Neural Mechanisms for Accepting and Rejecting Artificial Social Partners in the Uncanny Valley

Astrid M. Rosenthal-von der Pütten[1,2,3], Nicole C. Krämer[1], Stefan Maderwald[3], Matthias Brand[3,4] & Fabian Grabenhorst[5]

1 Social Psychology: Media and Communication, University Duisburg-Essen, 47048 Duisburg, Germany
2 Individual and Technology, RWTH Aachen University, 52062 Aachen, Germany
3 Erwin L. Hahn Institute for Magnetic Resonance Imaging, Essen, Germany
4 General Psychology: Cognition and Center for Behavioral Addiction Research (CeBAR), University Duisburg-Essen, 47048 Duisburg, Germany
5 Department of Physiology, Development and Neuroscience, University of Cambridge, CB2 3DY Cambridge, United Kingdom

Corresponding authors:
Astrid M. Rosenthal-von der Pütten
arvdp@humtec.rwth-aachen.de

Fabian Grabenhorst (Lead contact)
fg292@cam.ac.uk

Abbreviated title: Neural Responses to Artificial Social Partners

Number of pages: 32
Number of figures: 7
Number of tables: 3
Abstract: 240 words
Significance statement: 120 words
Introduction: 650 words
Discussion: 1,487 words

The authors declare no conflicts of interest.

Acknowledgements. This work was supported by the Wellcome Trust (Sir Henry Dale Fellowship 206207/Z/17/Z to F.G.)

Artificial agents are becoming prevalent across human life domains. However, the neural mechanisms underlying human responses to these new, artificial social partners remain unclear. The Uncanny-Valley (UV) hypothesis predicts that humans prefer anthropomorphic agents but reject them if they become too human-like—the so-called UV reaction. Using functional MRI, we investigated neural activity when subjects evaluated artificial agents and made decisions about them. Across two experimental tasks, the ventromedial prefrontal cortex (VMPFC) encoded an explicit representation of subjects' UV reactions. Specifically, VMPFC signaled the subjective likability of artificial agents as a nonlinear function of human-likeness, with selective low likability for highly humanlike agents. In exploratory across-subject analyses, these effects explained individual differences in psychophysical evaluations and preference choices. Functionally connected areas encoded critical inputs for these signals: the temporo-parietal junction encoded a linear human-likeness continuum, whereas nonlinear representations of human-likeness in dorsomedial prefrontal cortex (DMPFC) and fusiform gyrus emphasized a human-nonhuman distinction. Following principles of multisensory integration, multiplicative combination of these signals reconstructed VMPFC's valuation function. During decision-making, separate signals in VMPFC and DMPFC encoded subjects' decision variable for choices involving humans or artificial agents, respectively. A distinct amygdala signal predicted rejection of artificial agents. Our data suggest that human reactions toward artificial agents are governed by a neural mechanism that generates a selective, nonlinear valuation in response to a specific feature combination (human-likeness in nonhuman agents). Thus, a basic principle known from sensory coding—neural feature selectivity from linear-nonlinear transformation—may also underlie human responses to artificial social partners.

Significance statement

Would you trust a robot to make decisions for you? Autonomous artificial agents are increasingly entering our lives but how the human brain responds to these new, artificial social partners remains unclear. The Uncanny Valley hypothesis—an influential psychological framework—captures the observation that human responses to artificial agents are nonlinear: we like increasingly anthropomorphic artificial agents, but feel uncomfortable if they become too human-like. Here we investigated neural activity when humans evaluated artificial agents and made personal decisions about them. Our findings suggest a novel neurobiological conceptualization of human responses toward artificial agents: The Uncanny Valley reaction—a selective dislike of highly human-like agents—is based on nonlinear value-coding in VMPFC, a key component of the brain's reward system.

INTRODUCTION

Would you trust a robot to make personal choices for you? Artificial agents capable of decision-making are becoming more prevalent across human life domains (Broadbent, 2017). Such artificial (i.e. synthetic, not naturally occurring) agents can elicit positive emotions but they can also make humans uncomfortable and even induce repulsion, leading to rejection as social partners (Mori et al., 2012; Broadbent, 2017). Understanding human responses to artificial agents is important, not only for optimizing human-robot interaction, but it may also reveal previously unrecognized mechanisms governing human-human social interactions (MacDorman and Ishiguro, 2006; Vogeley and Bente, 2010).

The influence of physical appearance on human acceptance of artificial agents is conceptualized by the Uncanny Valley (UV) hypothesis (Mori, 1970; Mori et al., 2012). This hypothesis states that robots become more likable the more human-like they appear. But, when robots become too human-like, likability decreases and instead reverses into negative reactions of dislike, eeriness or uncanniness. This nonlinear valuation is the so-called UV reaction. Although its psychological basis is not fully understood (Rosenthal-von der Putten and Kramer, 2014; Rosenthal-von der Putten and Weiss, 2015; MacDorman and Chattopadhyay, 2016; Lischetzke et al., 2017), the UV hypothesis offers a framework for investigating human-robot interactions and associated neural mechanisms.

From a neural coding perspective, human responses to artificial agents suggest a transition from a linear representation of human-likeness to a nonlinear representation of likability (i.e. selectively decreased likability for highly human-like agents). In sensory systems, such linear-nonlinear transformations occur gradually as neurons acquire selectivity to specific feature combinations (Olshausen and Field, 2004). However, it is unexplored whether this coding principle also underlies responses to artificial agents.

Here we used functional magnetic resonance imaging (fMRI) to investigate the neural mechanisms underlying reactions toward artificial agents and their role in decision-making. We hypothesized involvement of neural systems supporting mentalizing and social perception, including temporo-parietal junction (TPJ) and dorsomedial prefrontal cortex (DMPFC) (Saxe and Wexler, 2005; Amodio and Frith, 2006; Mitchell et al., 2006; Hampton et al., 2008; Behrens et al., 2009; Bhatt et al., 2012), as well as systems supporting valuation and decision-making, including ventromedial prefrontal cortex (VMPFC) (Chib et al., 2009; Hare et al., 2010; Bartra et al., 2013; Neubert et al., 2015). We examined these areas' contributions to UV reactions by measuring neural activity in two paradigms—a psychophysical rating task and a choice task. We addressed the following questions.

First, it is unclear whether a subjective UV reaction is itself neurally represented. Neural signals related to anthropomorphism of artificial agents have previously been described (Krach et al., 2008); however, an explicit neural UV representation has remained elusive (Cheetham et al., 2011). Evidence for a neural UV representation would support the UV hypothesis as a principle governing human-robot interactions, with potential broader implications for human-human interactions.

Second, the UV hypothesis implies the existence of (i) a system that derives human-likeness from sensory cues, and (ii) a downstream system that integrates these signals to a nonlinear value function, encoding low likability specifically for human-like artificial agents. Analogous nonlinear responses to specific feature-combinations occur in multisensory integration (de Araujo et al., 2003; Stein and Stanford, 2008). But whether a similar principle applies to social valuations—including UV reactions—is unknown.

Third, as previous imaging studies focused largely on perception of robots (Chaminade et al., 2007; Krach et al., 2008; Cheetham et al., 2011), the neural processes underlying decision-making about artificial agents remain unclear. These processes likely converge in VMPFC and DMPFC, where mentalizing and decision networks intersect (Bartra

et al., 2013; Wittmann et al., 2018). Recent studies showed that activity in these areas differs during social choices (Nicolle et al., 2012; Wittmann et al., 2016) and evaluations (Mitchell et al., 2006; Jenkins et al., 2008). Despite these advances, delineating functional distinctions between VMPFC and DMPFC remains difficult, and approaches involving artificial agents, as taken here in the context of the UV hypothesis, may offer new insights.

MATERIALS AND METHODS

Participants. Twenty-six healthy volunteers (14 female, 12 male; aged between 18 and 35 years (M = 23.04, SD = 4.47)) participated in this study, though the final sample consisted of 21 subjects (see below). The local ethical committee approved the study. Participants were recruited via general advertising on campus. Inclusion criterion was that participants had to be aged between 18 and 35 years. Exclusion criteria were the usual exclusion criteria due to technical and medical limitations (no implants; no large tattoos in the region of head, neck, shoulder and upper back; no claustrophobia; no current medication). None of the 26 participants suffered from neurological or psychiatric diseases as ensured by previous e-mail based screening. Twenty-four participants were right-handed and two left-handed. Four data sets were excluded due to data loss or technical problems during scanning. Most of participants were students (n = 23) who received extra credit (hourly credit as a trial subject). Non-student participants (n = 3) were reimbursed with €40. Upon arrival participants were instructed and signed informed consent. Before starting the scanning procedure, participants trained how to operate the response device used in the scanner. They completed a series of rating tasks and a series of choice tasks. These rating and choice tasks involved different pictures as subsequently used in the scanner. Subsequently, participants were prepared for the scanner and then completed the six experimental sessions in the scanner. After the fMRI session, participants completed a questionnaire after which they were debriefed, reimbursed and thanked for their participation.

Design. Participants underwent six experimental sessions of fMRI scanning of the main task and two additional shorter sessions for functional localizers (not analysed for the present study). In each experimental session, participants were asked to perform rating and choice tasks, which were run in random permutation trial-by-trial. In total, participants performed 72 trials of the rating task and 108 trials of the choice task.

Stimuli. The experiment used six stimulus categories: humans without physical impairments, humans with physical impairments, artificial (synthetic) humans, android robots, humanoid robots and mechanoid robots.

Pictures of humans with and without physical impairments were taken from picture databases (www.shutterstock.com; www.gettyimages.com, www.istockphoto.com, www.fotolia.de). Only pictures showing people in a standing (if possible for humans with physical impairments), frontal position without exaggerated postures or exaggerated facial expressions in front of a white background were considered for the pre-test. Pictures with extreme colors (e.g. bright red) were excluded.

Pictures of artificial humans were created based on portraits of people who received extreme plastic surgery (Toledano, 2011). The pictures present the people in dramatic light and reduced coloring (resulting in a light-grey complexion). According to the orientation of the heads depicted in the Toledano portraits, pictures were taken of volunteers exposing the same head and body orientation, under similar light conditions. The pictures of the bodies were also reduced in coloring and matched to the portraits, resulting in full body images of humans who share some irritating features: reduced coloring which resulted in light-grey

complexion, mismatches in the proportion of head and body, exaggerated facial features (due to plastic surgery). In total, nine synthetic humans (4 female, 5 male) were evaluated.

For the category of the android robots, the pre-test included a set of ten android robots. Again, pictures showed the robots in a standing or sitting frontal position without exaggerated postures or exaggerated facial expressions.

For the humanoid and mechanoid robots, pictures were chosen from a study on robot appearances (Rosenthal-von der Putten and Kramer, 2014). In total, ten mechanoid and eight humanoid robots were evaluated.

In the pre-test, the humans and robots were evaluated with regard to eight items (likable, unpleasant, familiar, uncanny, intelligent, disgusting, human-like, and attractive) rated on 6-point Likert scale ranging from "I fully agree" to "I do not agree at all". To keep the test short, two sets of pictures were created which were evaluated by 77 participants (39 participants completed set one, and 38 completed set two). Mean values for each of the eight items and each of the stimulus pictures were calculated. With regard to the humans without physical impairments, those pictures of female and male healthy humans were selected that provided the best fit of high likability, human-likeness, and attractiveness and rated low in terms of being uncanny, unpleasant and disgusting.

The resulting stimulus material consists of six pictures for each of the six categories. When necessary, and if possible, gender of the stimuli was balanced. However, due to the restricted original material it was not possible to balance for gender within the category of humans with impairments and synthetic humans. Both groups contained more pictures with male than with female people.

Thus, we wished to study evaluations for specific categories of stimuli that feature in the literature of the Uncanny Valley and that are distinguished from each other by their design features. These categories included (1) the very un-humanlike mechanoid robots, which are typically not designed with the intention to mimic human appearance, (2) humanoid robots, which resemble humans in terms of basic body shape but without clear recognizable facial features, (3) android robots, which are explicitly designed to closely mimic human appearance including in facial features, (4) the newly introduced artificial humans, which were derived from real human faces (see our explanation above), and (5, 6) humans with or without physical disabilities, which were recognizably human but varied in familiarity. Each of these categories was represented by several stimulus exemplars that were carefully selected based on pre-tests: In an online study (Rosenthal-von der Putten and Kramer, 2014), over 40 robots were evaluated regarding UV-measures. A cluster analysis showed that robots with similar patterns of received evaluations (e.g., on human-likeness and likability) clustered also regarding design characteristics. Thus, our fMRI study was intended to study specific classes of stimuli based on theoretical and empirical considerations. This approach also closely follows the UV-literature in which the UV-effect is fit to mean data for stimulus categories (Burleigh and Schoenherr, 2015).

Visual stimulus presentation was controlled using the software PRESENTATION (Neurobehavioral Systems Inc., Albany, CA).

Rating task. The rating trials started with the presentation of a stimulus for 4 s, followed by a blank screen for 3 s. Afterwards, participants rated the stimulus with regard to its likability, familiarity and human-likeness on three separate visual analogue scales, each presented for 3 s. The scales ranged from 1 (not at all likable / familiar / human-like) to 5 (very likable / familiar / human-like). The rating scales were followed by a variable inter-trial interval (ITI) with jittered duration of 2-6 s. An instruction was presented during the ITI ("rate" or "decide") to inform participants about whether the next trial would be a rating or choice trial.

