Journal of Experimental Psychology:
Human Perception and Performance
2017, Vol. 43, No. 4, 651–666

0096-1523/17/$12.00 http://dx.doi.org/10.1037/xhp0000304

Is That a Human? Categorization (Dis)Fluency Drives Evaluations of Agents Ambiguous on Human-Likeness

Evan W. Carr, Galit Hofree, Kayla Sheldon, and Ayse P. Saygin
University of California San Diego

Piotr Winkielman
University of California San Diego, University of Warwick, and SWPS University of Social Sciences and Humanities

A fundamental and seemingly unbridgeable psychological boundary divides humans and nonhumans. Essentialism theories suggest that mixing these categories violates "natural kinds." Perceptual theories propose that such mixing creates incompatible cues. Most theories suggest that mixed agents, with both human and nonhuman features, obligatorily elicit discomfort. In contrast, we demonstrate top-down, cognitive control of these effects—such that the discomfort with mixed agents is partially driven by disfluent categorization of ambiguous features that are pertinent to the agent. Three experiments tested this idea. Participants classified 3 different agents (humans, androids, and robots) either on the human-likeness or control dimension and then evaluated them. Classifying on the human-likeness dimensions made the mixed agent (android) more disfluent, and in turn, more disliked. Disfluency also mediated the negative affective reaction. Critically, devaluation only resulted from disfluency on human-likeness—and not from an equally disfluent color dimension. We argue that negative consequences on evaluations of mixed agents arise from *integral disfluency* (on features that are relevant to the judgment at-hand, like ambiguous human-likeness). In contrast, no negative effects stem from *incidental disfluency* (on features that do not bear on the current judgment, like ambiguous color backgrounds). Overall, these findings support a top-down account of why, when, and how mixed agents elicit conflict and discomfort.

Public Significance Statement
People have always been fascinated with hybrid, mixed creatures. In antiquity, these were chimeras and griffins, but in modern times, these are androids (or robots with human-like features). However, such creatures (including androids) are often disliked. Dominant explanations claim that this occurs because they mix seemingly unbridgeable core "essences" or create early perceptual conflicts. In contrast, our experiments show that the dislike of androids can come from the aversive mental effort of categorizing them as human versus non-human. Critically, this effort (and the resultant dislike) is not inevitable, but instead depends on the perceiver's flexible categorization mindset. This suggests that higher-order cognitive processes can override seemingly "automatic" reactions to incompatible cues.

Keywords: emotion, categorization, judgment and decision making, cognitive processing, human-computer interaction

A key psychological distinction is the one that divides human and nonhuman. Treating an agent as human (or "human-like") fundamentally changes how individuals perceive, interpret, behave, communicate, or empathize (Dennett, 1971). Children and adults consider human-likeness to be a deep, unchangeable trait—a type of psychological essentialism (Medin & Ortony, 1989; Prentice & Miller, 2007). Such essentialist beliefs arise early in life (Gelman, 2004), structure social categories and stereotypes (Bas-

This article was published Online First January 26, 2017.

Evan W. Carr and Galit Hofree, Psychology Department and Cognitive Science Department, University of California San Diego; Kayla Sheldon, Psychology Department, University of California San Diego; Ayse P. Saygin, Cognitive Science Department, University of California San Diego; Piotr Winkielman, Psychology Department, University of California San Diego, Behavioural Science, Warwick Business School, University of Warwick, and SWPS University of Social Sciences and Humanities.

Evan W. Carr is now at Columbia Business School (Management Division), Columbia University. Galit Hofree is now employed with MATLAB (MathWorks).

Evan W. Carr conducted this research with government support under and awarded by the United States Department of Defense (DoD) and Army Research Office (ARO), via the National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. Piotr Winkielman was sponsored through the UCSD Academic Senate Grant and NSF BCS-1232676. Ayse P. Saygin was supported through NSF CAREER BCS-1151805, Kavli Institute for Brain and Mind, and the Qualcomm Institute (formerly Calit2). We thank H. Ishiguro and the Intelligent Robotics Laboratory at Osaka University for help in stimulus preparation. We also appreciate valuable feedback from Dave Barner and Christopher Oveis. Finally, we are grateful to the many research assistants that helped to run the current experiments, including (but not limited to) Kevin Wright, Lorelei Himlin, and Jason Herbert.

Correspondence concerning this article should be addressed to Evan W. Carr, Columbia Business School, 3022 Broadway, 7L Uris, New York, NY 10027. E-mail: ewcarr@ucsd.edu

tian & Haslam, 2006; Haslam, Bastian, Bain, & Kashima, 2006; Haslam, Rothschild, & Ernst, 2000; Howell, Weikum, & Dyck, 2011), drive attention (Bastian & Haslam, 2007), and guide automatic motor responses (Bastian, Loughnan, & Koval, 2011).

This human/nonhuman boundary is often investigated with agents that mix human and nonhuman features. It has long been noticed that mixed stimuli (e.g., chimeras, griffins, hybrids, mannequins, human dolls, etc.) generally elicit a sense of weirdness and discomfort (Frenkel-Brunswik, 1949; Jentsch, 1906). This issue gained renewed importance with the recent proliferation of bionic humans and androids (i.e., robots with human-like features and behaviors; Ishiguro, 2007; Mori, MacDorman, & Kageki, 2012). Here, the key psychological question is what causes such discomfort to mixed agents.

Extant theories propose a variety of processes. From the aforementioned essentialism perspective, individuals perceive the human/nonhuman boundary as fundamentally unbridgeable, and negative reactions spontaneously arise from the inappropriate blending of two different "natural kinds" or separate "essences" (see Demoulin, Leyens, & Yzerbyt, 2006). Proponents of essentialism argue that certain properties are intrinsic and immutable traits of the agent in-question, and perceivers can essentialize a variety of dimensions (e.g., gender, race, sexual orientation, etc.). Critically though, essentialized categories are usually said to have some key defining characteristics: clear and discrete boundaries from other categories, involuntary and unchanging membership, and observable features that reflect something about the underlying function of the agent (Prentice & Miller, 2007). While theoretically distinct (but still related), perception research tends to focus on "mismatches," or conflicting cues in visual, auditory, and motion processing of mixed agents (Katsyri, Forger, Makarainen, & Takala, 2015; MacDorman, Green, Ho, & Koch, 2009; Mitchell et al., 2011; Seyama & Nagayama, 2007). Other proposals highlight the potential role of conflict between the apparent lack of conscious experience paired with perceived agency (Waytz, Gray, Epley, & Wegner, 2010). This idea is related to suggestions that negative reactions to mixed agents reflect low-level mechanisms involving disease avoidance (MacDorman & Ishiguro, 2006). Overall, most (if not all) theories suggest that mixed agents (with both human and nonhuman features) spontaneously elicit conflict and discomfort.

In contrast to these assumptions, we propose that the relative dislike for mixed agents can be modified by contextual factors—providing a major theoretical qualification to these earlier claims. Specifically, the current article argues that reactions to mixed agents involve an interaction between top-down higher-order cognitive processes and bottom-up perceptual factors. Basically, we suggest that the sense of "weirdness" is not inherent to the perception of mixed agents, but rather is generated when people classify such agents into human versus nonhuman categories—resulting in the experience of categorization disfluency. This disfluency triggers negative affect, which generalizes to agent evaluations (as we explain next). Note that our top-down framework is not simply another level of analysis for essentialist or perceptual conflict theories. While there are some versions of bottom-up perceptual theories that could be considered compatible with our fluency account, these frameworks still differ on the type of role disfluency serves in generating negative affect (i.e., either as a key component or a mere byproduct). We will return to these theoretical distinctions in the Discussion. Generally, we posit that the subjective boundary between human and nonhuman entities can be markedly reconstructed via categorization processes, which has downstream consequences on judgments, evaluations, and attitudes.

Our proposal builds on several lines of previous research related to *fluency*—or changes in processing speed and effort (Schwarz, 1998). Here, much of the original research has focused on perceptual fluency (manipulated by enhancing low-level "surface" features, like clarity, contrast, readability, typicality, etc.; e.g., Carr, Rotteveel, & Winkielman, 2016; Reber & Schwarz, 1999; Reber, Winkielman, & Schwarz, 1998) or conceptual fluency (manipulated by facilitating the processing of stimulus meaning, as with semantic priming; e.g., Rajaram & Geraci, 2000; Whittlesea, 1993). Evidence shows that processing ease (fluency) increases evaluations, whereas *dis*fluency lowers them, as reflected in self-reports and physiological measures (Winkielman & Cacioppo, 2001). According to the *hedonic fluency model*, easy processing elicits positive affect, which is then (mis)attributed to the target stimulus (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). This positive affect presumably emerges because fluency reflects (or probabilistically signals) lower conflict and greater coherence in stimulus processing.

Importantly, such effects also extend to categorization fluency—or the effort needed to determine category membership (Halberstadt & Winkielman, 2013). Note that categorization fluency differs from perceptual fluency in that stimulus features remain unchanged. Rather, categorization fluency is ultimately task-dependent, and processing difficulty instead depends on which (un)ambiguous feature dimensions are highlighted by the current task. In other words, if a stimulus is ambiguous on some dimension, it will elicit *dis*fluency (and negative affect), but only in contexts requiring categorization on that particular dimension. To illustrate, Owen, Halberstadt, Carr, and Winkielman (2016) had participants categorize morphed male–female faces either on the central ambiguous dimension (gender) or an auxiliary unambiguous dimension (race). Devaluation for mixed-gender faces only occurred in the gender-categorization condition, when the gender-ambiguous faces were made disfluent (i.e., difficult to categorize). Similar effects have also been shown for biracial faces (Halberstadt & Winkielman, 2014) and those displaying mixed emotions (Winkielman, Olszanowski, & Gola, 2015).

