Cognitive Conflict as Possible Origin of the Uncanny Valley

Patrick P. Weis & Eva Wiese
George Mason University

In social robotics, the term *Uncanny Valley* describes the phenomenon that linear increases in human-likeness of an agent do not entail an equally linear increase in favorable reactions towards that agent. Instead, a pronounced dip or 'valley' at around 70% human-likeness emerges. One currently popular view to explain this drop in favorable reactions is delivered by the Categorical Perception Hypothesis. It is suggested that categorization of agents with mixed human and non-human features is associated with additional cognitive costs and that these costs are the cause of the Uncanny Valley. However, the nature of the cognitive costs is still matter of debate. The current study explores whether the cognitive costs associated with stimulus categorization around the Uncanny Valley could be due to cognitive conflict as evoked by simultaneous activation of two categories. Using the mouse tracking technique, we show that cognitive conflict indeed peaks around the Uncanny Valley region of human-likeness. Our findings lay the foundation for investigating the effects of cognitive conflict on positive affect towards agents of around 70% human-likeness, possibly leading to the unraveling of the origins of the Uncanny Valley.

INTRODUCTION

The Uncanny Valley (UV) hypothesis is a frequently discussed phenomenon in social robotics. It refers to the observation that human-like robotic agents are perceived more negatively than both completely human or less human-like robotic agents. In other words, human-likeness of an agent is not linearly related with its likability but shows a pronounced dip or 'valley' at around 70% human-likeness (Mori, MacDorman, & Kageki, 2012). While there is growing empirical evidence for the existence of the UV (e.g. Mathur & Reichling, 2016), there is no consensus about its theoretical underpinnings (Kätsyri, Förger, Mäkäräinen, & Takala, 2015; for a review). One suggestion is that very human-like but not perfectly human agents are likely to be confused with morbid or dead bodies, and are hence evaluated less favorably than less human-like or perfectly human-like agents (*Morbidity Hypothesis*; Mori et al., 2012). Another suggestion is that the UV is caused by expectations about the arrangement of features on a human face or body, which are violated by mixed-in non-human perceptual features in most artificial agents (*Perceptual Mismatch Hypothesis*; MacDorman, Green, Ho, & Koch, 2009). Lastly, it was suggested that agents with mixed human and non-human features are hard to categorize and that additional cognitive costs associated with this categorization are what causes the UV (*Categorical Perception Hypothesis;* Cheetham, Suter, & Jäncke, 2011*)*. While being closely related, the categorical perception hypothesis attributes the UV to a resource-draining categorization process and not to more general violations of perceptual expectations like the perceptual-mismatch hypothesis does. Currently, empirical evidence favors the perceptual-mismatch and categorical-perception hypotheses over the morbidity hypothesis (Kätsyri et al., 2015).

The categorical-perception hypothesis is based on the assumption that every time we are exposed to a face-like stimulus, we automatically ask ourselves whether this face represents a human or non-human being. Support for this assumption comes from evolutionary biology, linking the human tendency to categorize to survival, and the failure to categorize a stimulus to negative emotional responses (Burleigh & Schoenherr, 2015). Thus, the assumption underlying the categorical-perception hypothesis is that the harder it is to categorize a stimulus as human versus non-human, the more strongly exposure to this stimulus should evoke negative emotional responses.

One way to investigate categorical perception is by using morphs: a picture of category A is morphed into a picture of another category B, resulting in a sequence of stimuli that gradually decrease in "A-likeness" and increase in "B-likeness". Participants then have to categorize these stimuli as belonging either to category A or to category B. From morphing studies using "human" as category A and "robot" as category B, we know that perception of humanness indeed follows a categorical pattern, with the categorical boundary located at around 63% humanness (Cheetham et al., 2011; Martini, Gonzalez, & Wiese, 2016). This means that pairs of stimuli straddling the boundary between human and non-human are easier to discriminate than equally similar pairs located on the same side of the boundary (Cheetham et al.,

Copyright 2017 by Human Factors and Ergonomics Society. DOI 10.1177/1541931213601763

2011). It however also means that stimuli located around the category boundary are harder to categorize (i.e. longest reaction times: Mathur & Reichling, 2016), with potentially negative effects on performance on tasks that include these stimuli due to increased cognitive load.

