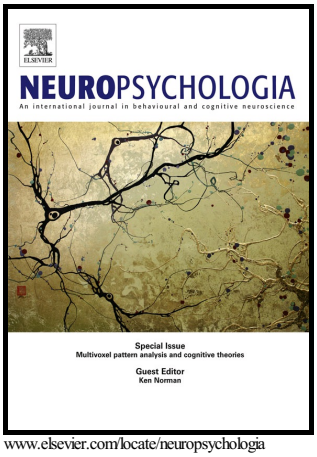
Uncanny valley as a window into predictive processing in the social brain

Burcu A. Urgen, Marta Kutas, Ayse P. Saygin



www.elsevier.com/locate/neuropsychologia

PII: S0028-3932(18)30168-4
DOI: https://doi.org/10.1016/j.neuropsychologia.2018.04.027
Reference: NSY6770

To appear in: *Neuropsychologia*

Received date: 29 September 2017
Revised date: 23 April 2018
Accepted date: 24 April 2018

Cite this article as: Burcu A. Urgen, Marta Kutas and Ayse P. Saygin, Uncanny valley as a window into predictive processing in the social brain, *Neuropsychologia,* https://doi.org/10.1016/j.neuropsychologia.2018.04.027

Uncanny valley as a window into predictive processing in the social brain

Burcu A. Urgen[1,2], Marta Kutas[1, 3], Ayse P. Saygin[1, 3]
[1]Department of Cognitive Science, UC San Diego, 9500 Gilman Drive, 92093, La Jolla, CA
[2] Department of Medicine and Surgery, University of Parma, Via Volturno 39, 43125, Parma, ITALY
[3] Neurosciences Program, UC San Diego, 9500 Gilman Drive, 92093, La Jolla, CA

*Corresponding author: Burcu A. Urgen, Via Volturno 39, Department of Medicine and Surgery, University of Parma, 43125, Parma, ITALY. Telephone: +39 345 0840758. burcu.urgen@gmail.com

ABSTRACT

Uncanny valley refers to humans' negative reaction to almost-but-not-quite-human agents. Theoretical work proposes prediction violation as an explanation for uncanny valley but no empirical work has directly tested it. Here, we provide evidence that supports this theory using event-related brain potential recordings from the human scalp. Human subjects were presented images and videos of three agents as EEG was recorded: a real human, a mechanical robot, and a realistic robot in between. The real human and the mechanical robot had congruent appearance and motion whereas the realistic robot had incongruent appearance and motion. We hypothesize that the appearance of the agent would provide a context to predict her movement, and accordingly the perception of the realistic robot would elicit an N400 effect indicating the violation of predictions, whereas the human and the mechanical robot would not. Our data confirmed this hypothesis

suggesting that uncanny valley could be explained by violation of one's predictions about human norms when encountered with realistic but artificial human forms. Importantly, our results implicate that the mechanisms underlying perception of other individuals in our environment are predictive in nature.

Keywords:
uncanny valley, predictive processing, N400, action perception, social neuroscience

1. INTRODUCTION

Our social milieu has changed tremendously in recent years, exposing us to social partners that are dramatically different from those the human brain has evolved with over many generations. Specifically, from guiding students in learning math and science, to helping children with autism and stroke survivors in their exercises, artificial human forms such as robots are rapidly becoming participants in our lives. The introduction of such artificial forms into our lives has in turn allowed us to study the fundamentals of human social cognition, similar to the use of artificial stimuli to learn about the fundamentals of human perception (Gregory, 1980; Rust and Movshon, 2005).

Uncanny valley is a phenomenon that refers to humans' response to artificial human forms, which possess almost human-like characteristics. In describing the phenomenon, Mori (1970), who introduced the term, proposes that the relationship between humanlikeness and humans' response to non-human agents is not a linear one. According to his framework, the increasing humanlikeness of an agent elicits positive responses from humans only up to a certain point, where increasing humanlikeness begins to elicit negative responses, thereby forming a deep valley (Figure 1).

Furthermore, it has been suggested that if the agent is moving, the responses will be more pronounced compared to the static form of the agent. Behavioral studies with humans have provided empirical evidence for the hypothetical curve in Figure 1 (MacDorman et al. 2009; Thompson et al., 2011; Matsuda et al., 2012; Poliakoff et al. 2013; Cheetham et al. 2013; Piwek et al. 2014; Macdorman and Chattopadhyay, 2016), and studies with non-human primates suggest that it has evolutionary origins (Steckenfinger and Ghazanfar, 2009).

