Commentary

The Uncanny Valley Phenomenon: Where Is the Categorical Boundary Between Categorization Difficulty and Categorization Failure?

Yuki Yamada[1], **Kyoshiro Sasaki**[2]

1. Faculty of Arts and Science, Kyushu University, Japan; 2. Faculty of Informatics, Kansai University, Japan

The uncanny valley phenomenon has been widely discussed in relation to human-like entities, but some studies suggest it also applies to inanimate stimuli. Recently, Sasaki et al.[1] used abstract figures as stimuli and argued that categorization failure, rather than difficulty, underlies the uncanny valley phenomenon. While we appreciate their interesting proposal, we clarify that categorization difficulty and failure are not mutually exclusive accounts. We critically examine the findings of Sasaki et al.[1], questioning the lack of direct evidence for categorization failure and their reliance on non-significant results. Furthermore, we propose that the (difficulty-originated) "stranger-avoidance" hypothesis remains a viable alternative, suggesting that categorization difficulty leads to negative responses. Future research should integrate these perspectives to achieve a more comprehensive understanding. Our commentary highlights the need for collaboration and theoretical refinement in uncanny valley research.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

The uncanny valley phenomenon remains a topic of significant interest in cognitive science, psychology, and human-computer interaction. As technology continues to advance, the relevance of this phenomenon has grown, particularly with the widespread adoption of large language models (LLMs) capable of generating human-like conversation. These systems have made interactions with artificial agents more seamless, yet concerns persist about whether their human-like capabilities might evoke feelings of eeriness[2]. The continuing interest in the uncanny valley phenomenon reflects its relevance in understanding human reactions to artificial entities. As artificial agents play increasingly prominent

doi.org/10.32388/JFVYNO

roles in daily life, from customer service robots to virtual assistants, elucidating the mechanisms underlying uncanny experiences has become more pressing. The intersection of perception, categorization, and affective responses offers a rich cognitive framework for studying these phenomena, with implications for both theoretical models and practical applications in design and human-agent interaction.

Recent research highlights that the uncanny valley extends beyond physical human-like features to include abstract or geometric stimuli[1]. Notably, our previous research[3][4][5] has shown that uncanny valley-like effects can also emerge in non-human and non-animal contexts (*i.e.*, fruits), and the recent study further expanded the scope of this phenomenon. Namely, Sasaki et al.[1] explored how individuals respond to geometric stimuli, investigating whether hard-to-categorize stimuli induce their low likability. Thus, it has been suggested that the complexity of the uncanny valley phenomenon is not solely tied to animacy-based attributes but also involves cognitive and perceptual characteristics of abstract or geometric objects.

This commentary revisits the categorization-based accounts for the uncanny valley, focusing on two hypotheses: categorization difficulty and categorization failure. Both hypotheses agree that categorization difficulty can lead to negative impressions but propose partially different cognitive mechanisms. The categorization difficulty hypothesis suggests that such stimuli may be categorized into a "stranger" category, associated with negative valence[6]; this hypothesis was supported by the subsequent study[7]. In contrast, the categorization failure hypothesis posits that low likability arises when categorization is not completed. Consequently, these subtle differences need to be addressed, and this paper aims to provide a bridge towards the development of the categorization-based account in future uncanny valley research.

Overview of Sasaki et al.[1]

Sasaki et al.[1] investigated the role of categorization processes in uncanny valley-like effects, using abstract geometric stimuli. In the first experiment, participants were asked to categorize and evaluate morphed geometric shapes, such as blends between a circle and a square. The results showed that intermediate morphs, which elicited greater categorization difficulty as indicated by longer response times, were consistently rated as less likable. These results supported the hypothesis that categorization challenges are closely linked to affective discomfort. The second experiment focused on perceptual

doi.org/10.32388/JFVYNO

fluency by introducing priming, where participants were exposed to identical stimuli immediately before evaluation. Contrary to expectations, priming did not improve the likability of difficult-to-categorize stimuli. This suggested that perceptual fluency alone might not fully explain the discomfort associated with categorization difficulty.

In their third experiment, cognitive fluency was manipulated by providing label cues (e.g., "circle" or "square") before stimulus presentation as text priming. These primers improved the likability of some stimuli but had limited effects on those with the highest categorization difficulty. The original authors claimed that certain stimuli might resist categorization entirely. The final experiment introduced a dual-task paradigm to impose cognitive load, requiring participants to memorize numbers while evaluating the stimuli. Cognitive load reduced the likability of easily categorized stimuli but did not significantly affect those that were difficult to categorize. The original authors interpreted this as evidence for categorization failure—the inability to assign stimuli to any category.

