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

Chemnitz University of Technology

Formal publication / citation:

Stein, J.-P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, 160, 43–50. doi: 10.1016/j.cognition.2016.12.010

Acknowledgement. This study was funded by the German Research Foundation (DFG) under grant 1780 ("CrossWorlds: Connecting Virtual and Real Social Worlds"). We thank Alexandra Jost for her help in the collection of data.

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.

Keywords: uncanny valley, theory of mind, social cognition, anthropocentrism, artificial intelligence, virtual reality

Venturing into the Uncanny Valley of Mind—The Influence of Mind Attribution on the Acceptance of Human-Like Characters in a Virtual Reality Setting

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).

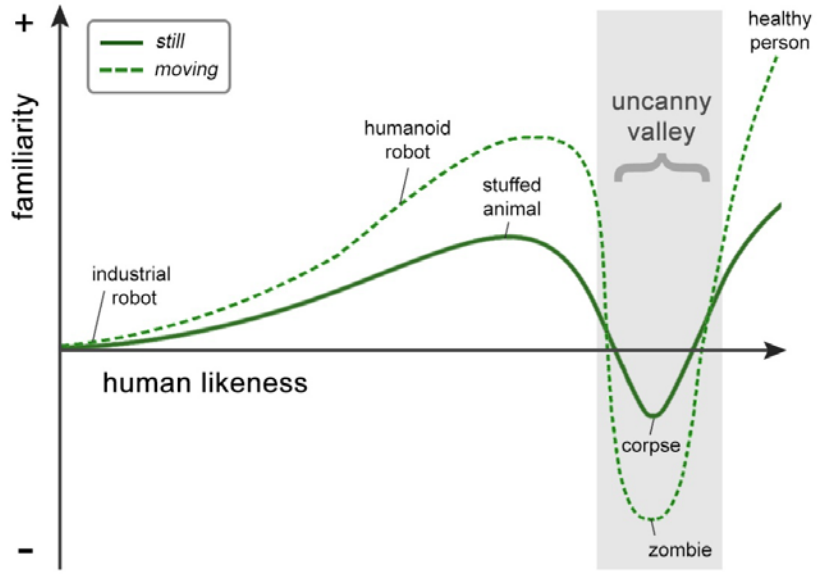

Figure 1. Uncanny valley model (redrawn from Mori, 1970)

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 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).

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, 2005). 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.

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).

However, not only the possession of mental states, but also the ability to ascribe them to oneself and others—known as *social cognition* or having a *theory of mind* (Premack & Woodruff, 1978)—has been discussed as essential difference between humans and other creations (Adolphs, 1999; Gallagher & Frith, 2012; Pagel, 2012; Vogeley & Bente, 2010). Although studies continue to present evidence for a basic theory of mind in some animal species (Call & Tomasello, 2008; Tomonaga & Uwano, 2010), the declaration of humans as "pride of creation" due to abilities like perspective-taking and empathic processing remains widespread. More recently, research on the role of a mirror neuron system in human neurophysiology has offered scientific footing to the anthropocentric idea of human uniqueness (Azar, 2005; Iacoboni, 2009), albeit not without controversy (Spaulding, 2013). From a more philosophical standpoint, the idea of superior human minds might even emerge as a species-related form of narcissism, which arguably presents itself in the careless

destruction of other creations. But no matter if mental states and the ability to ascribe them are interpreted as a natural product of neurological processes, spiritual privilege or *the* essence of human exceptionalism, it seems likely that the human identity would suffer severe consequences if virtual entities demonstrated their own, sophisticated theory of mind. Even more than two decades ago, when artificial intelligence was far less refined than it is today, scientists worried about diffusing the long-standing dichotomy of man and machine, and advised caution in the development of new human-like features (Nass, Lombard, Henriksen, & Steuer, 1995).

The Current Study

To explore the presented reasoning, this paper focuses on the perception of emotions and social cognition in a human replica as primary cause for an uncanny valley response. Following the theoretical groundwork, we devised an experiment that manipulated the mind attribution to human-like characters, while keeping constant their visual appearance and verbal expressions. Several groups of participants observed the same friendly and empathic dialogue scene of two virtual characters, but received different instructions as we claimed the 3D models to be either human-controlled "avatars" or computer-controlled "agents". Secondly, we manipulated the alleged autonomy of the characters, stating the dialogue to be either a work of the controller's "own imagination" or an intensely prepared script. In summary, this resulted in a 2×2 factorial design with the conditions "human, scripted", "human, autonomous", "computer, scripted" and "computer, autonomous".

