Mind-Reading Machines: Distinct User Responses to Thought-Detecting and Emotion-Detecting Robots

Andrea Grundke*, Jan-Philipp Stein, and Markus Appel

Psychology of Communication and New Media, Julius-Maximilians-Universität Würzburg, Germany

Manuscript accepted for publication in the journal

Technology, Mind, and Behavior (2021/09/10)

*Andrea Grundke

Julius-Maximilians-Universität Würzburg

Oswald-Külpe-Weg 82

97074 Würzburg (Germany)

Email: andrea.grundke@uni-wuerzburg.de

Word Count Abstract: 174

Word Count Text: 7,686 (including Tables and Figures)

Short Title: Mind-Reading Machines

Author Note: no conflicts of interest; data, materials, and supplement publicly available under https://doi.org/10.17605/OSF.IO/U52KM

Abstract

Human-like robots and other systems with artificial intelligence are increasingly capable of recognizing and interpreting the mental processes of their human users. The present research examines how people evaluate these seemingly mind-reading machines based on the well-established distinction of human mind into agency (i.e., thoughts and plans) and experience (i.e., emotions and desires). Theory and research that applied this distinction to human-robot-interaction showed that machines with experience were accepted less and were perceived to be eerier than those with agency. Considering that humans are not yet used to having their thoughts read by other entities and might feel uneasy about this notion, we proposed that thought-detecting robots are perceived to be eerier and are generally evaluated more negatively than emotion-detecting robots. Across two pre-registered experiments (N_1 = 335, N_2 = 536) based on text vignettes about different kinds of mind-detecting robots, we find support for our hypothesis. Furthermore, the observed effect remained independent of the six HEXACO personality dimensions, except for an unexpected interaction with conscientiousness. Implications and directions for future research are discussed.

Keywords: uncanny valley; mind perception; detector robots; personality; human-robot interaction

Mind-Reading Machines: Distinct User Responses to Thought-Detecting and Emotion-Detecting Robots

Thoughts are free, who can guess them?
They fly by like nocturnal shadows.
No person can know them, no hunter can shoot them
with powder and lead: Thoughts are free!

First verse of the German folk song *The thoughts are free [Die Gedanken sind frei]*

Since antiquity, humans have found relief in knowing that our cognitions cannot be accessed by anyone but ourselves (e.g., Cicero, ca. 52 B.C.E. /1977). Due to the constantly advancing development of artificial intelligence, however, this freedom of thoughts (as expressed in the German folk song *Die Gedanken sind frei*) is in peril. Likewise, artificial intelligence is increasingly used to evaluate human emotions. How do humans respond to these mind-reading technologies?

Human (and non-human) mind can be distinguished into agency (thoughts and plans) and experience (emotions and desires, Gray et al., 2007), a distinction that has recently been applied to human-machine-interaction (Appel et al., 2020; Gray & Wegner, 2012; Taylor et al., 2020). The respective studies show that machines with experience are less well-accepted and often perceived to be eerier than those with agency. Yet, it remains unclear how people react to robots who do not express their own mental states but instead detect the mind of the human user. In two pre-registered experiments, we apply the agency–experience distinction to juxtapose robots that can detect thoughts (thought detectors) with those that can detect emotions (emotion detectors).

Contrary to the effects for self-expressing machines, we propose an opposite effect for mind detection: Thought-detecting robots are expected to be eerier than emotion-detecting robots. Additionally, our second experiment applies the HEXACO model of personality in order to examine whether individual differences moderate this effect.

Humanoid Robots and the Uncanny Valley

The production and diversification of service robots is on the rise. The COVID-19 pandemic led to an increased demand for cleaning and disinfection robots, food and medication delivery robots, and edutainment and interaction robots (International Federation of Robotics, 2020). At the same time, a multi-wave international study showed that attitudes towards robots have become more negative over the last years (Gnambs & Appel, 2019). Faced with observations such as these, people may turn to scientific evidence to look for explanations.

A popular framework underlying negative responses to robots is the *uncanny valley* model (Mori, 1970; Mori et al., 2012; for reviews see Kätsyri et al., 2015; Wang et al., 2015; Złotowski et al., 2015). It states that responses to human-like entities such as robots or digital animations get more positive with increasing human likeness until a steep drop is observed for highly (but not perfectly) human-like entities. Whereas traditional uncanny valley research manipulated the human likeness of entities such as robots by changing their visual appearance (MacDorman & Ishiguro, 2006; Mathur & Reichling, 2009, 2016; Seyama & Nagayama, 2007), more recent research focused on functional features of the respective technologies, as well as user variables and context factors (e.g., Broadbent, 2017; Lischetzke et al., 2017; MacDorman & Entezari, 2015; Mara & Appel, 2015; Piwek et al., 2014; Rosenthal-von der Pütten & Weiss, 2015; Tu et al., 2020). Also, adhering to a psychological viewpoint rather than merely focusing

on visuals, the ascribed mind of robots could be a key to understanding negative responses to robots (e.g., Gray et al., 2007).

Ascribing Mind to Machines

Theory and research suggest that negative responses to human-like robots may depend strongly on the perception of a human-like mind in a machine (Gray & Wegner, 2012; Hegel et al., 2008; Stein & Ohler, 2017; Wegner & Gray, 2016). Indeed, at the age of nine, children already classify robots as more or less scary depending on whether they attribute a human-like mind to them (Brink et al., 2019).

As an underlying framework for this line of research, the mind perception dichotomy by Gray et al. (2007) has gained a lot of attention in recent years. In their initial research, Gray and colleagues asked participants to describe the extent to which different types of people, animals, God, and a robot possessed specific mental capacities. Based on these data, a principal component factor analysis revealed that mental capacities might be categorized into *experience* (i.e., the ability to feel emotions, have a personality and a consciousness) and *agency* (i.e., self-control, morality, memory, planning, communication, and thought). According to further research, it is especially experience that seems to be a fundamental part of how people conceptualize the *human* mind and therefore humanness in general (Gray et al., 2011; Haslam et al., 2005; Knobe & Prinz, 2008).

Considering this paradigm, as well as some alternative theoretical approaches (e.g., Malle, 2019; Weisman et al., 2017), the notion of mind perception has become increasingly relevant in the field of human–robot interaction. For instance, Gray and Wegner (2012) combined the uncanny valley hypothesis with the mind perception dichotomy and showed that machines equipped with experience were rated as much more discomforting and uncannier than

those demonstrating agency. In a similar vein, it has been shown that participants rather assigned agency characteristics than experience characteristics to robots (Brink et al., 2019; Gray et al., 2007; Wegner & Gray, 2016). Further building upon the work by Gray and Wegner (2012), Appel and colleagues (2020) presented evidence that a robot with experience was perceived to be eerier than a robot with agency, followed by a robot who merely served as a tool. Indicating notable generalizability, this finding was conceptually replicated for smart speakers in a recent study (Taylor et al., 2020).

Mind Detection by Machines

The mind perception literature has profoundly advanced the scholarly understanding of how people evaluate autonomous technology. However, we note that the scholarly interest in this regard has mainly revolved around the perception of (artificial) minds in machines—yet hardly looked at the other direction, that is, user evaluations of machines analyzing the human mind. Arguably, while this idea might have been dismissed as technically impossible a couple of decades ago, recent technological advancements have turned mind detection by robots into an imminent reality.

By now, advanced software that allows social robots and other technical devices to recognize the emotions of human users can reach impressive levels of accuracy (e.g., Affectiva, 2018; Alonso-Martin et al., 2013; Chen et al., 2020; Microsoft Azure, 2018), leading to an increased scientific interest in digital forms of emotional recognition and mind perception (Banks, 2019; Bianco & Ognibene, 2019; Dissing & Bolander, 2020; Gray & Wegner, 2012; Kang & Sundar, 2019; Stein et al., 2020). Along these lines, it has been suggested that machines might even become able to detect not only human emotions but also human thoughts in the future—a feat that would reach clearly beyond the capabilities of their human creators. In fact,

current-day technology already heralds the rise of these possibilities, as machines have been able to deduce internal thought from eye movements (Huang et al., 2019), create their own Theory of Mind for humans via computational models (Breazeal et al., 2009; Brooks & Szafir, 2019; Dissing & Bolander, 2020), or use language processing to identify political views (Colleoni et al., 2014), and suicidal intentions (Walsh et al., 2020).