There are several theories that attempt to explain uncanny valley including disease or threat avoidance and mate selection (Macdorman et al., 2009) but these theories lack the potential for scientific testability and are short on providing a mechanistic account of the phenomenon. One other hypothetical mechanism is Bayesian estimation or predictive coding, which is linked to a more general description of neural computational properties of the brain (Rao and Ballard, 1999; Friston, 2010, Moore, 2012), and therefore is a scientifically testable framework. According to predictive coding, the uncanny valley is related to violation of expectations in neural computing when the brain encounters almost-but-not-quite-human agents. A growing body of work has associated Mori's hypothetical curve to the processing of conflicting perceptual or cognitive cues, in which the stimuli are compatible with the elicited expectations or are in violation of them (Ho and MacDorman, 2008; Yamamoto et al., 2009; Mitchell et al., 2011; Cheetham et al., 2011; Saygin et al., 2012; Nie et al., 2012 Cheetham et al., 2013).

Here, we tested the predictive coding theory and its application in action perception (Kilner et al., 2007; Friston, 2010) as an underlying mechanism for uncanny valley. Accordingly, we hypothesize that upon exposure to a human-like form, our brains

predict human-like behavior, in specific human-like (biological) movement based on life-long experiences with conspecifics. Uncanny valley occurs when those predictions are not met, such as when faced with agents having human-like forms but non-human-like movements, a hypothesis that has been postulated by Saygin et al. (2012) previously. No empirical work to date has directly tested this theory of prediction violation.

In the present study, we used well-controlled stimuli, which did and did not violate appearance-motion predictions, together with electroencephalography (EEG) and a remarkable biomarker of human information processing, the event-related brain potential (ERP) N400 component to directly test this theory. N400 is the human brain's response to any meaningful stimulus. It is a negative-going event-related brain potential, which peaks around 400 ms after stimulus onset and is maximal in fronto-central regions of the human scalp to pictorial stimuli (Kutas and Federmeier, 2011). Its amplitude is relatively greater for items that violate one's predictions than for items that do not. Thus, it has been linked to the pre-activation of semantic knowledge during comprehension of meaningful stimuli including meaningful actions (Kutas and Federmeier, 2011; Amoruso et al. 2013).

We presented agents of varying humanlikeness in static and dynamic forms as EEG was recorded from human subjects. The stimuli consisted of a real human agent with human-like appearance and motion (Human), a realistic robot agent with human-like appearance and non-human-like motion (Android), and a mechanical robot with non-human-like appearance and motion (Robot) (Figure 2A). In this stimuli set, the real and mechanical agents (Human and Robot) had *congruent* appearance and motion whereas the realistic agent (Android) had *incongruent* appearance and motion. In this setting, the

appearance of the agent provides a context for the subsequent perception of the agent and activates world-knowledge (Metusalem at al., 2012) about agents that have that type of appearance. We hypothesized that the realistic agent (Android) would elicit a greater N400 response in dynamic form than the static form as its human-like appearance would lead to the prediction that it would move in a human-like way based on our world-knowledge but when it did not, it would violate that prediction. On the other hand, we hypothesized that the N400 amplitude for the static and dynamic forms would not differ for Human and Robot since both possess appearance-motion congruence (Human looks human-like, moves in a human-like way; Robot looks non-human-like, moves in a non-human way). Such a pattern of activity would provide direct empirical evidence for the prediction violation theory of uncanny valley.

2. MATERIALS AND METHODS:

2.1 Participants

Twenty right-handed adults (10 females; mean age = 23.8; SD = 4.8) from the student community at University of California, San Diego participated in the study. They had normal or corrected-to-normal vision, and no history of neurological disorders. Informed consent was obtained in accordance with the university's Human Research Protections Program. Participants were paid $8 per hour or received course credit. One subject's data was excluded due to high noise during EEG recording.