According to the interpretation of social cognition as distinct human privilege, we expected an alleged artificial intelligence ("computer, autonomous") that shows awareness of another character's emotions to strongly violate category expectations. Specifically, we theorized that only such an autonomous agent—but not a scripted one—would be attributed the mental processes underlying its empathic behavior, and that this *digital social cognition*

would appear uncannily human. At the same time, we suspected a human avatar that only acts as a "vessel" for scripted content to cause more eeriness than an autonomously acting one, as it approaches the category border between human and non-human from the other side of the uncanny valley.

H1a: *People will perceive autonomous virtual agents that display emotions and social cognition as more eerie than scripted virtual agents.*

H1b: *People will perceive autonomous human avatars that display emotions and social cognition as less eerie than scripted human avatars.*

Taking inspiration from the work by Gray and Wegner (2012), our assumptions intended to advance their notion of uncanny minds by turning the emotional computer into a *social entity*—a machine that perceives emotions, interprets them, and adapts its behavior accordingly. As we expected a particularly strong aversion against such a category-defying creation, we further hypothesized that autonomous virtual agents would receive the lowest attractiveness rating due to the subconscious impulse to avoid further contact.

H2: *People will perceive autonomous virtual agents that display emotions and social cognition as more human-like than scripted virtual agents.*

H3: *People will perceive autonomous virtual agents that display emotions and social cognition as least attractive among the four groups.*

Methods

Although the method of juxtaposing human avatars with virtual agents has been well-known to gaming research (Weibel, Wissmath, Habegger, Steiner, & Groner, 2008; Lim & Reeves, 2010) and persuasion studies (Fox et al., 2014; Guadagno, Blascovich, Bailenson, & McCall, 2007; Patel & MacDorman, 2015), we chose to apply it to a social media scenario,

assuming that most participants would be familiar with this specific domain. Also, with social network services turning into one of the most influential media branches (Ngai, Tao, & Moon, 2015), they have emerged as a particularly relevant platform for future uncanny valley occurrences. To increase the study's ecological validity even further, we chose to make use of head-mounted display (HMD) technology, which has remained in the focus of software and hardware developers in recent years (Benner & Wingfield, 2016; Zuckerberg, 2014). Since stereoscopic HMDs can visualize a highly immersive 3D environment, we hoped that participants would be more susceptible to our deceptive instructions than they would have been in front of a 2D screen, especially in the artificial intelligence condition.

Participants

We recruited 97 students at a German university (30 male, 67 female; age: M = 23.6 years, SD = 3.32) as participants for our 30 minute experiment. Although the recruitment process targeted students from a variety of study programs, those willing to participate were predominantly enrolled in media communication and education studies. With the exclusion of two participants who reported moderate simulator sickness from the applied VR technology, as well as three participants who failed the final manipulation check, the final sample consisted of 92 students (29 male, 63 female; age: M = 23.6 years, SD = 3.41). Participants received €5 or partial course credits for taking part in the experiment. We assigned them to one of the four experimental groups by means of block randomization and provided them with extensive consent forms at the beginning of their appointments. Since our study design featured several deceptive elements, each participation ended with a thorough debriefing about the real nature of the used materials and the aim of the study.

Stimuli

To create a customizable platform for the experiment, we programmed our own fictitious chat software with the game engine Unity (Version 5.2.0, Unity Technologies) and

its built-in VR support. While not featuring any actual chat functionality, the program contained several mock-up elements such as login and notification sounds, as well as eight different virtual characters. Apart from displaying it on a 2D computer screen, our application could be accessed through the HMD "Oculus Rift DK2", which completely blocks out peripheral visual information and records the user's head movement in order to provide them with 360-degree gaze orientation.

In its first stage, the chat program displayed a mostly empty urban plaza with only a couple of non-interactive characters walking around in a distance. After a 30-second waiting period, which allowed participants to adjust to the immersive VR environment, two high-resolution character models—one female, one male, both middle-aged—appeared in front of the participant's point of view and approached each other. Constituting the main part of the experiment, the two characters would then engage in a casual dialogue, which we had completely scripted beforehand. Participants were told to observe the conversation silently and not to interfere for standardization reasons. Although the final version of our software allowed for two-directional movement, we did not provide mouse or keyboard to prevent motion sickness and to restrict differences in the visual stimuli to basic head movements.