At the same time, it remains unclear how people react to these emerging technologies. Human behavior, appearance, and skills are often used as a reference point when designing modern-day technology (e.g., Eyssel et al., 2012; Huang & Mutlu, 2013; Niculescu et al., 2013; Salem et al., 2011), but users do not always appreciate impressions of humanness in their machines. Indeed, several studies showed that once new technologies threaten human uniqueness, they are typically met with strong aversion (e.g., Müller et al., 2020; Złotowski et al., 2017). Even more so, social cognitive abilities such as mind-reading might play a particular role in this regard (Stein & Ohler, 2017), as our ability to infer and analyze the emotions of those around us has long served as a distinct advantage to our species (Darwin, 1872/2009; Nesse, 1990). Considering this fear of losing our distinctiveness to machines, it appears likely that people might be wary of robots that detect others' emotions—or even surpass this ability with the possibility to "read" cognitions as well.

To this day, however, only a few psychological studies have actually examined user responses to mind-detecting technology in an empirical manner. Kang and Sundar (2019) found that a robot was evaluated more negatively if it correctly interpreted humans' sarcasm than if it failed to recognize this aspect of human behavior. Similarly, research by Stein and colleagues (2019) suggested that an artificial intelligence capable of analyzing participants' personality traits might be seen as threatening. Yet, previous efforts such as these were clearly limited by the

fact that they either focused only on emotional aspects of mind or kept the scope of the detection abilities ambiguous (e.g., Kang & Sundar, 2019; Stein et al., 2019; Stein & Ohler, 2017). Therefore, a structured exploration of user reactions to distinct forms of mind detection by machines is all but needed to close an important research gap in the field of human–computer interaction.

The Current Research

We assumed that—unlike the previously documented responses to robotic agency vs. experience (e.g., Appel et al., 2020; Gray & Wegner, 2012)—user evaluations might turn out quite differently for the *detection* of *human* agency vs. experience by social robots. More specifically, we expected a reversed effect: A robot's ability to analyze human experience should be perceived as less threatening and less uncanny than a robot's ability to analyze users' agency.

In their daily life, humans are generally quite used to other communicators detecting their emotions (Darwin, 1872/2009; Nesse, 1990), whereas precise thought detection is an ability largely unknown from the realm of human-to-human interaction. In turn, people are used to controlling their emotional displays and they have learned to deal with the unintentional communication of emotions (Tamir, 2016), yet they are much less experienced in controlling their thoughts or in coping with the unintentional communication of thoughts and plans. To illustrate this argument, one may consider the embarrassment that people tend to experience when human communication partners detect and interpret a Freudian slip, revealing supposedly true yet hidden thoughts and plans. Based on the large number of studies that have emphasized perceived control as a fundamental prerequisite of positive human-machine interactions (Kang, 2009; Roubroeks et al., 2010; Stein et al., 2019; Sundar, 2020; Zafari & Koeszegi, 2020;

Złotowski et al., 2017), we therefore expected a clear advantage of emotion-detecting over thought-detecting machines in participants' evaluations.

Apart from our main outcome variable eeriness (Gray & Wegner, 2012), which remains one of the most well-established ways of operationalizing robot acceptance (Diel et al., 2022), we used two additional dependent variables to get a more general overview of participants' assessment of this type of robotic technology. First, we focused on concerns about human identity, which emerged as a meaningful predictor of technology-related experience in previous research (Stein et al., 2019). More specifically, this variable assesses the extent to which users consider a machine as a symbolic threat to the distinctiveness of the human species (i.e., their uniquely human identity)—an impression that has, in turn, been linked to the unwillingness to further interact with technology (e.g., Kang & Sundar, 2020; Stein et al., 2019; Złotowski et al., 2017). As we presented emotion detectors (which have the same abilities as humans) and thought detectors, whose capabilities even exceed those of humans, we assumed that traditional human–machine boundaries could become blurred, resulting in a meaningful effect expressed by this variable. Second, the general evaluation of the new technology was assessed (Appel et al., 2019), in order to observe reactions towards the presented robots in a more generalizable way.

To implement the desired manipulation of robot characteristics, we used vignette texts—as previous work in the field of mind perception (Appel et al., 2020; Gray & Wegner, 2012; Swiderska & Küster, 2020; Ward et al., 2013) showed that this method can be an internally valid and efficient means to convey specific technological possibilities. In our first experiment, descriptions of an innovative robot able to analyze humans' agency or to analyze humans' experience were presented. As a control group, we presented a description of a robot who merely

served as a tool without any sophisticated analysis abilities. Based on the theory and research outlined above, the following hypotheses guided Experiment 1:

H1: The thought detector robot will evoke higher eeriness than the emotion detector robot (H1a), whereas the robot without analysis abilities will evoke the least eeriness (H1b).

H2: The thought detector robot will evoke stronger concerns about human identity than the emotion detector robot (H2a), whereas the robot without analysis abilities will evoke the least concerns (H2b).

H3: The thought detector robot will yield a more negative general evaluation than the emotion detector robot (H3a), whereas the robot without analysis abilities will yield the most positive general evaluation (H3b).

In addition to providing a replication of the effects tested in Experiment 1 (by using the same vignette texts), the second experiment examined the influence of users' individual differences on the acceptance of detector robots using the well-established HEXACO model of personality (Ashton et al., 2004). The hypotheses addressing the role of the users' personality will be introduced after the discussion of Experiment 1. Both experiments were pre-registered, with changes in the hypothesis numbering and exclusion criteria being documented in the online supplement. The pre-registrations, data, codes, and an online supplement can be found at Grundke et al. (2021, https://doi.org/10.17605/OSF.IO/U52KM).

Experiment 1

Method

Participants

A power analysis with G*Power (Faul et al., 2007) recommended at least 200 participants assuming a small to medium effect size of $f = .20$ (with alpha-error probability = .05, and power = .80) for the two-group fixed effect expected in Hypothesis 1a. Another 100 participants constituted the control condition, resulting in 300 participants. We invited 450 U.S.-American residents from the MTurk online participant pool (hit approval rate > 97%, hits > 1000), in order to have a buffer if careless responding occurred. Of the 443 completions, 44 participants did not have sufficient English skills, as indicated by two control questions, and were therefore not included in our statistical analyses (Kennedy et al., 2020). One additional participant failed an included attention check item and another three participants had large (> ±3 years) deviations when asked twice about their age. Moreover, 21 participants were excluded because their participation time was lower than 100 seconds ($n = 4$) or higher than 920 seconds ($n = 17$). Another 39 participants interchanged the thought detector robot and the experience detector robot in the manipulation check and were excluded (see online supplement for additional information). As such, the final sample consisted of 335 participants (154 female, 176 male, 5 non-binary or no answer) with an average age of 39.33 years ($SD = 12.00$, ranging from 21 to 75 years). Exploratory analyses revealed that age and gender did not moderate the influence of the robot manipulation on the dependent variables (see additional analyses on gender and age for both experiments in the online supplement).

Procedure

We asked participants to give informed consent before starting the online experiment. Following their random assignment to one of the three conditions, participants were presented with the respective vignette text matching their group. Subsequently, we asked them to fill in the chosen user evaluation questionnaires. Sociodemographic information and questions to identify

careless responding and low English proficiency followed (Kennedy et al., 2020; Meade & Craig, 2012; see online supplement for details), before participants were debriefed about the background of the experiment. Participants took on average 290.61 seconds (*SD* = 156.00) to complete the questionnaire, with a mean time of 42.67 seconds (*SD* = 49.78) spent on the page that presented the experimental stimulus. We complied with APA ethical standards in the treatment of our sample.

