Categorization-based stranger avoidance does not explain the uncanny valley effect

Karl F. MacDorman*[1] and Debaleena Chattopadhyay[2]

[1] Indiana University School of Informatics and Computing, 535 W. Michigan St., Indianapolis, IN 46202, USA
[2] Department of Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607, USA

Abstract. The uncanny valley hypothesis predicts that an entity appearing almost human risks eliciting cold, eerie feelings in viewers. Categorization-based stranger avoidance theory identifies the cause of this feeling as categorizing the entity into a novel category. This explanation is doubtful because *stranger* is not a novel category in adults; infants do *not* avoid strangers while the category *stranger* remains novel; infants old enough to fear strangers prefer photographs of strangers to those more closely resembling a familiar person; and the uncanny valley's characteristic eeriness is seldom felt when meeting strangers. We repeated our original experiment with a more realistic 3D computer model and found no support for categorization-based stranger avoidance theory. By contrast, realism inconsistency theory explains cold, eerie feelings elicited by transitions between instances of two different, mutually exclusive categories, given that at least one category is anthropomorphic: Cold, eerie feelings are caused by prediction error from perceiving some features as features of the first category and other features as features of the second category. In principle, realism inconsistency theory can explain not only negative evaluations of transitions between real and computer modeled humans but also between different vertebrate species.

Keywords: Anthropomorphism, Computer animation, Face perception, Novelty, Stranger avoidance

1. *3D computer model* is a distinct and familiar category

The crux of the commentators' argument is twofold. First, they claim categorization-based stranger avoidance explains our data and, thus, the uncanny valley effect. Their theory is that stranger avoidance "is triggered when an object has an improbable appearance and is therefore categorized into a novel class" (Kawabe, Sasaki, Ihaya, & Yamada, 2016). Second, they claim our experiment, which used stimuli varying from a person's 3D computer modeled face to a photograph, is not valid because *3D computer model* is not a distinct category; like the real person in the photograph, it is just another instance of the category *human being,* though an instance with computer modeled features. Therefore, they conclude that our stimuli do not transition between two different categories and advise us to "use a stimulus category dimension with a wider range containing a nonhuman entity, an ambiguous entity, and a human being."

Their argument is logically inconsistent. If our experiment did not entail a category transition

* Corresponding author.
E-mail address: kmacdorm@indiana.edu (K.F. MacDorman).

(second claim) because a 3D computer model of a human being is still perceived as an instance of the category *human being*, there would be no "novel class" and, hence, per their theory, no uncanny valley effect (first claim).

We disagree that a hand-drawn cartoon or a doll are instances of *nonhuman* categories while a 3D computer model and a real person are instances of the category *human*. (In their experiment, Yamada, Kawabe, and Ihaya, 2013, created morphs from the face of the Charlie Brown character to the face of a Japanese man.) A 3D computer model, hand-drawn cartoon, and doll can all be used to depict humans—in which case they are all depictions within the category *human.*

Within the category *human,* in our original experiment, we instructed participants to distinguish the *real* from the *computer animated*. In the demographics survey, our US participants reported watching films, videos, and television programs with 3D computer-animated human characters 3.65 hours per week on average (SD = 6.07, n = 365). They also reported playing videogames with 3D computer-animated human characters 4.54 hours per week (SD = 7.54) and having played them for 7.33 years (SD = 5.96). Clearly our participants' long-term exposure should have been sufficient to establish *3D computer model* as a distinct category.

In our experiment, the 3D computer model (0% real) was eeriest, which we attributed to prediction error caused by its features being perceived as features of different categories (MacDorman & Chattopadhyay, 2016; cf. Moore, 2012); eeriness declined as the stimuli transition to 100% real. The commentators claim their theory explains this decline because as stimuli appear more human they "can be better categorized into a familiar class." This explanation is contradicted, however, by the fact that the 100% 3D computer model was categorized with the greatest certainty and rapidity (greater even than 100% real, figure 4 and 7, top left). This result indicates *3D computer model* is a familiar category with a probable appearance, distinct from *real*. Further evidence is the logistic, nearly symmetrical curve with tight confidence intervals for percentage categorized as *real* (figure 4, top left). This pattern is consistent with a transition from one known category to a different known category (figure 6a of Feldman, Griffiths, & Morgan, 2009; figure 19.1 of Harnad, 1987). In fact, if the labels were removed, it would be impossible to deduce whether the transition were from *3D computer model* to *real* or vice versa.