This theoretical approach raises an important (and previously unexplored) possibility that changing someone's categorization mindset can alter negative responses to mixed agents, which contain supposedly "unbridgeable" human and nonhuman features. If so, their categorization on the human-likeness dimension should elicit disfluency (and devaluation), but this should be reduced during classification on an alternative feature dimension that is still social, yet unambiguous (e.g., gaze orientation cues). Further, as we explain next, to produce devaluation, perceivers must not only experience disfluency when processing the agent, but this disfluency must be derived from an integral (essential) rather than incidental (nonessential) feature of the agent.

Another contextual factor that dictates fluency-devaluation effects is the underlying relevance of the feature in-question. Some features can be *integral* (or pertinent to the judgment at-hand) while others can be *incidental* (or peripheral to the current task), as discussed by Bodenhausen (1993). Crucially though, evaluative consequences of (dis)fluency depend on the perceived relevance of the experience for the judgment at-hand (see Schwarz, 2010, for a review). For example, disfluency in categorizing someone else's emotional expression may lead the participant to judge the target as less trustworthy (i.e., since emotion is a key factor in trust

judgments; Winkielman, Olszanowski, & Gola, 2015). However, one can speculate that an equally difficult categorization experience on a secondary dimension (e.g., ambiguous hair color) will not lower trustworthiness judgments, given that hair color does not bear on trust. Consequently, for the current experiments, devaluation effects may only follow from disfluency that occurs in response to an agent's integral features, rather than an equally disfluent experience on incidental features. We expected integral disfluency (i.e., ambiguous human-likeness) to have downstream negative consequences on evaluations of mixed agents, but no effects to occur from incidental disfluency (i.e., ambiguous color backgrounds). We will return to this issue in the Discussion.

In short, we propose that affective responses to mixed agents are partially driven by top-down mechanisms. Categorization difficulty should trigger negative reactions to agents with ambiguous features when the current categorization task focuses on the ambiguous dimension (and only when the dimension is integral to the nature of the agent).

Current Research

We investigated our predictions in three experiments. In each study, participants saw and rated three different agents—one that was clearly human (a human), one that was clearly not human (a robot), and one that had both human and nonhuman features (an android). Participants rated these agents under two different conditions that varied on categorization requirements. Some participants categorized the agents as "human or nonhuman"—a selectively difficult task for the android agent—while others performed a control task on which the mixed agent was not selectively difficult.

To preview the results, categorization fluency impacted evaluative ratings of the different agents. Androids were devalued more so in the human-classification condition—both compared with a control task of speeded stimulus detection (Experiment 1) and an alternative task of social gaze categorization (Experiment 2). Using data from Experiments 1 and 2, multilevel mediation analyses demonstrated that categorization effort mediated the relationship between agent "mixed-ness" and weirdness judgments, but only for the human-classification condition.

Critically, the results show that these effects cannot be explained by simple misattribution of incidental task effort. This is because devaluation effects for mixed agents were eliminated in Experiment 3, which elicited disfluency by having subjects view androids with ambiguous color. These results suggest that disfluency translates into devaluation only when generated by ambiguous dimensions that are integral (i.e., human-likeness classification) rather than those that are incidental (i.e., color-classification) to the evaluative judgment.

As such, these results challenge the assumption that human and nonhuman categories are "unbridgeable"—where mixing produces obligatory conflict and devaluation. Instead, they argue for a more cognitively flexible, task-sensitive link between ambiguity, task-relevant fluency, and evaluation.

Experiment 1

Experiment 1 tested whether categorization disfluency impacted evaluative ratings of ambiguous nonhuman agents, using highly controlled images that placed androids toward the middle of the human-likeness continuum. We focused on weirdness ratings, given that

Figure 1. Example stimuli from Experiments 1 and 2 (grayscale images; top row) and Experiment 3 (blue/green images; bottom row). Each image depicted one of three agent types (i.e., human [left column], android [middle column], or robot [right column]), doing one of eight actions. Some actions showed the individual with their head oriented toward the camera (e.g., "talking" on the top row) while others showed the individual with their head oriented away from the camera (e.g., "turning" on the bottom row). See the online article for the color version of this figure.

dimensions of "eeriness" and "strangeness" are deemed important in many studies on nonhuman agents (Ho & MacDorman, 2010).

Method

Participants and stimuli. Fifty-two undergraduates (M_{age} = 21.00 years, SD_{age} = 2.44 years; 42 females) at the University of California, San Diego (UCSD) participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program.

Our stimuli were still images taken from the Saygin-Ishiguro Action Database (SIAD), which includes actions performed by human, android, and robot agents (Saygin & Stadler, 2012). For the android agent, these stimuli featured Repliee Q2 (see Figure 1, middle images), which was developed at Osaka University in collaboration with Kokoro Inc. More important, with brief exposures, people can mistake Repliee Q2 for a human being (Ishiguro, 2006). Repliee Q2 is an advanced humanoid robot that has 42 degrees of freedom and can make head and upper body movements. Because Repliee Q2 can make head and body movements, the images displayed both the head and upper body of the agents. For the human condition, the stimuli featured the female adult on whom Repliee Q2's appearance was modeled after (see Figure 1, left images). For the robot condition, the surface elements of the Repliee Q2 android were removed and stripped of human-like features, to reveal the underlying materials (e.g., wiring, metal limbs, and joints; see Figure 1, right images). For the purpose of the current experiments, to further match and standardize these images on perceptual cues (e.g., coloring, clothing, etc.), we edited them using Adobe PhotoShop CS2 to maximally control for each agent's general physical appearance (see Figure 1).

To create the stimuli, Repliee Q2 was photographed performing eight different actions (nudging, grasping, drinking, waving, talking, turning, wiping, and lifting), both with and without its original human-like surface features (i.e., android and robot agent conditions, respectively). Critically, the female model for Repliee Q2 then naturally performed the same actions several times, and the version of those actions that most closely matched that of Repliee Q2 were selected for the stimuli (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012; Saygin & Stadler, 2012). In short, the stimuli were highly controlled images of human, android, and robot actions, which mainly varied on the dimension of the agents' human-likeness.

The stimulus images were grayscale and cropped to 400 × 400 pixels. We used these grayscale images in Experiments 1 and 2, but note that the coloring of the images was edited for theoretical reasons in Experiment 3. Figure 1 shows examples of the different stimuli. All agents were photographed in the same room, with the same background, lighting, and camera settings.

Design and procedure. Participants were randomly assigned to one of two classification conditions (human-classification or no-classification). In the human-classification condition, participants were instructed to judge whether or not individuals in the different pictures were "human or nonhuman." In the no-classification condition, participants were instead told to "hit the spacebar as fast as possible, once the picture appears on the screen" (see Figure 2).

Next, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials), each of which asked for a

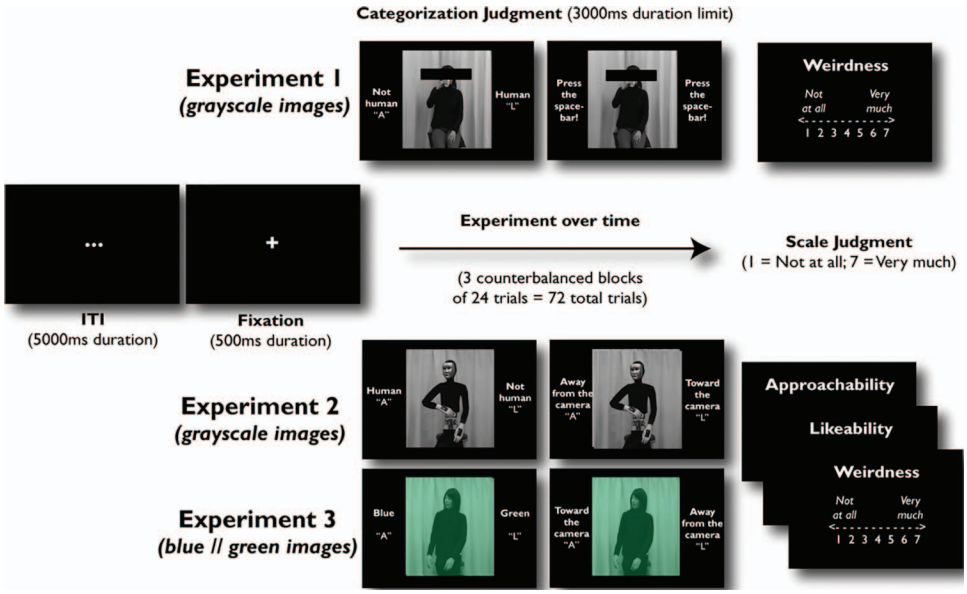

Figure 2. Design and procedure for Experiments 1, 2, and 3. Experiment 1 used grayscale images of all agent actions (first row), with a human-classification condition and no-classification condition (human/"toward" action example shown). Experiment 2 used grayscale images of all agent actions (second row) and used both a human-classification condition and orientation-classification condition (robot/"away" action example shown). Experiment 3 (bottom row) used 100% blue, 100% green, and 50/50 blue-green images of agent actions, with both a color-classification condition and orientation-classification condition (android/"away" action example shown). Note that Experiment 1 only involved weirdness ratings (all 72 trials), while Experiments 2 and 3 measured approachability, likability, and weirdness (24 trials each; 72 trials total). See the online article for the color version of this figure.

weirdness rating of the agent in the video. After each trial, participants gave the rating using a 1 (*not at all*) to 7 (*very much*) scale. Before rating each image, human-classification participants were told to categorize, "as quickly and accurately as possible, whether the agent in the picture was human or nonhuman," using the 'A' and 'L' keys on the keyboard (response labels were randomized across trials). No-classification participants only used the spacebar to indicate the onset of the image. Each trial began with a 500 ms fixation, followed by the 3,000 ms stimulus image (as soon as the participant responded, the image was replaced with the rating scale). The intertrial interval (ITI) was 5,000 ms (see Figure 2). Therefore, on each trial, we logged the participants' reaction time (RT) to classify the image and their weirdness ratings.