These studies show that the categorical boundary for perceiving humanness is located in proximity to the UV. However, what has not been shown is whether categorization difficulty indeed causes a cognitive conflict associated with increased cognitive load. Indirect measures of cognitive conflict (i.e., reaction times) were not able to provide evidence that categorizing agents falling in the UV induce cognitive conflicts (Mathur & Reichling, 2016). However, while reaction times provide good estimates for the general difficulty of a task, they are not able to directly capture whether or not task performance is associated with conflict processing. A better measure to capture conflict processing during categorization is mouse tracking (e.g., Freeman & Ambady, 2010). The method allows assessments of cognitive conflict processing via three different variables: area under the curve (AUC), maximum perpendicular deviation (MD), and the number of reversals of direction along the axis of decision (x-flips). All three variables are based on the trajectories of a mouse cursor that is used to categorize a stimulus. The variables are explained in detail in Freeman & Ambady (2010). For a brief explanation of x-flips, the only of these variables used in the current paper, see the section *Task and Mouse-tracking*.

The current study explores whether stimulus categorization leads to heightened cognitive conflict for agents falling in the UV. Our analysis focuses on x-flips, as they are the most reliable measure of stability or instability of category activation dynamics (Freeman & Johnson, 2016), and hence the most direct measure of cognitive conflict as evoked by resource-draining categorization processing. We expect to find the strongest cognitive conflict processing in the ultimate proximity of the UV. In particular, we expect to find a higher number of x-flips for agents falling in the UV compared to agents that are located further away from the UV on the morph spectrum. In addition, we explore whether individual tendencies to treat non-human agents as human-like (i.e., anthropomorphism; Epley, Waytz, & Cacioppo, 2007 ; for a review) modulates the degree of cognitive conflict agents falling in the UV induce. The hypothesis is that participants with a high tendency to anthropomorphize should be more willing to treat a wide range of stimuli as human and therefore should show a less pronounced cognitive conflict. Individual tendencies to treat non-human agents as human-like are measured using the IDAQ questionnaire (i.e., *Individual Differences in Anthropomorphism Questionnaire*; Waytz, Cacioppo, & Epley, 2010).

METHODS & MATERIALS

Participants

15 students (2 male; *M* = 20.9 years, *SD* = 2.9, range: 18-28) from George Mason University participated in this study in partial fulfillment of a course requirement. All participants reported English language proficiency and normal or corrected-to-normal visual acuity. The experimental session took twenty minutes and was approved by the local Institutional Review Board.

Stimuli

We created nine different robot-human spectra using the morphing software FantaMorph, each consisting of pictures set apart by 5% morphing steps, resulting in 21 stimuli for each spectrum (see **Figure 1** for an example spectrum) or a total of 189 stimuli. Each spectrum was based on one photograph of the face of a unique human and one photograph of the face of a unique robot. Robot photographs were obtained from Mathur and Reichling (2016), and followed-up by a Google image search. Only photographs from robots with human-like faces (i.e., having eyes and nose) were included in the morph spectra. After deciding on the robot photographs, we matched them on apparent gender, head orientation, and facial features with a photograph from the MUCT human face database (Milborrow, Morkel, & Nicolls, 2010). All pictures were cropped to the same aspect ratio and rescaled to 450 x 450 pixels.

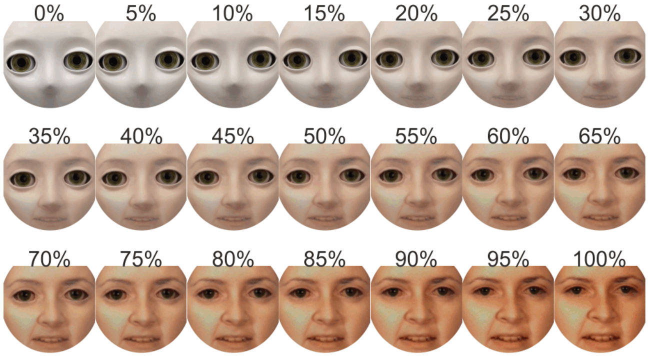

Figure 1. Stimuli: Example of a robot-human spectrum. A robot image (top left) was morphed into a human image (bottom right) in steps of 5%, resulting in a set of 21 stimuli per spectrum.

Task and Mouse-tracking

Participants were asked to categorize the morph images as human versus non-human (one picture at a time; order randomized throughout the experiment). To do so, participants had to click a start button located in the bottom center of the screen to make the image appear, and then move the mouse cursor from the bottom center of the screen to the respective answer box (human versus non-human); see **Figure 2**. Clicking the chosen answer box concluded the trial, which was marked by a blank screen presented for 1000 ms (i.e., inter trial interval, ITI). The answer boxes were always shown from the beginning of a trial, while the agent image appeared after the start button was pressed to assure that participants started with mouse movement right after image presentation.