These experiments are quite interesting because they collectively show that uncanny valley-like phenomena are not limited to stimuli such as living things and humans, but can also occur with abstract forms. In addition, the various cognitive manipulations that have not been used in previous uncanny valley research are innovative. Sasaki et al. argue that the categorization process, and in particular the interaction between difficulty and failure, plays a central role in understanding these phenomena. This research criticized simple fluency-based explanations and suggested the need for further research to elucidate the mechanisms that cause emotional responses to ambiguous stimuli.

Critical analysis

We acknowledge that Sasaki et al.[1] is a remarkable study that uses classical cognitive psychological techniques to theoretically examine the uncanny valley phenomenon. Despite their significant contributions, the interpretation of their findings raises several concerns. A central issue lies in the distinction between categorization difficulty and categorization failure. Unfortunately, the original authors dedicate little space to explaining their main argument, the categorization failure hypothesis. Based on what we can understand from the paper, our personal communication with them (it should be noted that one of the authors of this commentary and the original authors, Sasaki, K., are different individuals), and some inferences from the context, we would like to summarize their argument as follows:

1. In the categorization failure hypothesis, categorization is abandoned for the most categorization-difficult stimulus. On the other hand, in the processing fluency hypothesis categorization is merely delayed for this stimulus, and is still executed.
2. According to the categorization failure hypothesis, the likability of objects that could not be categorized decreases significantly, independent of the deterioration due to low processing fluency. According to the processing fluency hypothesis, the degree of categorization difficulty "alone" determines the likability of objects.
3. Priming operations increase processing fluency as long as categorization is executed, but if categorization fails (in the categorization failure hypothesis), there is no categorization processing to be facilitated, so they have no such effect.

Simply put, the categorization failure is an additional hypothesis that fills in the gaps in situations that cannot be explained by the processing fluency hypothesis. Therefore, it is not a "rather than" account as in Sasaki et al.[1]'s title, but instead exactly a supplementary hypothesis that coexists with the processing fluency hypothesis, based on the premise that the degree of difficulty in categorization affects likeability; indeed, Sasaki et al.[1] seem to assume the coexistence of these (see page 9 of their article).

The categorization failure hypothesis potentially provides important insights into affective reactions to hard-to-categorize objects. However, their findings that seem to demonstrate the categorization failure hypothesis leave some questions and will call for further investigation and discussion.

Firstly, the observed effects could also be attributed to extreme categorization difficulty rather than a complete failure to categorize. The fact that participants provided categorization responses indicates that some form of categorization process was engaged, even for the most ambiguous stimuli. This suggests that rather than being entirely abandoned, categorization may have been incomplete or prolonged, complicating the interpretation of failure. Without direct evidence of cognitive disengagement, such as self-reported confidence levels or neural measures of categorization activity, the distinction between difficulty and failure remains speculative.

Secondly, there is no direct evidence showing the priming effects on categorization. In their Experiments 2 and 3, the priming effect on likeability was examined by comparing the results with those of Experiment 1. However, it is not clear why the response times were not compared in this way. The categorization failure hypothesis predicts that the priming effect is ineffective for stimuli that cannot be categorized, and this should require a comparison of response times in the categorization task. Because

doi.org/10.32388/JFVYNO

the data set was not open, we used WebPlotDigitizer (https://automeris.io/) to examine the mean response time for the 30% stimulus[1], and found that it was 897 ms for Experiment 1, 882 ms for Experiment 2, and 1128 ms for Experiment 3. According to the categorization failure hypothesis, stimuli that cannot be categorized are not facilitated, so it would be predicted that there is no difference between Experiments 1 and 2, or between Experiments 1 and 3, for this stimulus. In the comparison between Experiments 1 and 2, they do not appear to differ greatly. However, the reaction times for that stimulus in Experiments 1 and 3 are obviously different. Furthermore, the primed stimulus in Experiment 3 was actually more difficult to categorize. Taking these results as a whole, at the moment there seems to be little evidence to support the claim that the priming effect does not affect the processing of hard-to-categorize stimuli.