Stimuli

Participants read a short text about an innovative robot named Ellix. Based on our between-subject design, three versions of this vignette text were prepared. In the first condition, Ellix was introduced as a thought detector robot. In the second condition, Ellix was supposedly able to detect humans' emotions. In the third condition, the robot did not have any advanced analysis abilities, merely serving as a daily life tool. The descriptions were based on extracts of the mind perception classification by Gray et al. (2007); however, we made sure to highlight that the robot was not able to *feel/think* as was the focus of previous work (Appel et al., 2020; Gray & Wegner, 2012) but to *recognize* thinking or feeling on the human users' side. The stimuli texts were as follows (thought detector condition, emotion detector condition, control condition):

> *Ellix, a robot that can read your thoughts*
>
> Ellix is a social robot, i.e., a robot that is meant to interact with humans. Ellix is equipped with over 100 sensors and an advanced artificial intelligence system to make sense of the data it receives from its surroundings. It observes the human iris, facial expressions, voice patterns, and micro-movements of the head. It further studies the posture and movement of all other parts of the body. With decades worth of psychological insight stored in its

algorithms, as well as machine learning procedures that make the system smarter with each use, Ellix is able to analyze human interaction partners. More specifically, Ellix possesses the constantly advancing ability to detect what humans think, for example which actions they wish to execute and whether or not they know the answer to a question.

Ellix, a robot that can read your emotions

Ellix is a social robot, i.e., a robot that is meant to interact with humans. Ellix is equipped with over 100 sensors and an advanced artificial intelligence system to make sense of the data it receives from its surroundings. It observes the human iris, facial expressions, voice patterns, and micro-movements of the head. It further studies the posture and movement of all other parts of the body. With decades worth of psychological insight stored in its algorithms, as well as machine learning procedures that make the system smarter with each use, Ellix is able to analyze human interaction partners. More specifically, Ellix possesses the constantly advancing ability to detect what humans feel, for example which feelings they wish to act upon and whether or not they feel anxious when they answer a question.

Ellix, a robot with 100 sensors

Ellix is a social robot, i.e., a robot that is meant to interact with humans. Ellix is equipped with over 100 sensors and an advanced artificial intelligence system to make sense of the data it receives from its surroundings. It observes the human iris, facial expressions, voice patterns, and micro-movements of the head. It further studies the posture and movement

of all other parts of the body. By these means, the system is equipped with the most recent technology to be useful as a daily-life tool.

Measures

Eeriness. The first dependent variable asked about users' feelings of eeriness in response to the robot and was measured with the help of three items ("uneasy", "unnerved", "creeped out") based on previous research (Gray & Wegner, 2012). A 7-point scale ranging from *not at all* (1) to *extremely* (7) was provided ($\alpha = .90$, $M = 3.61$, $SD = 1.83$).

Concerns about human identity. This dependent variable was a composite of the repulsion scale (Kamide et al., 2012, two items) and three items of the concerns about human identity scale by Stein et al. (2019). These five items (e.g. "I think that humans will be dominated by this robot before long") were presented on a 7-point scale ranging from *strongly disagree* (1) to *strongly agree* (7), $\alpha = .91$, $M = 2.93$ ($SD = 1.59$).

General evaluation. The third dependent variable consisted of three bipolar items ("hate it – love it"; "negative – positive"; "repulsive – attractive", Appel et al., 2019), which were presented on a 7-point scale ranging from –3 to +3, $\alpha = .97$, $M = 0.43$ ($SD = 1.67$).

Manipulation check. We asked participants to select the robot's ability that was introduced in the text describing the robot Ellix. Participants had to choose one of three options reflecting the description of the robot (see online supplement for details).

Results

All *p*-values in this manuscript are based on two-tailed testing. Omnibus tests for the effects of the experimental manipulation on the three outcome variables were conducted. Pillai's Trace showed that the general linear model combining all three dependent variables did not reach

statistical significance, $V = 0.03$, $F(6, 662) = 1.89$, $p = .081$, $\eta_p^2 = .02$. On closer inspection, between-subject tests showed a significant group difference for the dependent variable eeriness, $F(2, 332) = 3.60$, $p = .028$, $\eta_p^2 = .021$. Concerns about human identity, $F(2, 332) = 1.27$, $p = .282$, $\eta_p^2 = .008$, and participants' general evaluation of the robots, $F(2, 332) = 2.56$, $p = .079$, $\eta_p^2 = .015$, on the other hand, appeared to be unaffected by the treatment (see Table 1).

Table 1

Descriptive Results of Experiment 1

Variable	Thought Detector		Emotion Detector		Tool Robot	
	M	*SD*	*M*	*SD*	*M*	*SD*
Eeriness	4.01	1.95	3.37	1.68	3.50	1.81
Concerns about human identity	3.05	1.60	2.73	1.51	3.00	1.64
General evaluation	0.12	1.81	0.50	1.63	0.60	1.57

Note. Sample sizes: Thought detector: $n = 101$, Emotion Detector: $n = 105$, Tool Robot: $n = 129$

To test our specific hypotheses, planned contrasts were performed. As expected in Hypothesis 1a, the thought detector robot evoked higher eeriness than the emotion detector robot, $t(332) = -2.53$, $p = .012$, $d = 0.35$. The eeriness scores in response to the robot without analysis abilities (tool robot) were lower than the eeriness scores in the response to the thought detector, $t(332) = 2.10$, $p = .036$, $d = 0.28$, but they did not differ significantly from the emotion detector robot, $t(332) = -0.56$, $p = .576$, $d = 0.07$. Thus, the findings provide mixed support for Hypothesis 1b. An analysis contrasting the thought detector with both other conditions, $t(332) = -2.65$, $p = .008$, $d = 0.31$, underscores this pattern of results, indicating that the thought detector robot was perceived to be particularly eerie whereas the difference between the emotion detector robot and the control condition remained negligible.

As indicated by the omnibus ANOVA, concerns about human identity were not affected by the experimental manipulation. The largest difference between the groups—which emerged between thought detector and emotion detector robot—did not reach statistical significance, $t(332) = -1.44$, $p = .150$, $d = 0.20$. Thus, no support was found for Hypotheses 2a and 2b.

Similarly, we note that the general evaluation of the thought detector robot did not differ significantly from the emotion detector robot, $t(332) = 1.66$, $p = .097$, $d = 0.23$ (Hypothesis 3a). While the robot without analysis abilities was evaluated more positively than the thought detector robot, $t(332) = -2.19$, $p = .030$, $d = 0.29$, it did not differ significantly from the emotion detector robot, $t(332) = -0.45$, $p = .657$, $d = 0.06$. As such, our results offer mixed support for Hypothesis 3b. When contrasting the general evaluation of the thought detector with both other conditions, a significant effect emerged $t(332) = 2.19$, $p = .029$, $d = 0.26$.

Discussion

The results of this experiment show that a thought detector robot evokes less favorable responses than a robot that can detect human emotions or serves as a simple tool, particularly in terms of higher eeriness. Eeriness has been described as a reaction to something that seems unfamiliar, an entity that eludes the world we know and feel comfortable with (e.g., Jentsch, 1906/1997; Mori, 1970). As humans are not yet used to the notion of having their thoughts and plans read, this detection ability might indeed push a machine right into the uncanny valley. In contrast, an emotion-detecting robot was perceived to be as harmless as a simple tool in our study; participants felt mostly at ease with this hypothetical machine. In our interpretation, this may be explained by people's familiarity with the respective recognition processes—as well as participants' confidence that emotional displays can be regulated and coped with and, thus, remain fully under their control.

In a critical reflection on our study, we note that the manipulation check—despite being successful—indicated that several members of the control group had experienced difficulties identifying their condition. Furthermore, more than three dozen participants interchanged the description of the thought detector robot with the description of the experience detector robot. As a takeaway from these observations, we adapted the materials for our follow-up research by highlighting the important parts of the descriptions in a bold font (see online supplement). Since the evaluation of the emotion detector robot had not differed significantly from the tool robot, we further omitted the tool condition in our second study. Moreover, we advanced the current project by focusing on interindividual differences as an important influence on users' reactions to mind-reading machines.