Cognitive conflict processing was recorded via mouse movements (Freeman & Ambady, 2010), in particular x-flips (i.e., number of reversals of direction along the axis of decision; see red circles in **Figure 2**). X-flips have been shown to be the most reliable measure of stability or instability of category activation dynamics (Freeman & Johnson, 2016). In our setup, the axis of decision is the horizontal x-axis since the participant's categorization process (non-human = left; human = right) is based on the horizontal position of the mouse cursor; see **Figure 2.**

Design and Procedure

The experiment followed a one-factorial design with the within-participants factor *physical humanness*. Humanness was manipulated on a spectrum from robot-like to human-like in steps of 5%, resulting in a total of 21 levels of humanness per humanness spectrum. To increase external validity and to minimize artifacts coming from specific human or robot images, nine different robot-human spectra were created. Images were presented at the bottom center of the screen until participants made a decision (human versus non-human). Each image was presented once per participant, with the order of the images being randomized across the experiment.

At the beginning of the experiment, participants were seated in front of a computer and signed the informed consent form before being instructed to decide if a picture is either human or non-human by clicking on the respective word as fast as possible. After participants had read the instructions, they were asked to perform three practice trials to familiarize themselves with the mouse-tracking procedure. For the practice trials, morphed images were used that were not part of any of the experimental morph spectra. Upon completion of the practice trials, the main experiment began during which participants categorized all 189 agents as either human or non-human using the computer mouse. The main task took about 15 minutes to complete. Upon completion of the main experiment, participants were asked to fill out the IDAQ (Waytz et al., 2010), which measured the participant's individual inclination to attribute a mind to something non-human. After having completed the questionnaire, participants were informed about the purpose of the experiment and received their course credit before the session concluded.

Analysis

Trials with extreme reaction times deviating more than 2.5 standard deviations from the individual mean were excluded from analysis, which led to an exclusion of 0.3% of all trials. To investigate whether categorizing agents along the given spectra resulted in a categorical pattern, a three-parameter logistic function was fitted to the data of each participant and a one-sample t-test on the growth parameter was employed to test deviation from linearity (a procedure that has already been employed by other authors; Cheetham et al., 2011).

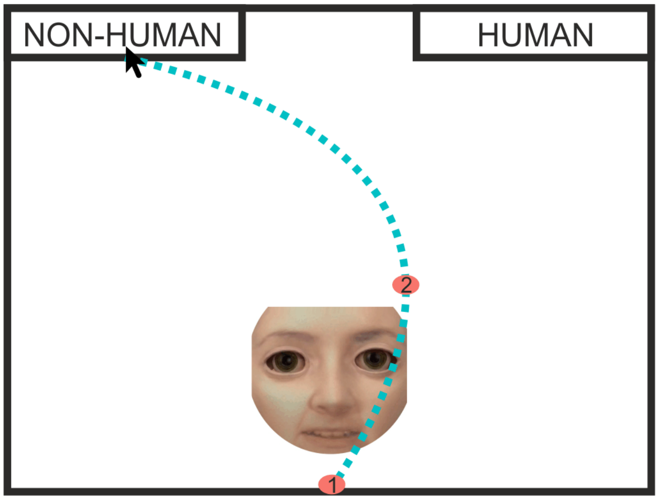

Figure 2. Example Trial: After pressing the start button, an agent image appeared on the screen (center, bottom) and participants were asked to categorize the image as either human or non-human by moving their mouse cursor to one of the two answer boxed (top left and right). The blue line shows an example trajectory. The red circles represent the locations of abrupt changes in the horizontal direction (x-flips), with higher number of x-flips indicating higher cognitive conflict processing.

To investigate whether categorizing agents falling in the UV resulted in stronger conflicts than categorizing agents located further away from the UV, a two-step procedure was employed: first, a one-factorial repeated measures ANOVA with the factor physical humanness and the dependent variable number of x-flips (as a measure of cognitive conflict) was used to investigate if physical human-likeness affected cognitive conflict processing. Second, a correlational analysis on individual categorical boundaries and individual maxima of cognitive conflict was conducted to support our hypothesis that maximal cognitive conflict tends to occur around each participant's categorical boundary. Lastly, to explore the effect of trait anthropomorphism, we conducted a median split on the IDAQ scores and plotted the resulting classification function for both groups. Testing differences between low and high anthropomorphism groups with inferential statistics is planned after doubling our sample size.