Results

Analysis strategy. All RTs and ratings were analyzed using trial-level data with multilevel models (MLMs; via restricted maximum likelihood estimation). MLMs more effectively handle hierarchical and unbalanced data with missing observations, relying on fewer assumptions regarding covariance structures and increasing parsimony and flexibility between models (Bagiella, Sloan, & Heitjan, 2000). Note that while we report MLM results here, because of the advantages over traditional analysis of variance (ANOVA) methods, all reported effects still replicate when using these traditional approaches.

MLMs were built with the *lme4* (Bates, Mächler, Bolker, & Walker, 2014) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) packages in R, using the maximal random-effects structure appropriate for the data (Barr, Levy, Scheepers, & Tily, 2013). Such a strategy strikes a balance in reducing possible Type 1 errors, while also avoiding overparameterization in the MLM (Bates, Kliegl, Vasishth, & Baayen, 2015). To obtain p value estimates for fixed-effects, we used Type 3 Satterthwaite approximations, which can sometimes result in decimal df, based on the number of observations (West, Welch, & Galecki, 2014).

Across all experiments, subjects who performed at ≤50% accuracy during the main task were removed from all analyses, and for the remaining subjects, error trials were excluded. For Experiment 1, one subject performed at ≤50% accuracy and another subject did not adhere to task instructions—therefore, these two subjects were excluded from the total sample, leaving a final $n = 50$.

We focused on nonerror trials to ensure that participants recognized the true human/nonhuman nature of the agent (i.e., whether it was actually human vs. nonhuman). We were also primarily interested to what extent participants' evaluations reflect the sheer effort of processing, as opposed to possible categorization errors. However, all analyses for error trials can be found in Footnotes 1 (Experiment 1),[1] 2 (Experiment 2),[2] and 3 (Experiment 3).[3]

Response RTs. Following previous fluency studies (Winkielman, Halberstadt, Fazendeiro, & Catty, 2006), we excluded trials with extremely fast (less than 200 ms) and slow (greater than 3,000 ms) RTs, and the remaining RTs were $\log_{10}$-transformed to normalize the response distribution. Next, we created an MLM with a Condition (2: human-classification, no-classification) × Agent (3: human, android, robot) fixed-effects structure.[4]

We observed strong evidence for all RT effects in Experiment 1. Critically, we detected the predicted Condition × Agent interaction, $F(2, 99.61) = 3.55$, $p = .03$. Follow-up tests demonstrated that only the human-classification subject group took longer to respond to the android, both compared with the human agent, $b = .05$, $SE = .01$, $t = 4.42$, $p < .001$, $d_Z = .92$, and robot agent, $b = .05$, $SE = .01$, $t = 4.00$, $p < .001$, $d_Z = .83$. The no-classification group showed no differences between android and human RTs, $b = .01$, $SE = .01$, $t = .82$, *ns*, $d_Z = .16$, and a smaller difference between android and robot RTs, $b = .02$, $SE = .01$, $t = 2.10$, $p = .04$, $d_Z = .40$. Neither condition differed in response RTs between the human and robot agents (see Figure 3).

Note that we also observed both main effects of Condition, $F(1, 49.92) = 57.56$, $p < .001$, and Agent, $F(2, 99.61) = 11.18$, $p < .001$. Human-classification subjects had longer RTs, and the android took the most time to categorize across conditions.

Overall, the android was selectively disfluent (compared with both of the other agents), and this occurred only within the human-classification condition.

[1] We could only analyze error trials in the human-classification condition for Experiment 1 (because there were no incorrect responses in the no-classification condition). To do this, we created a new MLM for only the human-classification condition, with Agent (3: human, android, robot) as the only fixed-effect factor. As expected, a main effect of Agent, $F(2, 38.90) = 6.53$, $p = .004$, showed that within the human-classification condition, participants had lower accuracy in response to mixed (android) agents—both compared with the human agent, $b = -.29$, $SE = .08$, $t = -3.61$, $p = .002$, $dz = .75$, and the robot agent, $b = -.28$, $SE = .08$, $t = -3.50$, $p = .002$, $dz = .73$.

[2] On Experiment 2, we analyzed accuracy using the same methods as reaction times (RTs; because both classification conditions could have correct and incorrect responses). As expected, human-classification participants showed a greater number of errors selectively in response to the mixed agent (android). A Condition × Agent interaction, $F(2, 167.80) = 43.26$, $p < .001$, demonstrated that human-classification participants had lower accuracy for androids—both compared with the human agent, $b = -.41$, $SE = .03$, $t = -12.54$, $p < .001$, $dz = 1.36$, and the robot agent, $b = -.41$, $SE = .03$, $t = -12.43$, $p < .001$, $dz = 1.35$. Furthermore, there was no difference between human and robot accuracy within the human-classification condition, $b = .01$, $SE = .01$, $t = .56$, *ns*, $dz = .06$. Importantly, for the orientation-classification condition, accuracy for the android did not significantly differ from the human agent, $b = -.06$, $SE = .03$, $t = -1.76$, *ns*, $dz = .19$, or the robot agent, $b = .01$, $SE = .03$, $t = .34$, *ns*, $dz = .04$. Both main effects were also significant. The main effect of Condition, $F(1, 167.14) = 28.97$, $p < .001$, demonstrated that orientation-classification participants were more accurate overall than human-classification participants. The main effect of Agent, $F(2, 167.80) = 54.33$, $p < .001$, demonstrated that participants were overall less accurate in response to the android compared with the other agents.

[3] We could not do the same analysis as Experiment 2 for accuracy in Experiment 3, because the android images in the color-classification condition were exactly 50/50 between blue and green (and thus, no response on those trials could be counted as correct or incorrect). However, we did analyze accuracy performance for all agent types in the orientation-classification condition (using similar methods as Experiment 1, by creating a new MLM only for the orientation-classification condition, with Agent [3: human, android, robot] as the only fixed-effect factor). We also checked overall accuracy for the human and robot images in the color-classification condition. First, on the orientation-classification condition, overall accuracy across agent types was high ($M = 90.87\%$, $SD = 28.80\%$). We did observe a main effect of Agent from the MLM, $F(2, 111.90) = 12.14$, $p < .001$, which showed that orientation-classification participants were more accurate in responding to the human agent—both compared with the android agent, $b = .04$, $SE = .01$, $t = 4.16$, $p < .001$, $dz = .52$, and the robot agent, $b = .05$, $SE = .01$, $t = 4.29$, $p < .001$, $dz = .54$. There were no accuracy differences between the android and robot agents, $b = .01$, $SE = .01$, $t = 0.69$, *ns*, $dz = .09$. Second, within the color-classification condition, accuracy performance was comparably high for both the human agent ($M = 95.71\%$, $SD = 20.27\%$) and robot agent ($M = 96.17\%$, $SD = 19.20\%$). Once again, note that we could not code accuracy for the android agent in the color-classification condition, because they were exactly 50% blue-green.

[4] Note that all of the reported effects for Experiment 1 and 2 reaction times (RTs) still hold, both with and without error trials on the android agent (where participants classified the android as "human").

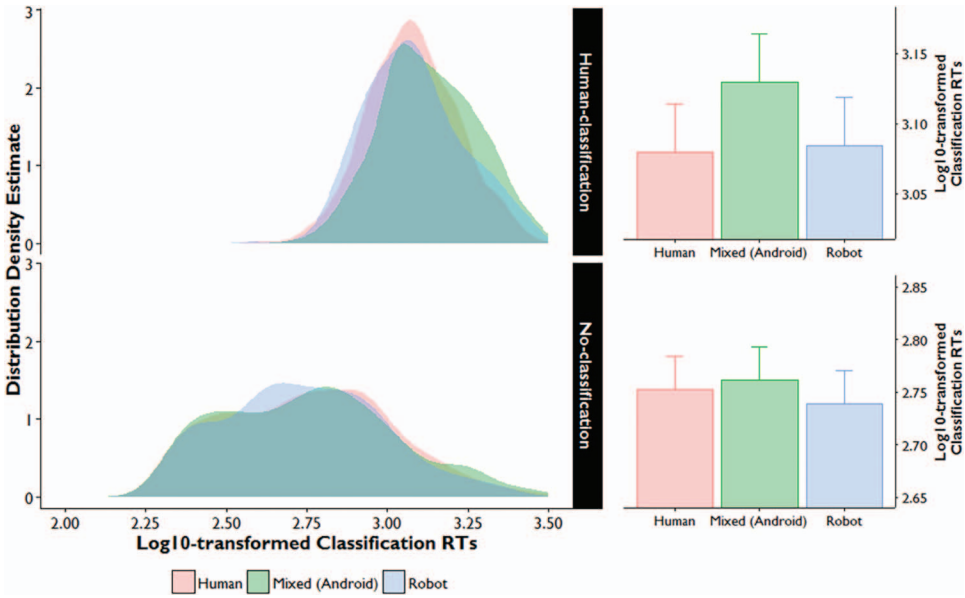

Figure 3. Density distributions and means/*SEM*s for $\log_{10}$-transformed reaction times (RTs) by classification condition (top row = human-classification; bottom row = no-classification) and agent type (indicated by colors) for Experiment 1. See the online article for the color version of this figure.

Weirdness ratings. We analyzed weirdness ratings similar to the RTs, using an MLM with a Condition (2: human-classification, no-classification) × Agent (3: human, android, robot) fixed-effects structure.

Crucially, we found evidence for a Condition × Agent interaction, $F(2, 99.06) = 5.42$, $p = .006$. Post hoc tests revealed that human-classification participants rated the android higher on weirdness (compared with the no-classification group; see Figure 4), $b = .79$, $SE = .32$, $t = 2.49$, $p = .01$, $d = .53$. Within the no-classification condition, participants rated the android as weirder than the human, $b = 2.36$, $t = 8.99$, $p < .001$, $d_Z = 1.73$, and the robot as weirder than the android, $b = 1.30$, $SE = .26$, $t = 4.97$, $p < .001$, $d_Z = .96$. Within the human-classification condition, participants still rated the android as weirder than the human, $b = 3.40$, $SE = .30$, $t = 11.45$, $p < .001$, $d_Z = 2.39$, but there was no difference between the android and robot agents, $b = .09$, $SE = .30$, $t = .31$, *ns*, $d_Z = .06$.

