Feeling robots and human zombies: Mind perception and the uncanny valley

Kurt Gray, Daniel M. Wegner

Department of Psychology, University of North Carolina, Chapel Hill, Davie Hall, CB # 3270, Chapel Hill, NC 27599, United States
Department of Psychology, Harvard University, Cambridge, MA, 02138, United States

Article info
Article history:
Received 7 October 2011
Revised 11 June 2012
Accepted 14 June 2012
Available online 9 July 2012

Keywords:
Artificial intelligence
Uncanny valley
Robots
Turing test
Humanness
Human–computer interaction

Abstract
The uncanny valley—the unnerving nature of humanlike robots—is an intriguing idea, but both its existence and its underlying cause are debated. We propose that humanlike robots are not only unnerving, but are so because their appearance prompts attributions of mind. In particular, we suggest that machines become unnerving when people ascribe to them experience (the capacity to feel and sense), rather than agency (the capacity to act and do). Experiment 1 examined whether a machine's humanlike appearance prompts both ascriptions of experience and feelings of unease. Experiment 2 tested whether a machine capable of experience remains unnerving, even without a humanlike appearance. Experiment 3 investigated whether the perceived lack of experience can also help explain the creepiness of unfeeling humans and philosophical zombies. These experiments demonstrate that feelings of uncanniness are tied to perceptions of experience, and also suggest that experience—but not agency—is seen as fundamental to humans, and fundamentally lacking in machines.

1. Introduction

People have long been fascinated with the idea of creating humans. In ancient Jewish mythology, a rabbi creates a Golem—an animated creature of clay and string—to protect the townspeople (Rosenberg, 2008). In Mary Shelley's Frankenstein (2004), the good doctor takes cadaver parts and, with the help of high voltage current, creates a living being. More recently, books and movies have imagined a time when computers and robots will be fully human—our friends and enemies, our lovers and therapists. Despite the diversity in these imaginings of things made human, there is a commonality underlying many of them, an undercurrent of apprehension or unease—the uncanniness of the inanimate made living (Kang, 2011). Now, with modern robotics capable of making more and more lifelike entities (Breazeal & Scassellati, 2002; Brooks, 2002), the possibility an artificial human is no longer science fiction.

The question is whether such entities are unnerving—and if they are—why?

The unsettling nature of humanlike robots was first suggested by Mori (1970) who thought that an increasingly humanlike appearance would lead to increased liking up to a point, after which robots appeared too human and became unnerving; he called this dip in liking the "uncanny valley." The uncanny valley has captured the imagination of scientists and laypeople alike (Wayne & Pasternack, 2011), but has received relatively little empirical attention. The few studies that test this idea find only mixed support for its existence, and in cases where the uncanny valley is documented, there are conflicting accounts of why it occurs (Hanson, Olney, Pereira, & Zielke, 2005; Ho, MacDorman, & Pramono, 2008; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2011; Seyama & Nagayama, 2007; Walters, Syrdal, Dautenhahn, te Boekhorst, & Koay, 2008). Some suggest humanlike robots may be unnerving because they remind people of death (MacDorman & Ishiguro, 2006), while others suggest that the unnervingness of humanlike robots may stem from abnormal facial features (Seyama &...
1.1. Mind perception

The hallmark of humanity is our minds. Many thinkers have emphasized the uniqueness of the human mind (Aristotle, BC350; Descartes, 1641), and some believe that artificial agents can never possess genuine human mental capacities (Searle, 1990). Perhaps what is unsettling about a human-looking robot, then, is that it appears to be human-minded. If the uncanny valley stems from perceiving mind in robots, the question is exactly what kind of mind drives feelings of unease? Research suggests that people perceive mind along the two independent dimensions of agency and experience (Gray, Gray, & Wegner, 2007; see also Fiske, Cuddy, & Glick, 2007; Gray, Jenkins, Heberlein, & Wegner, 2011a; Haslam, 2006; Knobe & Prinz, 2008; Robbins & Jack, 2006). Agency is the capacity to do, to plan and exert self-control; experience is the capacity to feel and to sense. Adult humans are seen to possess both agency and experience, and some have suggested that agency is the essential feature of human minds because it separates us from animals (e.g., Aristotle, BC350; Descartes, 1641).