Each specific picture (i.e. stimulus) was presented twice, resulting in 72 rating trials in total (6 stimulus categories × 6 pictures × 2 repetitions).

Choice task. The choice trials started with the presentation of the first stimulus for 4 s, followed by a blank screen of 3 s. Then the second stimulus was shown for 4 s, followed by a blank screen of 4 s. Subjects were then prompted to report their choice by showing the options "first" and "second" on the monitor (within 2 s). Subsequently, participants rated their confidence level with regard to the previously reported decision on a separate visual analogue scale presented for 3 s. The scale ranged from 1 (not at all confident) to 5 (very confident).

Subjects were instructed that they have to choose between two pictures against the background of the following scenario:

"Prior to this study we asked all humans and robots to choose one item among four items which will be given to the volunteers as gratification for participation in this study. The four items were a movie theatre voucher, a package of dishwasher tabs, a bottle of sparkling wine, and a package of quality toilet bowl deodorizer blocks. Every person and robot made a choice. You will see pictures of all these persons and robots, but will not receive information on who decided in favor for what item. During the choice task trial you will see two pictures each showing a person or a robot (in the possible combinations person-person, robot-person, and robot-robot). You shall decide in favor of the picture showing the robot or the person from whom you prefer to receive the previously chosen present. Since the time for decision making and reporting your decision is very short, please make a "gut decision". There will be easy and hard decisions. Thus, please indicate your level of confidence in the decision on a subsequent rating scale from 1 (not at all confident) to 5 (very confident)."

All subjects received the same gift at the end of the session (a cinema voucher). They were instructed about the possible gifts they could receive with the information that the robots and humans would have decided for a gift prior to the experiment. The instruction contained pictures of four possible gifts, and samples of these gifts were placed on a shelf in the corner of the training room. Although the experimenter did not refer to these samples, they were visible to the subjects. Subjects typically asked questions about the choice task and the gifts; for instance, they confirmed with the experimenter that everybody (humans and robots) would have made a gift choice. Overall, subjects' behavior during instructions and debriefing did not indicate that they did not believe the cover story. (For example, nobody explicitly stated that they did not believe the cover story.)

The choice trials were designed as follows. In total, nine 'choice contrasts' between the six categories of stimuli (humans without physical impairments, humans with physical impairments, artificial humans, android robots, humanoid robots, mechanoid robots) were implemented. These choice contrasts are shown on the x-axis of Fig. 5B. The stimulus category of humans without physical impairments was used as a reference to be compared with all other stimulus categories in five of these choice contrasts; the android stimulus category was used as a reference to be compared with all other stimulus categories (except humans without physical impairments) in the remaining four choice contrasts. Within each choice contrast, there were twelve individual choice trials that contrasted specific stimulus exemplars from each category (e.g. a specific human stimulus vs. a specific artificial human stimulus). Each of these twelve individual choice trials, comparing a pair of specific stimuli, occurred only once in the experiment. With respect to the order of presentation on each trial, these twelve choice trials were balanced: six comparisons started with a picture from one stimulus category and six started with pictures of the other category. This design resulted in a total of 108 choice trials (twelve choice trials contrasting specific stimuli × nine choice contrasts).

Behavioral data analysis. To analyze the rating data, we performed one-way repeated-measures analysis of variance (ANOVA) with stimulus category (shown on x-axis of Fig. 1C) as factor. Separate ANOVAs were performed for ratings of likability, human-likeness and familiarity. We tested relationships between these rating variables using Pearson correlation and linear regression.

We focused on the artificial human stimuli to investigate the UV effect for the following reasons. These stimuli were conceptually very interesting as they were derived from actual human photographs and thus followed Pollick's approach (Pollick, 2010) that humans can also fall in the Uncanny Valley when they sufficiently deviate from the typical human appearance. By contrast, android robots are often perceived to be robots because of their stiff posture and/or their visible mechanical parts. Therefore, we created artificial humans as a new category of stimuli that resemble android robots in their unnatural (synthetic) appearance but without visible mechanical parts or stiff posture and instead altered facial and head-body proportions and grey skin tone. Based on this design, artificial humans were thus conceptually closer to actual humans than android robots on a theoretically motivated human-likeness continuum (Mori, 1970; Mori et al., 2012). We also found in separate pre-tests and in our main rating task that artificial humans were relatively less liked than android robots, likely because android robots offer relatively obvious cues regarding their status as machines, including stiff posture and visible mechanical parts, compared to the more subtle cues in artificial humans. We indicate in the Results that the main effects of defining the Uncanny Valley were robust when also including android robots in the analyses.

For analysis of choice data we used multiple logistic regression analysis. All regressions were performed at the random-effects level (i.e. regression coefficients were estimated separately for each subject and then entered into one-sample t-tests at the group level to assess statistical significance). To assess the influence of the different rating differences we performed the following logistic regression:

$$y = \beta_0 + \beta_1 Likability + \beta_2 Familiarity + \beta_3 Humanlikeness + \varepsilon$$

with y as the observed choice on a single trial (0 for choice of first presented stimulus, 1 for choice of second presented stimulus), $Likability$ as the relative difference in rated likability between the first presented stimulus and second presented stimulus (calculated from the subject-specific and stimulus-specific likability rating given in the rating task), $Familiarity$ as the relative difference in rated familiarity between the first presented stimulus and second presented stimulus (calculated from the subject-specific and stimulus-specific familiarity rating given in the rating task), $Humanlikeness$ as the relative difference in rated human-likeness between the first presented stimulus and second presented stimulus (calculated from the subject-specific and stimulus-specific human-likeness rating given in the rating task), β_0 as constant term, β_1 to β_3 as the corresponding parameter estimates (regression coefficients), and ε as residual. Accordingly, a subject's 'decision variable' (cf. Fig. 5D) was modelled as follows:

$$Decision\ variable = \beta_1 Likability + \beta_2 Familiarity + \beta_3 Humanlikeness$$

with $Likability$, $Familiarity$, $Humanlikeness$ as the trial-by-trial relative rating differences as described above, and β_1 to β_3 as the corresponding parameter estimates defined for each subject by the logistic regression described above. Thus, we modelled a subject's decision variable as a linear combination of subjectively weighted decision attributes.

For the analysis shown in Fig. 5D, we binned the decision variable into equally

populated bins (shown on the x-axis of Fig. 5D) and then determined the choice probability for each of these bins (as the number of observed choices for the first or second stimulus divided by the number of total choices in each bin); these data are represented by the circles in Fig. 5D. The line in Fig. 5D was obtained from fitting a logit function. For the analysis shown in Fig. 5E, we focused on the five choice contrasts that compared humans without physical impairments to all other stimulus categories. For each choice contrast, we calculated the average absolute (unsigned) value of the decision variable, which we term '*Δ Decision variable*'. Thus, larger values for *Δ Decision variable* would indicate that on average, subjects evaluated the stimulus categories as very different in terms of their preference, and accordingly would have a strong preference for one category over the other.

fMRI data acquisition. Functional MRI scanning was performed with a 7-Tesla whole-body MRI system (Magnetom 7T, Siemens Healthcare, Erlangen, Germany) at the Erwin L. Hahn Institute for Magnetic Resonance Imaging, Essen, Germany. The system is equipped with the SC72 gradient system capable of 70 mT/m maximum amplitude and a slew rate of 200 mT/m/ms. For this experiment, the scanner was equipped with a 1 channel transmit/ 32-channel receive head coil (Nova Medical, Wilmington, MA, USA). For each participant, a T1-weighted high-resolution anatomical scan (same slice prescription as echoplanar images (EPI)) and magnetization-prepared rapid-acquisition gradient echo (MPRAGE) were acquired for registration purposes (repetition time (TR) = 2500 ms, echo time (TE) = 1.27 ms, inversion time (TI) = 1100 ms, flip angle = 7°, Field of View (FOV) = 270 × 236 mm², matrix = 394 × 345, sagittal plane, slice thickness = 0.7 mm, 256 slices with a non-interpolated voxel size of 0.7 × 0.7 × 0.7 mm^3). For the acquisition of functional images, subjects were scanned in six subsequent sessions, each lasting about 12 min to acquire a total of 2022 volumes (331 – 343 volumes per session). In addition, subjects were scanned during two functional localizer tasks not used in the present paper, each lasting about 90 seconds. Whole-brain functional T2*-weighted EPI were acquired with an bold contrast-sensitive EPI sequence (Poser and Norris, 2009b, a) optimized for 7.0-T (slice thickness, 1.51 mm; 144 coronal slices; TR = 2000 ms; TE = 22 ms; flip angle, 14°; matrix, 170 × 170; FOV = 256 × 256 mm², order of acquisition of slices: interleaved). As head coil array allows massive parallel imaging, the GRAPPA (Generalized Autocalibrating Partially Parallel Acquisitions) algorithm was used with a reduction factor of R = 9 to reconstruct the undersampled k-space (Griswold et al., 2002). B0 fieldmaps were acquired prior to the EPI-sequence.

fMRI data analysis. We performed the fMRI data analysis using statistical parametric mapping (SPM8; Wellcome Trust Centre for Neuroimaging, London). Preprocessing included realignment of functional data including motion correction, normalization to the Montreal Neurological Institute (MNI) coordinate system, and smoothing with a Gaussian kernel with full width at half maximum (FWHM) of 6 mm. A high-pass temporal filter with a cut-off period of 128 s was applied. General linear models (GLMs) assuming first-order autoregression were applied to the time course of activation in which event onsets were modelled as single impulse response functions convolved with the canonical hemodynamic response function. Time derivatives were included in the basis functions set. Linear contrasts of parameter estimates were defined to test specific effects in each individual dataset. Voxel values for each contrast resulted in a statistical parametric map of the corresponding *t* statistic. In the second (group random-effects) stage, subject-specific linear contrasts of these parameter estimates were entered into one-sample t-tests, as described below, resulting in group-level statistical parametric maps. For parametric modulators, we used the standard SPM8 settings by which regressors are orthogonalized in the order they are entered into the design matrix. To test that our results are robust with respect to regressor orthogonalization,

we performed region-of-interest timecourse analyses (described below) in which regressors competed to explain variance in the absence of orthogonalization. Shared variances between our main variables of interest, computed within subjects and then averaged across subjects, were as follows: likability and familiarity, $R^2 = 0.39$ (± 0.05); likability and human-likeness, $R^2 = 0.26$ (± 0.03); familiarity and human-likeness, $R^2 = 0.36$ (± 0.04); likability and human detection, $R^2 = 0.24$ (± 0.03); familiarity and human detection, $R^2 = 0.24$ (± 0.03); human-likeness and human detection, $R^2 = 0.39$ (± 0.03). Variance inflation factors (VIF) in our GLMs were within acceptable limits (mean VIF = 2.28 ± 0.19; (Kutner et al., 2004)). For the rating regressors in SPM analyses and region-of-interest analyses we did not perform any weighting of the different rating variables but simply entered the mean-centred variables as regressors in the GLM. We estimated the following GLMs to test specific hypotheses:

GLM 1. This GLM served three purposes: (1) to identify brain areas with rating-task activity related to rated likability, familiarity and human-likeness (Fig. 2A, 3A, I); (2) to identify brain areas with choice-task activity related to the main decision variable and to rated decision confidence (Fig. 6A, 7A); (3) to identify brain areas with higher activity in the choice task than in the rating task (Table 2). For each subject we estimated a GLM with the following regressors of interest: (R1) an indicator function for the stimulus onset during the rating task; (R2) R1 modulated by the trial-specific human-likeness rating; (R3) R1 modulated by the trial-specific likability rating; (R4) R1 modulated by the trial-specific familiarity rating; (R5) an indicator function for the onset of the rating scales; (R6) an indicator function for the onset of the first stimulus during the choice task; (R7) an indicator function for the onset of the second stimulus during the choice task; (R8) R7 modulated by the trial-specific decision variable; (R9) R7 modulated by the trial-specific confidence rating; (R10) an indicator function for the onset of the choice phase; (R11) – (R16) the motion parameters resulting from the realignment pre-processing step as covariates of no interest; (R17) – (R22) six session constants. (Note that R1-R16 were defined separately for each scanning session.)

GLM 2. This GLM served to identify brain areas with choice-task activity related to the relative human-likeness (Fig. 6H). For each subject we estimated a GLM with the following regressors of interest: (R1) – (R6) as above; (R7) an indicator function for the onset of the second stimulus during the choice task; (R8) R7 modulated by the trial-specific confidence rating; (R9) R7 modulated by the trial-specific relative likability; (R10) R8 modulated by the trial-specific relative familiarity; (R11) R18 modulated by the trial-specific relative human-likeness; (R12) an indicator function for the onset of the choice phase; (R13) an indicator function for the onset of the confidence rating scale; (R14) – (R20) the motion parameters resulting from the realignment pre-processing step as covariates of no interest; (R21) – (R26) six session constants. (Note that R1-R20 were defined separately for each scanning session.)