The conversation scene, which lasted for 140 seconds, was composed to include expressions about the general mood ("I'm feeling droopy"), temperature sensitivity ("Today is really hot") and hunger ("Yeah, I'm quite hungry, too") of the two dialogue partners. In order to operationalize their ability for social cognition, we scripted both characters to acknowledge each other's statements and to empathize with them (e.g., "That must be annoying for you."). As we intended to use the same scene in all four experimental conditions, we presented the spoken dialogue with a slight electronic distortion that could appear as either artificial voice (of a virtual agent) or inferior sound quality (of a human speaker controlling an avatar). Additionally, both characters displayed a small set of gesture animations during their speech,

which we had implemented using a license-free sample of motion capturing data (Carnegie Mellon University Graphics Lab, 2015). This body language included hand movements such as waving or the wiping of sweat (Fig. 2), as well as simple head movements. To avoid drawing too much attention to the details of our deception, we applied the animations in a restrained manner and told participants that the software could only provide a small amount of expressions, triggered by the human user's press of a button or the agent's programming, respectively. At the end of the scene, the characters would decide on visiting a virtual café, walk past the observer's point of view and fade out acoustically.

Figure 2. Screenshot from the presented VR scene

Procedure

According to our study design, the instructions given prior to the 3D scene provided the only, yet crucial, manipulation in the experiment. As such, we facilitated the introduction to the different conditions with several steps. After each participant had received the same technical briefing about our VR software, which "could be used to meet friends and chat with

them in real-time", we told them to observe either two human confederates or a new set of virtual agents—both framed as a basic "beta test" of the platform's appeal. With the additional variation of the characters' alleged autonomy, this resulted in four different instruction narratives (Fig. 3).

		alleged autonomy	
		scripted	autonomous
alleged identity	human	*"… a chat our confederates will present according to script, word by word"*	*"…a chat our confederates will improvise"*
	computer	*"…basic agents programmed with an example script"*	*"… intelligent agents with randomized emotional parameters and content processed in real-time"*

Figure 3. The study's 2 × 2 factorial design and excerpts from the corresponding narratives

In order to prime participants for the deceptive scenarios, we then asked them to read a single-page newspaper article about virtual reality and its applications while we pretended to log into our software's web server. Only those in the "computer, autonomous" condition received the article with three additional lines of text, telling them about recent breakthroughs in the field of artificial intelligence, neural network technology and machine learning. Apart from the statement that "AI systems could now process dialogue in real-time, using word databases and emotional algorithms", we also drew attention to their ability for social cognition, emphasizing that modern systems could interpret emotional cues and demonstrate "surprising levels of empathy".

In the conditions with supposedly human-controlled characters, we instead facilitated our deception by presenting a rigged intercom system. Doing so, we were able to stage the

coordination with two confederates, who were supposed to control the avatars from another laboratory.

Measures

We asked participants to fill out a 15-minute questionnaire immediately after the 3D scene had ended. The first part of the questionnaire comprised the *eeriness* (eight items, $\alpha = .85$), *human likeness* (six items, $\alpha = .84$) and *attractiveness* (five items, $\alpha = .81$) scales by Ho and MacDorman (2010), which feature 19 semantic differentials developed specifically for research on the uncanny valley. For sufficient differentiability, we presented the items (e.g., "reassuring—eerie" and "artificial—lifelike") in a 7-point answer format. We asked participants not to rate each character separately, but to report their combined impression of both characters in order to average possible gender biases.

Subsequent to the emotional evaluation, participants completed the 16 items of the Simulator Sickness Questionnaire (SSQ; Kennedy, Lane, Berbaum, & Lilienthal, 1993), a well-established instrument to register technology-induced symptoms of nausea, disorientation, and oculomotor dysfunction. Using the suggested cutoff value for "severe discomfort", we ensured that participants with strong physiological reactions to the HMD technology were excluded from the final set of data, as their ratings might have been confounded with unpleasant somatic side effects.

Further questions examined previous experience with social media and VR to monitor these variables for potential outliers, as well as a short socio-demographic inquiry. Lastly, we conducted a manipulation check that assessed the perception of the characters' behavioral autonomy ("The chat partners act on their own accord"), emotional autonomy ("The chat partners possess their own feelings"), and social competence ("The chat partners are socially competent") on 5-point scales. As a final question, participants had to answer explicitly who

they thought had controlled the two dialogue partners; choosing the option that did not match our instructions led to the exclusion from the study.

Results

Manipulation Check

After the exclusion of two people who had experienced unpleasant side effects from the VR display, as well as three participants who had not confirmed the alleged identity in our final question, we calculated separate one-way ANOVAs for the three manipulation check items.

