New Ideas in Psychology 50 (2018) 38–47

Contents lists available at ScienceDirect

New Ideas in Psychology

journal homepage: www.elsevier.com/locate/newideapsych

Wrong outside, wrong inside: A social functionalist approach to the uncanny feeling

Antonio Olivera-La Rosa[a,b,*]

[a] Department of Psychology and Social Sciences, Universidad Católica Luis Amigó, Transversal 514A #67B 90, Medellín, Colombia
[b] Human Evolution and Cognition Group, Associated Group to IFISC (University of the Balearic Islands – CSIC), Carr. de Valldemossa, km 7,5, 07122, Palma de Mallorca, Spain

ARTICLE INFO

Keywords:
Uncanny valley
Morality
Uncanny feeling
Dehumanization

ABSTRACT

The "uncanny valley" hypothesis (Mori 1970/2005) states that a near-human looking entity can engender negative feelings in an observer. I analyze the phenomenology of the uncanny feeling, which is largely understudied despite being the dependent variable in empirical studies. Next, I introduce a social functionalist account to the uncanny valley research. I propose that the uncanny feeling is a social response triggered by the perception that something is ambiguously wrong with the "humanness" of the human-like stimuli, and therefore needs to be avoided. By doing so, the uncanny feeling functions as a "wrong outside, wrong inside" heuristic with central moral connotations. I conclude that rethinking the uncanny feeling as a social response helps to integrate controversial findings within the field.

> *Climbing a mountain is an example of a function that does not increase continuously: a person's altitude does not always increase as the distance from the summit decreases owing to the intervening hills and valleys. I have noticed that, as robots appear more human-like, our sense of their familiarity increases until we come to a valley. I call this relation the "uncanny valley"*. (Mori, 1970/2005, p.33, p.33)

Since Mori's (1970/2005) seminal definition, the uncanny valley (UV) hypothesis has transcended its original focus on robotics to be embraced by psychological research. Indeed, the claim that any kind of human-likeness manipulation in a robot (or other entities) will trigger a negative affective response (e.g., a sense of unease or eeriness) at close-to-realistic levels has an undeniable psychological appeal.[1] Despite increasing interest in empirical approaches to this phenomenon, contradictory findings raised academic concerns about its scientific explanation and even its mere plausibility. In this article, I discuss the phenomenology of the uncanny feeling (UF), which is largely understudied in psychological research on the UV. I argue that a characterization of the UF that goes beyond its hedonic component is crucial to integrate existing data with theoretical explanations on the phenomena. In order to contextualize the characterization of the UF, I briefly review the main claims and limitations of the most influential accounts of this phenomenon. I do not, however, claim to offer a comprehensive state of the art of all the psychological research on the UV (for comprehensive theoretical revisions on this topic, see Kätsyri, Förger, Mäkäräinen, & Takala, 2015; Wang, Lilienfeld, & Rochat, 2015).

In the second part of this article, I propose a new theoretical account to the UF, in which it is understood as a social response directed to avoid interactions with morally ambiguous human-like entities. I believe that rethinking the UF as a social response may shed new light into its nature and overcome some limitations of previous work. In particular, I argue that a social functionalist approach to the "complex" phenomenology of the UF (which involves both basic affective dimensions and elaborated cognitions) may clarify mixed evidence, by highlighting that some hypotheses account for *some* basic mechanisms involved in the UF, but fail to capture the whole picture. Further, I propose that the perception of "twisted" (i.e., atypical/weird) facial features in human-like entities influences moral appraisals (i.e., subjective evaluations). I discuss this hypothesis in terms of well-established psychological mechanisms, such as face discrimination (perceptual narrowing), the automaticity of moral judgments, dehumanization, and the *Beauty-is-Good* stereotype.

Given this background, the two main goals for this paper are: a) to discuss the phenomenology of the UF, by focusing on the distinction between the affective and "cognitive" components of the emotional response; and, b) to propose a social functionalist approach to the role of the UF in the moral domain.

* Department of Psychology and Social Sciences, Universidad Católica Luis Amigó, Transversal 514A #67B 90, Medellín, Colombia.
E-mail addresses: antonio_olr@outlook.es, acensulay@yahoo.es.

[1] With regard to this claim, I agree with Wang and Rochat (2017, p. 15) that: "the precise shape of the uncanny valley graph depicted by Mori (1979/2005) should not be taken literally as the criteria for detecting the uncanny phenomenon."

https://doi.org/10.1016/j.newideapsych.2018.03.004
Received 30 January 2018; Received in revised form 24 March 2018; Accepted 25 March 2018

1. Psychological explanations of the UV

As mentioned above, psychological research on the UV has proven to be a dynamic field. Recent reviews addressed the main hypotheses concerning the explanation of the UV, on a deep level, (Kätsyri et al., 2015; Wang et al., 2015), making it pointless to offer a new comprehensive literature review on this problem. Congruently, in this section I only consider those studies that, from my point of view, are more influential in bolstering claims for the two dominant broader accounts of the UV: the "Cognitive-Processing" account and the "Humanness-Processing" account. I am aware, however, that the proposed distinction relies on practical reasons (e.g., dealing with non-integrative literature), and that there are alternative ways to group these hypotheses (see MacDorman, Green, Ho, & Koch, 2009; Wang et al., 2015). Indeed, both accounts are not mutually exclusive (i.e., the proposed psychological mechanisms may be simultaneously active in perceiving human-looking forms; see MacDorman et al., 2009) and share crucial aspects (e.g., deviations from human appearance produces prediction errors in brain regions associated with the perception of human faces; Chattopadhyay & MacDorman, 2016).