Furthermore, we are concerned when the commentators write, "stranger avoidance is not driven simply by categorization difficulty that can be quantified by measuring categorization latency." In their earlier paper, categorization difficulty was operationalized as categorization latency, and their result—that the least likable stimulus in the transition was also the stimulus with the longest latency—was interpreted as supporting their theory (experiment 1 and 2 in Yamada, Kawabe, & Ihaya, 2013). This repudiates their own experimental methodology used to support categorization-based stranger avoidance theory. Thus, the commentators should propose another way to test their theory's predictions, assuming their theory is falsifiable.

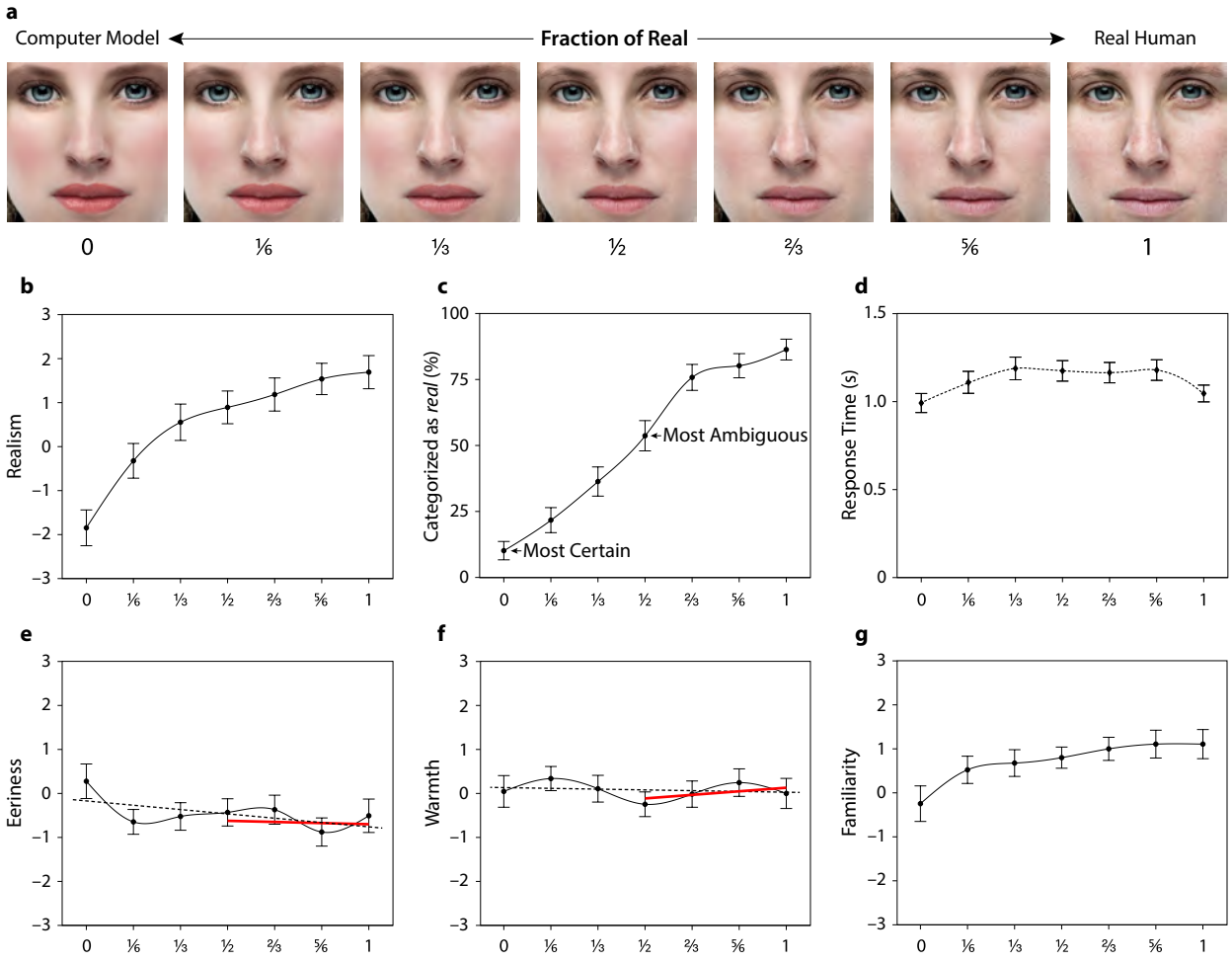

Fig. 1. Stimuli ratings of subjective realism, eeriness, warmth, and familiarity are plotted against their fraction of real as are percentage categorized as *real* and response time. For eeriness and warmth, the regression line for ½ to 100% real is shown in red and for 0% to 100% real is dashed. Error bars indicate the 95% confidence interval.