Experiment 2

The first aim of Experiment 2 was to replicate our main result of Experiment 1: We expected that a thought detector robot would again be perceived to be eerier than an emotion detector robot. Additionally, we decided to focus on the potential influence of dispositional factors regarding user responses to mind-reading robots. Previous work showed that stable individual differences can explain eeriness as a response to humanoid robots (e.g., Lischetzke et al., 2017; MacDorman & Entezari, 2015; Rosenthal-von der Pütten & Weiss, 2015). Therefore, we developed several hypotheses based on the HEXACO model of personality—one of the most often used models of basic personality structure (Moshagen et al., 2019), which consists of the factors honesty-humility, emotionality, extraversion, agreeableness, conscientiousness, and openness to experience.

Extraversion

Extraverted people feel positive about themselves, enjoy leading groups and social interactions, and they experience positive feelings of enthusiasm and energy (Lee & Ashton, 2009). Prior research showed that high extraversion predicted positive responses to robots (Esterwood & Robert, 2020; Mou et al., 2020; Santamaria & Nathan-Roberts, 2017). Given these results, we assumed that extraversion predicted more positive responses to detector robots as well. No differences between thought detector and emotion detector robots were formulated.

H4: Being extraverted is associated with weaker feelings of eeriness evoked by mind-detecting robots.

Openness to Experience

People who are open to experience take an interest in unusual ideas, become absorbed in the beauty of art and nature, and are interested in various domains of knowledge (Lee & Ashton, 2009). Openness was a predictor for the acceptance of new technologies in general (Korukonda, 2007; Nov & Ye, 2008), and some research showed that this trait predicted positive responses to robots (Conti et al., 2017; Morsünbül, 2019; Rossi et al., 2018, 2020, but see Müller & Richert, 2018). We therefore hypothesize that openness to experience predicts more positive responses to detection robots. No differences between thought detector and emotion detector robots were formulated.

H5: Being open to experience is associated with weaker feelings of eeriness evoked by mind-detecting robots.

Emotionality

Emotionality is described by the extent to which people experience fear of physical danger, experience anxiety in potentially stressful situations, need emotional support from others and feel empathy for others (Lee & Ashton, 2009). Some research in the context of social

robotics has dealt with the conceptually related factor of neuroticism. Neuroticism correlated with a more negative attitude towards a robot (Müller & Richert, 2018). These findings suggest that emotionality would predict higher aversion against supposedly mind-reading robots. No differences between thought detector and emotion detector robots were formulated.

H6: Being emotional is associated with stronger feelings of eeriness evoked by mind-detecting robots.

Agreeableness

People scoring high on this dimension tend to forgive wrongs that they suffered, are able to control their temper and are willing to compromise and cooperate with others (Lee & Ashton, 2009). Agreeableness was a predictor of trust in an autonomous security robot (Lyons et al., 2020) and was associated with higher trust in machines in general (Chien et al., 2016). Moreover, a higher score on agreeableness correlated with keeping a lower interpersonal distance to robots (Takayama & Pantofaru, 2009). Based on these results, a negative relationship with eeriness was expected for both detector robots. No differences between thought detector and emotion detector robots were formulated.

H7: Being agreeable is associated with weaker feelings of eeriness evoked by mind-detecting robots.

Conscientiousness

Conscientious persons organize their surroundings, are disciplined, and strive for perfection in their tasks (Lee & Ashton, 2009). No correlation between conscientiousness and the attitude towards robots was found in previous research (Müller & Richert, 2018). However, more conscientious people rated robot motion more negatively than less conscientious persons (Bodala et al., 2020) and preferred a text interface compared to a virtual character (Looije et al., 2010).

Given these few and mixed findings, we formulated no formal hypothesis and also no assumptions regarding differences between thought detector and emotion detector robots.

Honesty-Humility

The dimension Honesty-Humility is pronounced for people who avoid manipulating others for personal gain, who do not enjoy breaking rules and are uninterested in luxuries (Lee & Ashton, 2009). Special focus was put on the moderating role of the trait honesty-humility in our study. We assumed that people scoring high in the honesty-humility dimension would be less opposed to thought detection, as their overt behavior tends to be in line with their thoughts and plans. The latter is shown by negative correlations between honesty-humility and cheating behavior (Hilbig & Zettler, 2015; Kleinlogel et al., 2018, Moshagen et al., 2018; Pfattheicher et al., 2019). In human-robot interaction, cheating was negatively correlated with honesty-humility when a robot gave instructions (Petisca et al., 2019). Based on this line of argumentation, an interaction hypothesis was put forward.

H8: Scoring low in the honesty-humility dimension increases the difference of eeriness evoked by the thought detector robot and the emotion detector robot.

Method

Participants

An a-priori power analysis with G*Power and considerations regarding power of moderation effects (Giner-Sorolla, 2018; Simonsohn, 2014) yielded an aspired sample size of 500 participants. We invited 600 people of the MTurk participant pool (US residence, hit approval rate > 98%, hits > 1000) to participate in our online experiment to have a buffer if careless responding occurred. Of the 602 completions, 20 participants did not have sufficient English skills and were therefore not included in the analyses (Kennedy et al., 2020). Five

additional participants failed at least one attention check item and another eight participants had large (> ± 3 years) deviations when asked twice about their age. Moreover, 16 participants were excluded because their participation time was lower than 200 seconds (n = 10) or higher than 2800 seconds (n = 6). Seventeen participants interchanged the thought detector robot and the emotion detector robot, failing the manipulation check. The remaining sample consisted of 536 participants (238 female, 291 male, 7 non-binary or no answer) with an average age of 40.35 years (SD = 11.96, ranging from 19 to 79 years). Exploratory analyses revealed that age and gender did not moderate the influence of the robot manipulation on eeriness (see online supplement).

Procedure

Again, we asked participants to give informed consent before starting the online experiment. Questions that allow conclusions to be drawn about data quality were included in a similar manner than in the first experiment (see online supplement). Participants were randomly assigned to read a text about one of two robots: a thought detector robot or an emotion detector robot. The same stimuli as in Experiment 1 were used, albeit with a slight variation, we highlighted the manipulated parts of the descriptions in bold font (see online supplement). As an improved manipulation check, participants had to select the abilities of the robot about which they had been informed immediately after reading the robot descriptions. Subsequently, the participants filled in the eeriness and HEXACO measures, followed by the negative attitude towards a robot scale (Nomura et al., 2006) which was used in an exploratory analysis (see online supplement). The survey ended with sociodemographic questions, an opportunity to leave comments, and a debriefing. It took participants an average of 662.09 seconds (SD = 1063.97) to complete the questionnaire, including a mean duration of 65.24 seconds (SD = 106.44) spent on

the page that presented the experimental stimulus. Again, we complied with APA ethical standards in the treatment of our sample.

Measures

Eeriness. Eeriness was measured with the three items used in Experiment 1, resulting in a mean of $M = 3.72$ ($SD = 1.93$), $\alpha = .91$.