2.2 Stimuli

Stimuli consisted of video clips of actions performed by the humanoid robot Repliee Q2 (in Robotic and Human-like appearance) and by the human 'master', after whom Repliee Q2 was modeled (Figure 2A, also see Saygin et al. 2012 and Urgen et al. 2013 for details about the stimuli). We refer to these agents as the Robot, the Android (realistic robot), and the Human conditions. Note that the former two are in fact the same robot. Repliee Q2 has 42 degrees of freedom and can make face, head and upper body movements. However, the robot's movements did not match the dynamics of biological motion; it is mechanical or "robotic". The same body movements were videotaped in two appearance conditions. For the Robot condition, Repliee Q2's surface elements were removed to reveal its wiring, metal arms and joints, etc. The silicone 'skin' on the hands and face and some of the fine hair around the face could not be removed but was covered. It is important to note that the movement kinematics of the Android condition was identical to that of the Robot. The silicone skin on the hand or face did not affect the movement kinematics for the Android condition since the performed actions largely included arm and upper torso movements rather than fine detailed finger movements of the hand or face, and the skin was only 1.5 mm and tightly attached to the hand or face. For the Human condition, the female adult whose face was molded and used in constructing Repliee Q2 was videotaped performing the same actions. She was asked to watch each of Repliee Q2's actions and perform the same action naturally. All agents were videotaped in the same room with the same background. Video recordings were digitized, converted to grayscale and cropped to 400x400 pixels. Videos were clipped such that the motion of the agent began at the first frame of each video.

2.3 Procedure

Since prior knowledge can affect judgments of artificial agents differentially (Saygin and Cicekli, 2002), each participant was given exactly the same introduction to the study and the same exposure to the videos. Before starting EEG recordings, participants were shown each video and told whether each agent was a human or a robot, and the name of the action. Participants went through a practice session before the experiment. EEG was recorded as participants watched the images or video clips of the three agents performing eight different upper body actions (drinking from a cup, examining an object with hand, handwaving, turning the body, wiping a table, nudging, introducing self, and throwing a piece of paper). The videos were presented in two modes that we call *motion alone* and *still-then-motion*. In the *motion-alone* condition, 2-second videos were presented. In the *still-then-motion* condition, the first frame of the video was presented for 600-1000 ms (with a uniform probability jitter), and then the full video was played. The experiment consisted of 15 blocks. In each block, the eight videos of each agent were presented once in the *motion-alone* condition, and once in the *still-then-motion* condition. Stimuli were presented in a pseudo-randomized order ensuring that a video was not repeated on two consecutive trials. Each participant experienced a different pseudo-randomized stimuli sequence.

Stimuli were displayed on a 19” Dell Trinitron CRT monitor at 90 Hz using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). To prevent an augmented visual evoked potential at the beginning of video onset that might occlude subtle effects between conditions, we displayed a gray screen with a white fixation cross before the start of the video clip or still frame on each trial. Participants were instructed to fixate the

fixation cross at the center of the screen for 900-1200 ms (with a uniform probability jitter). A comprehension question was displayed every 6-10 trials, asking participants a true/false question about the action in the just seen video (e.g. Drinking?), after which they responded with a manual key press (Yes/No response).

2.4 EEG Recording and Data Analysis

EEG was recorded at 512 Hz from 64 ActiveTwo Ag/AgCl electrodes (Brain Vision, Inc.) following the International 10/20 system. The electrode-offset level was kept below 25 k-Ohm. Two additional electrodes were placed above and below the right eye to monitor oculomotor activity (1 additional electrode was placed on the forehead as a ground of the eye electrodes). The data were preprocessed with MATLAB and the EEGLAB toolbox (Delorme and Makeig, 2004). Each participant's data were first high-pass filtered at 1 Hz, low-pass filtered at 50 Hz, and re-referenced to average mastoid electrodes behind the right and left ear. Then the data were epoched ranging from 200 ms preceding video or first frame onset to 700 ms after video onset, and were time-locked to the onset of the video clips (*motion-alone* condition, see Procedures) or the first frame (*still-then-motion* condition, see Procedures) to compare the motion and still forms of the agents (we refer to these as *motion* and *still* conditions). Atypical epochs of electromyographic activity were removed from further analysis by semi-automated epoch rejection procedures (kurtosis and probability-based procedures with standard deviation = 6). After preprocessing, grand average event-related brain potentials (ERP) and scalp topographies were computed and plotted for each condition using Brain Vision Analyzer. If our manuscript is accepted, we will share the data and the code used in the analysis.

2.5 Statistical Analysis

The time window between 370-600 ms was considered for N400 analysis based on the grand average ERPs across all conditions. The area under curve measure was used to extract the N400 values for each agent under both motion and still condition for each subject in frontal channels (AF3, AFz, AF4, Fz, F1, F2, F3, F4, F5, F6) since N400 for pictorial stimuli has a more frontal distribution (Kutas and Federmeier, 2011). After preprocessing, data were exported to ERPLAB (http://erpinfo.org/erplab) and area under curve measures were extracted by using this toolbox. We then applied paired t-tests on the average frontal channel activity to compare the motion and still conditions for each agent (Robot, Android, Human). Since we expected motion condition to be greater than the still condition for Android (and no effect for Human and Robot), our t-tests were one-tailed. Although our hypotheses were best addressed by pair-wise t-tests, we also included an omnibus 3x2 ANOVA with factors Agent (Robot, Android, Human) and Mode (Static, Dynamic) for completeness.