Adult humans do have significantly more agency than robots, but there is also a striking experience gap: humans are perceived to be significantly more capable of experiencing emotion and sensation than are robots and other machines (Gray et al., 2007; Huebner, 2009). Furthermore, experience often accounts for more variance than agency in overall ascriptions of mind (Gray et al., 2007; see also Gray, Gray, & Wegner, 2008), and it is also more essentialized than agency (Haslam, Bain, Douge, Lee, & Bastian, 2005). This latter finding is especially important, because essentialized qualities are seen as fixed and unaltered by changes in appearance (Gelman, 2004; Haslam, 1998). If machines are held to be essentially lacking experience, then an appearance which suggests this capacity—i.e., humanlike eyes that convey emotion (Adolphs et al., 2005)—could conflict with this expectation and therefore be unsettling. Indeed, research suggests that the violations of deep-rooted expectancies generate unease across a variety of domains (Olson, Roese, & Zanna, 1996).

If perceptions of experience underlie the uncanny valley, it suggests a number of hypotheses. First, the unnerving nature of humanlike robots should be at least partially mediated by perceptions of experience, but not agency. Second, a machine perceived to have experience, but not agency, should be unnerving even without a humanlike appearance. Third, the uncanny valley should apply beyond machines, to any entities that violate expectancies of experience. In particular, because humans are fundamentally expected to have experience (Gray et al., 2011a; Knobe & Prinz, 2008), a person perceived to lack experience, but not agency, should be seen as unnerving. This would not only explain the unnervingness of those who seem incapable of fear or love (e.g., psychopaths), but also the strangeness of the idea of philosophical zombies—otherwise normal people without conscious experience

(Chalmers, 2003). Thus, the uncanny valley may apply to both humanlike robots and robotic humans.

1.2. The present research

In three studies, we investigate the uncanny valley and explore whether it can be explained as a more general phenomenon of mind perception. First, we measure whether a humanlike robot is felt to be unnerving, and whether these feelings are linked to perceptions of experience but not agency (Experiment 1). Second, we examine feelings of unease after directly increasing the amount of experience and/or agency perceived in a machine without a humanlike appearance (Experiment 2). Finally, we measure unease after decreasing the amount of experience and/or agency perceived in a human (Experiment 3).

2. Experiment 1: The uncanny valley and mind perception

This experiment tested whether the uncanny valley—the unnerving nature of humanlike robots—occurs, and whether it is tied to perceptions of experience. Participants saw a brief video of either a humanlike robot or a more mechanical robot and then rated their feelings of unease and attributions of agency and experience. We predicted that the humanlike robot would be more unnerving than the mechanical robot, and that perceptions of experience would help explain this effect.

2.1. Method

Participants (43 female, 62 male, 15 unspecified; M_age = 25) were recruited from subway stations and campus dining halls in Cambridge, MA. Experimenters, who were not blind to condition, randomly assigned participants to view one of two videos of Kaspar, a lifelike robot developed at the University of Hertfordshire (http://kaspar.feis.herts.ac.uk/; Blow, Dautenhahn, Appleby, Nehaniv, & Lee, 2006). This robot has been effective in helping children with autism (Robins & Dautenhahn, 2010), but may be somewhat unsettling to adults.

In the mechanical condition, participants watched 12 s of Kaspar filmed from behind, so that only its wiring and electrical components could be seen (Fig. 1). In the lifelike condition, participants saw 12 s of video in which the robot’s humanlike face was clearly visible (Fig. 1). After watching Kaspar simply move around, participants rated the extent to which they felt “uneasy,” “unnerved,” and “creeped out,” on 5-point scales from “not at all” (1) to “extremely” (5). These terms were used in place of “uncanny,” because they are more frequently used and less ambiguous.

Participants used the same five point scales to rate perceived experience—“This robot has the capacity to feel pain” and “This robot has the capacity to feel fear,” and agency—“This robot has the capacity to plan actions” and “This robot has the capacity to exercise self-control.” These specific capacities are taken from previous research and load highly on the general factors of experience and agency (Gray et al., 2011a).
2.2. Results and discussion

The 3 affective items were correlated, mean $r(110) = .67$, $p < .001$, and were averaged for an uncanny index. Similarly, the items assessing experience were averaged to form an experience index, $r(111) = .82$, $p < .001$, and the agency items were averaged to form an agency index, $r(111) = .74$, $p < .001$.

