for Burleigh and Schoenherr's (2014) explanation of UCV-like phenomena. Specifically, high cognitive and affective uncertainty is observed for stimuli located between the two categories due to stimulus ambiguity: stimuli located near the category boundary share more features with a contrasting category and are therefore more confusable. In contrast, stimuli located at the ends of the stimulus continuum are associated with low cognitive uncertainty because of their remoteness relative to the category boundary. Their high affective uncertainty must therefore be a response of infrequent exposure to stimuli during training. Thus, while these stimuli are unambiguously members of their respective categories, they are unfamiliar. This lack of familiarity results in negative affect.

In terms of a simple model of affective uncertainty, we found that using the outputs of a GLC in terms of 1) categorization accuracy and 2) response frequency can provide reasonable fits for the data obtained in experiments examining the UCV. Crucially, the GLC does this without invoking special learning and response mechanisms associated with "humanlikeness" (e.g., MacDorman & Ishiguro, 2006) as well as those that equate cognitive and affective uncertainty (e.g., Cheetham et al. 2011). We therefore suggest that UCV-like phenomenon simply reflect patterns that have been identified in discussed in the preference literature (e.g., Bonanno, & Stillings, 1986; Borenstein, 1989). In short, the special status that is typically ascribed to the "human" category in UCV studies is likely the result of increased high frequency of exemplars (humans) within the environment and cognitive uncertainty of stimuli that shared features from two contrasting categories.

References

Alfonso-Reese, L. A. (2006). General recognition theory of categorization: A MATLAB toolbox. *Behavior Research Methods*, 38, 579-583

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U. & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, 14, 33-53.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154-179.

Bonanno, G. A., & Stillings, N. A. (1986). Preference, familiarity, and recognition after repeated brief exposure to random geometric shapes. *American Journal of Psychology*, 99, 403-415.

Bornstein, R. F. (1989). Exposure and affect: overview and meta-analysis of research, 1968-1987. *Psychoogical Bulletin*, 106, 265–289. doi: 10.1037//0033- 2909.106.2.265

Burleigh, T. J., & Schoenherr, J. R. (2014). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology*, 5:1488.

Burleigh, T. J., Schoenherr, J. R., and Lacroix, G. L. (2013). Does the uncanny valley exist? an empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers and Human Behaviour*, 29, 759–771. doi: 10.1016/j.chb.2012.11.021

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5*, 126. doi: 10.3389/fnhum.2011.00126

Edmunds, C. E. R., & Wills, A. J. (2016). Modeling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS. In A. Papfragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 69-74). Austin: Cognitive Science Society.

Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive Psychology and Information Processing*. Hillsdale: Erlbaum.

MacDorman, K. F. & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*(3), 297-337.

Moore, R.K. (2012). A Bayesian explanation of the 'uncanny valley' efect and related psychological phenomena. *Scientifc Reports*, 2(864), 1-5.

Mori, M. (1970). Bukimi no tani [The uncanny valley]. *Energy*, 7, 33-35.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53–79.

Paul, E. J., Boomer, J., Smith, J. D., & Ashby, F. G. (2011). Information–integration category learning and the human uncertainty response. *Memory Cognition*, 39, 536–554.

Pothos, E. M. & Wills, A. J. (2011). *Formal Approaches in Categorisation.* Cambridge: University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin , & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 397-420). New York: Cambridge University Press.

Ueyama Y (2015). A Bayesian model of the uncanny valley effect for explaining the effects of therapeutic robots in autism spectrum disorder. *PloS ONE*, 10:e0138,642