We also detected a main effect of Agent, $F(2, 99.06) = 190.51$, $p < .001$, showing that overall, robots were judged as weirder than both the android and human agents. More important, this main effect of Agent Type occurred in both the experimental and control conditions (*p*s < .001). This indicates that participants in both conditions paid enough attention to the differences between human and nonhuman agents to form discriminative evaluations.

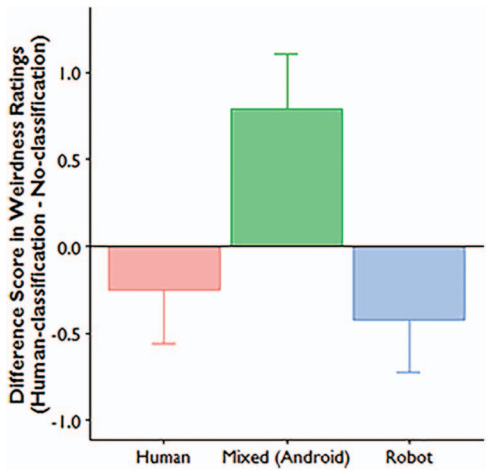

Figure 4. Weirdness difference scores by classification condition (human-classification—no-classification) across the different agent types (human, android, and robot) in Experiment 1. Error bars represent ±1 *SEM*. See the online article for the color version of this figure.

Experiment 2

Experiment 1 showed that the devaluation of mixed agents (androids) was selectively amplified in the human-classification condition (compared with the no-classification condition)—presumably because of disfluency caused by difficult categorization on the ambiguous human-likeness dimension.

We had three main goals for Experiment 2. First, we wanted to replicate the key effects from Experiment 1 with greater power and sample size. Second, we aimed to replace the no-classification control condition from Experiment 1 with a condition that was more closely matched to the experimental condition. To do this in Experiment 2, we used a social categorization control condition, which was designed to have a reasonable but equal difficulty for all agents, rather than selective difficulty for mixed agents (like in the experimental human-classification condition). Third, we added approachability and likability scales (along with the weirdness ratings from Experiment 1), to assess the generalizability of fluency-devaluation effects to other dimensions.

Method

Participants and stimuli. There were 170 UCSD undergraduates (M_{age} = 20.08 years, SD_{age} = 2.20 years; 114 females) who participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program. All experimental stimuli were the same as Experiment 1 (see Figure 1).

Design and procedure. The design for Experiment 2 was similar to Experiment 1, but with two main changes. First, the no-classification condition from Experiment 1 was replaced with the gaze orientation-classification condition in Experiment 2, which required detailed stimulus processing by asking participants to judge whether the agent's head was oriented "toward or away from the camera" (see Figure 2 and Li, Florendo, Miller, Ishiguro, & Saygin, 2015). The human-classification condition remained the same as Experiment 1.

Second, similar to Experiment 1, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials). However, in Experiment 2, each rating block of 24 trials was split by different rating dimensions (i.e., approachability, likability, *or* weirdness) to gauge the generalizability of the fluency effect. As before, after each trial, participants gave the rating using a 1 (*not at all*) to 7 (*very much*) scale, and both conditions made their respective classifications using the 'A' and 'L' keys on the keyboard (response labels were randomized across trials). All other timing and trial parameters were the same as Experiment 1 (see Figure 2).

Thus, in short, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials). Each block was randomly assigned to one of three judgments (approachability, likability, or weirdness; see Figure 2) and required categorization on the image stimuli from Experiment 1 (human-classification vs. orientation-classification). Three subjects performed at ≤50% accuracy and were excluded from the total sample, leaving a final n = 167.

Results

Categorization RTs. We analyzed RTs in the same way as Experiment 1, using an MLM according to a Condition (2: human-classification, no-classification) × Agent (3: human, android, robot) fixed-effects structure.

As with Experiment 1, all effects were significant. Importantly, we clearly replicated the Condition × Agent interaction, $F(2, 167.72) = 13.83$, $p < .001$. Human-classification participants took longer to categorize the android, both compared with the human agent, $b = .05$, $SE = .005$, $t = 9.92$, $p < .001$, $d_Z = 1.08$, and the robot agent, $b = .05$, $SE = .007$, $t = 7.84$, $p < .001$, $d_Z = .85$, but there was no difference between their human and robot RTs. Orientation-classification participants still took longer to classify the android compared with the human agent, $b = .02$, $SE = .005$, $t = 3.38$, $p < .001$, $d_Z = .37$, but not the robot agent, $b = .01$, $SE = .007$, $t = 1.10$, *ns*, $d_Z = .12$, and they also showed no differences between human and robot RTs (see Figure 5).

Also similar to Experiment 1, we detected some less theoretically important effects. Specifically, there were strong main effects of both Condition, $F(1, 167.02) = 16.59$, $p < .001$, and Agent, $F(2, 167.72) = 44.78$, $p < .001$. Generally, these effects showed that orientation-classification subjects had longer RTs, and the android took the most time to categorize when aggregating across conditions. In summary, once again, the android was selectively disfluent (compared with both the other agents)—but only in the human-classification condition.

Scale ratings. We analyzed all scale ratings in the same way as Experiment 1, using MLMs with Condition (2: human-

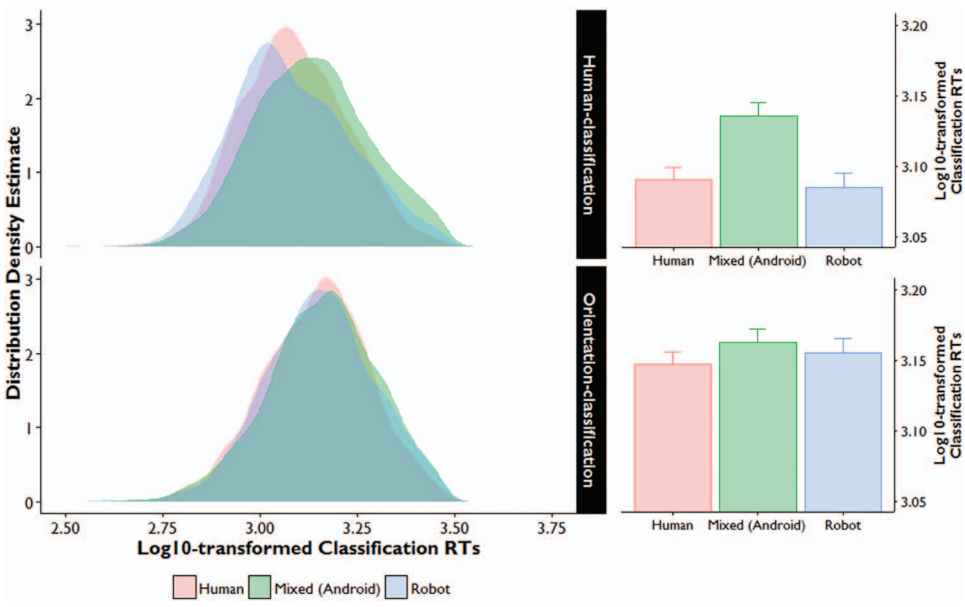

Figure 5. Density distributions and means/*SEM*s for $\log_{10}$-transformed reaction times (RTs) by classification condition (top row = human-classification; bottom row = orientation-classification) and agent type (indicated by colors) for Experiment 2. See the online article for the color version of this figure.

classification, no-classification) × Agent (3: human, android, robot) fixed-effects structures.

Approachability. On approachability, we detected a Condition × Agent interaction, $F(2, 159.94) = 11.24$, $p < .001$. Critically, human-classification participants rated the android lower in approachability (compared with the orientation-classification group), $b = -.64$, $SE = .19$, $t = -3.35$, $p = .001$, $d = .48$ (see Figure 6). Within the orientation-classification condition, participants rated the android as less approachable than the human, $b = -1.24$, $SE = .15$, $t = -8.55$, $p < .001$, $d_Z = .94$, and the robot as less approachable than the android, $b = -.45$, $SE = .16$, $t = -2.85$, $p = .005$, $d_Z = 0.31$. Within the human-classification condition, participants still rated the android as less approachable than the human, $b = -2.27$, $SE = .16$, $t = -13.96$, $p < .001$, $d_Z = 1.51$, but there was no difference between the android and robot agents, $b = .21$, $SE = .17$, $t = 1.23$, *ns*, $d_Z = 0.13$.

Note that we also observed a strong main effect of Agent, $F(2, 159.94) = 168.87$, $p < .001$, such that the robot was less approachable than the android, which in turn was less approachable than the human. This replicates the pattern of evaluative ratings from Experiment 1. This main effect of Agent Type occurred in both the experimental and control conditions (*p*s < .001), suggesting that participants in both conditions of Experiment 2 paid attention to the differences between human and nonhuman agents.