As expected, the humanlike robot ($M = 1.77$, $SD = .80$) was perceived as more uncanny than the mechanical robot ($M = 1.32$, $SD = .57$), $t(111) = 3.40$, $p < .01$. Pair-wise comparisons found that participants attributed similar agency to the humanlike ($M = 1.87$, $SD = .94$) and mechanical robots ($M = 2.00$, $SD = .97$), $t < 1$, but attributed greater experience to the humanlike ($M = 1.71$, $SD = .91$) than the mechanical robot ($M = 1.30$, $SD = .64$), $t(111) = 2.69$, $p < .01$. When ratings of agency and experience were simultaneously entered as predictors for unnervingness, a regression equation revealed that experience, $b = .42$, $t(111) = 4.30$, $p < .001$, but not agency, $b = -.05$, $t(111) = .46$, $p = .64$, significantly predicted uncanniness.

To determine whether perceptions of experience mediate feelings of uncanniness, a Sobel test (MacKinnon, Warsi, & Dwyer, 1995) was conducted with appearance as the predictor, uncanniness as the criterion, and perceptions of experience as the mediator. The test was significant, $Z = 2.28$, $p < .05$, suggesting that perceptions of experience at least partially mediate feelings of unease.

This experiment not only found support for the uncanny valley, but also provides an explanation for this odd effect. Robots with a humanlike appearance are attributed experience, which ostensibly violates expectancies. Despite the mediation analysis, however, this experiment is limited by the use of correlational methods—in the following experiment, we directly manipulate perceptions of experience.

3. Experiment 2: Machines with minds

The original conceptualization of the uncanny valley focused on the unnervingness of human-looking machines (Mori, 1970), but here we investigate whether feelings of unease can still occur when disconnected from appearance. In particular, this experiment tested whether a machine perceived to have experience, but not agency, would induce feelings of unease, even without a humanlike appearance. If so, it would suggest that any machine with perceived experience (e.g., a sophisticated chatbot that conveys emotions) may be unnerving. More generally, it would suggest that the uncanny valley stems from general cognitive expectations about what should or should not have a mind, and not simply odd appearances.

3.1. Method

Participants (22 female, 23 male, $M_{age} = 19$) were recruited as in Experiment 1 by experimenters who were blind to condition. Each participant was given a questionnaire describing “the Delta-Cray supercomputer,” which varied by condition. In the control condition, the computer was “like a normal computer, but much more powerful.” In the with-experience condition, the computer was described as able to feel some form of “hunger, fear and other emotions,” and in the with-agency condition, it was said to “independently execute actions” with “self-control and the capacity to plan ahead.” Participants rated their feelings of unease as in Experiment 1 and, as manipulation checks, also rated agency and experience as before.

To supplement these manipulation checks, a separate study asked 28 MTurk participants ($M_{age} = 35$, 16 female) to rate, on an 11-pt scale from “not at all” to “very much,” how much agency and experience terms (e.g., “hunger” or “self-control”) importantly characterize typical computers and humans.

3.2. Results and discussion

3.2.1. Manipulation checks

The questions assessing agency were correlated, $r(43) = .48$, $p < .001$, so were averaged to form an agency index. The experience questions were also correlated, $r(43) = .91$, $p < .001$, and were averaged to form an experience index. One-way analyses of variance (ANOVAs) revealed significant differences among conditions for both

![Humanlike Appearance](image1.png)  ![Mechanical Appearance](image2.png)

Fig. 1. Video stills from Experiment 1.
agency, \(F(2,42) = 36.90, p < .001, \eta^2 = .64\), and experience, \(F(2,42) = 612.50, p < .001, \eta^2 = .97\).