Most importantly, the reliance on non-significant results to support conclusions introduces a fundamental issue tied to the nature of null hypothesis significance testing (NHST). Non-significant findings merely indicate insufficient evidence to reject the null hypothesis, not evidence in favor of categorization failure, making any definitive claims based on such results problematic.

Therefore, the categorization failure hypothesis is still in the proposal stage, and much positive and direct evidence is needed to demonstrate it. But as we will briefly discuss later, this hypothesis is attractive insofar as it provokes discussion about the fate of uncategorized items in the cognitive and emotional processing.

Organizing the two hypotheses and findings

The fluency-based accounts, which posit that reduced processing fluency leads to negative evaluations, faces significant challenges in explaining uncanny valley(-like) phenomena. Although Sasaki et al. [1] report that perceptual and cognitive fluency manipulations had limited effects on improving likability, the conclusions rest on non-significant results, which complicates definitive interpretations. However, given the very small effect sizes for shape and text priming effects observed in Sasaki et al.'s experiments, the explanatory power of the fluency-based accounts may still be limited. Furthermore, as Reber et al.[8] argued, processing fluency enhances hedonic value and aesthetic appreciation but does not necessarily imply that reduced fluency diminishes these evaluations. Hence, we consider that fluency-based accounts cannot solely explain some experimental results[9], especially those in Sasaki et al.[1].

Our previous study, Kawabe et al.[6], proposed the stranger-avoidance hypothesis, which posits that categorization difficulty leads to the assignment of stimuli to a "stranger" category inherently associated with negative valence. This explanation is independent of the fluency hypothesis and was not directly tested by Sasaki et al.[1]. They possibly assume that the text priming used in their experiments facilitated categorization into a given set of geometric shape categories. Whether this assumption is true could indirectly be confirmed by examining the effect of the text priming on the latencies or the judgment proportion of categorization (i.e., comparing the results of the categorization tasks between Experiments 1 and 3). If the facilitation effect stemming from the text priming is observed in the categorization task, their assumption would be true, and their uncanny valley-like effect might be independent of the stranger-avoidance hypothesis. As we pointed out above, Sasaki et al.[1] did not examine the effect of the text priming on the categorization task, and their findings were unrelated to the "stranger" context. Thus, their findings could not rule out the stranger-avoidance hypothesis and their results could be explained by this hypothesis as well as the categorization failure hypothesis: morphed geometric objects would be categorized into novel/strange category classes, which would evoke negative reactions. From this position, Sasaki et al.'s findings intriguingly suggest that the concept of a "stranger" category might extend beyond humans to encompass unfamiliar or ambiguous objects more generally. This broader applicability raises the possibility that avoidance responses may reflect a generalized mechanism for identifying and reacting to suspicious or potentially harmful stimuli including novel foods[3][5].

Both the categorization difficulty and categorization failure hypotheses share a fundamental premise that high categorization difficulty is a prerequisite for eliciting uncanny valley-like effects. Attempts to refute categorization difficulty as a basis for these phenomena are therefore logically inconsistent. Rather, to be correct, what Sasaki et al.[1] rejected was the fluency-based account, and it is applicable to only the most hard-to categorize stimuli. Instead, the critical question we the categorization-based accounts researchers should tackle together is what cognitive processes are engaged when categorization becomes difficult. Multi-system theories of categorization suggest that categorization involves distinct but interacting processes—for instance, a fast, implicit system for perceptual grouping and a slower, explicit system for conceptual integration[10]. In this view, the stranger-avoidance hypothesis aligns with the explicit system, where stimuli are deliberately categorized into a negatively valenced 'stranger' category based on rule-based processing. Conversely, the categorization failure hypothesis may reflect disruptions in the implicit system, which is responsible for rapid perceptual grouping and associative processes. These two hypotheses may represent different facets of the categorization process: the stranger-avoidance hypothesis emphasizing explicit, conceptual mechanisms, and the failure hypothesis focusing on implicit, associative breakdowns in categorization. Examining these processes through multi-system theories allows for a more comprehensive understanding of how ambiguity in categorization contributes to negative reactions.