2.6 Localization of EEG Activity

For identifying the neural generators (sources) of the activity during the N400 period, we used the LORETA method (Pascual-Marqui et al., 1994). LORETA estimates the distributed neural activity in the cortex based on the scalp measurements of ERP differences. Localization of the EEG activity was as follows: First, we computed the N400 differences between the static and motion conditions of each agent (Robot, Android, Human), and then we took the grand average of the N400 differences. We then

applied LORETA to the grand average N400 difference waveform using the template brain in the time interval between 370-600 ms to estimate the distributed neural activity underlying N400.

3. RESULTS

Our results indicate that observation of all agents elicited an N400 component regardless of the presentation mode (static or dynamic) at frontal sites (AF3, AFz, AF4, Fz, F1, F2, F3, F4, F5, F6) on the human scalp (Figure 2B shows ERPs on a representative frontal channel Fz). As predicted by our hypothesis, the amplitude of N400 (measured with area under curve between 370-600 ms averaged across all frontal sites) in the dynamic form was significantly greater than the static form for Android ($t(18) = 2.401$, $p<0.05$). On the other hand, the static and dynamic forms did not differ either for Robot ($t(18) = 0.388$) or for Human ($t(18) = -0.346$) (Figure 2B for ERPs, Figure 2C for bar graphs). Figure 3 shows the topography of the N400 effect on the scalp.

Although our hypotheses were best tested with the pair-wise t-tests, for completeness we also reported the results of the omnibus ANOVA: The Agent x Mode ANOVA show a main effect of Agent ($F(2,36) = 8.56$, $p<0.05$), and a trend in the main effect of Mode ($F(1,18) = 3.87$, $p = 0.06$) and in the interaction of Agent x Mode ($F(2,36) = 2.84$, $p = 0.07$).

Our source analysis with LORETA suggests that the generator of the N400 component is a widely distributed network including the middle and superior temporal areas, temporal-parietal junction, and prefrontal areas (Figure 4). These areas align with the neural network that has been implicated for N400 with intracranial recordings and

MEG in humans (Kutas and Federmeier, 2011). The maximal source density of this network was identified as a region within the inferior parietal lobule (Brodman area 40; x = -59, y = -32, z = 29, MNI coordinates).

4. DISCUSSION

In conclusion, we aimed to "ground" uncanny valley, an esoteric yet compelling first-person experience, whose mechanisms are unknown, with an approach combining three "knowns": N400 (established dependent measure), controlled stimuli and experimental design, and a theory of neural computations. We offer direct empirical support for the prediction violation theory as an underlying mechanism for uncanny valley. In addition, with its excellent temporal resolution, EEG allowed us to characterize the time course of the activity that underlies the uncanny valley phenomenon. It seems that within 400 ms the brain has made predictions about the movement of the agent being observed, which in turn suggests that the mechanisms underlying perception of other individuals in our environment are predictive in nature.

Our study demonstrates the benefit of using neural dependent measures in testing hypotheses about uncanny valley, whose underlying mechanism has remained unknown. Previous research on uncanny valley has mainly focused on behavioral ratings (MacDorman et al., 2009; Thompson et al., 2011; Poliakoff et al., 2013; Piwek et al., 2014). While these efforts have been a good step to operationalize the uncanny valley, they fall short for a number of reasons. First, these studies usually ask for an explicit (or conscious) response such as humanlikeness, eeriness, or familiarity. However, explicit measures may be too restrictive or not be sufficient to characterize the reaction of the

human subjects for uncanny stimuli. Neuroimaging has the advantage to measure human responses implicitly without asking for a specific response. In the present study, N400 was used as such an implicit measure. Second, behavioral measures provide only the output of the system, which is not very informative about the course of processing. Neuroimaging provides a rich set of data, and especially the temporally sensitive methods such as EEG allow one to monitor the information processing during stimulus presentation. In addition, the use of well-established dependent measures, such as N400 in this study, help one to situate the uncanny valley in a well-studied cognitive domain.