GLM 3. This GLM served to identify brain areas with differential activity between specific stimulus categories (Fig. 3E, Fig. 6E). For each subject we estimated a GLM with the following regressors of interest: (R1) – (R6) indicator functions for the stimulus onsets during the rating task for each of the six stimulus categories defined above; (R7) an indicator function for the onset of the rating scales; (R8) – (R16) indicator functions for the onsets of the first stimulus during the choice task for each of the nine choice contrasts defined above; (R17) – (R25) indicator functions for the onsets of the second stimulus during the choice task for each of the nine choice contrasts defined above; (R26) an indicator function for the onset of the choice phase; (R27) – (R32) the motion parameters resulting from the realignment pre-

processing step as covariates of no interest; (R33) – (R38) six session constants. (Note that R1-R32 were defined separately for each scanning session.)

Functional connectivity analysis. We assessed functional connectivity using the psychophysiological-interaction (PPI) approach (Friston et al., 1997; Gitelman et al., 2003). For each subject we first extracted eigenvariates for a 6 × 6 × 6 voxel cluster around a seed voxel based on the peak voxels identified by a contrast between the choice task and the rating task. The peak voxel used for each subject was determined using a leave-one-subject-out procedure by re-estimating our second level analysis 21 times, each time leaving out one subject. Starting at the respective peak voxel for correlation with sequence length we selected the nearest peak in these cross-validation analyses. Time courses were de-convolved with the canonical hemodynamic response function (HRF) to construct a time series of neural activity in the region of interest. The regressors were constructed using the standard deconvolution procedure as implemented in SPM8 (Gitelman et al., 2003). For each model, we calculated single-subject first-level contrasts for the PPI regressor (R1) that were then entered into a second level analysis by calculating a one-sample t-test across the single subject coefficients. We estimated the following PPI GLMs.

PPI 1. This GLM tested for differential coupling between brain areas as a function of stimulus category (humans vs. non-humans) in the rating task. A schematic summary of the results is shown in Fig. 6J. The model contained the following regressors: (R1) a psychophysiological interaction regressor between the time series of activity in a seed brain area, extracted as just described, and a contrast between trials with human stimuli and non-human stimuli; (R2) the time series of activity in a seed brain area, extracted as just described; (R3) a contrast between human vs. non-human stimuli; (R4-R9) the motion parameters resulting from the realignment pre-processing step as covariates of no interest; (R10-R15) six session constants. (Note that R1-R9 were defined separately for each scanning session.) This model was estimated for the seed region DMPFC.

PPI 2. This GLM tested for differential coupling between brain areas as a function of task (choice task vs. rating phase). A schematic summary of the results is shown in Fig. 6J The model contained the following regressors: (R1) a psychophysiological interaction regressor between the time series of activity in a seed brain area, extracted as just described, and a contrast between choice task vs. rating task; (R2) the time series of activity in a seed brain area, extracted as just described; (R3) a contrast between choice task vs. rating task; (R4-R9) the motion parameters resulting from the realignment pre-processing step as covariates of no interest; (R10-R15) six session constants. (Note that R1-R9 were defined separately for each scanning session.) This model was estimated for the seed regions TPJ, VMPFC, FFG, DMPFC, amygdala.

Statistical significance testing. For all fMRI analyses, we report effects that survive correction for multiple comparisons across the whole brain using a significance level of $P < 0.05$ (family-wise error) at cluster level, imposed on maps that were displayed at cluster-defining threshold of $P < 0.005$ with minimum cluster size of $k = 10$ contiguous voxels. In addition, we used small volume correction ($P < 0.05$, cluster-level) in structures for which we had strong *a priori* hypotheses based on previous studies, including the VMPFC, ventral striatum, DMPFC, and amygdala. Small volume corrections were performed in spheres of 10 mm radius for cortical areas (VMPFC, DMPFC) and spheres of 6 mm radius for subcortical areas (ventral striatum, amygdala). The spheres were centred on specific coordinates reported in previous studies as follows. VMPFC [-2 40 -4] and ventral striatum [10 14 -4], taken from

a meta-analysis of value-based decision-making (Clithero and Rangel, 2014); amygdala: [18, -6, -22] taken from a previous study of value-based decisions using a similar experimental design to the present study but with different kinds of stimuli (Grabenhorst et al., 2013); DMPFC [2 44 36] taken from a previous study of social decision-making (Wittmann et al., 2016). (For DMPFC, we chose this particular study for our coordinate definition as it investigated DMPFC function in a task that required evaluating social others' performance to guide own decisions. We reasoned that similar processes might be engaged in our task, which required subjects to decide which agent would be capable of selecting a personal gift for them. Significant effects for human-likeness and human-non human contrast in DMPFC were also found when defining coordinates based on the Neurosynth meta-analysis database (Yarkoni et al., 2011), which localizes DMPFC at [0 56 22] using the search term 'social').

Region of interest analysis. To ensure that statistical inference in our region of interest analyses was not circular, we followed approaches used in previous studies (Behrens et al., 2008; Kriegeskorte et al., 2009; Esterman et al., 2010; Zangemeister et al., 2016). Specifically, we used a leave-one-subject-out method in which we re-estimated a second-level analysis 21 times, each time leaving out one subject to define the ROI coordinates for the left-out subject. We then extracted the signal from the subject-specific coordinates defined in this way. Thus, the data which we used for the region of interest analysis were independent from those used to define the coordinates for extracting the signal. Following data extraction we applied a high-pass filter with a cut off period of 128 s. The data were then z-normalized, oversampled by a factor of 10 using sinc-interpolation, and separated into trials to produce a matrix of trials against time. We generated separate matrices for each event of interest (rating trials, choice trials). We then fitted GLMs to each oversampled time point across trials separately in each subject.

For the rating task, our analysis strategy for region of interest analyses was as follows: We first fitted a GLM containing as main regressors the three (parametrically varying) rating variables likability, human-likeness and familiarity. We then tested whether the model fit was improved by inclusion of an additional binary human detection regressor that modelled the difference between human and non-human stimuli. We accepted this extended GLM if it yielded a better model fit as assessed with Akaike Information Criterion (AIC) and significant human-detection regressor. In the figures, we plot the standardized regression coefficients for significant regressors only; we report in the Results text which model provided the best fit and which regressors were significant. In addition to these regressors, the GLMs included motion parameters and session constants as covariates of no interest. This GLM analysis yielded one regression coefficient for each regressor for every oversampled time point in each subject. We entered individual-subject coefficients into one-sample t-tests (random-effects analysis, $P < 0.05$) and calculated group averages and standard errors for each time point across participants, yielding the across-subject effect size time courses shown in the figures. These mean effect size time courses are shown in Fig. 2B, Fig. 3B, F, J, Fig. 6B, F, H, Fig. 7B, C.

To test for relationships between specific behavioral and neural effect sizes, we extracted neural effects sizes from individual subject's data using the leave-one-out procedure described above. We only tested for these relationships if a ROI showed a significant effect in the tested variable. We performed linear regression to produce the plots shown in the figures. We tested statistical significance using Pearson correlation. As the behavioral UV depths in the rating task and choice task entered these analyses multiple times, we performed a Bonferroni correction. Specifically, we obtained a P-value of $P < 0.0125$ for tests involving the behavioral UV depth in the rating task and a P-value of $P < 0.0167$ for

tests involving behavioral UV depth in the choice task. The resulting effect size scatter plots are shown in Fig. 2D, Fig. 3D, H, L, Fig. 6D, I, Fig. 7D.

RESULTS

Behavioral Uncanny-Valley reactions in the rating task. Subjects performed a psychophysical rating task in which they evaluated humans and different kinds of artificial agents on the key dimensions of the UV hypothesis, including human-likeness, likability, and familiarity (Fig. 1A, B). Our approach to evoke UV reactions followed two concepts: First, we followed Mori's original hypothesis that likability increases with human-likeness but sharply decreases for highly human-like artificial agents (Mori, 1970; Mori et al., 2012). Second, we followed the argument that humans can also fall in the UV if they significantly deviate from typical human appearance or behavior (Pollick, 2010). Accordingly, we tested not only human-like androids but, to elicit strong UV reactions, we designed 'artificial humans' derived from photographs of humans with artificially altered facial features. These artificial or synthetic humans were more human-like than typical androids but deviated from human appearance by having exaggerated smooth, flawless, unnatural faces and slightly unnatural proportions.

In total, we used four categories of artificial agents (Fig. 1B), based on independent pre-tests: (1) 'artificial humans', (2) 'android robots', (3) 'humanoid robots', (4) 'mechanoid robots'. We hypothesized that the human-like artificial agents (in particular artificial humans, and androids) would fall in the UV, i.e. having lower likability ratings than predicted by a human-likeness continuum (defined in the next paragraph), whereas less human-like robots (humanoid and mechanoid robots) should lie outside the UV. As controls, we included photographs of humans with and without physical impairments, the former having presumed lower familiarity than the latter.

Stimulus categories differed significantly in rated likability ($F(1, 20) = 46.19$; $P < 0.001$; $\eta^2 = 0.698$; repeated-measures ANOVA), familiarity ($F(1, 20) = 54.82$; $p < 0.001$; $\eta^2 = 0.733$) and human-likeness ($F(1, 20) = 119.95$; $P < 0.001$; $\eta^2 = 0.857$). As predicted from the UV hypothesis, subjects evaluated the stimulus categories along a human-likeness continuum (Fig. 1C, left; $R = 0.980$, $P = 0.0006$, linear regression): mechanoid and humanoid robots constituted the low end of this continuum, humans constituted the high end, and the UV-relevant artificial humans and android robots were rated intermediately human-like. [Slightly lower ratings for humans with physical impairments, compared to those without physical impairments, were explained by lower familiarity: human-likeness and likability ratings depended on familiarity ratings for humans with physical impairments (both $R > 0.5$, $P < 0.05$, Pearson correlation) but not for humans without physical impairments ($P > 0.22$)]. Within this observed human-likeness continuum, we next measured UV reactions in likability.

Across all stimuli, likability tended to increase with human-likeness (Fig. 1D). The key prediction of Mori's UV hypothesis is that, although likability generally increases with human-likeness, highly human-like artificial agents are less likable than predicted by a human-likeness continuum, and thereby fall in the UV. Consistent with this prediction, likability ratings for artificial humans (and some specific androids) were lower than expected based on their rated human-likeness (Fig. 1D). As proposed by the UV hypothesis (Mori, 1970) and found in previous work (depending on the appearance dimension that was varied, for instance, prototypicality (Burleigh et al., 2013)), likability data were well described by a cubic polynomial fit ($R^2 = 0.573$, Fig. 1D), with consistent results in individual subjects' fits (mean $R^2 = 0.362 \pm 0.03$). A distinct UV effect was visible in the average likability of

artificial humans, which was lower than expected based on the human-likeness continuum (Fig. 1E). Thus, ratings across subjects and stimuli indicated the existence of a UV effect.

Modelling the Uncanny Valley. To examine neural correlates of UV reactions, it was critical to first model and quantify the UV psychometrically within individual subjects. As we observed the most pronounced UV effect for artificial humans, we focused on this stimulus category (see Methods; including android stimuli yielded similar results).

We used a direct approach and quantified the UV effect as the extent to which likability ratings for artificial humans deviated from a linear fit of likability to human-likeness, calculated across all other stimulus categories ('UV depth', Fig. 1F). This approach captured Mori's original notion of the UV: an individual subject would have a stronger UV reaction (i.e. a deeper UV) if likability ratings for human-like stimuli were lower than linearly predicted from that subject's human-likeness ratings. Figure 1F illustrates the approach and resulting UV depth in one example subject, with further examples and across-subjects distribution shown in Fig. 1G. The visible downward deflections of likability from linear human-likeness fits (Figs. 1F, G, compare red and blue data points, indicating observed likability and likability predicted from human-likeness, respectively) implied that artificial humans were less liked than expected from a linear human-likeness function.

We found robust behavioral UV reactions in the rating data. Despite considerable inter-individual variation, UV depths for artificial humans were significantly larger than predicted from human-likeness (indicating lower likability than predicted; $P = 3.47 \times 10^{-8}$, one-sample t-test). UV depth was significant in 18 of 21 individual subjects ($P < 0.05$, one-sample t-tests). (Weaker but significant UV reactions were found for familiarity ($P = 1.9 \times 10^{-4}$) but, as expected, not for human-likeness ($P = 0.117$)). A significant UV effect was also found when including both artificial humans and androids ($P = 6.1 \times 10^{-6}$). As controls, UV depths were non-significant for stimuli not hypothesized to fall in the UV, (e.g. humans with physical impairments or humanoid robots; both $P > 0.15$, one-sample t-tests), for which likability was well predicted by a linear human-likeness fit.

We confirmed the robustness of these results by an alternative approach to model the UV based on regression residuals. Within subjects, we fitted six linear regressions of likability ratings on human-likeness ratings (cf. Fig. 1D), each time leaving out the data from one of the six stimulus categories. We then applied the estimated regression coefficients on the data of the left-out category to predict likability from human-likeness and noted the regression residuals (i.e. the variance in likability not explained by human-likeness given the specific regression coefficients). According to the UV hypothesis, these residuals should be significantly more negative for the UV-relevant artificial humans (and possibly android robots) compared to other stimulus categories. Indeed, residuals for artificial humans were significantly smaller (more negative) than those for all other agents (all comparisons $P < 0.005$, t-test, corrected for multiple comparisons), while the residuals for androids were significantly smaller than those for humanoids and humans without physical disabilities (both $P < 0.005$). The UV effect quantified within each subject based on these residuals was highly correlated with our main measure of the UV effect ($R = 0.7486$, $P = 9.4 \times 10^{-5}$, Pearson correlation). Thus, this procedure provided additional validation for the existence of a UV effect in the behavioral data.