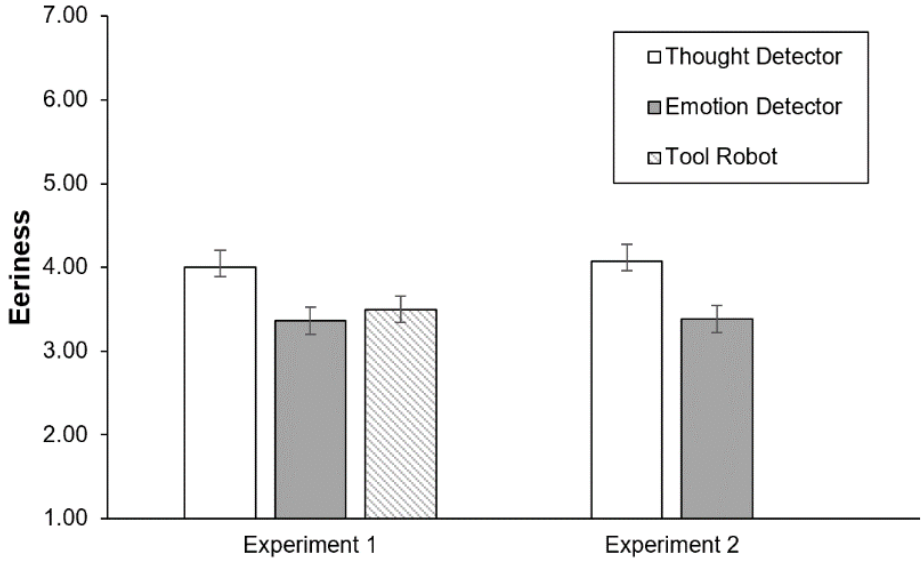
Personality. We used the HEXACO-60 questionnaire (Ashton & Lee, 2009), consisting of 60 items. Each dimension was measured through ten items on a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). All Cronbach's αs reached values of .72 or above. For detailed descriptive statistics see Supplement S6.

Results

In support of Hypothesis 1a and replicating the results of Experiment 1, the thought detector robot ($M = 4.08$, $SD = 1.87$) was perceived to be significantly eerier than the emotion detector robot ($M = 3.38$, $SD = 1.92$), $t(534) = 4.23$, $p < .001$, $d = 0.37$ (see Figure 1 eeriness results in both experiments).

Figure 1

Eeriness Means and Standard Errors in Both Experiments



The main effects and interactions of robot condition and HEXACO dimensions were analyzed by a hierarchical two-step regression. The results of the regression model are depicted in Table 2.

Table 2

Results of a Hierarchical Regression Analysis

Variable	*B*	95% CI for B		SE B	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.078	.078***
Constant	4.05***	3.82	4.28	0.12			
Condition [a]	−0.64***	−0.96	−0.33	0.16	−.17***		
Extraversion	−0.03	−0.22	0.16	0.10	−.02		
Openness to Experience	−0.16	−0.33	0.01	0.08	−.08		
Emotionality	0.09	−0.08	0.26	0.09	.05		
Agreeableness	−0.36***	−0.55	−0.17	0.10	−.19***		
Conscientiousness	0.12	−0.07	0.31	0.10	.06		
Honesty-Humility	0.00	−0.17	0.17	0.09	.00		
Step 2						.095	.016
Constant	4.03***	3.80	4.26	0.12			
Condition [a]	−0.64***	−0.96	−0.32	0.16	−.17***		
Extraversion	−0.05	−0.33	0.22	0.14	−.03		
Openness to Experience	−0.04	−0.29	0.20	0.12	−.02		
Emotionality	0.09	−0.15	0.33	0.12	.05		
Agreeableness	−0.35*	−0.62	−0.08	0.14	−.18*		
Conscientiousness	−0.14	−0.42	0.13	0.14	−.07		
Honesty-Humility	0.18	−0.09	0.44	0.13	.09		
Extraversion*Condition [a]	0.09	−0.29	0.47	0.19	.03		
Openness to E.* Condition [a]	−0.19	−0.53	0.14	0.17	−.07		
Emotionality* Condition [a]	0.00	−0.33	0.34	0.17	.002		
Agreeableness*Condition [a]	−0.06	−0.44	0.32	0.19	−.02		
Conscientiousness* Condition [a]	0.48*	0.10	0.86	0.19	.17*		
Honesty-Humility* Condition [a]	−0.28	−0.63	0.07	0.18	−.11		

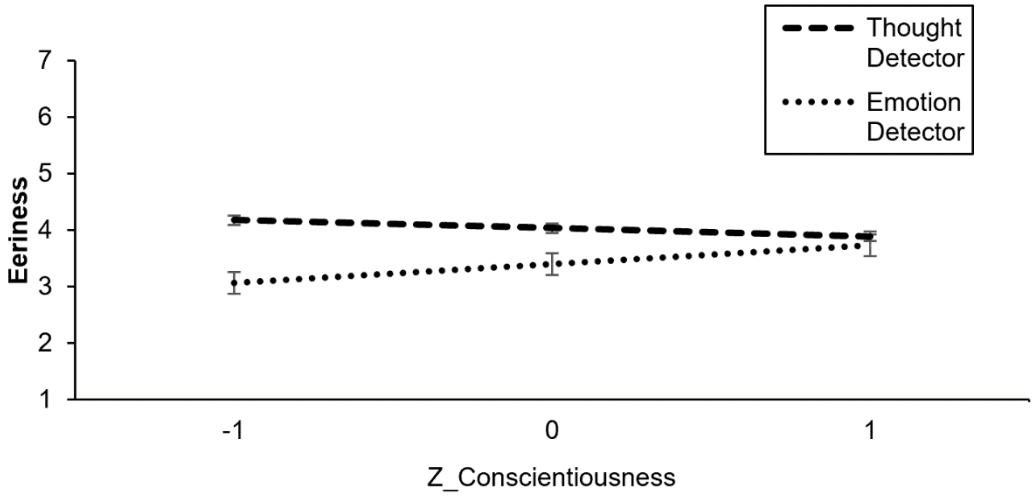
Note. All continuous predictors were z-standardized; $N = 536$; CI = Confidence Interval; *LL* = lower limit; *UL* = upper limit.
[a] dummy-coded (0 – thought detector robot; 1 – emotion detector robot).
*$p < .05$. *** $p < .001$.

In the first step of the hierarchical regression, all six HEXACO traits and the experimental factor were entered. In addition to the main effect of the experimental factor, a significant effect was found for agreeableness, $t(530) = -3.74$, $p < .001$. As expected in Hypothesis 7, being agreeable was associated with a lower level of eeriness evoked by the detector robots. None of the assumed remaining HEXACO effects reached statistical significance, so Hypotheses 4, 5, and 6 had to be rejected.

The second regression step—which also included interaction terms between the HEXACO dimensions and the assigned condition—revealed no interaction effect for honesty-humility, which led to a rejection of Hypothesis 8. However, unexpectedly, we observed a significant interaction between participants' conscientiousness and the robot condition, $B = .48$, $SE = 0.19$, $p = .014$, $\Delta R^2 = .01$ (see Figure 2), which was further examined using the SPSS-macro PROCESS (Hayes, 2012). Follow-up analyzes (Aiken & West, 1991) revealed that participants who were low in conscientiousness (−1 *SD*) perceived the thought detector robot to be significantly eerier than the emotion detector, $B = -1.11$, $SE = 0.25$, $t(524) = -4.45$, $p < .001$, 95% CI [−1.61, −0.62]. In contrast, the detector condition had no impact on participants who were high in conscientiousness (+1 *SD*), $B = -0.16$, $SE = 0.25$, $t(524) = -0.64$, $p = .519$, 95% CI [−0.66, 0.33]. According to the Johnson-Neyman technique, the manipulation of detecting abilities had a significant effect on participants' perceived eeriness for z-standardized values ≤ 0.54 of conscientiousness. About 69.59% of our participants fell into this significant region.

Figure 2

Interaction between Robot and Conscientiousness



Note. Error bars represent ± 1SE.

Discussion

Corroborating our results from Experiment 1, the thought detector robot was perceived as significantly eerier than the emotion detector robot. Moreover, a significant effect of agreeableness was found: Higher levels in this basic personality dimension were associated with less eeriness ascribed to the detector robots, matching the way this trait had affected user responses in prior human–robot studies (e.g., Chien et al., 2016; Lyons et al., 2020; Takayama & Pantofaru, 2009). As people high in agreeableness typically react in a tolerant and kind-mannered way to outside influences, it comes as little surprise that they also responded more positively to the presented detection robots. At the same time, we were surprised by a lack of noteworthy effects for the remaining HEXACO dimensions. Also, unlike expected, our data did not reveal a significant interaction of the dimension honesty-humility and the robot condition in our moderated regression analysis. Instead, the thought detector robot was generally evaluated as eerier than the emotion detector robot, regardless of participants' honesty-humility scores.