Concluding remarks

This commentary aimed to clarify the distinctions and overlaps between categorization-based accounts, particularly the categorization difficulty (now referred to as the stranger-avoidance) and categorization failure hypotheses. Importantly, these hypotheses should not be viewed as mutually exclusive, but rather as complementary perspectives that emphasize different facets of the cognitive process under conditions of categorization difficulty. While categorization-based accounts have made striking progress in explaining uncanny valley-like effects, future work must also consider their relationship to alternative theories, such as configural processing[11][12], atypicality[13][14], and perceptual mismatch[15][16][17] theories. Integrating these approaches may provide a more holistic understanding of how eeriness arises in response to ambiguous or unfamiliar stimuli. Achieving this integration will require open communication and collaboration among researchers on this topic. Such collaboration will be critical not only for theoretical or scientific advances but also for practical applications in design and technology, ensuring that artificial systems are better aligned with human perceptual and emotional expectations.

In conclusion, the uncanny valley remains fertile ground for exploration. By embracing the multifaceted nature of this phenomenon, building theories with avoiding mutual misunderstanding, and working together to address unresolved questions, researchers can deepen our understanding of human responses to ambiguity and improve human-artificial interaction in meaningful ways.

References

1. a, b, c, d, e, f, g, h, i, j, k, l, m, n, o*Sasaki K, Yonemitsu F, Ariga A (2025). "The uncanny valley phenomenon can be explained by categorization failure rather than categorization difficulty." Visual Cognition. 1–12. doi:10.1080/13506285.2024.2448697*
2. ^*Kjeldgaard-Christiansen J (2024). "What is creepiness, and what makes ChatGPT creepy?" Leviathan. 10: 1–15. doi:10.7146/lev102024144284*

3. ^a, b^ *Yamada Y, Kawabe T, Ihaya K (2012). "Can you eat it? A link between categorization difficulty and food likability." Advances in Cognitive Psychology. 8(3): 248–254.*
4. ^ *Yamada Y, Kawabe T, Ihaya K (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon: Categorization and evaluation. Jpn Psychol Res. 55(1):20-32. https://doi.org/10.1111/j.1468-5884.2012.00538.x.*
5. ^a, b^ *Yamada Y, Sasaki K, Kunieda S, Wada Y (2014). "Scents boost preference for novel fruits." Appetite. 81(C): 102–107.*
6. ^a, b^ *Kawabe T, Sasaki K, Ihaya K, Yamada Y (2017). "When categorization-based stranger avoidance explains the uncanny valley: A comment on MacDorman and Chattopadhyay (2016)." Cognition. 161: 129–131.*
7. ^ *Sasaki K, Ihaya K, Yamada Y (2017). "Avoidance of Novelty Contributes to the Uncanny Valley." Frontiers in Psychology. 8: 1792. doi:10.3389/fpsyg.2017.01792*
8. ^ *Reber R, Schwarz N, Winkielman P (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Pers Soc Psychol Rev. 8(4):364-82.*
9. ^ *MacDorman KF, Chattopadhyay D (2016). "Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not." Cognition. 146(C): 190–205.*
10. ^ *Minda JP, Roark CL, Kalra P, Cruz A (2024). "Single and multiple systems in categorization and category learning." Nature Reviews Psychology. 3(8): 536–551. doi:10.1038/s44159-024-00336-7*
11. ^ *Chattopadhyay D, MacDorman KF (2016). "Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley." Journal of Vision. 16(11): 7–25.*
12. ^ *Kätsyri J, de Gelder B, Takala T (2019). "Virtual faces evoke only a weak uncanny valley effect: An empirical investigation with controlled virtual face images." Perception. 48(10): 968–991. doi:10.1177/0301006619869134*
13. ^ *MacDorman KF, Green RD, Ho CC, Koch CT (2009). "Too real for comfort? Uncanny responses to computer generated faces." Computers in Human Behavior. 25: 695–710.*
14. ^ *Strait MK, Floerke VA, Ju W, Maddox K, Remedios JD, Jung MF, Urry HL (2017). "Understanding the uncanny: Both atypical features and category ambiguity provoke aversion toward humanlike robots." Frontiers in Psychology. 8: 1366. doi:10.3389/fpsyg.2017.01366*
15. ^ *Mitchell WJ, Szerszen KA Sr, Lu AS, Schermerhorn PW, Scheutz M, Macdorman KF (2011). "A mismatch in the human realism of face and voice produces an uncanny valley." I-Perception. 2(1): 10–12. doi:10.1068/i0415*

16. ^Moore RK (2012). "A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena." Scientific Reports. 2(1): 864. doi:10.1038/srep00864
17. ^Seyama J, Nagayama RS (2007). "The uncanny valley: Effect of realism on the impression of artificial human faces." Presence. 16(4): 337–351. doi:10.1162/pres.16.4.337

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.