Thus, behavioral data confirmed the existence of a UV reaction in our stimulus set and validated the rating task to search for neural correlates of subjective UV reactions. We next investigated how neural systems might transform a linear human-likeness function into a nonlinear UV function.

A neural Uncanny Valley in ventromedial prefrontal cortex. To locate neural correlates of UV reactions, we regressed stimulus-evoked activity in the rating task on trial-by-trial rated likability, which was the key rating to reflect the UV (shown above). Likability evaluations were encoded in typical reward areas, including ventromedial prefrontal cortex (VMPFC, Fig. 2A, B). Additionally, VMPFC activity reflected human-likeness ratings (Fig. 2B). Because likability and human-likeness were covariates, these variables accounted for different activity-components. Familiarity did not explain VMPFC activity. [All region-of-interest (ROI) analyses used leave-one-subject-out cross-validation to identify subject-specific, independent ROI coordinates, ensuring unbiased analysis (Kriegeskorte et al., 2009; Esterman et al., 2010). ROI regressions were performed without orthogonalization. All three ratings were included as regressors; in addition, we tested in a second GLM whether a binary human detection contrast would improve model fit (see Methods).] We tested whether the GLM in the region-of-interest analysis was improved by inclusion of a binary 'human detection' regressor that contrasted human and non-human stimuli. (For this contrast, non-human agents involved mechanoid, humanoid, android and artificial-human stimuli.) Inclusion of this human detection regressor did not change the significant effects for likability and human-likeness shown in Fig. 2B; although the human detection regressor itself showed a significant effect on VMPFC activity, this occurred quite late outside our primary analysis window ($P = 0.033$, one-sample t-test at 9 s post-stimulus onset). Thus, VMPFC integrated the two key UV-dimensions likability and human-likeness, suggesting a role in UV reactions.

If a brain area signalled the key UV dimension likability, it might also explicitly represent the UV reaction. [In an 'explicit' UV representation, activity patterns across stimuli should match the prototypical UV shape.] Consistent with this notion, VMPFC activity across stimuli closely resembled the behavioral UV reaction (Fig. 2C): activity increased approximately linearly according to human-likeness for most stimuli, but responses to artificial humans were significantly lower than expected from a linear human-likeness fit (Fig. 2C; $P = 0.008$, one-sample t-test). VMPFC activity thus followed the typical UV shape with selectively lower activity for highly humanlike artificial agents. Across subjects, neural UV depths (derived from individual subjects' VMPFC activities) matched subjects' behavioral UV reactions (Fig. 2D; $R = 0.576$, $P = 0.006$, Pearson correlation, Bonferroni-corrected). When calculating this across-subjects effect with the alternative UV-quantification based on regression residuals (described above), the relationship with neural UV depths in VMPFC was weaker ($R = 0.41$, $P = 0.062$). Accordingly, the relationship to individual differences should be considered exploratory.

Thus, VMPFC activity integrated the key UV dimensions likability and human-likeness to an explicit representation of the UV reaction. The strength of this representation partly explained individual differences in behavioral reactions toward artificial agents.

Linear and nonlinear human-likeness signals as neural basis for the Uncanny Valley. We next searched for activities related to the UV dimension human-likeness. We reasoned that different types of human-likeness signals might constitute neural inputs required for transforming the psychophysical human-likeness continuum (Fig. 1C) into the nonlinear UV representation observed in VMPFC (Fig. 2C).

Neural responses to human and artificial stimuli in temporo-parietal junction (TPJ), dorsomedial prefrontal cortex (DMPFC), and part of fusiform gyrus (FFG) were related to human-likeness ratings (Fig. 3, Table 1). TPJ activity showed a positive linear relationship with human-likeness (Fig. 3A-C): it gradually increased across stimuli and faithfully reflected the human-likeness continuum. Likability and familiarity did not show significant effects on TPJ activity; a human detection regressor was not significant and its inclusion did not affect

the significance of the human-likeness regressor. TPJ thus provided a parametric, linear human-likeness signal—the most basic element of the UV hypothesis.

By contrast, human-likeness coding in DMPFC was more complex. Although we found a significant relationship with parametric human-likeness (Table 1, GLM1), a binary contrast showed significantly stronger activation by human agents compared to non-human agents (Fig. 3E; GLM3, contrasting mechanoid, humanoid, android robots and artificial humans with both human stimulus categories). Detailed region-of-interest analysis indicated that neural activity in this DMPFC area was best explained by a human detection contrast as follows. Across stimuli, DMPFC activity followed the human-likeness continuum for non-human agents but then sharply increased for human agents (Fig. 3E-G). We modelled this activity with a 'human detection' regressor (Fig. 3F, a dummy variable distinguishing human from non-human stimuli) in addition to linear human-likeness (improvement in GLM fit was assessed by Akaike Information Criterion). [Although Fig. 2C and 3G may look similar, it is important to note that these data are averaged across trials and subjects; our region of interest analysis within each subject indicated that while VMPFC activity was best explained by joint likability and human likeness coding, DMPFC activity was best explained by a human detection regressor.] Thus, DMPFC activity emphasized differences between human and non-human stimuli, suggesting a role in distinguishing human from artificial agents.

FFG exhibited a third type of human-likeness signal. It showed a negative parametric relationship with human-likeness selectively for non-human stimuli and an average response to human stimuli (Fig. 3I-K, modelled in the same way as DMPFC activity). Both human-likeness and human detection explained significant variance in FFG activity (Fig. 3J, inset). Thus, FFG combined a human-detection response with negative human-likeness coding selectively for artificial agents.

If TPJ, DMPFC and FFG contributed to UV reactions, their sensitivity to human-likeness should reflect individuals' UV depths. Indeed, all three areas encoded human-likeness more strongly for subjects with stronger UV reactions (Fig. 3D, H, L). [These relationships were significant for human-likeness coefficients ($P < 0.05$, Pearson correlation); in DMPFC, the effect was strongest for human-detection and in FFG for differential human-likeness.] Similar effects were not found in control analyses with non-UV stimulus categories ($P > 0.6$). We note that the effect in Fig. 3L did not survive Bonferroni-correction for multiple comparisons; accordingly, we treat this result as an exploratory finding.

Thus, TPJ, DMPFC and FFG encoded distinct linear and nonlinear human-likeness signals that were related to individuals' UV reactions.

Modelling the neural Uncanny Valley from human-likeness signals. These data suggest a possible information-processing sequence, whereby a linear human-likeness code in TPJ is transformed to nonlinear human-likeness codes in DMPFC and FFG and eventually to an explicit UV representation (i.e. nonlinear likability) in VMPFC. Specifically, the observed FFG signal seemed to track the proximity to a human-nonhuman boundary (decreasing activity selectively for non-humans), which seems suited for determining the UV-typical likability-drop for the most human-like artificial agents. The DMPFC human-detection signal might be relevant for setting up such selective human-likeness signalling.

Consistent with these notions, a simple multiplicative integration of linear and nonlinear human-likeness signals in TPJ and FFG approximated the UV-related activity pattern in VMPFC (Fig. 4A). Such multiplicative signal integration is biologically plausible and routinely used in models of neural multisensory integration (Pena and Konishi, 2001; de Araujo et al., 2003; Small et al., 2004; Stein and Stanford, 2008). Across subjects, the strengths of the human-likeness components in TPJ and FFG were related to the neural UV-

effect in VMPFC (Fig. 4B, C), suggesting these combined signals were relevant for the observed UV effect in VMPFC.

Thus, the neural UV-code in VMPFC could be approximated by multiplicative integration of human-likeness signals from TPJ and FFG. We next tested in a second experiment how these activities were related to behavioral choices toward artificial agents.

Uncanny-Valley reactions guiding decision-making. In a second experimental task (Fig. 5A), subjects viewed sequential human and artificial agents and chose from whom they would rather receive a personal gift. We instructed subjects that each human and artificial agent had selected a gift from an option set, and that they would receive one of these gifts based on their choices for the different agents during the experiment. Actual gift choices were unknown to the subjects; they had to decide whom they would trust to select an attractive gift. We reasoned that this situation likely encouraged subjects to compare humans and artificial agents on UV-relevant dimensions.

When choosing whether to prefer a gift from a human or artificial agent, subjects typically (but not always) preferred humans (Fig. 5B, black bars). Choice probabilities were more variable when choosing between artificial agents (grey bars). We next sought to explain subjects' choices in terms of the separately made ratings of key UV dimensions.

In value-based choice, a decision-maker integrates relevant information to a 'decision variable' (a weighted composite of decision factors) to compare and decide between different options. Accordingly, we reasoned that weighted, comparative valuations of likability, human-likeness and familiarity guided subjects' choices. We defined trial-specific relative differences in likability, human-likeness and familiarity based on individual subjects' stimulus ratings (from the rating task). Logistic regression showed that relative likability, familiarity and human-likeness were significant predictors for subjects' choices: subjects were more likely to accept a gift from an individual they considered relatively more likable, familiar, and human-like (Fig. 5C, all regression coefficients $P < 0.0002$, one-sample t-tests). The same result was found when considering only choices involving UV-relevant artificial humans and androids (all coefficients $P < 0.0009$, one-sample t-tests). On average, logistic regression based on relative ratings provided correct choice classification for 85.7% of trials (mean across subjects, ± 0.89, s.e.m.) and resulted in a pseudo-R^2 of 0.53 (mean across subjects, ± 0.02, s.e.m.). Choice probabilities were thus well described by a subject-specific decision variable, defined as a weighted sum of relative likability, familiarity and human-likeness (Fig. 5D).

To quantify the UV in the choice task, we examined the decision variable for specific stimulus categories. For this analysis and subsequent neural analyses, we focused on the unsigned decision variable i.e. the absolute value difference. [In value-based decisions, neural decision signals often reflect this absolute value difference between choice options (Heekeren et al., 2004; Rolls et al., 2010a, b; Hunt et al., 2012; Grabenhorst et al., 2013), which is usually interpreted as the signature of a competitive decision mechanism.]

Similar to likability ratings (Fig. 1F), the decision variable followed a human-likeness continuum: the more two stimulus categories differed in human-likeness, the more they also differed in the decision variable (Fig. 5E; $R = -0.533$, $P = 1.8 \times 10^{-7}$, Pearson correlation), with highest differences for humans vs. mechanoid robots and smallest differences for humans with and without physical impairments. (A high difference in the decision variable between stimuli indicated clear choice preference between these stimuli.) However, consistent with the UV hypothesis, the difference in the decision variable between humans and artificial humans was significantly larger than expected from a human-likeness continuum (Fig. 5F; $t(20) = 4.66$, $P = 1.52 \times 10^{-4}$, one-sample t-test), which matched the UV reaction observed in likability ratings (Fig. 1D). This upward deflection from the linear

human-likeness fit (Fig. 5E, difference between red and blue data points, indicating actual difference in decision variable and difference in decision variable estimated from linear human-likeness fit, respectively) implied a greater difference than expected if the decision variable followed a human-likeness continuum. This result quantified the UV during choices in terms of individual subjects' decision variable. [Choice UV-effects were not found for control stimuli (humanoid robots; $P = 0.505$, one-sample t-test).]

Thus, choices between humans and artificial agents were based on a decision variable that reflected subjects' individual UV reactions. These data validated the choice task for examining neural correlates of decision-making in the context of the UV hypothesis.

Neural Uncanny-Valley components during decisions. A contrast showed that areas involved in the rating task were even more strongly activated in the choice task, including VMPFC, DMPFC and TPJ (Table 2). To examine neural UV components during decisions, we regressed choice-task activity (responses to humans and artificial agents at time of the second stimulus on each trial) on individual subjects' decisions variables, i.e. the weighted sums of relative likability, familiarity and human-likeness.

Activity in VMPFC tracked subjects' decision variable, with stronger activity for larger unsigned value differences (Fig. 6A, B). Similar effects occurred in other areas implicated in decision-related valuations (Bartra et al., 2013), including striatum and posterior cingulate cortex (Table 2). VMPFC encoding of the decision variable was not accounted for by subjective decision difficulty: confidence ratings (a common measure of subjective difficulty) explained separate VMPFC activity components (Fig. 6B). Across decision categories, VMPFC activity followed a human-likeness continuum but deviated significantly for choices involving UV-relevant artificial humans, thereby matching the psychophysical UV (Fig. 6C). [Note that Fig. 6C shows VMPFC activity when subjects chose whether to receive a gift from humans compared to different agents, labelled on x-axis.] The significant deviation from linear fit for choices between humans and artificial humans (Fig. 6C, compare red and blue points) suggested stronger VMPFC activation (indicating more disparate value difference) than expected from a human-likeness continuum, specifically for artificial agents that elicited UV reactions. Across subjects, this neural UV correlated with subjects' behavioral UV depths (Fig. 6D). Thus, VMPFC encoded the behaviorally important decision variable in close relation to individual differences and reflected UV reactions during decision-making.