As a main result of our second experiment, we therefore note that people's evaluation of detector robots appears to be mostly unaffected by their fundamental personality traits. Arguably, this implies that the notion of sophisticated analysis robots may cause unease in a rather universal way, emerging as a strong challenge to people's idea of a good, unthreatening machine.

It should be noted, however, that our data yielded an unexpected interaction effect regarding another HEXACO trait: The higher participants scored in conscientiousness, the smaller was the difference between the eeriness ratings for the two detector robots. In our interpretation, this might be explained by the specific characteristics of highly conscientious individuals, who tend to put a strong emphasis on (cognitive) achievement and performance, while considering overt emotions as detrimental for success (Witteman et al., 2009; for an overview of the interplay of conscientiousness and negative affect see Fayard et al., 2012; Javaras et al., 2012). Further research is needed to find out how human conscientiousness influences interactions with robots—and to scrutinize the robustness of the uncovered interaction effect.

General Discussion

Robots and artificial intelligence are considered key technologies for the societies of today–even if not all prophecies made in science fiction have materialized (yet). User responses to these advanced technologies are of basic and applied relevance. Connecting the mind perception literature (Gray et al., 2007) and the uncanny valley hypothesis (Jentsch, 1906/1997; Mori, 1970), research on human-machine-interactions has demonstrated that robots who are ascribed human mind elicit negative responses such as eeriness (e.g., Stein & Ohler, 2017). Importantly, machines with emotions (experience) were found to be more aversive (Appel et al., 2020; Gray & Wegner, 2012; Taylor et al., 2020) than machines with thoughts and plans

(agency). Unlike previous research that was primarily focused on user responses to mind in a machine, we focused on a reversed perspective—the evaluation of machines capable of reading the human mind. Following our data analysis, we report that our main assumption held true across two experiments: In the realm of *mind-reading* machines, a thought detector is perceived as eerier than an emotion detector. With this fascinating outcome, we suggest that our results clearly advance the investigation of the *uncanny valley of mind* (Kang & Sundar, 2019; Stein & Ohler, 2017), both by shifting its overarching perspective and by introducing an important cognitive component. Offering further support for this main result, our second experiment showed that the stronger aversion against thought-detecting machines remained independent of several basic HEXACO personality dimensions. To us, this suggests that being apprehensive towards the concept of thought detection connects most humans regardless of their personality dispositions.

Proceeding to a psychological interpretation of our findings, we suggest that the need to perceive oneself as being in control is as important for human-robot interactions as it is for human-human interactions; potentially even more so. This desire for control, however, may be harmed by robots that appear able to look into the human mind. While we are used to sharing (and hiding) our emotions during many daily life interactions, it turns into a much more delicate matter if robots or other AI-based systems start to correctly infer what its user is thinking; in a dystopian scenario, this information could quickly be used against the human user in question, for instance in a job assessment or law-related context. Considering that the fear of artificial intelligence turning against humans has been named as a central caveat of human-computer interaction research (Cave & Dihal, 2019), even the most pessimistic imaginations should probably be kept in mind when designing detector robots. Based on our findings, we recommend

that developers of robotic and AI systems strive for absolute transparency regarding the capabilities of their created products and machines. Privacy guidelines should always be incorporated to make sure that the detecting entity does not share the results of its analysis with third parties; in all likability, this will help to alleviate the apprehension among potential users.

Limitations and Future Work

We note several limitations of the current experiments, which might also offer inspiration for future work. First, the observed mean eeriness ratings ranged between 3 and 4 on a 7-point scale, implying that the robot descriptions did not elicit particularly strong eeriness among participants. We assume that the online survey methodology paired with written text manipulations increased participants' psychological distance to our stimuli, thus preventing stronger emotional reactions. Similarly, since we (purposely) did not offer any information about the robots' appearance, some participants might have imagined a very friendly-looking or cute machine, which might have "softened" the eeriness evoked by our mind manipulation.

Second, we did not specify which emotions or thoughts could be detected by the robot. Emphasizing the detection of *negative* feelings or cognitions, for example, could have increased eeriness ratings in a notable manner, as participants might see it as more discomforting to have their sadness, anxiety, or anger discovered. A similar notion concerns the reading of thoughts, as it appears highly likely that some cognitions might be more sensitive or confidential for us than others. Hence, future research is encouraged to examine differences in users' experience and evaluations in response to robots detecting different thoughts and emotions.

Lastly, we believe that the methodological approach of using written vignette texts as stimuli deserves particular attention. While we still consider it as a very useful way of putting the mental abilities of a machine front and center, it might be worth considering to also show

pictures or even focus on live interactions with real robots in order to advance the discussed line of research. Doing so, fascinating interaction effects between the robots' mental capacities and its specific embodiment could be found, as suggested by another recent study (Stein et al., 2020). Building upon this, the influence of thought detection or emotion detection could also be explored in very different contexts: For instance, we strongly believe that a robot's capability to detect aspects of human mind will be evaluated differently in court cases, a therapeutic setting, nursing scenarios or smart homes (Thakur & Han, 2020). This way, evidence on the generalizability of the reported main effect could be gathered. Along the same lines, it should be explored whether the stronger aversion against a thought-detecting machine also persists in other cultures, as all participants taking part in our online experiments were recruited in the United States. Specifically, it might make sense to focus on participants from more collectivistic societies in future efforts, as the stronger social interdependence in the respective countries might also modulate the desire to avoid having one's mind read by another entity.

Conclusion

As cherished in the German folk song mentioned at the beginning of this paper (*Die Gedanken sind frei*), humans seem to truly appreciate the fact that their thoughts may roam free, without the risk of insulting others or having to admit one's secret desires. Accordingly, we found that the concept of thought-detecting machines—a hypothetical notion that does not seem so far removed from reality anymore, considering current technical developments—elicits significantly more unease than the concurrent idea of a robot analyzing human feelings. While this psychological observation may give developers pause or make them question the ethical boundaries of their innovations, it may also be possible to pave a path for well-accepted thought

detectors; as long as control perceptions are kept in mind, people might get used to this novel experience after all.

References

Affectiva. (2018). *Solutions.* https://www.affectiva.com/what/products/

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Sage.

Alonso-Martin, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., & Salichs, M. A. (2013). A multimodal emotion detection system during human–robot interaction. *Sensors*, *13*(11), 15549-15581. https://doi.org/10.3390/s131115549

Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, *102*, 274-286. https://doi.org/10.1016/j.chb.2019.07.031

Appel, M., Marker, C., & Mara, M. (2019). Otakuism and the appeal of sex robots. *Frontiers in Psychology*, *10*, Article 569. https://doi.org/10.3389/fpsyg.2019.00569

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*, 340-345. https://doi.org/10.1080/00223890902935878

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology*, *87*(5), 707-721. https://doi.org/10.1037/0022-3514.87.5.707

Banks, J. (2019). Theory of mind in social robots: Replication of five established human tests. *International Journal of Social Robotics, 12,* 403-414. https://doi.org/10.1007/s12369-019-00588-x

Bianco, F. & Ognibene, D. (2019, November 26-29). Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics. In M. A Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, A. Castro-González, & H. He (Eds.), *ICSR 2019: Social Robotics* (pp. 77-78). Springer. https://doi.org/10.1007/978-3-030-35888-4

Bodala, I. P., Churamani, N., & Gunes, H. (2020). Creating a robot coach for mindfulness and wellbeing: A longitudinal study. *arXiv preprint.* Advance online publication. arXiv:2006.05289.