2. Experiment

The commentators' theory would predict that an improbable appearance elicits the uncanny valley effect (hypothesis 1) and the effect disappears as the appearance becomes more probable (hypothesis 2). They identified this trend along the control transition in our original experiment, with eeriness decreasing from 3D computer model (0% real) to photograph (100% real). However, we assume this trend was caused by inconsistency in the realism of the features of the 3D computer models. This is because some features were harder to model and, therefore, appeared less realistic than others. To address the issue of inconsistent feature realism, we created a more realistic 3D computer model for this experiment.

To test the above hypotheses, we recruited 74 undergraduate and graduate students (37% female) from a Midwestern university (M_{age} = 23.8, SD = 5.1) in September 2016 and conducted

a within-group experiment following our prior methodology (section 2.4–2.6 of MacDorman & Chattopadhyay, 2016), though starting with image ratings, followed by a demographics questionnaire, and finishing with the categorization task. Each participant rated and categorized seven images of a woman varying by sixths from 0% real (3D computer model) to 100% real (photograph) as either *computer animated* or *real* (Figure 1a).

Each image was categorized four times—twice in the first block and twice in the second—with presentation order randomized in each block. Each block started with two unrelated warm-up images and included 20 other images. "Categorize the face below as quickly and as accurately as you can" appeared above the image. The anchors *computer animated* and *real* appeared on opposite sides of the image, left–right counterbalanced between blocks. "Press *e*" appeared below the left anchor and "Press *i*" below the right.

The independent variable (IV) was the woman's fraction of real in the image (0, ⅙, ⅓, ½, ⅔, ⅚ 1). All changes in the IV were of equal size. Dependent variables for each image were percentage categorized as *real,* response time, and ratings on the 7-point semantic differential scales for realism (*computer-animated–real, replica–original,* and *digitally copied–authentic*), familiarity (*rarely seen–common, unfamiliar–recognizable,* and *unique–familiar*), eeriness (*ordinary–creepy, plain–weird,* and *predictable–eerie*), and warmth (*cold-hearted–warm-hearted, hostile–friendly,* and *grumpy–cheerful*).

A manipulation check confirmed that realism increased with fraction of real (Figure 1b). The 3D computer was also rated higher in realism ($M = –1.82$, $SD = 1.62$, $n = 74$) than the six human models of the original experiment ($M = –2.59$, $SD = 1.03$, $n = 651$). Certainty decreased from 0% to ½ real and then increased to 100% real (c). The most certain image was 0% real, which was categorized as *computer animated* 90% of the time (c) and had the fastest response time (d); the most ambiguous image was ½ real, which was categorized as *real* 54% of the time (c). This is the most improbable image.

Contrary to hypothesis 1, the most improbable image was not the eeriest; instead, the most certain image was the eeriest (c). Contrary to hypothesis 2, eeriness did not decrease as the appearance increased from the most improbable to 100% real (e). In fitting a regression line from the most improbable image (½ real) to 100% real, there was no significant change in eeriness (adj. $R^2 = .006$, $F[3, 289] = 1.62$, $p = .184$, e) or warmth (adj. $R^2 = .004$, $F[3, 289] = 1.40$, $p = .244$, f). Across the full range from 0 to 100% real, eeriness increased significantly, though with a negligible effect size (adj. $R^2 = .042$, $F[6, 506] = 4.72$, $p < .001$); warmth (i.e., likeability) did not change significantly (adj. $R^2 = .004$, $F[6, 506] = 1.32$, $p = .245$).

Increases in fraction of real produced the largest increase in realism (b) and in familiarity (g) between 0% and ⅙real and the largest decrease in eeriness. As in this experiment, elsewhere we

have found the 3D computer model (0% real) to be subjectively rated as least familiar (Chattopadhyay & MacDorman, figure 6). We propose this finding is explained but by *perceptual narrowing:* Human infants learn to discriminate human faces better than faces of other species (Pascalis, de Haan, & Nelson, 2002). Perceptual narrowing causes a human face that varies from human norms to appear unfamiliar. We found reducing consistency in feature realism elicits cold, eerie feelings only when it also reduces familiarity, which we interpret as supporting the role of perceptual narrowing in the uncanny valley effect (Chattopadhyay & MacDorman, 2016).