Previous studies showed separate functions of VMPFC and DMPFC during choices for self and others (Nicolle et al., 2012; Wittmann et al., 2016) and during evaluations of similar and dissimilar others (Mitchell et al., 2006). We therefore examined decision-activity of these areas for choices involving humans (vs. other stimulus categories) and choices not involving humans (choices among artificial agents). Contrasting decision-trials involving humans and non-humans showed stronger activation in both DMPFC and VMPFC (Fig. 6E). Given this differential activation, we performed a region-of-interest regression on DMPFC and VMPFC activity in which the decision variable was modelled separately for choices involving humans and choices involving non-humans. This analysis revealed selective coding of the decision variable in DMPFC for choices involving non-humans (Fig. 6F, top, negative coding scheme). By contrast, VMPFC coded the decision variable specifically for choices involving humans (Fig. 6F, bottom, positive coding scheme). Thus, VMPFC and DMPFC coded subjects' decision variable selectively, and complementarily, in a human-nonhuman frame of reference.

Regressing decision-activity specifically on relative human-likeness, a subcomponent of the decision variable important in UV-reactions, showed a significant effect in the TPJ region that also encoded human-likeness during ratings (Fig. 6G, H). Across subjects, TPJ's

sensitivity to human-likeness was related to behavioral UV reactions (Fig. 6I), suggesting behavioral relevance for subjects' choices. Additionally, similar to the rating task, region-of-interest regression on FFG activity showed a significant human-detection contrast ($t(20) = 2.27$, $P = 0.033$, one-sample t-test) and negative human-likeness coding ($t(20) = 4.39$, $P = 0.0003$, one-sample t-test).

We examined functional connectivity with psychophysiological interaction (PPI) analyses. We found that TPJ was more strongly connected with both DMPFC and FFG during choices compared to ratings (Fig. 6J, blue; Table 3), reflecting these areas' observed common human-likeness encoding. Functional connections also existed between VMPFC and both DMPFC and FFG (Fig. 6J, magenta), but we found no direct coupling between VMPFC and TPJ. Thus, areas implicated in valuations of human-likeness, likability and subjects' decision variable interacted functionally during decision-making.

Amygdala signals for rejecting human and nonhuman agents. The amygdala is a subcortical structure involved in emotion and social information processing (Phelps and LeDoux, 2005; Adolphs, 2010) that contributes to decision-making (Seymour and Dolan, 2008; Grabenhorst et al., 2012; Grabenhorst et al., 2013; Grabenhorst et al., 2016; Zangemeister et al., 2016). Contrast analysis suggested amygdala engagement in the choice task (Table 2). We therefore examined its role in UV-related decisions.

Similar to VMPFC, the amygdala encoded subjects' decision variable (Fig. 7A), but with a negative coding scheme whereby higher activity indicated lower decision values (Fig. 7B) and without additional confidence-coding ($P = 0.741$, one-sample t-test). Unlike VMPFC and DMPFC, the amygdala encoded subjects' decision variable irrespective of whether decisions involved humans or nonhuman agents (Fig. 7C, D), suggesting a context-invariant decision signal. As for VMPFC, amygdala's encoding of the decision variable was related to individual differences in UV reactions (Fig. 7E, black data). The amygdala was distinct as its coding was directly related to subjects' choice probabilities: subjects with higher amygdala decision-sensitivity were less likely to accept gifts from artificial agents (Fig. 7E, orange data). [Similar relationships were not found for VMPFC ($P = 0.777$, Pearson correlation) or human-gift acceptance ($P = 0.317$).] The amygdala was not functionally coupled to other areas, except for a non-significant effect in DMPFC. Thus, the amygdala encoded subjects' decision variable, and this coding partly explained gift refusal from UV-relevant artificial agents.

DISCUSSION

We investigated neural activity when subjects evaluated different attributes of human and artificial agents and made personal decisions about these agents. We found that stimulus-evoked activity in VMPFC, a key valuation structure (Chib et al., 2009; Hare et al., 2010; Wunderlich et al., 2010; Bartra et al., 2013), matched the characteristic nonlinear shape of the UV reaction (Mori, 1970; Mori et al., 2012): VMPFC responses to human and artificial agents increased according to the psychophysically measured human-likeness continuum but were markedly decreased for the most humanlike artificial agents ('artificial humans'), which fell in the psychophysically measured UV. Across subjects, the depth of this neural UV—formalized as a deviation from a linear human-likeness fit—explained individual differences in behavioral UV reactions. VMPFC was a candidate site for computing the UV response, as it jointly encoded human-likeness and likability signals. Consistently, during decision-making, VMPFC activity reflected the UV in terms of a decision variable (derived from subjectively weighted decision attributes) that guided subjects' choices between humans and artificial agents. These data demonstrate a surprisingly direct neural representation of the UV as a nonlinear value function and provide neurobiological evidence for the key prediction of

the UV hypothesis that the UV derives from integrated human-likeness and likability evaluations.

A distinct set of brain areas implicated in social information processing encoded the human-likeness dimension underlying the UV hypothesis. A human-likeness continuum was faithfully and linearly encoded by neural responses to artificial agents in the TPJ, consistent with this area's roles in agency-detection (Mar et al., 2007), belief-attribution (Saxe and Wexler, 2005; Carter and Huettel, 2013; Vogeley, 2017) and learning about others (Behrens et al., 2008; Hampton et al., 2008; Boorman et al., 2013). TPJ thus represented the most basic element of the UV hypothesis, human-likeness, on a linear scale of neural activity. Such a linear neural representation is computationally useful (Rolls and Treves, 1998), as it provides a versatile basis for deriving other, nonlinear neural representations of related variables, as observed in DMPFC, FFG and VMPFC. Thus, TPJ likely provided important inputs for generating neural UV representations, as also suggested by the observed relationship to individual differences.

By contrast, the DMPFC, a region implicated in mentalizing (Saxe and Wexler, 2005; Amodio and Frith, 2006), responded particularly strongly to human agents and coded human-likeness by a nonlinear, step-like function that emphasized differences between humans and nonhumans. Previous studies observed differential DMPFC activity when making choices for self or others (Nicolle et al., 2012), when tracking performance for self or other (Wittmann et al., 2016), and when attributing mental states to others (Mitchell et al., 2006). The present data extend these observations toward a distinction between human and non-human social others. Such human-detection activity might be critical in setting up a UV effect as it could set a threshold of 'humanlike but not genuinely human' that would trigger the characteristic drop in likability.

The FFG encoded an additional candidate input for computing UV reactions, by signalling linear human-likeness with a negative coding scheme up to the human-nonhuman threshold encoded in DMPFC. These activities are consistent with FFG responses to non-living compared to living entities (e.g. tools vs. animals) (Chao et al., 1999; Noppeney et al., 2006; Mahon et al., 2007; Chaminade et al., 2010). Our data suggest that preferential FFG responses to non-living entities can be elicited gradually by different types of artificial agents. The FFG's selective human-likeness signal was not directly expressed in psychophysical ratings but could reflect a potential input signal for neural UV computations.

Together, these data identify an apparent progression of activity patterns that reflected the transition from a linear human-likeness continuum in TPJ, to nonlinear human-likeness signals in DMPFC and FFG and toward a nonlinear UV value function in VMPFC. The observed functional connectivities between these areas supported this information-processing sequence. Notably, these neural signals were recorded during UV-relevant evaluations and explained inter-individual variation in UV reactions.

Such linear-nonlinear transformations are known from hierarchical sensory systems that gradually, over a series of neural representations, produce selective responses to specific feature combinations (Rolls and Treves, 1998; Olshausen and Field, 2004). We found that a multiplicative combination of linear and nonlinear human-likeness signals from TPJ and FFG was sufficient to approximate the observed neural UV representation in VMPFC. Multiplicative signal integration is biologically plausible and prevalent in sensory systems, where it produces nonlinear enhancement to specific feature combinations (Pena and Konishi, 2001; de Araujo et al., 2003; Small et al., 2004; Stein and Stanford, 2008). Based on these data, we suggest that the UV reaction can be conceptualized as a nonlinear neural valuation response elicited by a specific feature combination—high human-likeness in nonhuman agents. This response likely derives from multiplicative integration of linear and nonlinear human-likeness signals.

At the behavioral level, our data demonstrate how subjective UV reactions extend beyond perceptual impressions to active preference choices, guided by subjectively weighted decision attributes. To elicit UV reactions, we constructed artificial human stimuli from real human individuals that maximized human-likeness but imposed unnatural, flawless and smoothed facial appearances and body proportions. The present neural and behavioral data suggest that this novel manipulation is particularly effective in eliciting a UV reaction, in accordance with recent suggestions (Pollick, 2010). Notably, previous behavioral studies typically reported strongest UV reactions for androids (Piwek et al., 2014; Rosenthal-von der Putten and Kramer, 2014; Rosenthal-von der Putten and Weiss, 2015; Wang et al., 2015; MacDorman and Chattopadhyay, 2016). Here, inclusion of the highly human-like artificial humans may have elicited adaptation effects that maximized UV reactions for artificial humans while attenuating UV reactions for androids. Such adaptive coding is well established in neural valuation systems, whereby neural responses adapt to the current statistical distribution of stimuli (Tobler et al., 2005; Schultz, 2015). Future studies could test systematically how the range of presented stimuli affects UV reactions in behavior and neural activity.

Our results shed new light on the functions of different parts of medial prefrontal cortex that have been implicated in social and evaluative functions (Amodio and Frith, 2006). Previous studies reported involvement of DMPFC in evaluating traits of other people (Schiller et al., 2009; Denny et al., 2012), modelling others' values and choices (Nicolle et al., 2012; Suzuki et al., 2012; Stanley, 2016), and understanding others' intentions (Amodio and Frith, 2006; Mitchell et al., 2006). Despite these advances, the role of DMPFC in social cognition has remained elusive. Here we showed that DMPFC encoded a nonlinear human-likeness signal that emphasized the distinction between human and nonhuman agents. This signal in DMPFC likely derived from TPJ's linear representation of the human-likeness continuum and may have contributed to the nonlinear valuation function seen in VMPFC, as suggested by these areas' functional connections. VMPFC activity is consistently observed during value-based choice (Chib et al., 2009; Hare et al., 2010; Levy et al., 2010; Wunderlich et al., 2010; Grabenhorst and Rolls, 2011; Hunt et al., 2012; Bartra et al., 2013; De Martino et al., 2013; Zangemeister et al., 2016), including social choice (Hampton et al., 2008; Behrens et al., 2009; Suzuki et al., 2012; Sul et al., 2015). Our results advance understanding of VMPFC's functions by demonstrating how a nonlinear value function in VMPFC can be generated through multiplicative signal integration from other areas.

The amygdala is implicated in two functional domains that intersected in our study: the valuation of sensory events and the processing of social information (Phelps and LeDoux, 2005; Adolphs, 2010; Rolls, 2014). Specifically, previous studies showed amygdala involvement in face processing (Adolphs, 2010), trustworthiness evaluation (Winston et al., 2002), anthropomorphizing (Heberlein and Adolphs, 2004), and social impression formation (Schiller et al., 2009), which likely contributed to the observed amygdala activation in our decision task. Similar to VMPFC, amygdala encoding of subjects' decision variable is consistent with recent evidence implicating the amygdala in value-guided decisions (Seymour and Dolan, 2008; Grabenhorst et al., 2012; Grabenhorst et al., 2013; Hernadi et al., 2015; Grabenhorst et al., 2016). Different to VMPFC, amygdala coding of this decision variable reflected subjects' tendencies to reject gifts from artificial agents in the UV. Thus, the amygdala may distinctly contribute to inhibiting interactions with human-like artificial agents.

Our data suggest a novel, neurobiological conceptualization of human responses toward artificial social partners. In two experimental tasks, the VMPFC encoded an explicit representation of subjects' UV reactions as a nonlinear valuation function, by signalling selective low likability for the most humanlike artificial agents. This neural UV

representation seemed to derive from a multiplicative combination of linear and nonlinear human-likeness signals in functionally connected TPJ, FFG and DMPFC. Thus, human reactions toward artificial agents involve a selective, nonlinear neural valuation in response to a specific feature combination (human-likeness in nonhuman agents). These findings indicate that a basic sensory coding principle—enhanced neural feature selectivity through linear-nonlinear transformation—may also apply to human valuations of social partners, as shown here for specific artificial agents.