Likability. For likability ratings, we observed very similar effects to approachability. We once again detected a Condition × Agent interaction, $F(2, 165.29) = 5.43$, $p < .01$. As with approachability ratings, human-classification participants gave lower likability scores to androids (compared with the orientation-classification condition), $b = -.48$, $SE = .19$, $t = -2.46$, $p = .02$, $d = .39$ (see Figure 6). Within the orientation-classification condition, participants rated the android as less likable than the hu-

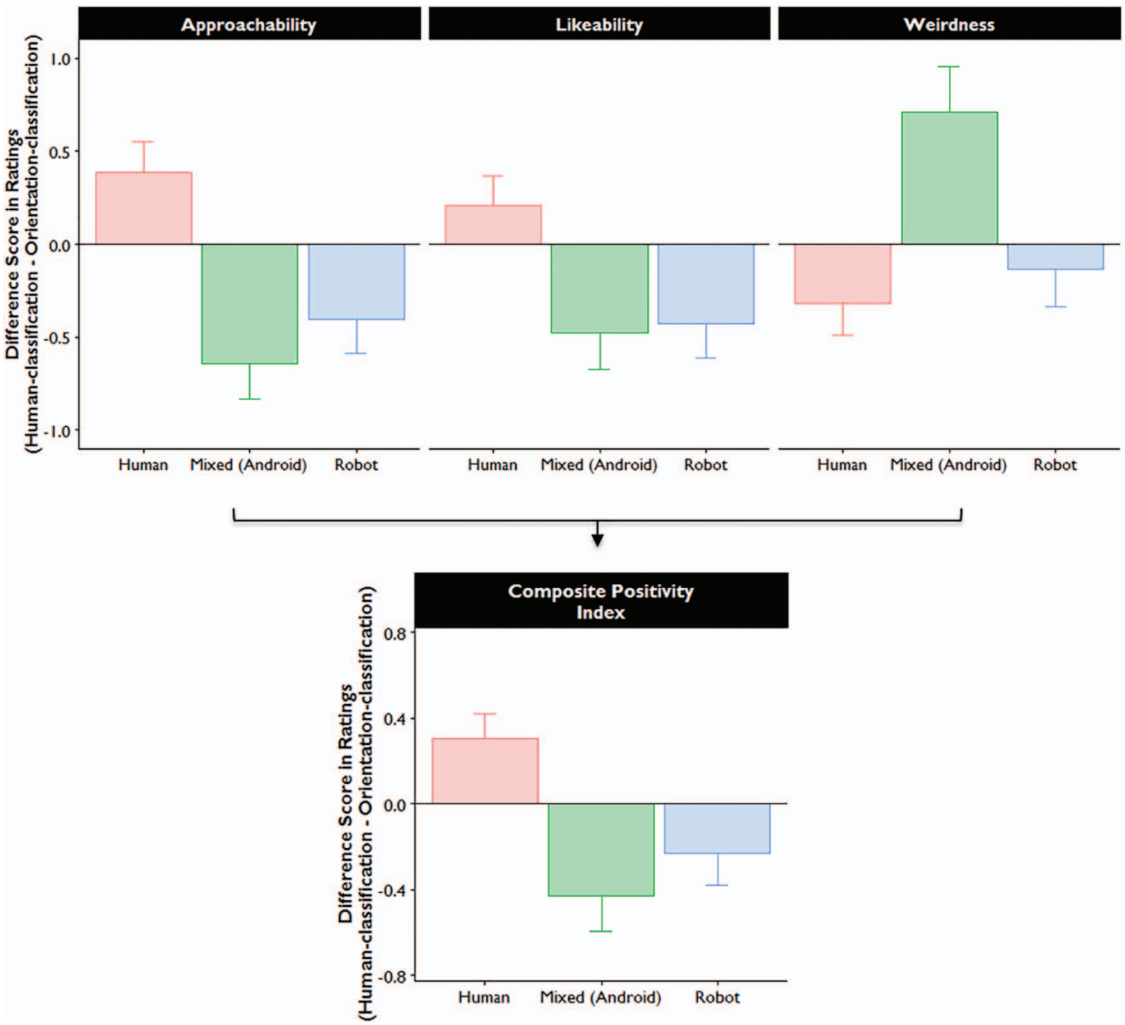

Figure 6. Difference scores by classification condition (human-classification—orientation-classification) on scale ratings for the different agent types (human, android, and robot; indicated by bar colors) in Experiment 2. Individual rating dimensions are shown (approachability, likability, and weirdness), along with the composite positivity index (average of approachability, likability, and reverse-coded weirdness scores). All graphs plot least squares means, along with *SE*s. See the online article for the color version of this figure.

man, $b = -1.39$, $SE = .16$, $t = -8.68$, $p < .001$, $d_Z = .96$, and the robot as less likable than the android, $b = -.31$, $SE = .15$, $t = -2.07$, $p = .04$, $d_Z = .23$. Within the human-classification condition, participants still rated the android as less likable than the human, $b = -2.08$, $SE = .18$, $t = -11.70$, $p < .001$, $d_Z = 1.27$, but there was no difference between the android and robot agents, $b = .26$, $SE = .17$, $t = 1.54$, *ns*, $d_Z = .17$.

Also similar to approachability, there was again significant evidence for a main effect of Agent, $F(2, 165.29) = 181.00$, $p < .001$, where the robot was less likable than the android, which in turn was less likable than the human. This main effect of Agent Type was present in both the experimental and control conditions ($ps < .001$).

Weirdness. With weirdness ratings, the central effects from Experiment 1 replicated. There was a basic main effect of Agent, $F(2, 166.89) = 253.97$, $p < .001$, with participants rating the robot as weirder than the android, which was rated weirder than the human. This main effect occurred in both conditions ($ps < .001$).

Crucially, we also observed a Condition × Agent interaction, $F(2, 166.89) = 7.12$, $p = .001$. Post hoc breakdowns of this interaction revealed that human-classification participants rated the android higher on weirdness (compared with the orientation-classification group), $b = .71$, $SE = .25$, $t = 2.86$, $p = .005$, $d = .51$ (see Figure 6). Within the orientation-classification condition, participants rated the android as weirder than the human, $b = 1.94$, $SE = .19$, $t = 10.18$, $p < .001$, $d_Z = 1.12$, and the robot as weirder than the android, $b = .83$, $SE = .19$, $t = 4.30$, $p < .001$, $d_Z = .47$. Within the human-classification condition, participants still rated the android as weirder than the human, $b = 2.96$, $SE = .21$, $t = 13.88$, $p < .001$, $d_Z = 1.51$, but there was no difference between the android and robot agents, $b = .02$, $SE = .21$, $t = 0.10$, *ns*, $d_Z = .01$.

Composite positivity index. We also created a composite rating by averaging approachability, likability, and reverse-coded weirdness scores. This yielded a general positivity index for each participant, toward each agent. This allowed us to gauge how (and to what magnitude) fluency effects on evaluation generalize more broadly, to overall positive and negative dimensions. Note that while we still used similar MLM methods for the composite rating, this MLM was run on subject means instead of trial-level data (because we needed to obtain a single composite score for each subject by averaging their responses across the other rating dimensions).

We detected clear evidence for a Condition × Agent interaction, $F(2, 159.08) = 7.19$, $p = .001$, and this interaction demonstrated a parallel pattern to the individual rating dimensions. Human-classification participants responded more negatively to androids (compared with orientation-classification participants), $b = -.43$, $SE = .16$, $t = -2.63$, $p = .01$, $d = .40$ (see Figure 6). Finally, we also observed a main effect of Agent, $F(2, 159.08) = 310.69$, $p < .001$, where the robot received lower ratings than both the android and human. This main effect occurred in both conditions ($ps < .001$), showing that participants in both conditions paid attention to the differences between human and nonhuman agents.

Multilevel Mediation Across Experiments 1 and 2

To gauge whether fluency actually drives changes in weirdness evaluations (aside from merely accompanying them), we collapsed the categorization RTs ($\log_{10}$-transformed) and weirdness rating data from both Experiments 1 and 2 and applied multilevel mediation analyses for each condition (human-classification vs. orientation/no-classification), via the *mediation* package in R (R Core Team, 2014; Tingley et al., 2013). While the human-classification condition was exactly the same in both Experiments 1 and 2, note that the alternative conditions were different (i.e., in Experiment 1, no-classification participants only had a simple RT task, but in Experiment 2, this was changed to an orientation-classification task). For simplicity, these conditions were collapsed in the main mediation analyses, but separate analyses for each condition also did not yield any effects. Furthermore, we only used weirdness ratings for mediation because Experiment 1 did not have any approachability or likability ratings (compared with Experiment 2, which had all three dimensions).

Essentially, these methods allow for model-based estimation of the average total, direct, and indirect mediation effects using hierarchical data structures. Such a strategy is appropriate for repeated-measures designs to account for observations nested within subjects (Tingley et al., 2013). Our main predictor was agent mixed-ness, which was dummy-coded as either 0 (not mixed [human and robot]) or 1 (mixed [android]). Our main DV was weirdness ratings, and our mediator was $\log_{10}$-categorization RT. To conduct the multilevel mediation analyses for Experiments 1 and 2, mixed-effects models were constructed for each of the mediation paths, using by-subject random effects parameters. All simulations from the *mediation* package in R were based on 5,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects.

We observed evidence for mediation only within the human-classification subject group. Agent mixed-ness was a significant predictor of $\log_{10}$-RTs (*a*-path: $b = .05$, $SE = .01$, $t = 6.12$, $p < .001$), and $\log_{10}$-RTs were a significant predictor of weirdness ratings (*b*-path: $b = 3.21$, $SE = 1.02$, $t = 3.16$, $p = .002$). The total effect was significant (*c*-path: $b = 1.44$, 95% confidence interval [CI] [.99, 1.91], $p < .01$), as was the average direct effect of agent mixed-ness on weirdness ratings (*c'*-path: $b = 1.31$, 95% CI [.85, 1.80], $p < .01$). And critically, the average causal mediation effect was also significant ($b = .13$, 95% CI [.03, .25], $p = .02$), demonstrating $\log_{10}$-categorization RTs as a mediator.

Note that when these same analyses were done for the orientation/no-classification subject groups, we observed no evidence of mediation. Agent mixed-ness did not predict $\log_{10}$-RTs (*a*-path: $b = .005$, $SE = .006$, $t = .85$, *ns*), and $\log_{10}$-RTs did not predict weirdness ratings (*b*-path: $b = -.07$, $SE = .41$, $t = -.18$, *ns*). The total effect was still significant (*c*-path: $b = .55$, 95% CI [.14, .95], $p = .01$) as was the average direct effect (*c'*-path: $b = .55$, 95% CI [.15, .96], $p = .01$), but there was no average causal mediation effect ($b < .001$, 95% CI [−.01, .01], *ns*).