The use of event-related brain potentials in the current study provides a confirming evidence for a hypothesis that was proposed in our previous fMRI study of action perception that used the same stimuli. Saygin et al. (2012) found differential activity in parietal cortex for the android compared with the human and robot, which was *interpreted* as supporting evidence for the hypothesis that uncanny valley might occur due to the incongruity of appearance and motion in the action processing network. The N400 effect for the android we found in the present study corroborates this interpretation.

The characteristics of the N400 component in the present study was consistent with those in the literature (Kutas and Federmeier, 2011). Not only the topography is the topography similar to the previous N400 studies with pictures and videos (Sitnikova et al. 2008; Bach et al. 2009; Shibata et al. 2009; Proverbia and Riva, 2009), but the estimated neural generators of the N400 effect were consistent with previous reports. Our source analysis with LORETA (data not shown) confirmed that the N400 component in the present study was generated by a widely distributed network including the middle and superior temporal areas, temporal-parietal junction, and prefrontal areas consistent with

the neural network that has been implicated for N400 with intracranial recordings and MEG in humans (Kutas and Federmeier, 2011). The maximal source density of this network was identified as a region within the inferior parietal lobule, considered to be an area that integrates sensory, motor, and conceptual information during action perception (Amoruso et al. 2013), and therefore is likely to be sensitive to prediction violations.

We performed source localization for the N400 effect in our data 1) to see whether the estimated generators in our study were consistent with the broader literature on the component, and 2) to compare with results from a separate study using the same stimuli, along with a neuroimaging method with much more precise resolution (Saygin et al., 2012). By themselves, they do not constitute a precise localization of the N400 effect due to the inherent nature of source localization methods. Without subject specific measurements, these results should be interpreted with caution given limitations of our approach (i.e. projecting group averaged data onto a template brain).

It is important to note that N400 effect, initially discovered in the linguistic domain, has been shown with non-linguistic stimuli as well including pictures and videos (Sitnikova et al. 2008; Bach et al. 2009; Shibata et al. 2009; Proverbia and Riva, 2009). Therefore, it is thought to reflect a generic semantic process regardless of the stimulus type (Federmeier and Kutas, 2011). In the present study, we suggest that the form (appearance) of the visually presented agent provided a context from which the subject inferred how the agent would move over time (e.g. a mechanical/robotic appearance would activate the semantic network that includes motion information associated with robotic appearances), the same way a preceding word group provides a context for the upcoming word in a sentence (e.g. I take coffee with cream and _____ (“sugar” instead of

"dog"), and activates the relevant semantic network associated with the word group. Thus, using EEG has allowed us to link the uncanny valley phenomenon to general cognitive processing using the well-established dependent measure N400.

The present study also provides a potential link between the N400 component and the predictive coding theories of perception of other individuals (Kilner et al., 2007a; 2007b). The N400 component has long been used to study the predictive mechanisms of the human brain in many domains (e.g. language). On the other hand, predictive coding theory has been proposed as a mechanism of perception of other individuals but it has never been tested in an empirical setting directly. Although further work at a physiological level is needed to link the N400 with the theoretical constructs of the predictive coding theory (e.g. prediction error signals), the current study together with existing empirical and theoretical work (Amoruso et al. 2013; Rabovsky and McRae, 2014) implicate the N400 as a potential dependent measure for testing predictive coding theories of human cognition.

Furthermore, the use of N400 opened a new avenue of research to better characterize the uncanny valley, and guide the design of future artificial agents. For instance, appearance-motion incongruence is one particular instance where expectations are violated in agent perception. A broader range of expectation violations such as visual-auditory cues or task-relevant contextual violations (Ho and MacDorman, 2008; Yamamoto et al. 2009; Mitchell et al. 2011; Nie et al. 2012;) likewise can be studied to better understand the phenomenon and aid the design process in the artificial agent technology by means of integrating knowledge from cognitive sciences (Norman, 2013), which is important for critical application domains such as education and healthcare.

Acknowledgements: This research was supported by NSF (CAREER BCS1151805), DARPA, Kavli Institute for Brain and Mind, and Qualcomm Institute (Calit2). We thank Hiroshi Ishiguro and Intelligent Robotics Laboratory at Osaka University for the preparation of the stimuli; Alvin Li, Akila Kadambi, Wayne Khoe, Edward Nguyen for assistance in data collection, Markus Plank in data analysis, and Chris Berka for her feedback on the design.