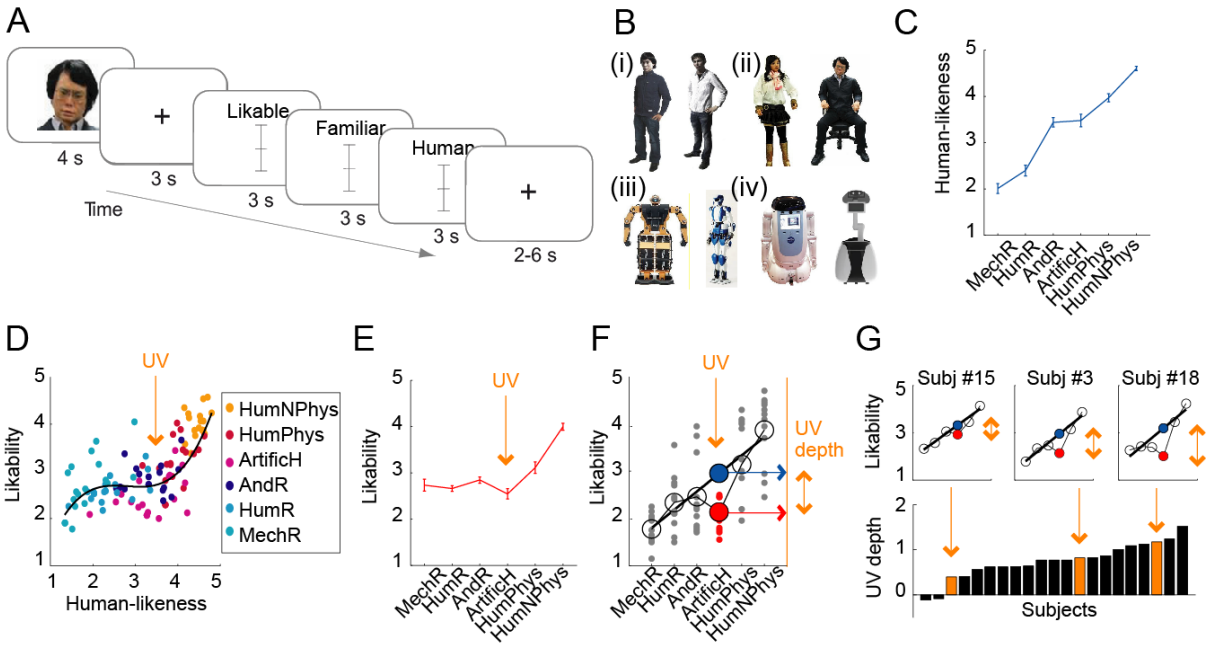

Fig 1. Behavioral Uncanny-Valley reactions. (A) Rating task. Subjects evaluated UV-relevant dimensions of humans and artificial agents. (B) Example stimuli. (i) Artificial humans; (ii) androids; (iii) humanoid robots; (iv) mechanoid robots. (Human examples not shown). (C) Human-likeness ratings for stimulus categories (MechR: mechanoid robots; HumR: humanoid robots; AndR: android robots; ArtificH: artificial humans; HumPhys: humans with physical impairments; HumNPhys: humans without physical impairments). [Relatively lower ratings for HumNPhys compared to HumPhys due to lower familiarity, see text.] (D-G) UV effect in likability. (D) Likability ratings for all stimuli. Black line: third-order polynomial fit (R^2 = 0.573). (E) Likability for artificial humans showed the UV-characteristic drop, indicating deviation from linear relationship between human-likeness and likability. (F) Quantifying UV depth. UV-definition in single subject's likability ratings. Black line: linear regression-fit of likability to human-likeness continuum from all stimuli (grey data) except UV-relevant artificial humans (red data). Blue point: predicted likability for artificial humans from linear fit; red point: observed likability. UV depth defined as difference between predicted and observed likability. (G) UV depths across subjects. The procedure in (F) defined UV depths, illustrated for three individual subjects (top) and all subjects (bottom; $P = 3.47 \times 10^{-8}$, one-sample t-test). UV effects were not found in several controls (e.g. humanoid robots; $P > 0.15$, one-sample t-test). Error bars: s.e.m.

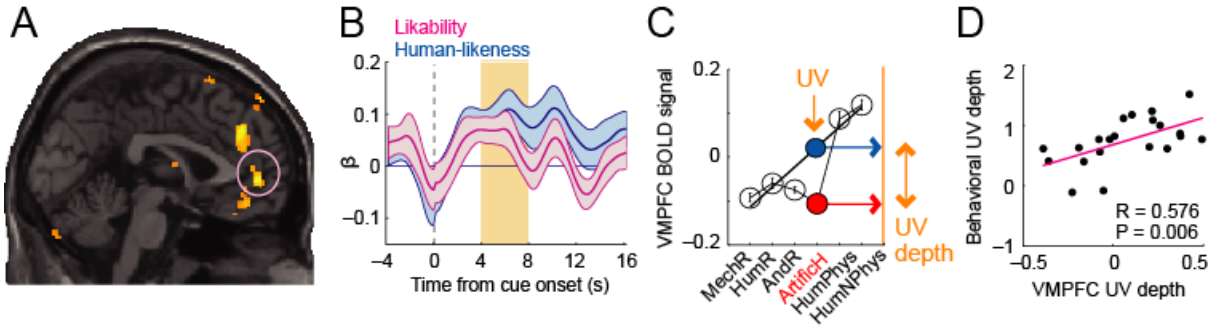

Fig. 2. Uncanny Valley in ventromedial prefrontal cortex. (A) VMPFC signalled subjective likability. Activity during rating task reflected trial-by-trial likability ratings ($P < 0.05$, whole-brain corrected at cluster level; all statistical maps thresholded at $P < 0.005$, uncorrected for display purposes). (B). Integration of likability and human-likeness. Region-of-interest regression of VMPFC activity on likability ($t(20) = 2.93$, $P = 0.008$; one-sample t-test in random-effects analysis) and human-likeness ($t(20) = 3.45$, $P = 0.003$). (C). VMPFC activity encodes UV reactions. As for likability ratings, VMPFC responses to artificial humans were lower than expected based on human-likeness continuum. Black line: linear regression-fit of VMPFC activity to human-likeness continuum from all stimuli except UV-relevant artificial humans ($R = 0.268$, $P = 0.008$, Pearson correlation). Blue point: predicted VMPFC response for artificial humans from linear fit; red point: measured response. UV depth defined as difference between predicted and measured response (deviation from linear fit: $t(20) = -2.97$, $P = 0.008$, one-sample t-test). (D) Neural UV depths matched behavioral UV depths. Linear regression of behavioral UV depths (Fig. 1F) on neural UV depths in VMPFC activity ($P = 0.006$; significant robust regression; mean-centred neural UV depths extracted from independently defined coordinates using leave-one-subject-out cross-validation).

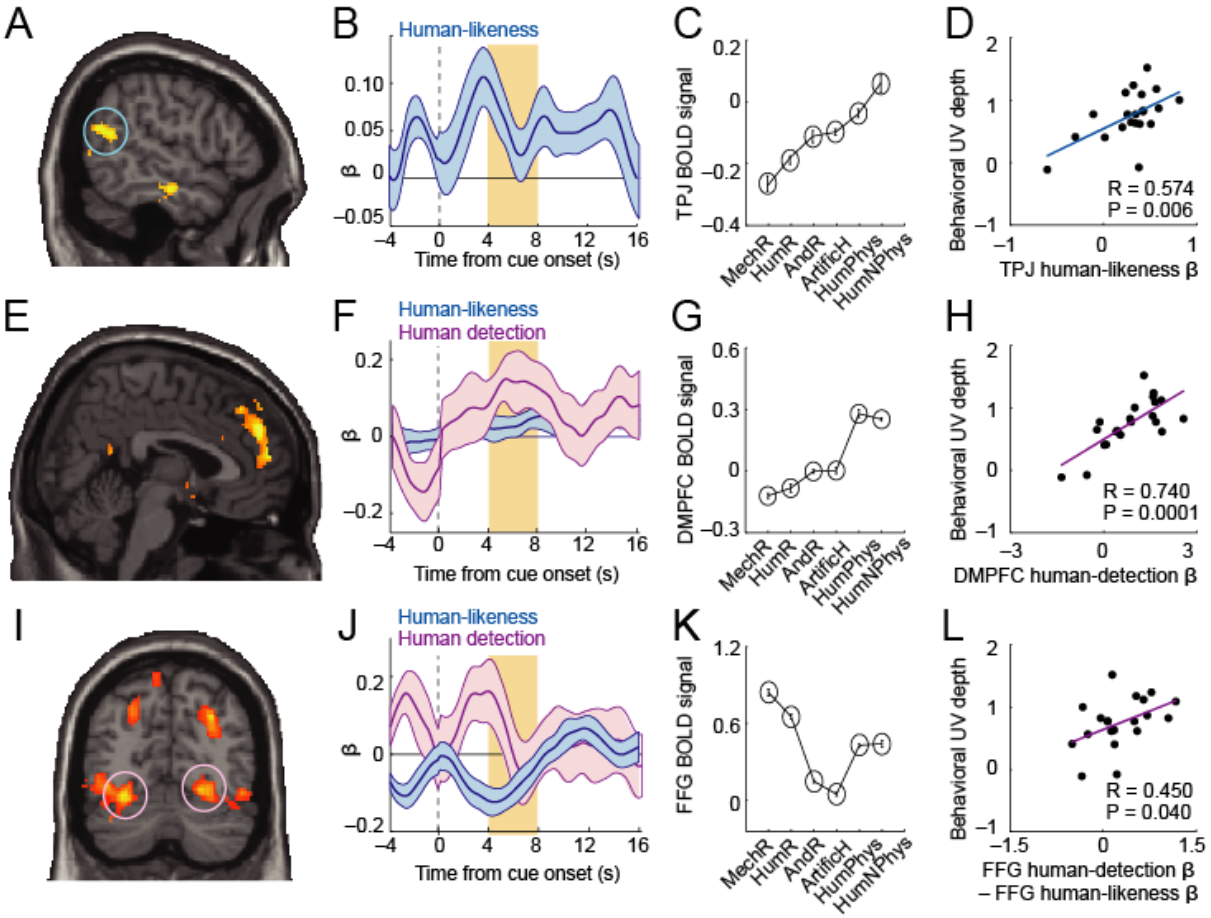

Fig. 3. Linear and nonlinear human-likeness coding. (A) Linear coding of subjective human-likeness in TPJ. Activity reflected trial-by-trial human-likeness ratings ($P < 0.05$, whole-brain corrected). (B) Region-of-interest regression of TPJ activity on human-likeness (with likability and familiarity covariates; $t(20) = 3.47$, $P = 0.002$, one-sample t-test in random effects analysis). (C) TPJ activity closely followed the human-likeness continuum. (D) TPJ human-likeness sensitivity reflected subjects' behavioral UV depths. Linear regression of behavioral UV-depth (cf. Fig. 1F) on TPJ human-likeness βs (significant robust regression). (E) Stronger DMPFC-activation for humans vs. non-humans shown by contrast analysis ([10 48 14], z-score = 4.78, $P = 0.001$, whole-brain corrected). (F) Regression of DMPFC activity on human detection ($t(20) = 2.26$, $P = 0.036$) and human-likeness ($P = 0.603$; non-significant likability and familiarity covariates). (G) Selective DMPFC response to humans. (H) Linear regression of behavioral UV depths on DMPFC human-detection βs (significant robust regression). (I) Negative human-likeness coding in FFG ($P < 0.05$, whole-brain corrected). (J) Regression of FFG activity on human-detection ($t(20) = -2.26$, $P = 0.034$) and human-likeness ($t(20) = -3.56$, $P = 0.0019$). Inset: significant β-difference for human-likeness and human-detection. (K) FFG activity across stimulus categories. (L) FFG nonlinear human-likeness sensitivity reflected behavioral UV depths (significant robust regression).

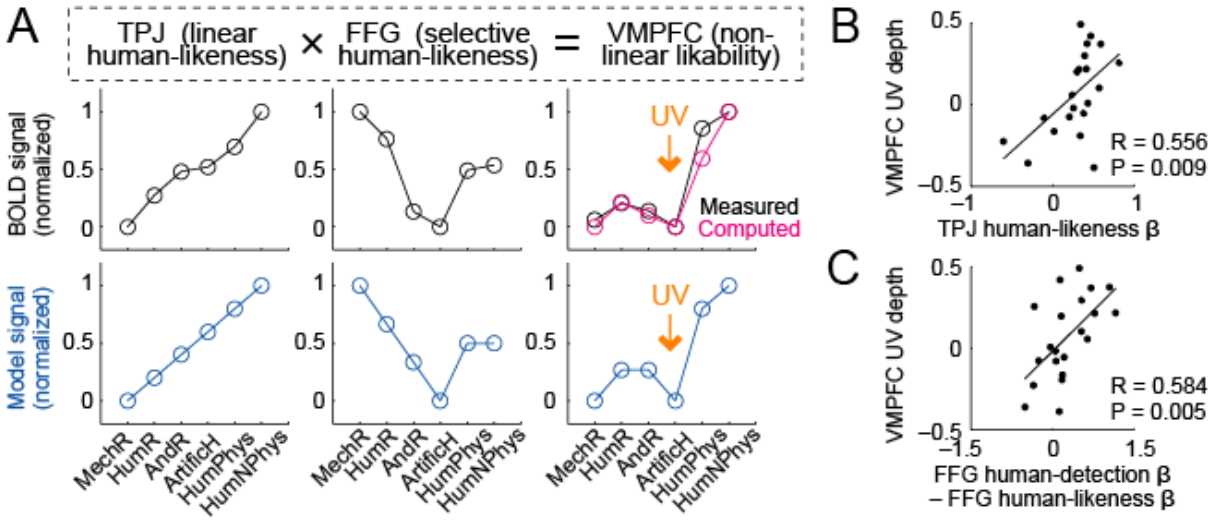

Fig. 4. Constructing the Uncanny Valley from linear and nonlinear human-likeness signals. (A) Combining human-likeness signals in TPJ and FFG to construct nonlinear likability in VMPFC. Upper panels: measured activity patterns (across subjects) in TPJ, FFG and VMPFC. Multiplicative combination of measured TPJ and FFG signals ('Computed') approximated the observed neural UV in VMPFC ('Measured'). Lower panels: Modelled signals. TPJ activity modelled by linear human-likeness function; FFG activity modelled as inverse, linear human-likeness function (negative linear relation across hypothesized human-likeness continuum), selectively for non-humans, with undifferentiated (average) response to human stimuli; VMPFC activity modelled as multiplicative combination of these signals. (B, C) Relation between neural human-likeness sensitivities and neural UV depth in VMPFC across subjects. (B) Linear regression of TPJ human-likeness on VMPFC UV depth (significant robust regression). (C) Linear regression of FFG human-likeness on VMPFC UV depth (significant robust regression).