Regarding the behavioral autonomy rating, our analysis resulted in significant differences between conditions, $F(3,88) = 5.90$, $p = .001$. Post-hoc LSD tests revealed that participants in the "computer, autonomous" group indeed perceived more freedom of action in the characters ($M = 2.64$, $SD = 1.00$) than those in the "computer, scripted" condition ($M = 1.83$, $SD = 0.72$), $p = .004$. Similarly, differences between the "human, autonomous" ($M = 2.83$, $SD = 1.03$) and "human, scripted" conditions ($M = 2.13$, $SD = 0.85$) turned out significant, $p = .010$. This means that, for both virtual agents and human avatars, participants perceived the characters' actions as more independent if our instruction had claimed so. We further note that the artificial intelligence ("computer, autonomous") was ascribed nearly as much behavioral autonomy as self-directed humans, indicating the success of this difficult deception.

For emotional autonomy, another one-way ANOVA revealed significant differences between groups, $F(3,88) = 6.31$, $p = .001$. Post-hoc testing showed that the "computer, autonomous" characters were attributed more own feelings ($M = 2.55$, $SD = 1.06$) than the "computer, scripted" characters ($M = 1.78$, $SD = 0.80$), $p = .007$. However, the difference between "human, autonomous" ($M = 2.91$, $SD = 0.95$) and "human, scripted" conditions ($M = 2.63$, $SD = 0.88$) did not emerge as significant, $p = .287$. Thus, participants perceived

emotional autonomy in human controllers irrespective of their alleged spontaneity, whereas only self-directed virtual agents were attributed their own feelings. Arguably, the finding that participants tended to ascribe some form of emotionality to humans, even within the constraints of a verbatim script, matches the idea of emotions as a very basic human quality.

Finally, we compared the perceived social competence between the four conditions. An analysis of variance did not result in significant group differences, $F(3,88) = 1.53$, $p = .212$. Moderate means in every condition—from "computer, scripted" ($M = 2.78$, $SD = 0.95$) and "human, autonomous ($M = 3.09$, $SD = 0.95$) to "computer, autonomous" ($M = 3.18$, $SD = 1.00$) and "human, scripted" ($M = 3.33$, $SD = 0.70$)—suggest that participants generally perceived the characters as empathic and friendly. Combined with the other two manipulation check items, this speaks to the successful manipulation of social cognition perceptions in our experiment. Although the characters appeared socially competent to all groups, only the supposedly autonomous entities were regarded as initiators of the displayed behavior. Therefore, we argue that participants only ascribed the reasons for the empathic conversation—i.e., social cognition—to intelligent virtual agents and human avatars, but not to scripted agents.

Uncanny valley indices

Table 1

Means and standard deviations on the three uncanny valley indices for each condition

	Attribution of mind							
	Human, scripted (n = 24)		Human, autonomous (n = 23)		Computer, scripted (n = 23)		Computer, autonomous (n = 22)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
eeriness	2.97	0.80	2.79	0.97	2.77	0.59	3.43	0.96
human likeness	3.10	0.90	2.96	1.26	3.11	1.04	3.46	1.08
attractiveness	4.86	0.79	4.77	0.88	4.57	0.78	4.98	0.90

Note. Indices range from 1 to 7.

To investigate if different attributions of mind corresponded to differences in our participants' emotional response, we calculated two-factorial analyses of variance for each index of the Ho and MacDorman questionnaire. Checking the requirements for the parametric procedure, the human likeness and attractiveness scores appeared normally distributed in Shapiro-Wilk tests, whereas the eeriness scores in one group ("human, autonomous") deviated slightly from a normal distribution. However, as we found homogeneity of variances for all variables, and every group featured more than 20 participants, we applied parametric tests nonetheless, as they prove robust against small requirement violations when investigating groups of sufficient size (Stevens, 1999).