Breazeal, C., Gray, J., & Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research, 28*(5), 656-680. https://doi.org/10.1177/0278364909102796

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development, 90*(4), 1202-1214. https://doi.org/10.1111/cdev.12999

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*, 627-652. https://doi.org/10.1146/annurev-psych-010416-043958

Brooks, C., & Szafir, D. (2019, November 7-9). *Building second-order mental models for human-robot interaction* [Paper presentation]. AAAI Fall Symposium Series, Arlington, VA, United States. https://arxiv.org/pdf/1909.06508.pdf

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence, 1,* 74-78. https://doi.org/10.1038/s42256-019-0020-9

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences, 509*, 150-163. https://doi.org/10.1016/j.ins.2019.09.005

Chien, S. Y., Sycara, K., Liu, J. S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 841-845. https://doi.org/10.1177/1541931213601192

Cicero. (1977). *Cicero: Selected political speeches* (M. Grant, Trans.). Penguin Classics. (Original work published ca. 52 B.C.E.)

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication, 64*(2), 317-332. https://doi.org/10.1111/jcom.12084

Conti, D., Commodari, E., & Buono, S. (2017). Personality factors and acceptability of socially assistive robotics in teachers with and without specialized training for children with disability. *Life Span and Disability, 20*(2), 251-272. http://shura.shu.ac.uk/id/eprint/18254

Darwin, C. (2009). *The expression of the emotions in man and animals* (P. Ekman, Ed.). Oxford University Press. (Original work published 1872)

Diel, A., Weigelt, S., & MacDorman, K. F. (2022). A meta-analysis of the uncanny valley's independent and dependent variables. *ACM Transactions on Human–Robot Interaction, 11*, 1.

Dissing, L., & Bolander, T. (2020, July 11-17). *Implementing theory of mind on a robot using dynamic epistemic logic* [Paper presentation]. International Joint Conference on Artificial Intelligence (IJCAI), Yokohama, Japan. http://www.imm.dtu.dk/~tobo/dissing2020implementing_proceedings.pdf

Esterwood, C., & Robert, L. P. (2020, November 10-13). *Personality in healthcare human robot interaction (H-HRI): A literature review and brief critique* [Paper presentation]. 8th International Conference on Human-Agent Interaction (HAI 2020), Sydney, Australia. https://doi.org/10.1145/3406499.3415075

Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (2012, March 5-8). *'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism* [Paper presentation]. 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston, MA, United States. https://doi.org/10.1145/2157689.2157717

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Fayard, J. V., Roberts, B. W., Robins, R. W., & Watson, D. (2012). Uncovering the affective core of conscientiousness: The role of self-conscious emotions. *Journal of Personality*, *80*(1), 1-32. https://doi.org/10.1111/j.1467-6494.2011.00720.x

Giner-Sorolla, R. (2018, January 24). Powering your interaction. *Approaching significance.* https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2

Gnambs, T., & Appel, M. (2019). Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in Human Behavior*, *93*, 53-61. https://doi.org/10.1016/j.chb.2018.11.045

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125-130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology, 101*(6), 1207–1220. https://doi.org/10.1037/a0025883

Haslam, N., Bain, P., Douge, L., Lee, M., & Bastian, B. (2005). More human than you: Attributing humanness to self and others. *Journal of Personality and Social Psychology*, *89*(6), 937-950. https://doi.org/10.1037/0022-3514.89.6.937

Hayes, A. F. (2012). *Process: A versatile computational tool for observed variable moderation, mediation, and conditional process modeling.* http://www.afhayes.com/public/process2012.pdf

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008, August 1-3). *Understanding social robots: A user study on anthropomorphism* [Paper presentation]. 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2008, Munich, Germany. https://doi.org/10.1109/ROMAN.2008.4600728

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, *57*, 72-88. https://doi.org/10.1016/j.jrp.2015.04.003

Huang, C., & Mutlu, B. (2013, June 24-28). *Modeling and evaluating narrative gestures for humanlike robots* [Paper presentation]. Robotics: Science and Systems, Berlin, Germany. http://www.roboticsproceedings.org/rss09/p26.pdf

Huang, M. X., Li, J., Ngai, G., Leong, H. V., & Bulling, A. (2019, October 21-25). *Moment-to-moment detection of internal thought during video viewing from eye vergence behavior* [Paper presentation]. Proceedings of the 27th ACM International Conference on Multimedia, Nice, France. https://doi.org/10.1145/3343031.3350573

International Federation of Robotics. (2020). *IFR Press Conference.* https://ifr.org/downloads/press2018/Presentation_WR_2020.pdf

Javaras, K. N., Schaefer, S. M., van Reekum, C. M., Lapate, R. C., Greischar, L. L., Bachhuber, D. R., Love, G. D., Ryff, C. D., & Davidson, R. J. (2012). Conscientiousness predicts greater recovery from negative emotion. *Emotion*, *12*(5), 875-881. https://doi.org/10.1037/a0028105

Jentsch, E. (1906/ 1997). On the psychology of the uncanny. *Angelaki: Journal of the Theoretical Humanities*, *2*(1), 7-16. https://doi.org/10.1080/09697259708571910 (Reprinted from "Zur Psychologie des Unheimlichen", 1906, *Psychiatrisch-Neurologische Wochenschrift, 8*[22-23], 195-198.)

Kamide, H., Mae, K., Shigemi, S., Arai, T. (2012, May 14-18). *A psychological scale for general impressions of humanoids* [Paper presentation]. IEEE International Conference on

Robotics and Automation, Saint Paul, MN, United States. https://doi.org/10.1109/ICRA.2012.6224790

Kang, J., & Sundar, S. S. (2019, June 26-28). *Social robots with a theory of mind (ToM): Are we threatened when they can read our emotions?* [Paper presentation]. Ambient Intelligence–Software and Applications–10th International Symposium on Ambient Intelligence, ISAml 2019, Avila, Spain. https://doi.org/10.1007/978-3-030-24097-4_10

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

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*, Article 390. https://doi.org/10.3389/fpsyg.2015.00390

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4), 614-629. https://doi.org/10.1017/psrm.2020.6

Kleinlogel, E. P., Dietz, J., & Antonakis, J. (2018). Lucky, competent, or just a cheat? Interactive effects of honesty-humility and moral cues on cheating behavior. *Personality and Social Psychology Bulletin*, *44*(2), 158-172. https://doi.org/10.1177/0146167217733071

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, *7*(1), 67-83. https://doi.org/10.1007/s11097-007-9066-y

Korukonda, A. R. (2007). Differences that do matter: A dialectic analysis of individual characteristics and personality dimensions contributing to computer anxiety. *Computers in Human Behavior*, *23*(4), 1921-1942. https://doi.org/10.1016/j.chb.2006.02.003

Lee, K., & Ashton, L. (2009). *Scale descriptions.* The HEXACO personality inventory – revised. http://www.hexaco.org/scaledescriptions

Lischetzke, T., Izydorczyk, D., Hüller, C., & Appel, M. (2017). The topography of the uncanny valley and individuals' need for structure: A nonlinear mixed effects analysis. *Journal of Research in Personality*, *68*, 96-113. https://doi.org/10.1016/j.jrp.2017.02.001

Looije, R., Neerincx, M. A., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, *68*(6), 386-397. https://doi.org/10.1016/j.ijhcs.2009.08.007

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020, September 7-9). *The role of individual differences as predictors of trust in autonomous security robots* [Paper presentation]. 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy. https://doi.org/10.1109/ICHMS49158.2020.9209544

MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, *16*(2), 141-172. https://doi.org/10.1075/is.16.2.01mac

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*(3), 297-337. https://doi.org/10.1075/is.7.3.03mac

Malle, B. F. (2019). How many dimensions of mind perception really are there? In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2268-2274). Cognitive Science Society.Mara, M., & Appel, M. (2015). Science fiction reduces the eeriness of android robots: A field experiment. *Computers in Human Behavior*, *48*, 156-162. https://doi.org/10.1016/j.chb.2015.01.007

Mathur, M. B., & Reichling, D. B. (2009, March 9-13). *An uncanny game of trust: Social trustworthiness of robots inferred from subtle anthropomorphic facial cues* [Paper presentation]. 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), La Jolla, CA, United States. https://doi.org/10.1145/1514095.1514192