Least significant difference (LSD) tests revealed that the with-agency computer was perceived to possess more agency (\(M = 3.97, SD = .67\)) than the control computer (\(M = 1.83, SD = .94\), \(p < .001\), and the with-experience computer (\(M = 1.83, SD = .73\), \(p < .001\), which did not significantly differ from each other, \(p = 1.00\). Similarly, LSD tests revealed that the with-experience computer was perceived to have more experience (\(M = 4.53, SD = .52\)) than both the control computer (\(M = 1.03, SD = .13\), \(p < .001\), and the with-agency computer (\(M = 1.03, SD = .13\), \(p < .001\), which did not differ from each other, \(p = 1.00\).

For the separate MTurk study, a 2 (Computer/Human) \(\times\) 2 (Agency/Experience) within-subjects ANOVA revealed two main effects. Humans were unsurprisingly seen to be better characterized by mind-related terms (i.e., both agency and experience; \(M = 7.20, SD = 1.71\)) than were computers (\(M = 1.26, SD = 1.97\), \(F(1,26) = 100.06, p < .001\). Typical humans and computers were both also generally characterized more by agency (\(M = 4.74, SD = 1.47\)) than by experience (\(M = 3.72, SD = 1.47\), \(F(1,26) = 6.59, p < .05\). The ANOVA revealed no interaction, \(F(1,26) = .02, p = .89\), suggesting that any link between perceived experience and uncanniness is not because experience terms typically apply disproportionately to humans than computers (relative to agency terms).

3.2.2. Uncanniness

The affective ratings were significantly correlated, mean \(r(43) = .60, p < .01\), and were averaged to form an uncanniness index. A one-way ANOVA revealed differences among conditions, \(F(2,42) = 102.08, p < .001, \eta^2 = .83\), and LSD tests showed that the experience condition (\(M = 3.27, SD = .61\)) was significantly more unnerving than either the control (\(M = 1.22, SD = .24\), \(p < .001\), or agency conditions (\(M = 1.36, SD = .39\), \(p < .001\), which did not significantly differ from each other, \(p = .41\)) (Fig. 2). A computer capable of experience is unsettling in a way that one with agency is not. This link between experience and uncanniness was further confirmed by regression analyses which found that increased experience, \(\beta = .94, t(42) = 13.55, p < .001\), but not increased agency, \(\beta = -.08, t(42) = 1.13, p = .26\), significantly predicted uncanniness.

Finding that a machine with experience is unnerving—even without a humanlike appearance—provides additional evidence that perceptions of experience may underlie the uncanny valley. That perceptions of agency do not evoke similar feelings is likely due in part to familiarity, since some machines have significant agency, but it also suggests that perceptions of experience may be more essential to the idea of humanness (Haslam et al., 2005). If this is the case, then a human without experience should be more unnerving than a human without agency.

4. Experiment 3: People without minds

This experiment examined how people perceive those who have lost significant amounts of mental capacity. We presented participants with a picture of a man described as being either normal, lacking agency, or lacking experience, and then assessed feelings of unnerviness. If perceptions of experience are linked to unnerviness, then the man without experience, but not agency, should be unnerving. This result would be especially noteworthy in light of the MTurk study reported in Experiment 2, which highlights the importance of agency in explicit ratings of humans. If experience does drive feelings of uncanniness, it suggests a disjunction between explicit ratings of the human mind that center on agency, and deeper affective expectancies that center on experience.

4.1. Method

Participants (16 female, 21 male, 7 unspecified, \(M_{age} = 19\)) were recruited as in Experiment 1. Each received a questionnaire showing a man and a description of his mental capacities. In the control condition, he was described as “quite normal.” In the agency-less condition, he was described as unable to “plan or make goals,” or “do things a normal person can do.” In the experience-less condition, he was described as unable to “feel pain, pleasure or fear or otherwise experience what a normal person can experience.” Manipulation effectiveness and affective reactions were assessed as in Experiment 1. Experimenters were blind to condition.

4.2. Results and discussion

4.2.1. Manipulation checks

Agency item ratings were correlated, \(r(42) = .86, p < .001\), and were averaged in an agency index. Experience items were also correlated, \(r(42) = .94, p < .001\), and were averaged for an experience index. One-way ANOVAs revealed significant differences among conditions for both agency, \(F(2,41) = 9.01, p < .01, \eta^2 = .31\), and experience, \(F(2,41) = 14.18, p < .001, \eta^2 = .41\).