REFERENCES

Amoruso, L., Gelormini, C., Aboitiz, F., Gonzales, M.A., Manes, F., Cardona, J.F., and Ibanez, A. (2013). N400 ERPs for actions: building meaning in context. *Frontiers in Human Neuroscience, 7*.

Bach, P., Gunter, T.C., Knoblich, G., Prinz, W., Friederici, A.D. (2009). N400-like negativities in action perception reflect the activation of two components of an action representation. *Soc. Neurosci.* 4:212–32.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision,* 10(4): 433-436.

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P. and Jancke, L. (2013). Category processing and the human likeness dimension of the Uncanny Valley Hypothesis: Eye-Tracking Data. *Frontiers in Psychology, 4*.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1): 9-21.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*:127-138.

Gregory, R.L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B*, 290 (1038).

Ho, C. and MacDorman, K.F. (2008). Human emotion and the uncanny valley: a GLM, MDS, and Isomap analysis of robot video ratings. *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, Amsterdam, The Netherlands, ACM.

Ishiguro, H. (2006). Android science: conscious and subconscious recognition. *Connection Science, 18*(4): 319-332.

Kilner, J.M., Friston, K.J., Frith, C.D. (2007) Predictive coding: an account of the mirror neuron system. *Cognitive Processing, 8*:159-166.

Kutas, M. and Federmeier, K.D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Reviews of Psychology*, 62: 621-647.

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

Macdorman, K.F. and Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition, 146*, 190-205.

Matsuda, Y. T., Okamoto, Y., Ida, M., Okanoya, K., Myowa-Yamakoshi, M. (2012). Infants prefer the faces of strangers or mothers to morphed faces: an uncanny valley between social novelty and familiarity. *Biologial Letters, 8*(5): 725-728.

Metusalem et al. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language, 66 (4),* 545-567.

Mitchell, W.J., Szerszen, K.A., Lu, A.S., Schermerhorn, P.W., Scheutz, M. and MacDorman, K.F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *Iperception,* 2(1): 10-12.

Moore, R.K. (2012). A bayesian explanation of the 'uncanny valley' effect and related psychologial phenomena. *Scientific Reports, 2* (864).

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

Nie, J., Park, M., Marin, A.L., Sundar, S.S. (2012). Can you hold my hand? Physical warmth in human-robot interaction. *Human-Robot Interaction*. Boston, Massachusetts, USA.

Norman, D. (2013). The Design of Everyday Things. New York: Basic Books. London: MIT Press (UK edition).

Pascual-Marqui, R.D., Michel, C.M. and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. International Journal of Psychophysiology, *18*(1), 49-65.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4): 437-442.

Piwek, L., McKay, L.S., and Pollick, F.E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition, 130*(3): 271-277.

Poliakoff, E., Beach, N., Best, R., Howard, T., and Gowen, E. (2013). Can looking at a hand make your skin crawl? Peering into the uncanny valley for hands. *Perception, 42*: 998-1000.

Proverbio, A.M. and Riva, F. (2009). RP and N400 ERP components reflect semantic violations in visual processing of human actions. *Neurosci. Lett.* 459:142–46.

Rabovsky, M. and McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based onnectionist attractor model of word meaning. *Cognition, 132*, 68-89.

Rao, R.P. and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2, 79–87.

Rust, N. and Movshon, T. (2005). In praise of artifice. *Nature Neuroscience, 8*, 1647-1650.

Saygin, A.P. and Cicekli, I. (2002). Pragmatics in human-computer conversations. *Journal of Pragmatics, 34*(3): 227-258.

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

Shibata, H., Gyoba, J., Suzuki, Y. (2009). Event-related potentials during the evaluation of the appropriateness of cooperative actions. *Neurosci. Lett.* 452:189–93.

Sitnikova, T., Holcomb, P.J., Kiyonaga, K.A., Kuperberg, G.R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *J. Cogn. Neurosci.* 20:1–21.

Steckenfinger, S.A. and Ghazanfar, A.A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences of the United States of America 106*(43): 18362-18366.

Thompson, J.C., Trafton, J.G. and McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception, 40*: 695–705.

Urgen, B.A., Plank, M., Ishiguro, H., Poizner, H., and Saygin, A.P. (2013). EEG Mu and theta oscillation during perception of human and robot actions. *Frontiers in Neurorobotics*, 9.

Yamamoto, K., Tanaka, S., Kobayashi, H., Kozima, H. and Hashiya, K. (2009). A non-humanoid robot in the "uncanny valley": Experimental analysis of the reaction to behavioral contingency in 2-3 year old children. *Plos One*, 4(9).