1.1. The "cognitive-processing" account

According to this account, the UF is caused by a domain-general cognitive mechanism, with no particular focus on the "humanness" component of the stimuli. In this context, recent academic debate has largely focused on what is the best cognitive explanation underlying the UF. Whereas some research argues that the UF is caused by the artificial-human categorical uncertainty (*Categorization Difficulty hypothesis*), other researchers claim that the main mechanism underlying the UF is the perceptual mismatch between incongruent features (e.g. artificial vs. human, *Perceptual Mismatch hypothesis*).

1.1.1. Categorization Difficulty hypothesis

This hypothesis claims that the UF is caused by the ambiguity in categorizing artificial (but highly realistic) entities as either human or artificial entities (Burleigh, Schoenherr, & Lacroix, 2013; Yamada, Kawabe, & Ihaya, 2013). Therefore, it is argued that those objects that are located at the category boundary between "artificial" and "human" are perceived as ambiguous and difficult to process (dissonant), thus triggering the experience of the UF. Although some authors have suggested alternative versions of this hypothesis (Burleigh & Schoenherr, 2014; Schoenherr & Burleigh, 2014), proponents of this framework agree on one fundamental claim: affective negative responses are explained as a function of stimulus distance from a category boundary (Cheetham, Suter, & Jäncke, 2011). This hypothesis is built on research into well-established psychological mechanisms such as cognitive dissonance (Elliot & Devine, 1994) and categorical perception (Goldstone & Hendrickson, 2010).

Although empirical evidence broadly supports the fundamental claim (Kätsyri et al., 2015), a certain number of critical aspects remain unclear (i.e., question the scope of some empirical findings supporting this hypothesis). From a methodological point of view, some authors argue that lower affective ratings at the category boundary might result from artifacts produced by human-likeness manipulations (such as image morphing artifacts[2]; see MacDorman & Chattopadhyay, 2016; Kätsyri et al., 2015). More crucial for the present review, ambiguity in the conceptualization of the dependent variable may played a large role in current controversial findings (see Section 2.1).

Crucially, the role of the humanness of stimuli remains unclear in this hypothesis. Strictly speaking, if a general-domain cognitive mechanism largely explains the UF, this affective pattern may result irrespective of the humanness of the stimuli (as suggested by Ramey, 2006). Few studies have tested whether ambiguity in stimulus categorization causes negative affinity for non-human stimuli. Results by Yamada et al. (2013), Ferrey, Burleigh, and Fenske (2015), and Ramey (2006) suggest that this is indeed the case.

1.1.2. Perceptual mismatch hypothesis

Other researchers argued that the UF is caused by an inconsistency between the human-likeness levels of specific cues (Kätsyri et al., 2015; MacDorman & Chattopadhyay, 2016; Seyama & Nagayama, 2007). For instance, mismatches induced by dissonant facial features (such as clearly artificial eyes on fully human-like face[3]) would violate *a priori* expectations of the perceiver, causing negative affect (i.e. negative hedonic tone). Indeed, there is evidence that a variety of cross-modal mismatches may be associated with the UF (Mitchell et al., 2011; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012). Some authors claim that the UF might be the by-product of heightened sensitivity to atypical features on humanlike characters (Brenton, Gillies, Ballin, & Chatting, 2005; MacDorman et al., 2009). From an evolutionary perspective, it is plausible that the human visual system has acquired expertise in detecting atypical facial features (Nesse, 2005). As a result, any atypical physical trait (e.g., grossly enlarged eyes) may violate our "natural" expectations, causing negative affect. In this vein, the possibility that the UF is explained as low attractiveness (instead of its lack of realism) of the human replica has also been noted by Hanson (2005).

There are, however, some critical issues that undermine the perceptual mismatch hypothesis. First, as Wang et al. (2015) claim, this hypothesis fails to consider the effects of positive violations of expectations (e.g., humor comprehension). Second, this explanation faces the same ambiguities as the Categorization Difficulty hypothesis when defining both the dependent variable (e.g., affinity/likability/eeriness) and the role of the humanness of the stimuli in the obtained effects. As highlighted by MacDorman and Ishiguro (2006, p. 301): "While many non-biological phenomena can violate our expectations, the eerie sensation associated with the uncanny valley may be peculiar to the violation of human-directed expectations, which are largely subconscious." I will address this point on a deeper level in the analysis of UF phenomenology (Section 2.2).

1.2. The "Humanness-Processing" account

In contrast to theories focusing on the cognitive mechanisms underlying the UF, a second group of hypotheses focuses on perceptual discrepancies related to the humanness of stimuli (understood as those qualities that define our species, Złotowski, Proudfoot, & Bartneck, 2013). Therefore, the UF is understood as largely automatic stimulus-driven processing that occurs at early stages in perception (MacDorman et al., 2009; Wang et al., 2015). This claim is supported by the human visual system's special sensitivity to perceiving anthropomorphic entities (Złotowski et al., 2013), and by the fact that perceiving small deviations from human appearance produces large prediction errors in brain regions associated with the recognition of human faces (Chattopadhyay & MacDorman, 2016).

It is notable that these claims are not irreconcilable with more specific versions of the "Cognitive-Processing" framework. Thus, the main distinction between proponents of this framework and the "Cognitive-Processing account" is in the focus on the humanness of stimuli. For instance, those results showing that negative affinity occurs despite the humanness of stimuli (Ferrey et al., 2015; Ramey, 2006) may be interpreted as a more general instance of stimulus devaluation, instead of as evidence of the UF *per se*.

[2] Image morphing technique is used to construct a sequence of gradual changes between two images (typically, computer graphics and human faces, see Kätsyri et al., 2015).

[3] That is, those faces that are perceived as having characteristic human facial patterns without, at the same time, pertaining to other clearly defined category (e.g., apes).