RESULTS

Participants exhibited a step-like, in contrast to a linear, function when categorizing robot-human morphs into non-human or human as shown by a one-sample t-test on the growth rate parameters of the individual three-parameter logistic fits (t(14) = 6.94, p < 0.001); see **Figure 3A**. Since the fitting procedure failed due to a non-sigmoid shape of the data for one participant, we used a growth rate of 0 for this participant.

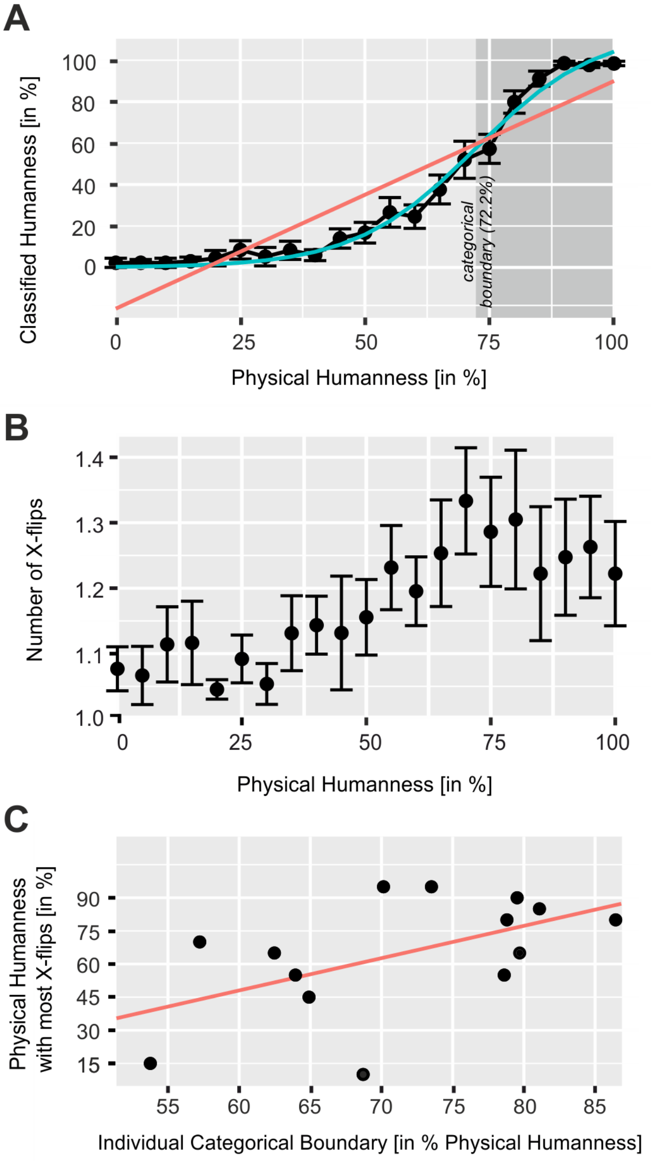

Figure 3. Categorization and Cognitive Conflict: (A) *Categorization data*. Grand average empirical data are depicted in black; a linear model fit is shown in red, a logistic fit in blue. The categorical boundary (i.e., the value of physical humanness associated with the steepest change in classified humanness) is depicted by a change in background brightness. Error bars depict standard errors. (B) *X-flip data*. Physical humanness alters cognitive conflict. The highest cognitive conflict is measured around the categorical boundary. Error bars depict standard errors. (C) *Relation between individual categorical boundaries and cognitive conflict maxima*. The higher the categorical boundary is on the spectrum of physical humanness, the higher the cognitive conflict maximum (measured by individual x-flip maxima); red line shows linear fit.

Physical humanness altered cognitive conflict as shown by a one-factorial repeated-measures ANOVA (*F*(20, 280) = 2.17, *p* = .003), with the factor physical humanness and the dependent variable number of x-flips; see **Figure 3B**. As hypothesized, the instability is, at least descriptively, the greatest in proximity of the uncanny valley (i.e. around 70%, compare to e.g., Cheetham, Pavlovic, Jordan, Suter, & Jancke, 2013). To further test this hypothesis, we extracted the categorical boundary as well as the point of maximal cognitive conflict and ran a correlational analysis; see **Figure 3C**. We defined the categorical boundary as the point in classified humanness where the sigmoidal fit reaches its maximal slope (compare **Figure 3A** for the grand average categorical boundary). The point of maximal cognitive conflict is defined as the point in classified humanness where the highest categorization instability as measured by x-flips was exhibited. Since the categorical boundary relies on a sigmoidal fit, the one subject that did not exhibit a sigmoidal fit was excluded from this particular analysis. The correlational analysis supported our hypothesis (Pearson's r = 0.54, *t*(12) = 2.19, *p* = 0.0482). In conclusion, both ANOVA and correlation analysis findings are in congruence with our hypothesis that the uncanny valley represents the area of maximal cognitive conflict as evoked by categorization instability.