Experiment 3

We replicated the key findings from Experiment 1 with Experiment 2 (where androids were selectively devalued in the human-classification condition) and this generalized to all evaluative dimensions (approachability, likability, and weirdness). Multilevel mediation analyses across data from both Experiments 1 and 2

revealed that categorization fluency ($\log_{10}$-RTs) mediated the effect between agent mixed-ness and weirdness ratings.

In Experiment 3, we wanted to investigate the specificity of the fluency-devaluation effects. One key question is whether similar devaluation effects emerge if androids are disfluent on a dimension that is *not* a key feature of the agent (an ambiguous dimension that is *not* human-likeness). This is theoretically important because it addresses a key theoretical (yet underexplored) distinction between disfluency that results from integral versus incidental ambiguity (Bodenhausen, 1993). If fluency-rating effects *do* emerge when the android is selectively disfluent but on incidental features (e.g., mixed color background cues, instead of human-likeness), this would argue that devaluation arises from general misattribution of task difficulty. If *not*, this would suggest that, to influence evaluations, the disfluency must be meaningfully connected to the underlying nature of the target. We return to these theoretical alternatives later in the Discussion.

Method

Participants and stimuli. There were 122 UCSD undergraduates (M_{age} = 20.25 years, SD_{age} = 1.49 years; 91 females) who participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program.

Design and procedure. The design for Experiment 3 was similar to Experiment 2, but with two central (and related) changes. First, we altered the image stimuli from Experiments 1 and 2 to be tinted across different shades of blue and green, but importantly, in a way that mirrored the "blending" of the agents themselves. More specifically, the human and robot images were tinted as either 100% blue or 100% green, respectively (image colors *not* mixed), while the android images were tinted to be exactly 50% blue-green (image colors mixed). This was done to create a situation where the level of ambiguity (and thereby, classification difficulty) was still tied to each individual agent, but in a manner that did not specifically relate to the human-likeness dimension (see Figures 1 and 2).

Second, using these color-modified images, we changed the human-classification conditions from Experiments 1 and 2 to a *color*-classification condition in Experiment 3. In Experiment 3, color-classification participants were instead instructed to categorize each of the stimulus images on whether or not they were "blue or green." Note that with this setup, the fluency structure of the human-classification conditions from the previous experiments is still preserved (i.e., robots and humans are easier to categorize, while the android is made selectively difficult) but refers to an incidental dimension (i.e., whether the individual images are "blue or green").

Finally, after the experiment, we also debriefed each participant and asked for their opinions on what they thought the study was investigating. None of the participants mentioned anything related to categorization difficulty impacting their ratings, based on the color or agents in the stimuli.

One subject performed at ≤50% accuracy and was thus excluded from the total sample, leaving a final n = 121. All other parameters and analysis procedures remained the same as Experiments 1 and 2 (see Figure 2).

Results

Categorization RTs. We analyzed RTs using the same MLM methods as Experiments 1 and 2. Aside from a main effect of Agent, $F(2, 121.15) = 104.14$, $p < .001$, we found the predicted Condition × Agent interaction, $F(2, 121.15) = 92.77$, $p < .001$.

Color-classification participants took longer to categorize the android, both compared with the human agent, $b = .08$, $SE = .005$, $t = 15.80$, $p < .001$, $d_Z = 2.09$, and the robot agent, $b = .09$, $SE = .005$, $t = 17.03$, $p < .001$, $d_Z = 2.26$, but there was no difference between their human and robot RTs. Orientation-classification participants took less time to categorize the human agent, both compared with the android agent, $b = -.01$, $SE = .005$, $t = -2.34$, $p = .02$, $d_Z = .29$, and the robot agent, $b = -.02$, $SE = .005$, $t = -3.24$, $p < .01$, $d_Z = .41$, while showing no differences between android and robot RTs, $b < .01$, $SE = .005$, $t = .94$, *ns*, $d_Z = .12$ (see Figure 7).

In summary, the pattern of RTs for the color-classification condition was similar to that of the human-classification conditions in Experiments 1 and 2. In Experiment 3, android images were selectively disfluent (i.e., took longer to categorize) only in the color-classification condition, whereas there were no consistent RT differences for android images in the alternative orientation-classification condition.

Scale ratings. We analyzed all scale ratings for Experiment 3 using the same MLM methods as Experiments 1 and 2.

Approachability. Intriguingly, with Experiment 3, we did *not* observe a Condition × Agent interaction, $F(2, 120.48) = 0.32$, *ns*. When the difficult human-classification condition was changed to a difficult *color*-classification in Experiment 3, the fluency effects on approachability ratings disappeared. Color-classification participants did not differ from the orientation-classification participants on approachability ratings for the android, $b = .19$, $SE = .19$, $t = 1.03$, *ns*, $d = .15$ (see Figure 8).

Furthermore, as was the case with Experiments 1 and 2, we observed a strong main effect of Agent, $F(2, 120.48) = 105.58$, $p < .001$, such that participants rated the robot as less approachable than the android, which was less approachable than the human (same as in Experiments 1 and 2). Once again, this main effect of Agent Type occurred in both the experimental and control conditions (ps $< .001$). This suggests that participants in both conditions of Experiment 3 noticed differences between human and nonhuman agents. Other ratings demonstrated this same pattern, as indicated below.

Likability. The effects on the likability ratings were very similar to those of the approachability dimension. Crucially, we also did *not* detect a Condition × Agent interaction on likability ratings, $F(2, 120.87) = .36$, *ns*. Once again, the color-classification group did not differ from the orientation-classification group for likability ratings on the android, $b = .16$, $SE = .18$, $t = .87$, *ns*, $d = .11$ (see Figure 8).

Note that once again, we also saw a main effect of Agent, $F(2, 120.87) = 120.80$, $p < .001$, where the robot was rated as less likable than the android, which was rated less likable than the human. This main effect was significant in both the experimental and control conditions (ps $< .001$).

Weirdness. With weirdness ratings, once again, we did *not* find a Condition × Agent interaction, $F(2, 120.50) = 1.14$, *ns*. The color-classification group did not differ from the orientation-

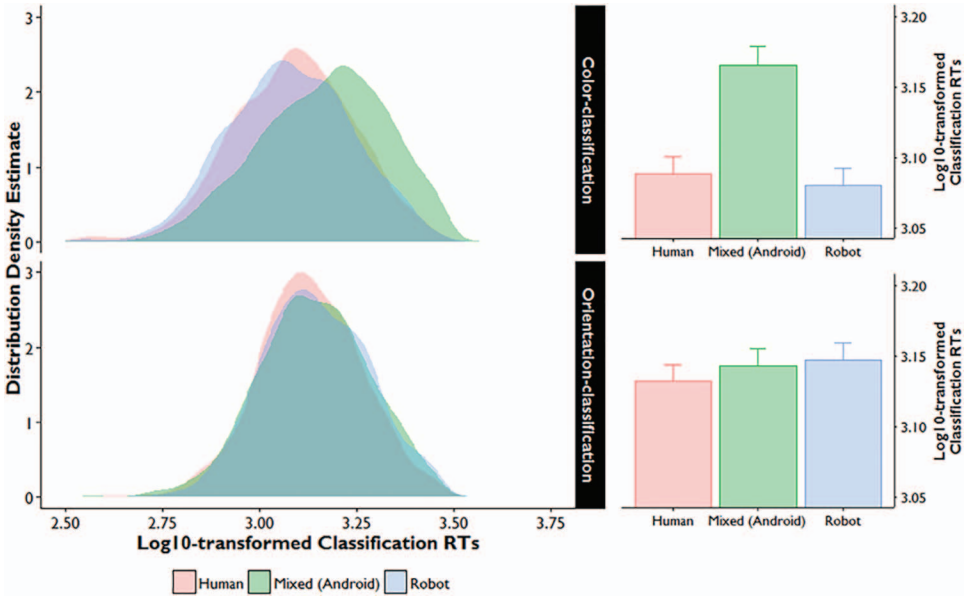

Figure 7. Density distributions and means/*SEM*s for $\log_{10}$-transformed reaction times (RTs) by classification condition (top row = color-classification; bottom-row = orientation-classification) and agent type (indicated by colors) for Experiment 3. See the online article for the color version of this figure.

classification group for weirdness ratings on the android, $b = .06$, $SE = .23$, $t = .28$, *ns*, $d = .03$ (see Figure 8).

Moreover, as with the other rating dimensions, we observed a similar main effect of Agent as Experiments 1 and 2, $F(2, 120.50) = 267.02$, $p < .001$, such that participants rated the robot as weirder than the android, which was rated as weirder than the human. This main effect was significant in both the experimental and control conditions (*p*s $< .001$).

Composite positivity index. As with Experiment 2, we constructed a composite positivity index by averaging approachability, likability, and reverse-coded weirdness scores. Once again, we found a main effect of Agent, $F(2, 121.00) = 231.80$, $p < .001$, where the robot was rated lower than both the android and human agents. This main effect occurred in both classification conditions (*p*s $< .001$). Importantly though, we did not detect any evidence for a Condition × Agent interaction, $F(2, 121.00) = .74$, *ns*, suggesting that differences in agent ratings were not influenced by the categorization condition (color-classification vs. orientation-classification; see Figure 8).

