Cognition 160 (2017) 43–50

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting

Jan-Philipp Stein *, Peter Ohler

Technische Universität Chemnitz, Germany

ARTICLE INFO

Article history:
Received 28 April 2016
Revised 18 December 2016
Accepted 23 December 2016
Available online 30 December 2016

Keywords:
Uncanny valley
Theory of mind
Social cognition
Anthropocentrism
Artificial intelligence
Virtual reality

ABSTRACT

For more than 40 years, the uncanny valley model has captivated researchers from various fields of expertise. Still, explanations as to why slightly imperfect human-like characters can evoke feelings of eeriness remain the subject of controversy. Many experiments exploring the phenomenon have emphasized specific visual factors in connection to evolutionary psychological theories or an underlying categorization conflict. More recently, studies have also shifted away focus from the appearance of human-like entities, instead exploring their mental capabilities as basis for observers' discomfort. In order to advance this perspective, we introduced 92 participants to a virtual reality (VR) chat program and presented them with two digital characters engaged in an emotional and empathic dialogue. Using the same pre-recorded 3D scene, we manipulated the perceived control type of the depicted characters (human-controlled avatars vs. computer-controlled agents), as well as their alleged level of autonomy (scripted vs. self-directed actions). Statistical analyses revealed that participants experienced significantly stronger eeriness if they perceived the empathic characters to be autonomous artificial intelligences. As human likeness and attractiveness ratings did not result in significant group differences, we present our results as evidence for an "uncanny valley of mind" that relies on the attribution of emotions and social cognition to non-human entities. A possible relationship to the philosophy of anthropocentrism and its "threat to human distinctiveness" concept is discussed.

1. Introduction

Computer systems have become inseparably entangled with people's daily lives, ever growing in complexity and sophistication. Apart from many beneficial effects, research has also explored unpleasant experiences that result from engaging advanced technologies. A prominent contribution to this field, the *uncanny valley* theory (1970) by Japanese robotics engineer Masahiro Mori illustrates how complex human-like replicas (such as robots and digital animations) can evoke strong feelings of eeriness if they approach a high level of realism while still featuring subtle imperfections (Fig. 1).

Although its basic assumptions have remained mostly unchanged for more than four decades, the model has not lost any relevance due to the continued success and advancement of digital technology. Even more so, the exploration of uncanny valleys has ceased to be a merely academic venture, as modern robotics keep unfolding their economic potential and big-budget entertainment media stand and fall with the perception of their virtual characters (Barnes, 2011; Tinwell & Sloan, 2014).

Traditionally, research on the uncanny valley effect has focused on an object's specific appearance or motion patterns to explore which features might come across as abnormal and unsettling (Bartneck, Kanda, Ishiguro, & Hagita, 2009; Hanson, 2006; Seyama & Nagayama, 2007). As numerous studies have succeeded in exposing such visual imperfections and connected them to negative evaluations, the phenomenon has been framed by theories such as pathogen avoidance (Ho, MacDorman, & Pramono, 2008), mortality salience (MacDorman & Ishiguro, 2006) or the fear of psychopathic individuals (Tinwell, Abdel Nabi, & Charlton, 2013). Pursuant to these evolutionary psychological approaches, the aversion against human-like entities with slight defects might serve as part of a behavioral immune system (Schaller & Park, 2011), shielding individuals against potential dangers to themselves or their progeny.

Concurrently, another research direction has put aside evolutionary factors in favor of an underlying cognitive dissonance effect

* Corresponding author.
E-mail address: jan-philipp.stein@phil.tu-chemnitz.de (J.-P. Stein).

http://dx.doi.org/10.1016/j.cognition.2016.12.010
0010-0277/

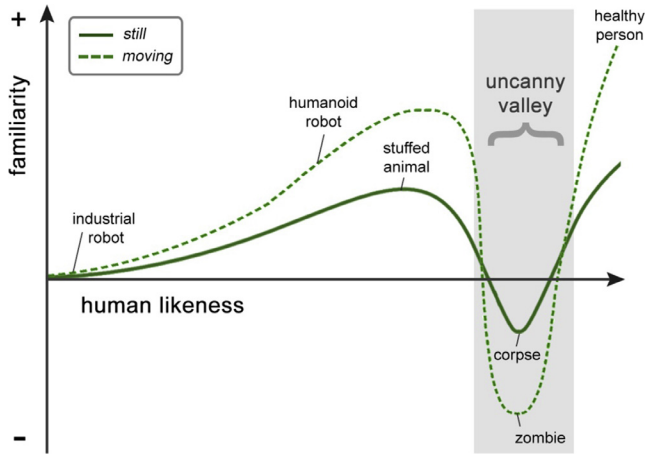

Fig. 1. Uncanny valley model (redrawn from Mori, 1970).