3. Stranger avoidance cannot explain why novelty should elicit cold, eerie feelings

The commentators explain the uncanny valley as a "general cognitive function to emotionally evaluate an object." However, they then conflate this general function with a *specific* phenomenon, stranger avoidance, and a *specific* mechanism, a danger avoidance system: "humans tend to avoid strangers who could potentially harm them physically or impair their genetic fitness" (Yamada, Kawabe, & Ihaya, 2013, p. 30). They then cite (indirectly through LeDoux) findings on identical twins (M_{age} = 22 mo.) indicating a fear of strangers has a genetic component (Plomin & Rowe, 1979). In this context, though, the fear is a heritable trait and not a general cognitive function.

Any experiment on stranger avoidance should include among its stimuli not only strangers but also familiar persons as controls; however, little research on the uncanny valley compares strangers with familiar persons, and none of the commentators' experiments have included familiar persons. Contrary to categorization-based stranger avoidance theory, Matsuda and colleagues (2012) found infants (7–12 mo.) in fact prefer a stranger's face to the same face morphed with their mother's. Beyond this finding, stranger avoidance theory may be countered by observing it does not relate to the uncanny valley effect: Meeting a new person seldom elicits cold, eerie feelings.

Avoidance of "an object categorized as a novel class" describes fear of novelty, not stranger avoidance. *Neophobia* could be as maladaptive as *neophilia* (e.g., sensation seeking) when taken to an extreme. A healthy organism must balance the exploration of the new with the exploitation of the already known in part because what is new can only be understood in relation to what is already known. For example, human infants must first be able to distinguish a stranger from their mother (3–4 mo.) before they can fear the stranger (7–9 mo.); an infant's failure to distinguish a stranger elicits no fear or aversion (Bronson, 1968). Stranger avoidance can only be elicited when the viewer categorizes a person into the known category *stranger;* it cannot explain why categorizing an entity into a novel category should elicit a negative evaluation.

As a general cognitive function, the relation between novelty and affect is not negative but an inverted *U* (Berlyne, 1970; Lang, Bradley, Sparks, & Lee, 2007; Zuckerman, 1976): Low levels of novelty elicit neutral or negative affect (e.g., boredom), moderate levels elicit positive affect

(e.g., curiosity), and high levels elicit negative affect, specifically, fear because unexpected outcomes can be dangerous. However, a general cognitive function cannot explain why the uncanny valley effect is most pronounced in viewing *anthropomorphic* stimuli, nor can it explain why its experiential quality of *eeriness* should be so uncommon yet so distinctive (Mangan, 2015).

Acknowledgements

This research was supported by the US National Institutes of Health (P20 GM066402).

References

Berlyne, D. E. (1970). Novelty, complexity and hedonic value. *Perception and Psychophysics, 8,* 279–286. doi:10.3758/BF03212593

Bronson, G. W. (1968). The development of fear in man and other animals. *Child Development, 39*(2), 409–431. doi:10.2307/1126955

Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision, 16*(11):7, 1–25. doi:10.1167/16.11.7

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752–782. doi:10.1037/a0017196

Harnad, S. (1987). Category induction and representation In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (Chapter 19, pp. 535–565). New York, NY: Cambridge University Press.

Lang, A., Bradley, S. D., Sparks, J. V., Jr., & Lee, S. (2007). The motivation activation measure (MAM): How well does MAM predict individual differences in physiological indicators of appetitive and aversive activation? *Communication Methods and Measures, 1*(2), 113–136. doi:10.1080/19312450701399370

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition, 146,* 190–205. doi:10.1016/j.cognition.2015.09.019

Mangan, B. B. (2015). The uncanny valley as fringe experience. *Interaction Studies, 16*(2), 193–199. doi:10.1075/is.16.2.05man

Matsuda, Y.-T., Okamoto, Y., Ida, M., Okanoya, K., & Myowa-Yamakoshi, M. (2012). Infants prefer the faces of strangers or mothers to morphed faces: An uncanny valley between social novelty and familiarity. *Biology Letters, 8,* 725–728. doi:10.1098/rsbl.2012.0346

Moore, R. K. (2012). A Bayesian explanation of the 'uncanny valley' effect and related psychological phenomena. *Scientific Reports, 2*(864), 1–5. http://dx.doi. org/10.1038/srep00864.

Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? *Science, 296*(5571), 1321–1323. doi:10.1126/ science.1070223.

Plomin, R., & Rowe, D. C. (1979). Genetic and environmental etiology of social behavior in infancy. *Developmental Psychology, 15*(1), 62–72. doi:10.1037/h0078078

Zuckerman, M. (1976). Sensation seeking and anxiety, traits and states, as determinants of behavior in novel situations. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and anxiety* (pp. 141–170). Oxford, England: Hemisphere.