FIGURE CAPTIONS

Figure 1. Hypothetical curves that depict the uncanny valley effect for static and moving agents in varying levels of humanlikeness.

Figure 2. Stimuli used in the ERP experiment, ERP plots for the N400 effect, and bar plot for the N400 effect. (A) Sample static frames from the movies used in the EEG experiment depicting the three agents: Robot, Android, Human. (B) ERP plots of a representative frontal site (Fz) for static and dynamic forms for each agent (Robot, Android, Human). N400 is greater for moving Android compared to static, whereas no such difference was found for Human or Robot. (C) Bar graphs representing the area under curve for N400 (370-600 ms) for each of the conditions. N400 is significantly greater for dynamic than static form for Android, whereas they did not differ for Robot or Human.

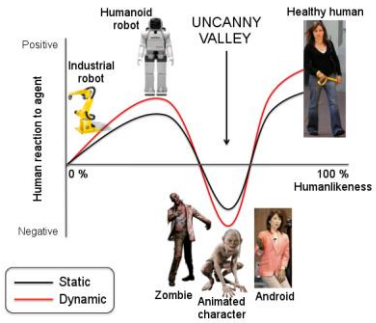
Figure 3. ERP scalp maps representing the difference between static and dynamic forms for each agent (Human, Android, Human) in the time interval of the N400 (370 ms – 600 ms).

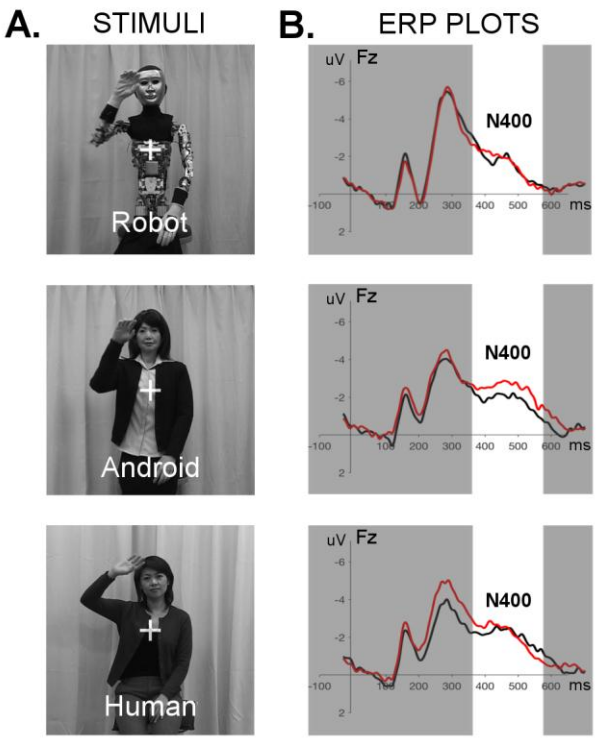
Figure 4. Source reconstruction analysis, all conditions (dynamic-static) collapsed. LORETA analysis in the N400 (370-600 ms) interval identified a distributed brain activity including middle and superior temporal areas (MTG and STG), tempora-parietal

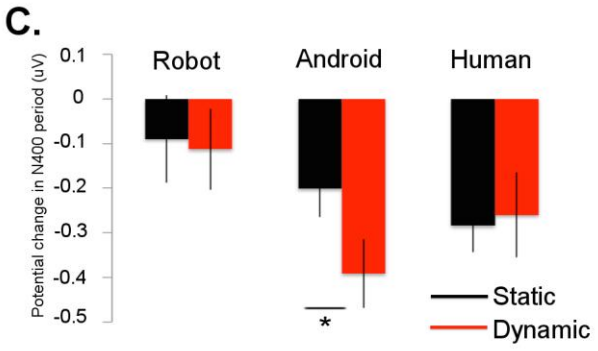
junction (IPL), and frontal areas (IFG, MFG, Medial FG), primarily in the left hemisphere (Difference waves of all conditions (Robot, Android, Human) are collapsed). Colorbar shows the source density. The maximal source density of this network is inferior parietal lobule (x = -59, y = -32, z = 29, MNI coordinates).