LSD tests revealed that the agency-less man was indeed perceived to have less agency (\(M = 2.30, SD = .99\)) than the control man (\(M = 4.14, SD = 1.31\), \(p < .001\), and the experience-less man (\(M = 3.50, SD = 1.25\), \(p < .01\), who did not significantly differ, \(p = .15\). Similarly, LSD tests revealed that the experience-less man was perceived with less capacity for experience (\(M = 1.70, SD = .96\)) than the control man.
(M = 3.68, SD = 1.73), p < .001, and the agency-less man (M = 4.10, SD = 1.17), p < .001, who did not significantly differ, p = .39.

4.2.2. Uncanniness

The affect variables were significantly correlated, mean r(42) = .47, p < .01, and were averaged to form an uncanny index. A one-way ANOVA revealed a significant effect of condition on this index, F(2,41) = 6.50, p < .01, η² = .49, and LSD tests showed the experience-less condition (M = 2.88, SD = 1.04) was significantly more uncanny than the control (M = 1.88, SD = .67), p < .01, and agency-less condition (M = 1.98, SD = .74), p < .01, which did not differ, p = .76 (Fig. 2).

It appears that a person without experience makes people uneasy in a way that someone without agency does not. Additional support for the role of experience in making an entity unnerving comes from a regression analysis in which reduced experience significantly predicted feelings of uncanniness, β = −.54, t(41) = 3.90, p < .001, but reduced agency did not, β = −.19, t(41) = 1.34, p = .19. Along with the previous studies, these results demonstrate that feelings of unease are linked to perceptions of the capacity for experience. Furthermore, because feelings of unease are linked to violations of fundamental expectancies (Olson et al., 1996), these data suggest that experience, and not agency, is fundamental to intuitive conceptions of the human mind.

5. Conclusions

Three studies demonstrate that perceptions of mind—in particular those of experience—are linked to feelings of unease. This link not only helps to explain the uncanny valley, but also the general aversion that people seem to have to machines attempting to pass the Turing (1950) test for the dissociation between explicit and implicit conceptions of the human mind (Gray, Young, & Waytz, 2012), one additional question is whether moral intuitions are tied to the uncanny valley. Indeed, perceptions of experience prompt ascriptions of moral rights (Gray, Knobe, Sheskin, Bloom, & Barrett, 2011b; Gray et al., 2007), and it seems strange to confer rights to robots.1

Although this research focused on strange minds—on feeling robots and unfeeling people—it speaks to the broader idea about what makes us human. Higher cognition may separate us from animals and feature prominently in explicit definitions of the human mind (Aristotle, BC350), but feelings of unease indicate that experience is implicitly viewed as more essential to humans. Additional evidence for the dissociation between explicit and implicit conceptions of the human mind comes from two follow up studies. In the first, participants explicitly rated agency as more characteristic of the human mind than experience. In the second, an implicit association test (IAT; Greenwald, Nosek, & Banaji, 2003) revealed that people more strongly associate the “human mind” with experience-related terms than with agency-related terms (Gray, 2010). Thus, despite humanity’s tendency to exult our agentic capacities (Shakespeare, 1992, Act II, Scene ii, 285–300), the deep-seated, implicit and intuitive essence of our minds is instead our hearts—our feelings and emotion. This in turn has implications for the advancement of machines. For example, programmers attempting to pass the Turing (1950) test should focus on conveying experience and not agency.

The idea of a fully human machine may be only an idea, but advancing technology suggests that there may come a time when we are swept away by deep poetry about the human condition, written not by flesh and blood, but by silicon and metal. The question is whether we will always be unnerved by that idea.

Acknowledgments

This research was supported by NSF Grant BCS-0841746 and by fellowships from the Social Sciences and Humanities Research Council of Canada and the Institute for Humane Studies. We thank for assistance Trevor Corbin, Kerstin Daunenhahn, Jesse Gray, Daniel Liss, Andrew Olson, Katherine Rippe, Dean Scaduto, Chelsea Schein, Claire Sheldon, Mike Spinelli, Liz Victor, and Adrian Ward.

References


1 Of course, the American Society for the Prevention of Cruelty to Robots (www.aspcr.com) might disagree.