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. https://doi.org/10.1016/j.cognition.2015.09.008

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455. http://dx.doi.org/10.1037/a0028085

Microsoft Azure. (2018). *Cognitive services.* https://azure.microsoft.com/en-us/services/cognitiveservices/face/

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

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98-100. https://doi.org/10.1109/MRA.2012.2192811

Morsünbül, Ü. (2019). Human-robot interaction: How do personality traits affect attitudes towards robot? *Journal of Human Sciences*, *16*(2), 499-504. https://doi.org/10.14687/jhs.v16i2.5636

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review, 125*(5), 656-688. https://doi.org/10.1037/rev0000111

Moshagen, M., Thielmann, I., Hilbig, B. E., & Zettler, I. (2019). Meta-analytic investigations of the HEXACO Personality Inventory (-Revised). *Zeitschrift für Psychologie, 227*, 186-194. https://doi.org/10.1027/2151-2604/a000377

Mou, Y., Shi, C., Shen, T., & Xu, K. (2020). A systematic review of the personality of robot: Mapping its conceptualization, operationalization, contextualization and effects. *International Journal of Human–Computer Interaction*, *36*(6), 591-605. https://doi.org/10.1080/10447318.2019.1663008

Müller, B. C. N., Gao, X., Nijssen, S. R. R., & Damen, T. G. E. (2020). I, robot: How human appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00663-8

Müller, S. L., & Richert, A. (2018, June 26-29). *The big-five personality dimensions and attitudes to-wards robots: A cross sectional study* [Paper Presentation]. 11th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greeece. https://doi.org/10.1145/3197768.3203178

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature, 1*(3), 261-289. https://doi.org/10.1007/BF02733986

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics, 5*(2), 171-191. https://doi.org/10.1007/s12369-012-0171-x

Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *AI & Society*, *20*(2), 138-150.

Nov, O., & Ye, C. (2008, January 7-10). *Personality and technology acceptance: Personal innovativeness in IT, openness and resistance to change* [Paper presentation]. 41st Annual Hawaii International Conference on System Sciences, Waikoloa, HI, United States. https://doi.org/10.1109/HICSS.2008.348

Petisca, S., Esteves, F., & Paiva, A. (2019, November 4-8). *Cheating with robots: How at ease do they make us feel?* [Paper presentation]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Macau, China. https://doi.org/10.1109/IROS40897.2019.8967790

Pfattheicher, S., Schindler, S., & Nockur, L. (2019). On the impact of honesty-humility and a cue of being watched on cheating behavior. *Journal of Economic Psychology*, *71*, 159-174. https://doi.org/10.1016/j.joep.2018.06.004

Piwek, L., McKay, L. S., & Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, *130*, 271-277. https://doi.org/10.1016/j.cognition.2013.11.001

Rosenthal-von der Pütten, A. M., & Weiss, A. (2015). The uncanny valley phenomenon: Does it affect all of us? *Interaction Studies*, *16*, 206-214. https://doi.org/10.1075/is.16.2.07ros

Rossi, S., Conti, D., Garramone, F., Santangelo, G., Staffa, M., Varrasi, S., & Di Nuovo, A. (2020). The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. *Robotics*, *9*(2), Article 39. https://doi.org/10.3390/robotics9020039

Rossi, S., Santangelo, G., Staffa, M., Varrasi, S., Conti, D., & Di Nuovo, A. (2018, August 27-31). *Psychometric evaluation supported by a social robot: Personality factors and technology acceptance* [Paper presentation]. 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China. https://doi.org/10.1109/ROMAN.2018.8525838

Roubroeks, M. A. J., Ham, J. R. C., & Midden, C. J. H. (2010). The dominant robot: Threatening robots cause psychological reactance, especially when they have incongruent goals. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.), *Persuasive Technology- Lecture Notes in Computer Science* (pp. 174-184). Springer. https://doi.org/10.1007/978-3-642-13226-1_18

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2011). Effects of gesture on the perception of psychological anthropomorphism: A case study with a humanoid robot. In B. Mutlu, C. Bartneck, J. Ham, V. Evers, & T. Kanda (Eds.), *ICSR 2011: Social Robotics* (pp. 31-41). Springer.

Santamaria, T., & Nathan-Roberts, D. (2017). Personality measurement and design in human-robot interaction: A systematic and critical review. *Proceedings of the Human Factors*

and Ergonomics Society Annual Meeting, *61*(1), 853-857. https://doi.org/10.1177/1541931213601686

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. https://doi.org/10.1162/pres.16.4.337

Simonsohn, U. (2014, March 12). [17] No-way interactions. *Data Colada.* http://datacolada.org/17

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. https://doi.org/10.1016/j.cognition.2016.12.010

Stein, J.-P., Appel, M., Jost, A., & Ohler, P. (2020). Matter over mind? How the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *International Journal of Human-Computer Studies*, Article 102463. https://doi.org/10.1016/j.ijhcs.2020.102463

Stein, J.-P., Liebold, B., & Ohler, P. (2019). Stay back, clever thing! Linking situational control and human uniqueness concerns to the aversion against autonomous technology. *Computers in Human Behavior*, *95*, 73-82. https://doi.org/10.1016/j.chb.2019.01.021

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74-88. https://doi.org/10.1093/jcmc/zmz026

Swiderska, A., & Küster, D. (2020). Robots as malevolent moral agents: harmful behavior results in dehumanization, not anthropomorphism. *Cognitive Science*, *44*(7), Article e12872. https://doi.org/10.1111/cogs.12872

Takayama, L., & Pantofaru, C. (2009, October 10-15). *Influences on proxemic behaviors in human-robot interaction* [Paper presentation]. IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, United States. https://doi.org/10.1109/IROS.2009.5354145

Tamir, M. (2016). Why do people regulate their emotions? A taxonomy of motives in emotion regulation. *Personality and Social Psychology Review*, *20*(3), 199-222. https://doi.org/10.1177/1088868315586325

Taylor, J., Weiss, S. M., & Marshall, P. J. (2020). “Alexa, how are you feeling today?”: Mind perception, smart speakers, and uncanniness. *Interaction Studies*, *21*(3), 329-352. https://doi.org/10.1075/is.19015.tay

Thakur, N., & Han, C. Y. (2018, November 1-3). *A hierarchical model for analyzing user experiences in affect aware systems* [Paper presentation]. 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, Canada. https://doi.org/10.1109/IEMCON.2018.8614787

Tu, Y. C., Chien, S. E., & Yeh, S. L. (2020). Age-related differences in the uncanny valley effect. *Gerontology*, *66*(4), 382-392. https://doi.org/10.1159/000507812

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry*, *59*(12), 1261-1270. https://doi.org/10.1111/jcpp.12916

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393-407. https://doi.org/10.1037/gpr0000056

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, *24*(8), 1437-1445. https://doi.org/10.1177/0956797612472343

Wegner, D. M., & Gray, K. (2016). *The mind club: Who thinks, what feels, and why it matters.* Viking.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people’s conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374-11379. https://doi.org/10.1073/pnas.1704347114

Witteman, C., van den Bercken, J., Claes, L., & Godoy, A. (2009). Assessing rational and intuitive thinking styles. *European Journal of Psychological Assessment*, *25*(1), 39-47. https://doi.org/10.1027/1015-5759.25.1.39

Zafari, S., & Koeszegi, S.T. (2020). Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics.* Advance online publication. https://doi.org/10.1007/s12369-020-00672-7

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics*, *7*(3), 347-360. https://doi.org/10.1007/s12369-014-0267-6

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies, 100*, 48-54. https://doi.org/10.1016/j.ijhcs.2016.12.008

Grundke, A., Stein, J.-P., & Appel, M. (2021). Mind-reading machines: Distinct user responses to thought-detecting and emotion-detecting robots. Open Science Framework. https://doi.org/10.17605/OSF.IO/U52KM