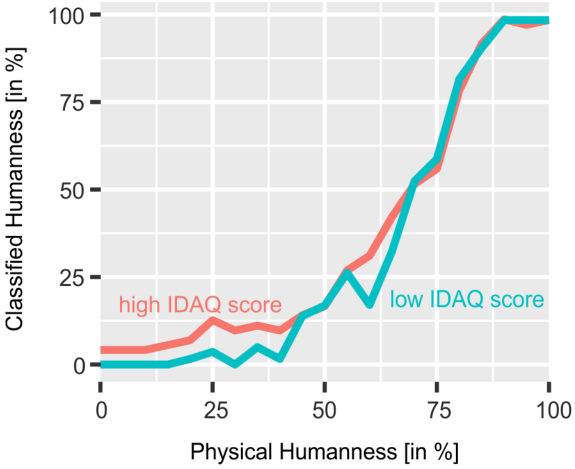

Figure 4. Humanness ratings by IDAQ groups: Eight subjects in high IDAQ group (>= median), seven in low IDAQ group (<median). IDAQ: Individual Differences in Anthropomorphism Questionnaire (Waytz et al., 2010).

Lastly, we are reporting the relation between the inclination to anthropomorphize as measured by the IDAQ and the humanness ratings; see **Figure 4**. Descriptively, from looking at the data, we suggest that the categorical boundary of individuals with high IDAQ scores is both shifted to the left and less pronounced (i.e., less steep) when compared to individuals with low IDAQ scores, as would be expected. We are planning on testing this hypothesis with inferential statistics after increasing sample size.

DISCUSSION

The current study explored whether the cognitive costs associated with stimulus categorization around the Uncanny Valley could be due to cognitive conflict as evoked by simultaneous activation of two categories. Using the mouse tracking technique, we showed that cognitive conflict indeed peaks around the Uncanny Valley region of human-likeness. A preliminary analysis additionally suggests that participants with high inclination to anthropomorphize show a less pronounced categorical boundary and thus less cognitive conflict than participants with low inclination to anthropomorphize.

Our results are in line with the hypothesis that the UV is caused by difficulties categorizing agents of around 70% human-likeness (Cheetham et al., 2011). This study is to our knowledge the first to directly measure the cognitive conflict resulting from an ambiguous categorization process, relying on the mouse tracking technique instead of on more indirect measures like reaction time. In other words, it was hitherto not clear if the categorical boundary observed in proximity of the uncanny valley (Cheetham et al., 2011) is accompanied by an increase in cognitive conflict.

While originally tailored to further investigate the categorical perception hypothesis, the current research is also well consistent with the perceptual mismatch hypothesis. Thus, it is not clear if the reported cognitive conflict maxima are indeed caused by high-level categorical processing or, as the perceptual mismatch hypothesis would claim, by lower-level effects due to co-occurring unexpected compositions of artificial and human-like facial features. Relatedly, the cognitive conflict maxima might also capture problems with fluent stimulus processing. Low processing fluency has shown to be related with negative evaluations of the respective stimulus (Winkielman et al., 2003) and could thus be similarly related to the negative evaluations in the UV.

In all cases however, the reported findings lay the foundation for investigating the effects of cognitive conflict on affective evaluations of agents at around 70% human-likeness. It thus is to be researched if the Uncanny Valley might be originating from negative emotions as evoked by cognitive conflict.

REFERENCES

Burleigh, T. J., & Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01488

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category Processing and the human likeness dimension of the Uncanny Valley Hypothesis: Eye-Tracking Data. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00108

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.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., & Johnson, K. L. (2016). More Than Meets the Eye: Split-Second Social Perception. *Trends in Cognitive Sciences*, *20*(5), 362–374.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4392592/

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior*, *25*(3), 695–710.

Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing Minds in Others – Can Agents with Robotic Appearance Have Human-Like Preferences? *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146310

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32.

Milborrow, S., Morkel, J., & Nicolls, F. (2010). The MUCT landmarked face database. *Pattern Recognition Association of South Africa*, *201*(0). Retrieved from http://www.dip.ee.uct.ac.za/~nicolls/publish/sm10-prasa.pdf

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Winkielman, P., Schwarz, N., Fazendeiro, T., Reber, R., Musch, J., & Klauer, K. C. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, 189–217.