Discussion

Our results suggest that negative evaluations of mixed agents can arise from the processing effort exerted to classify such agents on dimensions relevant to human features. Such disfluency and resulting devaluation occurred only when participants first categorized those agents along the human-likeness dimension on which they were ambiguous (Experiments 1 and 2). These effects did not occur when processing of mixed agents was measured using a generic stimulus detection RT task (Experiment 1) or when these agents were classified along a social orientation dimension on which they were *un*ambiguous (Experiment 2). These effects emerged even though participants devoted an overall comparable amount of time to processing the agents in the control and experimental conditions (Experiment 2). Consistent with this, mediation effects emerged only for participants in the human-classification condition (Experiments 1 and 2). These results cannot be a mere byproduct of general difficulty misattribution, because a color-classification task that made androids selectively disfluent did not yield similar patterns (Experiment 3). Note that our findings also cannot be explained by lack of attention to relevant agent features in the control condition. After all, in each classification condition across all experiments (including control conditions), participants were sensitive to relevant human/nonhuman features and adjusted their evaluative ratings accordingly.

These findings provide an important qualification to previous essentialist claims about the "unbridgeable" boundary between human and nonhuman entities. Recall that these essentialized properties are viewed as deep and immutable traits of the agent in-question. While the experience of being human is certainly one example, note that perceivers can also essentialize other social dimensions (e.g., gender, race, sexual orientation, etc.; Bastian & Haslam, 2006; Haslam, Rothschild, & Ernst, 2000; Haslam, Bastian, Bain, & Kashima, 2006; Howell, Weikum, & Dyck, 2011). Regardless of the specific dimension, essentialized categories carry with them a list of defining characteristics: clear and discrete boundaries from other categories, involuntary and unchanging membership, and observable features that reflect something about the underlying function of the agent (Prentice & Miller, 2007).

On this essentialism view, the spontaneous negative responses to mixed agents arise because of a violation caused by blurring different "natural kinds" for human and nonhuman (see Demoulin, Leyens, & Yzerbyt, 2006; Medin & Ortony, 1989; Prentice & Miller, 2007). While theoretically distinct, other frameworks similarly posit that "mismatches" spontaneously yield negative re-

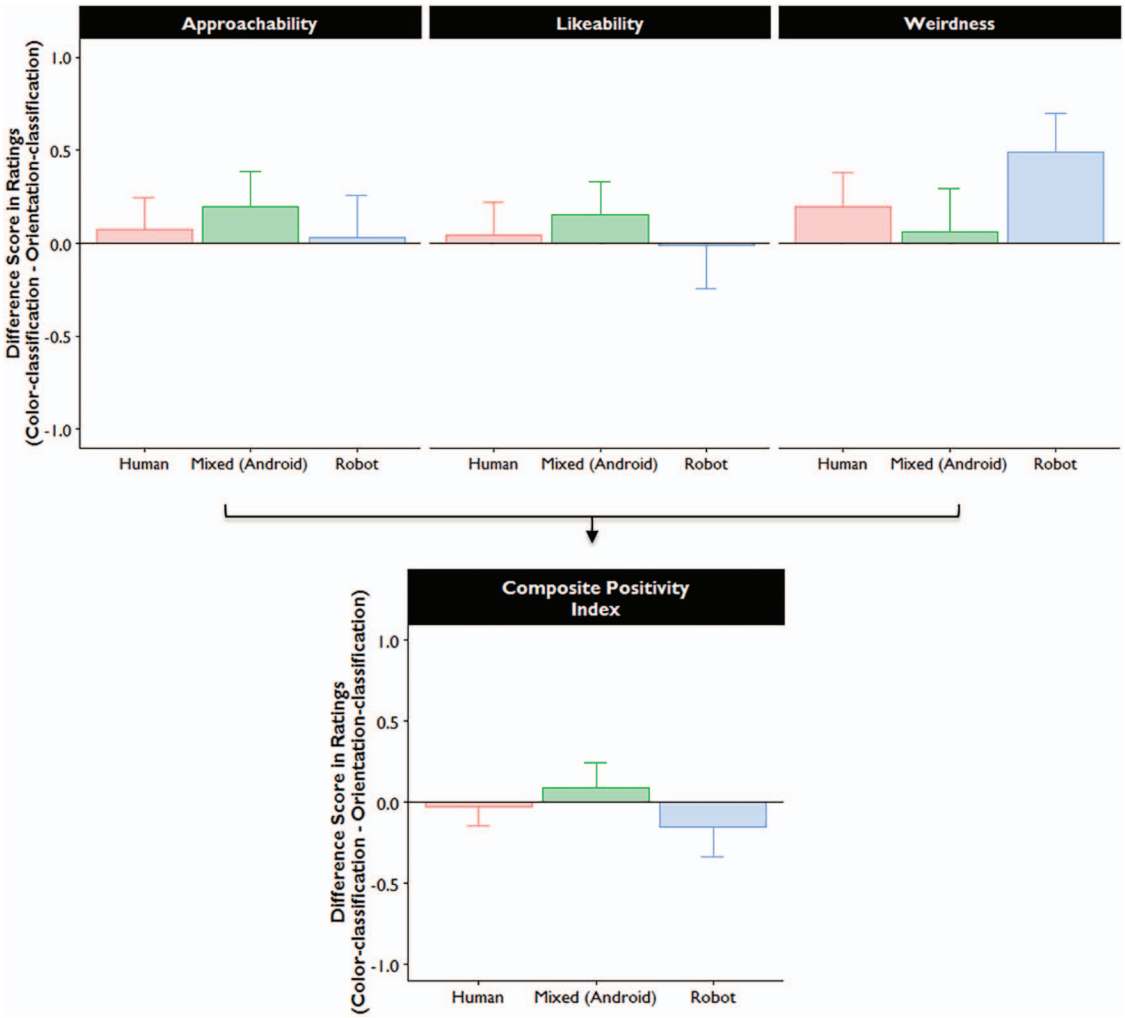

Figure 8. Difference scores by classification condition (color-classification—orientation-classification) on scale ratings for the different agent types (human, android, and robot; indicated by bar colors) in Experiment 3. Individual rating dimensions are shown (approachability, likability, and weirdness), along with the composite positivity index (average of approachability, likability, and reverse-coded weirdness scores). All graphs plot least squares means, along with *SE*s. See the online article for the color version of this figure.

sponses to mixed agents. These mismatches can be perceptual, resulting in conflicting cues in visual, auditory, and motion processing (Katsyri et al., 2015; MacDorman, Green, Ho, & Koch, 2009; Mitchell et al., 2011; Seyama & Nagayama, 2007). They can also be more conceptual, as with incompatible cues for mind perception (Waytz, Gray, Epley, & Wegner, 2010). Our findings challenge views that push for strong automaticity in negative responses to mixed agents—instead, we show context-sensitivity and top-down control of these effects. This nicely corroborates recent work showing that negative responses can be modified by situational factors (Pollick, 2010), because both behavioral and neural responses can vary based on depth of processing (Cheetham et al., 2013) and subjectivity in judging human-likeness (Cheetham, Suter, & Jancke, 2011). Furthermore, it is worth keeping in mind that the tendency toward essentializing categories is quite variable across different tasks and perceivers (Kalish, 2002). As Prentice and Miller (2007) state, "essentialism is not an all-or-none proposition, but rather is a matter of degree" (p. 203). Therefore, our results more so argue against "strong" versions of the theory that do not allow for (a) dependence of responses to these categories on the specific task and context settings or (b) flexibility in the construal of essential nature for human and nonhuman categories.

Critically, the current findings also offer an important theoretical extension of previous fluency models. Note that when substituting the human-classification condition for a *color*-classification task with the same "difficulty structure" (where 50/50 blue-green judgments made android images selectively disfluent), all devaluation effects dissipated. This is a key point, since it highlights the importance of the human-likeness dimension, and it demonstrates that devaluation effects do not simply result from any general categorization difficulty. One explanation for these color-classification results appeals to the distinction between integral versus incidental cues (Bodenhausen, 1993). Devaluation effects

may only follow from disfluency that occurs in response to an agent's key central features (i.e., *integral* human-likeness) rather than ancillary features with similar ambiguity (i.e., *incidental* colored backgrounds). More theoretically, we suggest that the evaluative consequences of (dis)fluency depend on the metacognitive processes that arise from monitoring that processing experience (see Schwarz, 2010, for a review). On one level, contextual variables can impact the processing experience itself, as with making information more or less difficult to process (similar to the different classification conditions in our experiments). However, on another level, contextual variables can further impact how the metacognitive experience of (dis)fluency is interpreted and used—and this can dictate how later judgments are influenced. In our experiments, even though human-classification and color-classification led to similar experiences of disfluent processing, the interpretation of those metacognitive experiences is likely what drove differences in the evaluation of mixed agents. If one experiences disfluency on an integral feature of the agent (e.g., human-likeness), this will likely have downstream negative consequences on judgment. However, similar disfluency on an incidental feature (e.g., color background) would not be deemed relevant, and thus "gated" from the evaluation. Note that the integral versus incidental distinction is different than the idea that subjects discount blatant cues of cognitive disfluency in their ratings. In our studies, the color-based and human-based disfluency were equally strong and salient. The key difference here lies in participants' beliefs about the relevance of fluency cues to the particular judgment (here, evaluative rating). This also corresponds well to ideas of feelings-as-information, where an experience is used as a cue in making judgments, but only when the experience is considered to be appropriate and relevant to the judgment at-hand (Schwarz & Clore, 2003).