as explanation for the uncanny valley (Ramey, 2005; Yamada, Kawabe, & Ihaya, 2013). This theory builds upon the paradigm that people use a combination of perceptual cues and former experiences to categorize a subject (e.g., as "human" or "robot") so that they can efficiently anticipate its behavior. Once they encounter an entity that violates their expectations, however, observers are likely to experience cognitive dissonance, which then manifests emotionally as uneasiness, disgust, or fear. Notably, this line of thought corresponds to one of the first definitions of the "uncanny" term by German psychologist Ernst Anton Jentsch, who coined it as an eerie sensation arising from "doubts about the animation or non-animation of things" (Jentsch, 1906, p. 204). More than a hundred years later, Jentsch's conceptualization has become firmly embedded in the natural sciences, as studies applying eye-tracking and neuroimaging methods continue to support the cognitive dissonance hypothesis (Cheetham, Pavlovic, Jordan, Suter, & Jancke, 2013; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012). At the same time, literature has remarked upon stimulus novelty as an essential factor for mental categorization conflicts (Grinbaum, 2015); given multiple interactions, people should be able to form new templates for elements that have repeatedly defied expectations, resulting in the "infill" of previously prevalent uncanny valleys. On the other hand, with analogue and digital human simulations advancing constantly, categorization conflicts might just shift to higher levels of realism, as people get increasingly sensitive in detecting visual flaws (Tinwell & Grimshaw, 2009).

1.1. Mind in a machine

Apart from the many studies on visual influences, a large body of research has demonstrated that the attribution of certain mental capacities (such as goal direction and interactivity) is also an important factor in the perception of an entity's animacy and therefore its categorization (Fukuda & Ueda, 2010; Tremoulet & Feldman, 2006). As most modern computers and robots can provide an animate impression by acting in seemingly goal-directed ways, people have been shown to "apply social rules and expectations" to them (Nass & Moon, 2000, p. 87), inferring ideas about a machine's "personality" or some form of *digital mind*. However, research has also indicated that people tend to attribute only one of two mind dimensions to non-human entities: Unlike *experience* (defined as the ability to feel), they merely ascribe *agency* (the ability to plan and act) to their technology, reserving the former as a distinctively human trait (Gray, Gray, & Wegner, 2007; Knobe & Prinz, 2008). Even more so, a pioneering experiment by Kurt Gray and Daniel Wegner has illustrated that blending this differentiation—by presenting a "feeling" computer system, even without mention of a human-like appearance—could lead to significant unease among participants (Gray & Wegner, 2012). In another study of the same paper, the authors found that a human subject bereft of any emotions was also rated as eerie, hinting at the possible uncanniness of emotional experience from the other side of the man-machine continuum. Following this groundwork, research about job replacements by robots has shown that people feel increased discomfort if they consider losing an emotion-related job to a machine, rather than one that relies on cognitive tasks (Waytz & Norton, 2014). In contrast to this, studies on embodied conversational agents in training contexts have indicated that people might actually prefer a digital character that expresses emotions to a neutral counterpart (Creed, Beale, & Cowan, 2014; Lim & Aylett, 2007). Recent findings from the field of social robotics even suggest that people may only rely on visual cues to assess a human-like entity, taking its presumed mental abilities into little consideration (Ferrari, Paladino, & Jetten, 2016).

Undoubtedly, the diversity of these results invites further investigation of the circumstances under which attributions of mind place a creation into an uncanny valley. It seems particularly necessary to explore different facets of artificial minds that eventually contribute to observers' discomfort. As research has indicated that people feel anxious about machines *expressing* their own emotional experience (Gray & Wegner, 2012), it stands to reason to focus next on machines that also *understand* emotional experience in others—considering that feelings are rarely confined to a single consciousness, but serve a social function between individuals (Frijda & Mesquita, 1994). Therefore, a scenario in which digital entities recognize emotional states and react to them in a socially aware manner should shed new light on the *uncanny valley of mind*—a phenomenon that might, after all, relate to a basic understanding of human uniqueness.

1.2. Threats to human distinctiveness

Throughout history, many cultures have regarded a consciousness enriched by emotional states as inherently human domain, closely related to philosophical concepts like a person's *spirit* or *soul* (Gray, 2010). Although theology, natural sciences and social studies vary in their understanding of artificiality and spiritual essence, the Cartesian interpretation of humans as "ghosts" in (bodily) machines has been a prominent philosophical consensus for many, especially Christian, civilizations (Fuller, 2014). Influenced by countless myths about golems, homunculi and other revolting creations, "the Western man puts all his pride in [a] *delta* which is supposed to be specifically human" (Kaplan, 2004, p. 477)—a mental (and, to some, spiritual) component that clearly distinguishes humans from other beings. Considering the long-standing prevalence of this worldview, it can be argued that many people would sense a fundamental threat to their identity—their differentia specifica—if previously "soulless" machines began to share their more complex mental abilities. In consequence of this *threat to human distinctiveness* hypothesis, the aversion against intelligent non-humans constitutes a sociocultural form of threat avoidance (MacDorman & Entezari, 2015), which serves to protect not only the individual, but also humanity in general. As culture studies reveal a more generous conceptualization of the "soul" in East Asian societies (Kaplan, 2004), this theory also accounts for the higher robot acceptance in countries like Japan; their inhabitants, influenced by everyday Buddhism and Shintoism, might simply be more accepting of "spirited" machines instead of feeling replaced or violated (Borody, 2013; Gee, Browne, & Kawamura, 2005).