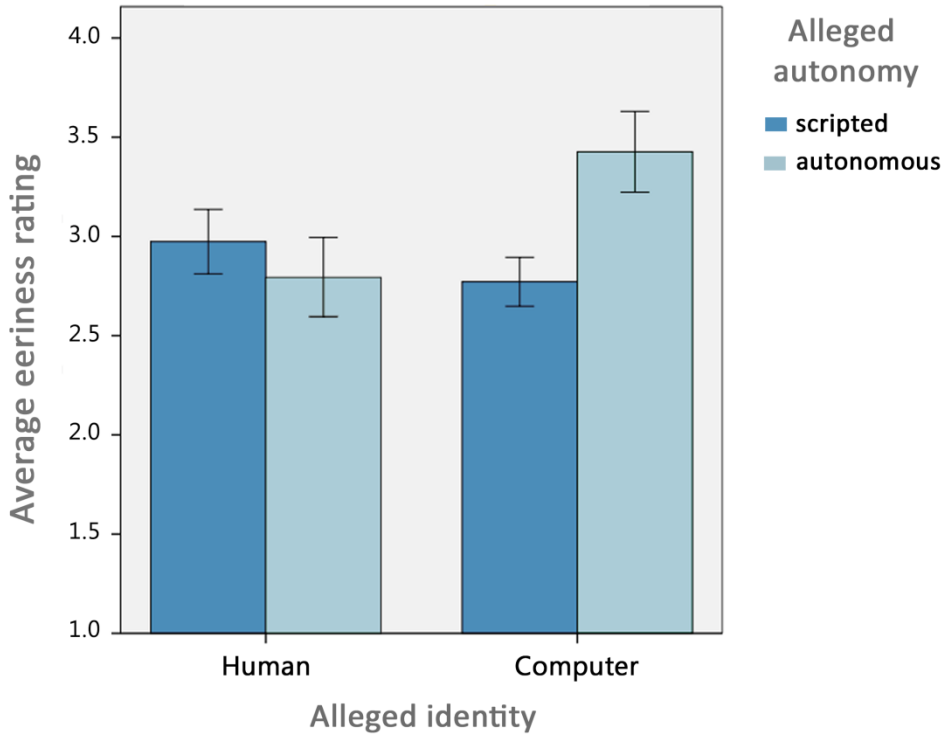

Figure 4. Average eeriness ratings for the different mind attributions (error bars reflect +/- 1 standard error of the means).

Eeriness scores. The two-way ANOVA conducted with participants' eeriness scores showed no significant main effect for the factor "alleged identity", $F(1,88) = 1.51$, $p = .223$,

as well as no significant main effect for the factor "alleged autonomy", $F(1,88) = 1.83$, $p = .180$. However, an interaction between both factors (Fig. 4) emerged as significant, $F(1,88) = 5.68$, $p = .019$ and $\eta_p^2 = 0.06$, constituting a moderate effect (Cohen, 1988). While spontaneously acting humans appeared as less eerie than those following a script, participants experienced it as more uncanny to watch allegedly autonomous agents than scripted ones. As such, we confirm our main hypotheses H1a and H1b.

Human likeness scores. Unlike the analysis of the eeriness ratings, a two-way ANOVA regarding human likeness yielded no significant results. There was no significant main effect for the factor "alleged identity", $F(1,88) = 0.17$, $p = .258$, no significant main effect for the factor "alleged autonomy", $F(1,88) = 1.64$, $p = .647$, and no significant interaction between both, $F(1,88) = 0.48$, $p = .266$. Although we did find slightly higher scores in the "computer, autonomous" than in the "computer, scripted" condition—aligning with the assumption of H2—we cannot support this hypothesis due to a lack of statistical significance.

Attractiveness scores. Similar to the findings for the human likeness ratings, the two-factorial analysis of variance did not produce any significant results for the attractiveness scores, neither a main effect for "alleged identity", $F(1,88) = 0.04$, $p = .847$, nor a main effect for "alleged autonomy", $F(1,88) = 0.81$, $p = .371$, or a significant interaction, $F(1,88) = 2.05$, $p = .156$. Despite our data revealing the highest attractiveness mean in the "computer, autonomous" condition—forming a numerical pattern that contradicts H3—the non-significant results again ask for a cautious interpretation.

Discussion

The presented study aimed at investigating the effects of mind attribution on the well-known uncanny valley phenomenon. In a virtual reality experiment, we equalized the factor *appearance* but manipulated the perception of *emotional and social abilities* in human-like

characters. Building upon previous research, we designed characters that not only expressed their own emotional experience, but also empathized with the mental states of each other. Following different introductions to our stimuli (as scripted virtual agents, autonomous artificial intelligence, or human avatars), we were able to compare the effects of different attributions of mind on observers' emotional response. Statistical analyses revealed a significant interaction of the factors "identity" and "autonomy", as participants found autonomous human avatars that showed signs of social cognition less eerie than scripted ones, but autonomous virtual agents significantly more eerie than their scripted counterparts. In light of the considerable effect size, we present this result as evidence for the uncanniness that arises from emotional computer systems turning into social beings with their own theory of mind.