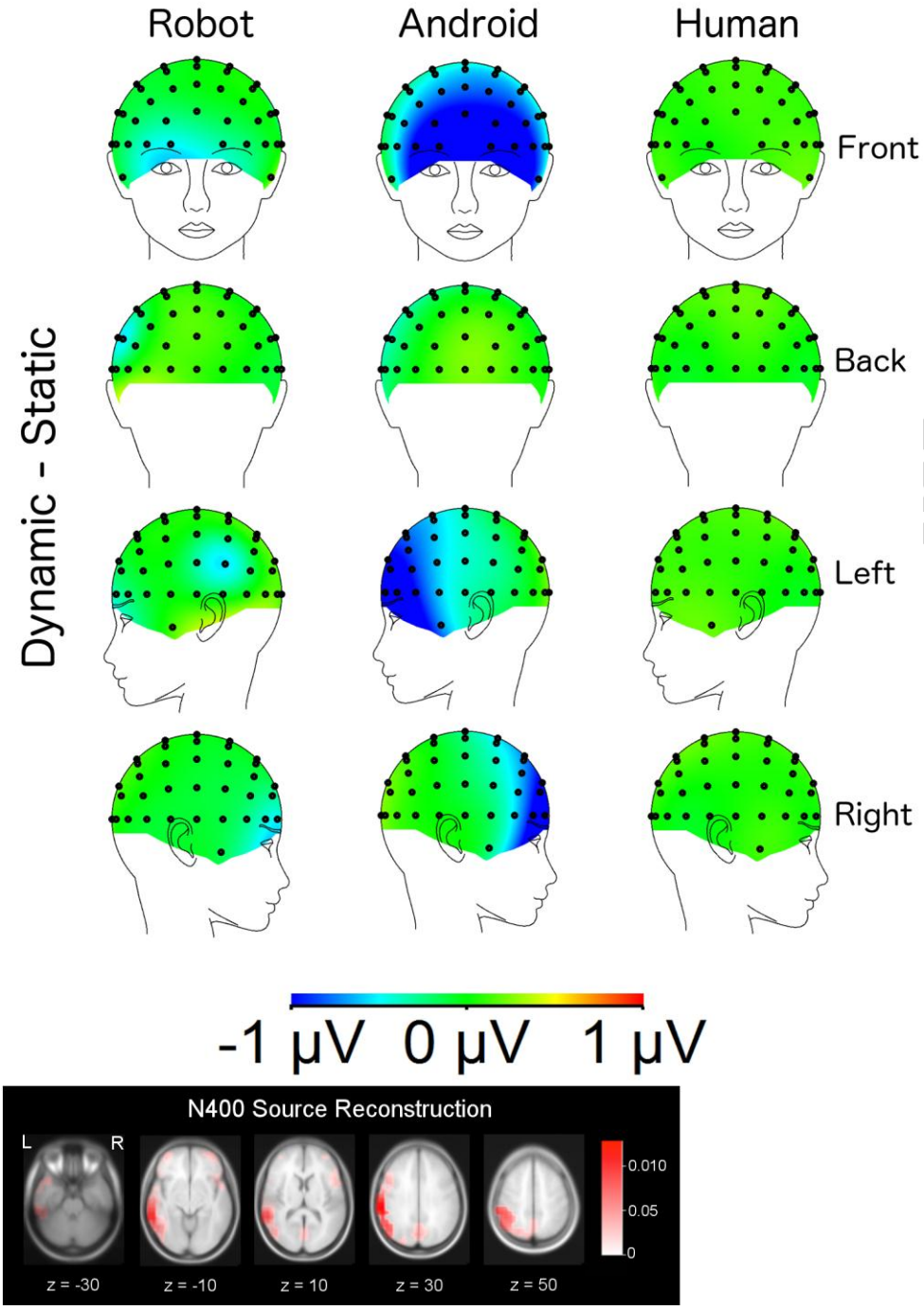
Highlights

- Uncanny valley can be explained by violation of one's predictions about human norms
- N400 ERP component can be used to assess the design quality of social robots
- Mechanisms underlying perception of other individuals are predictive in nature


Humanoid robot
UNCANNY VALLEY
Healthy human
Positive
Industrial robot
Human reaction to agent
0 %
100 %
Humanlikeness
Negative
Static
Dynamic
Zombie
Animated character
Android


A.
STIMULI
B.
ERP PLOTS
Robot
Android
Human
uV
Fz
N400
ms


C.
Potential change in N400 period (uV)
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
Robot
Android
Human
*
Static
Dynamic


Robot
Android
Human
Front
Back
Left
Right
Dynamic - Static
-1 μV 0 μV 1 μV
N400 Source Reconstruction
L
R
0.010
0.05
0
z = -30
z = -10
z = 10
z = 30
z = 50