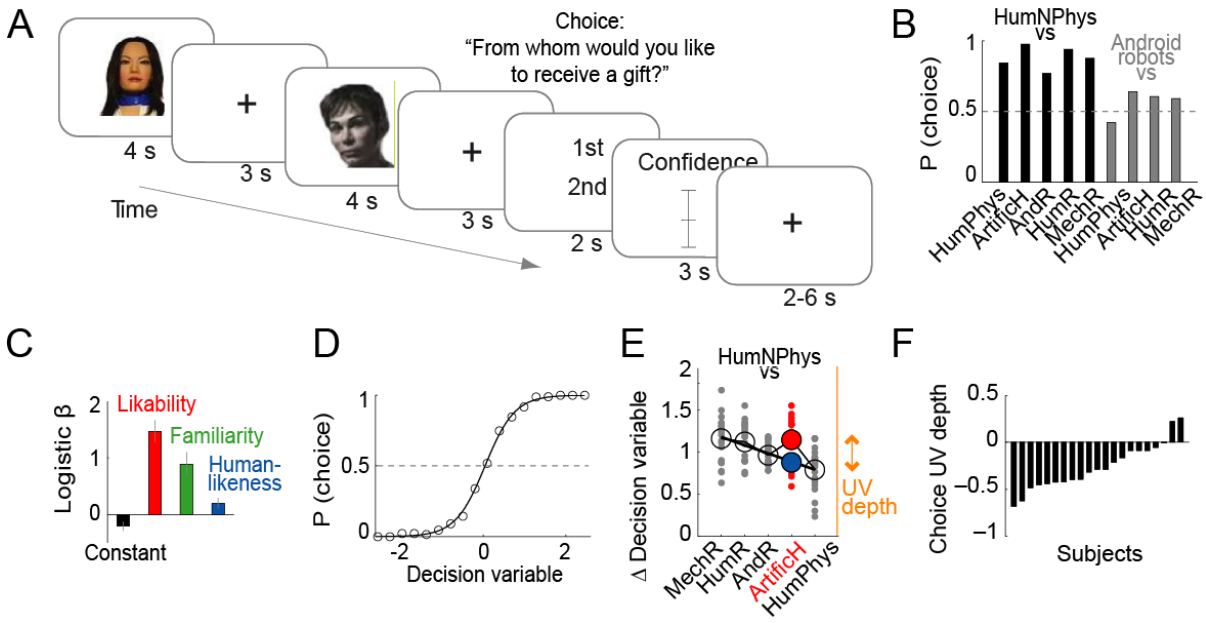

Fig. 5. Decision-making to accept or reject artificial agents. (A) Choice task. Subjects viewed sequential stimuli and decided from whom they would like to receive a gift (example from categories android robots (first screen) and artificial humans (second screen). (B) Choice behavior. Probability for gift-acceptance from humans (blacks) and androids (grey) over other stimulus categories. (C) Modelling choices with logistic regression. β weights of relative likability, familiarity and human-likeness (all $P < 0.002$, ones-sample t-tests). (D) Psychometric function: Relation between decision variable and choice probability. Relative differences in likability, familiarity and human-likeness were weighted for individual subjects, based on subject-specific logistic regression, and summed to form a decision variable. Choice probabilities calculated for equally populated decision-variable bins and fitted with logit function. (E) Defining UV depth for the choice task, based on decision variable. Unsigned (absolute) differences in decision variable between stimulus categories (Δ Decision Variable, derived from stimulus-specific ratings during rating task; data for all subjects). Black line: linear regression-fit to human-likeness continuum for all conditions (grey data) except choices involving artificial humans (red data). Blue point: predicted decision variable for artificial humans from linear fit; red point: observed decision variable. Choice UV depth defined as difference between predicted and observed decision variable. (F) Choice UV depth across subjects. The procedure in (F) defined UV depths for all subjects ($P = 1.1 \times 10^{-4}$, one-sample t-test).

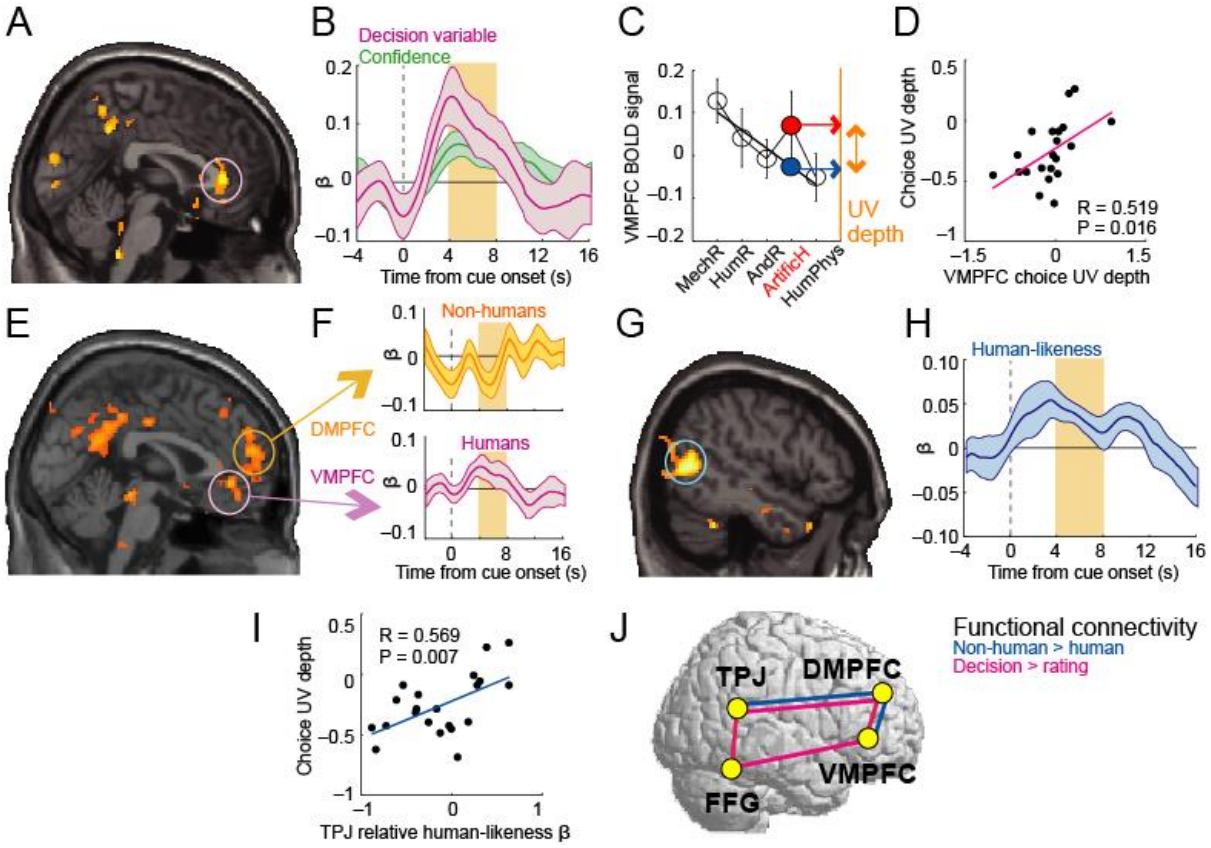

Fig. 6. Neural Uncanny-Valley components during decision-making. (A) VMPFC activity during gift-choices between humans and artificial agents coded subjects' decision variable ($P < 0.05$; small-volume correction). Decision variable used as parametric modulator was the sum of subjectively weighted relative likability, familiarity and human-likeness, which explained choices (cf. Fig. 5F, G). (B) Region-of-interest regression of VMPFC activity on decision variable ($t(20) = 2.25$, $P = 0.0356$; one-sample t-test, random-effects analysis) and rated confidence ($t(20) = 3.48$, $P = 0.002$). (C) VMPFC activity during choices reflects UV. VMPFC responses during choices involving artificial humans deviated from expected human-likeness continuum. Black line: linear regression-fit of activity on relative human-likeness for all stimuli except artificial humans ($R = -0.248$, $P = 0.023$, Pearson correlation). Blue point: predicted response for artificial humans from linear fit; red point: measured response. (D) Linear regression of behavioral choice UV depth (defined in Fig. 5E) on neural UV depth derived from VMPFC activity ($P = 0.016$; significant robust regression). (E) Stronger activation in VMPFC ([-6 40 -6], $z = 4.21$) and DMPFC ([0 56 18], $z = 3.81$) during choices involving humans contrasted with choices without humans (both $P < 0.001$, whole-brain corrected). (F) DMPFC coded decision variable specifically for choices involving non-humans ($t(20) = -2.45$, $P = 0.0294$, one-sample t-test); VMPFC coded decision variable specifically for choices involving humans ($t(20) = 2.139$, $P = 0.044$, one-sample t-test). (G) TPJ activity reflected the human-likeness component of the decision variable ($P < 0.05$, whole-brain corrected). (H) Region-of-interest regression of TPJ activity on relative human-likeness ($t(20) = 3.678$, $P = 0.0015$, one-sample t-test). (I) Behavioral UV depths matched TPJ human-likeness βs (robust regression). (J) Functional connectivity. Psychophysiological interactions identified pairs of brain regions with stronger activity-correlations during choices than ratings (magenta connections) or stronger activity-correlations for non-humans compared to humans (blue connections, rating task). Summary figure based on whole-brain-corrected statistical maps.

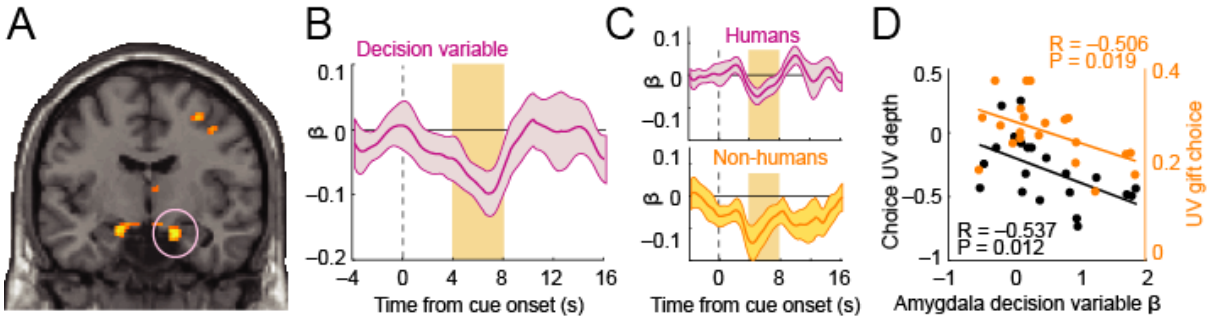

Fig. 7. Amygdala rejection-signals for humans and artificial agents. (A) Amygdala choice-activity coded subjects' decision variable ($P < 0.05$; small-volume correction on pre-defined amygdala coordinates). (B) Region-of-interest regression showed negative amygdala-coding of decision variable ($t(20) = -3.05$, $P = 0.006$, one-sample t-test). (C) Significant coding of decision variable for choices involving humans ($P = 0.0173$, one-sample t-test) and choices without humans ($P = 0.005$). (D) Amygdala sensitivity to decision variable predicted behavioral UV depths (black data) and gift rejections from artificial agents (orange data). Linear regressions of choice UV depths ($P = 0.012$, robust regression) and probability of gift acceptance from artificial humans and androids ($P = 0.019$, robust regression) on amygdala βs.

Table 1. Rating task analyses. Results are whole-brain corrected, P < 0.05, cluster level; maps thresholded at P < 0.005, extent threshold 10 voxels. PM: parametric modulator; C: contrast; PPI: psychophysiological interaction; sv: small-volume correction, based on pre-defined coordinates (see Methods). X, Y, Z: coordinates in MNI space.

Effect	Area	X	Y	Z	z-score	P
Human-likeness PM, positive	TPJ	56	-60	20	3.53	0.029
	DMPFC	4	40	42	3.14	0.018 sv
Human-likeness PM, negative	FFG	26	-66	-12	4.22	0.001
	Occipital gyri	-20	-102	2	5.54	0.001
	Middle frontal gyrus	-48	36	24	3.90	0.001
Likability PM, positive	VMPFC	12	48	8	4.00	0.001
	Striate area	30	-98	18	3.87	0.002
	Ventral striatum	12	12	-10	3.10	0.043 sv
Likability PM, negative	-					
Familiarity PM, positive	Ventral striatum	10	14	-4	3.40	0.025 sv
Familiarity PM, negative	-					

Table 2. Choice task analyses. Results are whole-brain corrected, P < 0.05, cluster level; maps thresholded at P < 0.005, extent threshold 10 voxels. PM: parametric modulator; C: contrast; PPI: psychophysiological interaction; sv: small-volume correction, based on pre-defined coordinates (see Methods). X, Y, Z: coordinates in MNI space.