Unlike our initial assumption, we found no differences for perceived human likeness between the four experimental conditions. We assume that participants have inferred this rating predominantly from visual features. In fact, one of the six items of the corresponding scale by Ho and MacDorman explicitly references the visual nature of the stimuli ("mechanical movement—biological movement"), prompting similar evaluations in all conditions. Another explanation might lie in specific features of our 3D scene, such as the subtly distorted voices or short delays after each spoken message; these elements may have seemed peculiar to users who expected a human avatar, preventing the emergence of significant results.

Similar to the human likeness scores, we found no significant difference between participants' attractiveness ratings. To our surprise, the scores on this index turned out high throughout all conditions (*M* ranging from 4.57 to 4.98 on a 7-point scale). This could be connected to the fact that participants often reported their amazement about the applied VR technology during and after their appointments. As the concept of attractiveness and the

applied measurement clearly depend on factors such as style and aesthetics (items include "crude—stylish" and "ugly—beautiful"), it seems likely that our visually impressive presentation influenced the answers on this index for all groups.

These observations notwithstanding, we note that the combination of our findings matches the suggested theory of an *uncanny valley of mind*. While all four groups ascribed the same visual appeal to the depicted characters, they reported different levels of eeriness depending on the attributed mind. With every other variable kept constant between conditions, this result indicates an aversion that is not softened by the "stylishness" of a human-like character, but might persist because of its unexpected emotional and social skills.

As laid out in the introduction of this paper, many cultures regard emotional experience as intrinsically human privilege. Similarly, the cognitive ability to attribute mental states to oneself and others (theory of mind) constitutes a central argument for many people's anthropocentric worldview, highlighting the superiority of human minds. Indeed, our results support the assumption that people react with increased caution if a virtual creation starts to resemble (or at least competently replicate) the prowess of a human brain. The revealed effect suggests that people prefer human-like replicas to be limited to a certain set of characteristics and might not appreciate them to behave in an empathic or social manner. It could be that they worry about losing their supremacy as humans—as suggested by the *threat to human distinctiveness* hypothesis—or even fear imminent harm from the sophisticated non-human creation. As several participants of our study reported their unease about not being able to anticipate the autonomous agents' next action, we note that their discomfort might indeed be connected to a perceived loss of control, which has been suggested as uncanny valley correlate in previous literature (Kang, 2009).

At the same time, one may consider an ethical component that contributes to the aversion against emotionally aware computers. Since ascribing mind to an entity also means

imposing moral responsibilities on it (Waytz, Gray, Epley & Wegner, 2010; Gray & Schein, 2012), the arrival of emotional and empathic machinery can hardly result in anything but technology-related skepticism. Digital systems that are able to reflect about mental states, or at least calculate some form of emotional consequence, will be given new ethical standards to abide by: Just like their human creators, "feeling" computers operate with moral gravity. Especially in contexts that blur the physical border between human and human-like simulation (e.g., robotics or immersive virtual reality), this sense of ethical responsibility might directly influence the comfort that people experience, as they wonder if the system's theory of mind matches their own—and what to expect if it does not. Ultimately, this relates to a basic principle that literature has suggested more than two decades ago: "Individuals reserve the right to decide which roles computers should fill" (Nass, Lombard, Henriksen, & Steuer, 1995, p. 237).

Limitations and Future Work

Our study has explored the uncanniness of human-like replicas that have the potential to conquer uniquely human domains in an uncontrollable way. However, we want to report several limitations of the conducted work. As most students responding to our recruitment process came from media and communication studies, it seems likely that the evaluation of the VR scenario was subject to a sampling bias. This is reflected by the moderate eeriness and high attractiveness ratings collected from our media savvy participants. Several students expressed their curiosity in the applied technology, which might have been less fascinating, or even explicitly unpleasant, to people from other areas of expertise. A similar limitation occurs due to the small age range of our sample, as younger people ("digital natives") can be much more accepting of new technology than older generations (Buckingham, 2013). Hence, we consider it possible that additional examinations among older participant samples could reveal an even stronger aversion against the idea of emotionally aware computer systems.

Lastly, the deceptive and persuasive elements of our manipulation leave room for methodological critique. While the final manipulation check only led to the exclusion of three people, it remains unclear how the remaining participants related to our complex instruction statements. As some of the recruited students reported to be well-accustomed to experimental studies, they might have been wary of deceptions, or have answered according to social desirability. Although our explicit manipulation check resulted in different emotional autonomy scores between conditions, we collected no qualitative data that could indicate whether participants attributed the same type of *genuine* emotion and empathy to the artificial intelligence as they would have ascribed to humans. The participants in the respective condition might have been completely persuaded by our introduction of the "intense emotional realism of neural network systems", or they may have considered the empathic behavior merely as well-executed pretense. A more extensive inquiry into the perceived authenticity of digital emotions and social cognition might help to improve the informative value of further studies. Still, as we have provided an enriched scenario and received only affirmative feedback during the debriefing sessions, we consider our manipulation as controlled as possible.

Considering the documented impact of individual beliefs on uncanny valley sensitivity (MacDorman & Entezari, 2015), an essential next step might be to collect data from different cultural backgrounds. As our study sample mainly consisted of atheistic students with a Western socialization, the examination of a more diverse group of participants should yield further insightful results. Also, with literature highlighting the influence of personality traits on the perceived uncanniness (MacDorman & Entezari, 2015; von der Pütten, Krämer, & Gratch, 2010), new studies could include personality assessments to be used as covariates. Additional objective measures (i.e., eye-tracking or neuroimaging methods) would help to produce results that go beyond people's conscious—and possibly biased—evaluations. By

this means, research might yield interesting insights into the biological processes that underlie the aversion against emotionally aware non-humans.

Finally, we recommend that future studies consider the manipulation of both appearance and mind when evaluation non-human characters, instead of equalizing one of these factors. Since previous research has indicated that the perceived personality of a robot is connected to its facial design (Broadbent et al., 2013), additional experiments should investigate if the attribution of social cognition interacts with certain visual features to create even more distinct forms of digital eeriness. Against the background of unstoppable technological progress, any effort to explore an *uncanny valley of mind* will hold great value, as it supports the harmonious co-existence of humans and their machinery in the long run.

References

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences, 3(12),* 469–479.

Azar, B. (2005). How mimicry begat culture. *Monitor on Psychology, 36(9),* 54.

Barnes, B. (2011, March 15). Many culprits in fall of a family film. *New York Times.* Retrieved from http://www.nytimes.com/2011/03/15/business/media/15mars.html

Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009). My robotic doppelganger – A critical look at the uncanny valley theory. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication,* 269–276.

Benner, K., & Wingfield, N. (2016, January 29). Apple sets its sights on virtual reality. *New York Times.* Retrieved from http://www.nytimes.com/2016/01/30/technology/apple-sets-its-sights-on-virtual-reality.html?_r=0

Borody, W. A. (2013). The japanese roboticist Masahiro Mori's Buddhist inspired concept of "The Uncanny Valley". *Journal of Evolution and Technology, 23(1),* 31–44.

Broadbent, E., Kumar, V., Li, X., Sollers, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality. *PLoS ONE, 8,* e72589.

Buckingham, D. (2013). Is there a digital generation? In D. Buckingham & R. Willett (Eds.), *Digital generations: Children, young people, and new media* (pp. 1-18). New York: Routledge.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? Thirty years later. *Trends in Cognitive Sciences, 12(5),* 187–192.

Carnegie Mellon University Graphics Lab (2015). *CMU Graphics Lab Motion Capture Database.* Retrieved from http://mocap.cs.cmu.edu

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*, 108.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Creed, C., Beale, R., & Cowan, B. (2014). The impact of an embodied agent's emotional expressions over multiple interactions. *Interacting with Computers, 27(2),* 172–188.

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics, 8(2),* 287–302.

Fox, J., Ahn, S. J., Janssen, J., Yeykelis, L., Segovia, K., & Bailenson, J. N. (2014). Avatars versus agents: A meta-analysis quantifying the effect of agency. *Human-Computer Interaction, 30(5),* 401–432.

Frijda, N. H., & Mesquita, B. (1994). The social roles and functions of emotions. In H. R. Markus & S. Kitayama (Eds.), *Emotion and Culture* (pp.51–87). New York: American Psychological Association.

Fukuda, H., & Ueda, K. (2010). Interaction with a moving object affects one's perception of its animacy. *International Journal of Social Robotics, 2(2),* 187–193.

Fuller, M. (2014). The concept of the soul: Some scientific and religious perspectives. In M. Fuller (Ed.), *The Concept of the Soul: Scientific and Religious Perspectives* (pp. 1–4). Newcastle upon Tyne: Cambridge Scholars Publishing.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7(2),* 77–83.

Gee, F. C., Browne, W. N., & Kawamura, K. (2005). Uncanny valley revisited. In *Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication,* 151–157.

Gray, A. J. (2010). Whatever happened to the soul? Some theological implications of neuroscience. *Mental Health, Religion & Culture, 13(6),* 637–648.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315,* 619.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology, 3(3),* 405–423.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125,* 125–130.