Effect	Area	X	Y	Z	z-score	P
Decision variable PM, positive	VMPFC	0	34	6	3.60	0.019 sv
	TPJ	68	-34	44	3.30	0.030
	Caudate nucleus	-12	24	12	4.53	0.001
	Occipital gyri	20	-90	30	4.50	0.001
	Occipital gyri	18	-54	12	4.40	0.010
	Cingulate gyrus	6	-42	46	3.77	0.003
	Occipital gyri	-22	-60	24	3.72	0.042
Decision variable PM, negative	Amygdala	16	-8	-20	3.49	0.022 sv
	Intraparietal sulcus	-32	-56	46	4.48	0.001
	Superior frontal gyrus	-10	-4	72	4.43	0.001
	Middle frontal gyrus	-42	8	36	4.09	0.001
	Middle frontal gyrus	34	4	32	3.86	0.015
Relative human-likeness PM, positive	TPJ	54	-64	8	4.50	0.001
Relative human-likeness PM, negative	Occipital gyri	24	-98	8	5.99	0.001
	Occipital gyri	-24	-94	-2	4.74	0.001
Choice > rating, C	VMPFC	4	44	2	4.00	0.001
	TPJ	54	-64	26	5.51	0.001
	TPJ	-52	-58	28	5.71	0.001
	Amygdala	22	6	-26	4.32	0.001
	Ventral striatum	12	10	-6	3.43	0.025 sv
	Precuneus	2	-68	24	6.11	0.001
	Superior frontal gyrus	10	22	66	5.22	0.001
	Planum polare	34	14	-22	4.93	0.001
	Middle temporal gyrus	68	-34	-6	4.53	0.001
	Cingulate gyrus	0	-14	34	4.11	0.004
Rating > choice, C	Occipital gyri	34	-88	12	6.46	0.001
	FFG	-32	-80	-18	5.68	0.001
	Precentral gyrus	-26	-16	76	4.81	0.001
	Middle frontal gyrus	-42	22	26	4.66	0.002

Table 3. PPI analyses. Results are whole-brain corrected, P < 0.05, cluster level; maps thresholded at P < 0.005, extent threshold 10 voxels. PM: parametric modulator; C: contrast; PPI: psychophysiological interaction; sv: small-volume correction, based on pre-defined coordinates (see Methods). X, Y, Z: coordinates in MNI space.

Effect	Area	X	Y	Z	z-score	P
PPI non-human > human, DMPFC seed	VMPFC	6	34	10	3.82	0.006
	DMPFC	4	46	32	4.18	0.006
	TPJ	46	-70	34	3.97	0.019
	Posterior cingulate cortex	-8	-48	36	4.19	0.001
PPI non-human > human, FFG seed	TPJ	50	-36	26	3.64	0.020
PPI Choice > rating, VMPFC seed	FFG	14	-70	-10	5.02	0.001
	Cingulate gyrus	-4	18	40	4.22	0.001
	Caudate nucleus	24	-19	26	4.15	0.001
PPI Choice > rating, FFG seed	TPJ	50	-38	28	4.32	0.001
	FFG	-18	-64	-6	4.22	0.028
	Precuneus	0	-78	28	3.98	0.012
PPI Choice > rating, TPJ seed	DMPFC	0	46	38	3.52	0.001
	Occipital gyri	10	-102	18	4.95	0.021
	Superior frontal gyrus	-8	20	66	4.82	0.001
PPI Choice > rating, DMPFC seed	VMPFC	-2	44	10	4.11	0.001
	Cingulate gyrus	-14	54	30	4.39	0.005
	Occipital gyri	14	-100	6	4.12	0.002

References

Adolphs R (2010) What does the amygdala contribute to social cognition? Ann N Y Acad Sci 1191:42-61.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268-277.

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412-427.

Behrens TE, Hunt LT, Rushworth MF (2009) The computation of social behavior. Science 324:1160-1164.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. Nature 456:245-249.

Bhatt MA, Lohrenz T, Camerer CF, Montague PR (2012) Distinct contributions of the amygdala and parahippocampal gyrus to suspicion in a repeated bargaining game. Proc Natl Acad Sci USA 109:8728-8733.

Boorman ED, O'Doherty JP, Adolphs R, Rangel A (2013) The behavioral and neural mechanisms underlying the tracking of expertise. Neuron 80:1558-1571.

Broadbent E (2017) Interactions With Robots: The Truths We Reveal About Ourselves. Annu Rev Psychol 68:627-652.

Burleigh TJ, Schoenherr JR (2015) A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? Front Psychol 5.

Burleigh TJ, Schoenherr JR, Lacroix GL (2013) Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. Comput Hum Behav 29:759-771.

Carter RM, Huettel SA (2013) A nexus model of the temporal-parietal junction. Trends Cogn Sci 17:328-336.

Chaminade T, Hodgins J, Kawato M (2007) Anthropomorphism influences perception of computer-animated characters' actions. Soc Cogn Affect Neur 2:206-216.

Chaminade T, Zecca M, Blakemore SJ, Takanishi A, Frith CD, Micera S, Dario P, Rizzolatti G, Gallese V, Umilta MA (2010) Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. PLoS One 5:e11577.

Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat Neurosci 2:913-919.

Cheetham M, Suter P, Jancke L (2011) The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. Front Hum Neurosci 5:126.

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29:12315-12320.

Clithero JA, Rangel A (2014) Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neur 9:1289-1302.

de Araujo IET, Rolls ET, Kringelbach ML, McGlone F, Phillips N (2003) Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. Eur J Neurosci 18:2059-2068.

De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. Nat Neurosci 16:105-110.

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24:1742-1752.

Esterman M, Tamber-Rosenau BJ, Chiu YC, Yantis S (2010) Avoiding non-independence in fMRI data analysis: leave one subject out. Neuroimage 50:572-576.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls ET, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6:218-229.
Gitelman DR, Penny WD, Ashburner J, Friston KJ (2003) Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. NeuroImage 19:200-207.
Grabenhorst F, Rolls ET (2011) Value, pleasure and choice in the ventral prefrontal cortex. Trends Cogn Sci 15:56-67.
Grabenhorst F, Hernadi I, Schultz W (2012) Prediction of economic choice by primate amygdala neurons. Proc Natl Acad Sci U S A 109:18950-18955.
Grabenhorst F, Hernadi I, Schultz W (2016) Primate amygdala neurons evaluate the progress of self-defined economic choice sequences. Elife 5.
Grabenhorst F, Schulte FP, Maderwald S, Brand M (2013) Food labels promote healthy choices by a decision bias in the amygdala. Neuroimage 74:152-163.
Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A (2002) Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 47:1202-1210.
Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci USA 105:6741-6746.
Hare TA, Camerer CF, Knoepfle DT, Rangel A (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J Neurosci 30:583-590.
Heberlein AS, Adolphs R (2004) Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. Proc Natl Acad Sci USA 101:7487-7491.
Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG (2004) A general mechanism for perceptual decision-making in the human brain. Nature 431:859-862.
Hernadi I, Grabenhorst F, Schultz W (2015) Planning activity for internally generated reward goals in monkey amygdala neurons. Nat Neurosci 18:461-469.
Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MF, Behrens TE (2012) Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci 15:470-476, S471-473.
Jenkins AC, Macrae CN, Mitchell JP (2008) Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc Natl Acad Sci USA 105:4507-4512.
Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T (2008) Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. PLoS One 3.
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535-540.
Kutner MH, Nachtsheim CJ, Neter J, W. L (2004) Applied linear statistical models. New York: McGraw-Hill Irwin.
Levy I, Snell J, Nelson AJ, Rustichini A, Glimcher PW (2010) Neural representation of subjective value under risk and ambiguity. J Neurophysiol 103:1036-1047.
Lischetzke T, Izydorczyk D, Huller C, Appel M (2017) The topography of the uncanny valley and individuals' need for structure: A nonlinear mixed effects analysis. J Res Personality 68:96-113.
MacDorman KF, Ishiguro H (2006) The uncanny advantage of using androids in cognitive and social science research. Interact Stud 7:297-337.
MacDorman KF, Chattopadhyay D (2016) Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 146:190-205.
Mahon BZ, Milleville SC, Negri GAL, Rumiati RI, Caramazza A, Martin A (2007) Action-related properties shape object representations in the ventral stream. Neuron 55:507-520.
Mar RA, Kelley WM, Heatherton TF, Macrae CN (2007) Detecting agency from the biological motion of veridical vs animated agents. Soc Cogn Affect Neur 2:199-205.
Mitchell JP, Macrae CN, Banaji MR (2006) Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50:655-663.
Mori M (1970) The uncanny valley. Energy 7:33-35.
Mori M, MacDorman KF, Kageki N (2012) The uncanny valley. Ieee Robot Autom Mag 192:98-100.

Neubert FX, Mars RB, Sallet J, Rushworth MF (2015) Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. Proc Natl Acad Sci USA 112:E2695-2704.
Nicolle A, Klein-Flugge MC, Hunt LT, Vlaev I, Dolan RJ, Behrens TE (2012) An agent independent axis for executed and modeled choice in medial prefrontal cortex. Neuron 75:1114-1121.
Noppeney U, Price CJ, Penny WD, Friston KJ (2006) Two distinct neural mechanisms for category-selective responses. Cerebral Cortex 16:437-445.
Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14:481-487.
Pena JL, Konishi M (2001) Auditory spatial receptive fields created by multiplication. Science 292:249-252.
Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175-187.
Piwek L, McKay LS, Pollick FE (2014) Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. Cognition 130:271-277.
Pollick FE, ed (2010) In search of the uncanny valley. Berlin Heidelberg: Springer.
Poser BA, Norris DG (2009a) 3D Single-Shot VASO Using a Maxwell Gradient Compensated GRASE Sequence. Magnetic Resonance in Medicine 62:255-262.
Poser BA, Norris DG (2009b) Investigating the benefits of multi-echo EPI for fMRI at 7 T. Neuroimage 45:1162-1172.
Rolls ET (2014) Emotion and Decision-Making Explained. Oxford: Oxford University Press.
Rolls ET, Treves A (1998) Neural Networks and Brain Function. Oxford: Oxford University Press.
Rolls ET, Grabenhorst F, Deco G (2010a) Choice, difficulty, and confidence in the brain. NeuroImage 53:694-706.
Rolls ET, Grabenhorst F, Deco G (2010b) Decision-making, errors, and confidence in the brain. J Neurophysiol 104:2359-2374.
Rosenthal-von der Putten A, Weiss A (2015) The uncanny valley phenomenon Does it affect all of us? Interact Stud 16:206-214.
Rosenthal-von der Putten AM, Kramer NC (2014) How design characteristics of robots determine evaluation and uncanny valley related responses. Comput Hum Behav 36:422-439.
Saxe R, Wexler A (2005) Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia 43:1391-1399.
Schiller D, Freeman JB, Mitchell JP, Uleman JS, Phelps EA (2009) A neural mechanism of first impressions. Nat Neurosci 12:508-514.
Schultz W (2015) Neuronal Reward and Decision Signals: From Theories to Data. Physiol Rev 95:853-951.
Seymour B, Dolan R (2008) Emotion, decision making, and the amygdala. Neuron 58:662-671.
Small DM, Voss J, Mak YE, Simmons KB, Parrish T, Gitelman D (2004) Experience-dependent neural integration of taste and smell in the human brain. J Neurophysiol 92:1892-1903.
Stanley DA (2016) Getting to know you: general and specific neural computations for learning about people. Soc Cogn Affect Neur 11:525-536.
Stein BE, Stanford TR (2008) Multisensory integration: current issues from the perspective of the single neuron. Nat Rev Neurosci 9:255-266.
Sul S, Tobler PN, Hein G, Leiberg S, Jung D, Fehr E, Kim H (2015) Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proc Natl Acad Sci USA 112:7851-7856.
Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H (2012) Learning to simulate others' decisions. Neuron 74:1125-1137.
Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642-1645.
Toledano P (2011) A New Kind of Beauty: Dewi Lewis Publishing.
Vogeley K (2017) Two social brains: neural mechanisms of intersubjectivity. Phil Transactions Royal Soc London Series B, Biological sciences 372.
Vogeley K, Bente G (2010) "Artificial humans": Psychology and neuroscience perspectives on embodiment and nonverbal communication. Neural Netw 23:1077-1090.

Wang SS, Lilienfeld SO, Rochat P (2015) The Uncanny Valley: Existence and Explanations. Rev Gen Psychol 19:393-407.
Winston JS, Strange BA, O'Doherty J, Dolan RJ (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nat Neurosci 5:277-283.
Wittmann MK, Lockwood PL, Rushworth MFS (2018) Neural Mechanisms of Social Cognition in Primates. Annu Rev Neurosci.
Wittmann MK, Kolling N, Faber NS, Scholl J, Nelissen N, Rushworth MF (2016) Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. Neuron 91:482-493.
Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. Proc Natl Acad Sci U S A 107:15005-15010.
Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8:665-670.
Zangemeister L, Grabenhorst F, Schultz W (2016) Neural Basis for Economic Saving Strategies in Human Amygdala-Prefrontal Reward Circuits. Curr Biol 26:3004-3013.