Grinbaum, A. (2015). Uncanny valley explained by Girard's theory. *IEEE Robotics & Automation Magazine, 22(1),* 152–150.

Guadagno, R. E., Blascovich, J., Bailenson, J. N., & McCall, C. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology, 10,* 1–22.

Hanson, D. (2006). Exploring the aesthetic range for humanoid robots. In *Proceedings of the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science,* 16–20.

Ho, C.-C., MacDorman, K. F., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction,* 169–176.

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26,* 1508–1518.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology, 60,* 653–670.

Jentsch, E. (1906). Zur Psychologie des Unheimlichen [On the psychology of the uncanny]. *Psychiatrisch-neurologische Wochenschrift, 22,* 195–205.

Kang, M. (2009). The ambivalent power of the robot. *Antennae, 1(9),* 47–58.

Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics, 1(3),* 465–480.

Kennedy, R. S., Lane, N. E., Berbaum, K. S. & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3(3),* 203–220.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences, 7,* 67– 83.

Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies, 68,* 57–68.

Lim, Y. M., & Aylett, R. (2007). Feel the difference: A guide with attitude! In *Proceedings of the 7th International Conference on Intelligent Virtual Agents,* 317–330.

MacDorman, K. F., & Entezari, S. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies, 16(2),* 141–172.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7(3),* 297–337.

Mori, M. (1970). The uncanny valley. *Energy, 7(4),* 33–35.

Nass, C. I., Lombard, M., Henriksen, L., & Steuer, J. (1995). Anthropocentrism and computers. *Behaviour & Information Technology, 14(4),* 229–238.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56(1),* 81-103.

Ngai, E. W. T., Tao, S. S. C., & Moon, K. K. L. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*, *35*, 33–44.

Pagel, M. (2012). Evolution: Adapted to culture. *Nature*, 482, 297–299.

Patel, H., & MacDorman, K. F. (2015). Sending an avatar to do a human's job: Compliance with authority persists despite the uncanny valley. *Presence*, *24(1)*, 1–23.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1(4)*, 515–526.

Ramey, C. H. (2005). The uncanny valley of similarities concerning abortion, baldness, heaps of sand, and humanlike robots. In *Proceedings of Views of the Uncanny Valley Workshop: IEEE-RAS International Conference on Humanoid Robots*, 8–13.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7(4)*, 413–422.

Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, *20(2)*, 99–103.

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments*, *16(4)*, 337–351.

Spaulding, S. (2013), Mirror neurons and social cognition. *Mind and Language*, *28(2)*, 233–257.

Stevens, J. (1999). *Intermediate Statistics. A Modern Approach.* London: Erlbaum.

Tinwell, A., Abdel Nabi, D., & Charlton, J. (2013). Perception of psychopathy and the uncanny valley in virtual characters. *Computers in Human Behavior*, *29*, 1617–1625.

Tinwell, A., & Grimshaw, M. (2009). Bridging the uncanny: An impossible traverse? In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, 66–73.

Tinwell, A., & Sloan, R. J. S. (2014). Children's perception of uncanny human-like virtual characters. *Computers in Human Behavior*, *36*, 286–296.

Tomonaga, M., & Uwano, Y. (2010). Bottlenose dolphins' (Tursiops truncatus) theory of mind as demonstrated by responses to their trainers' attentional states. *International Journal of Comparative Psychology*, *23(3)*, 386–400.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, *68(6)*, 1047–1058.

Unity [Computer software]. (2015). Retrieved from https://unity3d.com/

Vogeley, K., & Bente, G. (2010). "Artificial humans": Psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Networks*, *23*, 1077–1090.

von der Pütten, A., Krämer, N. C., & Gratch, J. (2010). How our personalities shape our interactions with virtual characters. Implications for research and development. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (eds.), *Intelligent Virtual Agents* (pp. 208–221). Berlin: Springer.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*, 383–388.

Waytz, A. & Norton, M. I. (2014). Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking – not feeling – jobs. *Emotion*, *14*, 434–444.

Weibel, D., Wissmath, B., Habegger, S. Steiner, Y., & Groner, R. (2008). Playing online games against computer versus human controlled opponents: effects on presence, flow, and enjoyment. *Computers in Human Behaviour*, *24*, 2274–2291.

Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the “uncanny valley” phenomenon. *Japanese Psychological Research*, *55*, 20–32

Zuckerberg, M. (2014). *Acquiring Oculus Rift*. Retrieved from https://www.facebook.com/zuck/posts/10101319050523971