More broadly, our work also relates to findings on inhibitory devaluation and stimulus-category competition (Fenske & Raymond, 2006; Raymond, 2009). These models argue that the discomfort with mixed agents is a more specific example of cognitive interference, which emerges from resolving multiple competing stimulus representations via inhibition (Ferrey, Burleigh, & Fenske, 2015). This inhibition leads to negative evaluation, which has been shown with human faces and bodies (Fenske et al., 2005; Ferrey, Frischen, & Fenske, 2012) and nonhuman entities (Griffiths & Mitchell, 2008). Some of these studies found that negative evaluation of stimuli on a categorical boundary can occur without explicit categorization (e.g., Ferrey et al., 2015). Our results differ from these findings. First, our experiments show strong task sensitivity—that is, in our experiments, the devaluation effects were more pronounced with categorization. Second, our Experiment 3 demonstrates that difficult color-categorization did not lead to devaluation of ambiguous images. One interpretation of this difference is that our stimuli are richer and more complex, thus leading participants' construal of those stimuli to be more dependent on categorical processes. In fact, some other work suggests that with highly similar and familiar stimuli, categorization conflict may spontaneously "pop out" without any categorization task, leading to devaluation (Halberstadt, Pecher, Zeelenberg, Ip Wai, & Winkielman, 2013). Further, notice here that our fluency perspective is theoretically distinct from models of inhibition. While fluency theories focus on category processing effort, inhibition theories focus on resolving cognitive conflict (often by attaching inhibitory "tags" to distractors). These theoretical perspectives should be investigated further, along with other frameworks linking emotion and categorization to judgments of human and nonhuman agents (e.g., Burleigh & Schoenherr, 2015; Cheetham et al., 2013, 2011).

Limitations and Future Directions

There are also some important limitations to consider for the current experiments. First, we used longer RTs as our main operationalization of disfluency (i.e., greater RTs indicates greater processing difficulty). Note, however, that while processing difficulty would certainly yield longer RTs, longer RTs do not necessarily have to index disfluency (e.g., longer RTs could also emerge from reduced motivation, increased curiosity toward the stimulus, etc.). Thus, in future studies, alternative fluency measures to RTs should also be incorporated to extend our findings.

Second, our stimuli were highly controlled images of human, android, and robot agents doing different actions (Saygin & Stadler, 2012), but our stimulus set only contained one specific example for each agent type. While our experiments were not designed to investigate subtle gradations in human-likeness between human and robot, our android agents were likely not exactly in the middle of this continuum (the perception of which also probably varies across participants). Therefore, future research may also want to include multiple exemplars of human, android, and robot agents, which may be able to offer more precise degrees of human-likeness (e.g., human/nonhuman morphs; Mathur & Reichling, 2016; Powers, Worsham, Freeman, Wheatley, & Heatherton, 2014).

Third, the directionality of our fluency-devaluation effect on mixed agents remains unclear. More specifically, explicit categorization on the human-likeness dimension (as with Experiments 1 and 2) may amplify disfluency and negative attitudes for mixed agents. Another possibility is that perceivers rapidly and implicitly categorize the agents on the human-likeness dimension as the "default," and forced categorization on an alternative dimension (e.g., orientation-classification) then *reduces* negative attitudes for mixed agents. Our results from Experiment 1 might suggest that perceivers do not spontaneously categorize the agents on human-likeness, since human-classification participants judged the android to be selectively weirder than participants with only a detection RT task (see Figure 4). However, note that participants had much faster RTs without categorization (around 500–600 ms) compared with those in the human-classification condition (around 1,200–1,400 ms; see Figure 3). Implicit categorization on the human-likeness dimension may take longer to emerge, or the no-classification task in Experiment 1 might have distracted participants enough such that spontaneous human-likeness categorization could not occur. This would be interesting to examine in future studies, considering previous papers reporting "spontaneous" negative responses to mixed agents (Mitchell et al., 2011; Tinwell, Grimshaw, Nabi, & Williams, 2011; Złotowski et al., 2015). It may be possible that spontaneous "pop-out" effects from mixed agents are at least partially because of implicit and disfluent categorization on the central human-likeness dimension (also see Burleigh & Schoenherr, 2015; Cheetham et al., 2013, 2011; Ito & Cacioppo, 2000; Wiese, Schweinberger, & Neumann, 2008).

Notwithstanding, our results do not rule out bottom-up effects in processing mixed agents. This seems evident from our RT results in Experiment 1 (see Figure 3) and Experiment 2 (see Figure 5), where participants still took longer to respond to androids even during the alternative control tasks (albeit the differences were much smaller than human-classification conditions). Furthermore, keep in mind that some versions of bottom-up perceptual theories can be considered compatible with fluency frameworks (as when incompatibility is detected early and leads to obligatory difficulty in feature processing).

Finally, some theories hold that disfluency itself is not the key driver of negative evaluation, but rather it is the implications of such disfluency. As examples, disfluency or inconsistency might only matter if they signal a prediction error (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012), a gap in knowledge (Kruglanski, 2013; Roets et al., 2015; Viola et al., 2015), or a collapse in the sense of meaning (Proulx & Inzlicht, 2012). Therefore, we argue that negative responses to mixed agents involve an *interaction* between bottom-up perceptual factors and top-down categorization processes.

Conclusion

In summary, the current findings underscore the broader theoretical point that higher-order processes can modify how we evaluate nonhuman agents. These influences, which we explored here with "hot" evaluative judgments, are also likely to bear on "cold" cognitive assessments of mind perception, agency, and intentionality (Gray & Wegner, 2012; Waytz, Gray, Epley, & Wegner, 2010). Essentially, categorization (dis)fluency can modify the impact of this fundamental boundary between human and nonhuman, where we can dehumanize the living (Haslam, 2006) or anthropomorphize the artificial (Chandler & Schwarz, 2010; Epley, Waytz, & Cacioppo, 2007).

References

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37,* 13–20. http://dx.doi.org/10.1111/1469-8986.3710013

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42,* 228–235. http://dx.doi.org/10.1016/j.jesp.2005.03.003

Bastian, B., & Haslam, N. (2007). Psychological essentialism and attention allocation: Preferences for stereotype-consistent versus stereotype-inconsistent information. *The Journal of Social Psychology, 147,* 531–541. http://dx.doi.org/10.3200/SOCP.147.5.531-542

Bastian, B., Loughnan, S., & Koval, P. (2011). Essentialist beliefs predict automatic motor responses to social categories. *Group Processes & Intergroup Relations, 14,* 559–567. http://dx.doi.org/10.1177/1368430210385258

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv,*1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48.

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 13–37). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-0-08-088579-7.50006-5

Burleigh, T. J., & Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology, 5,* 1488. http://dx.doi.org/10.3389/fpsyg.2014.01488

Carr, E. W., Rotteveel, M., & Winkielman, P. (2016). Easy moves: Perceptual fluency facilitates approach-related action. *Emotion, 16,* 540–552. http://dx.doi.org/10.1037/emo0000146

Chandler, J., & Schwarz, N. (2010). Use does not wear ragged the fabric of friendship: Thinking of objects as alive makes people less willing to replace them. *Journal of Consumer Psychology, 20,* 138–145. http://dx.doi.org/10.1016/j.jcps.2009.12.008

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category processing and the human likeness dimension of the uncanny valley hypothesis: Eye-tracking data. *Frontiers in Psychology, 4,* 108. http://dx.doi.org/10.3389/fpsyg.2013.00108

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5,* 126. http://dx.doi.org/10.3389/fnhum.2011.00126

Demoulin, S., Leyens, J. P., & Yzerbyt, V. (2006). Lay theories of essentialism. *Group Processes & Intergroup Relations, 9,* 25–42. http://dx.doi.org/10.1177/1368430206059856

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68,* 87–106. http://dx.doi.org/10.2307/2025382

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114,* 864–886. http://dx.doi.org/10.1037/0033-295X.114.4.864

Fenske, M. J., & Raymond, J. E. (2006). Affective influences of selective attention. *Current Directions in Psychological Science, 15,* 312–316. http://dx.doi.org/10.1111/j.1467-8721.2006.00459.x

Fenske, M. J., Raymond, J. E., Kessler, K., Westoby, N., & Tipper, S. P. (2005). Attentional inhibition has social-emotional consequences for unfamiliar faces. *Psychological Science, 16,* 753–758. http://dx.doi.org/10.1111/j.1467-9280.2005.01609.x

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: A novel account of the uncanny valley. *Frontiers in Psychology, 6,* 249. http://dx.doi.org/10.3389/fpsyg.2015.00249

Ferrey, A. E., Frischen, A., & Fenske, M. J. (2012). Hot or not: Response inhibition reduces the hedonic value and motivational incentive of sexual stimuli. *Frontiers in Psychology, 3,* 575.

Frenkel-Brunswik, E. (1949). Intolerance of ambiguity as an emotional and perceptual personality variable. *Journal of Personality, 18,* 108–143. http://dx.doi.org/10.1111/j.1467-6494.1949.tb01236.x

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences, 8,* 404–409. http://dx.doi.org/10.1016/j.tics.2004.07.001

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

Griffiths, O., & Mitchell, C. J. (2008). Negative priming reduces affective ratings. *Cognition and Emotion, 22,* 1119–1129. http://dx.doi.org/10.1080/02699930701664930

Halberstadt, J., Pecher, D., Zeelenberg, R., Ip Wai, L., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty in averageness appear and reverse. *Psychological Science, 24,* 2343–2346. http://dx.doi.org/10.1177/0956797613491969

Halberstadt, J., & Winkielman, P. (2013). When good blends go bad: How fluency can explain when we like and dislike ambiguity. In C. Unkelbach & R. Greisfelder (Eds.), *The experience of thinking: How*

feelings from mental processes influence cognition and behavior (pp. 133–151). New York, NY: Psychology Press.

Halberstadt, J., & Winkielman, P. (2014). Easy on the eyes, or hard to categorize: Classification difficulty decreases the appeal of facial blends. *Journal of Experimental Social Psychology, 50,* 175–183. http://dx.doi.org/10.1016/j.jesp.2013.08.004

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10,* 252–264. http://dx.doi.org/10.1207/s15327957pspr1003_4

Haslam, N., Bastian, B., Bain, P., & Kashima, Y. (2006). Psychological essentialism, implicit theories, and intergroup relations. *Group Processes & Intergroup Relations, 9,* 63–76. http://dx.doi.org/10.1177/1368430206059861

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39,* 113–127. http://dx.doi.org/10.1348/014466600164363

Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26,* 1508–1518. http://dx.doi.org/10.1016/j.chb.2010.05.015

Howell, A. J., Weikum, B. A., & Dyck, H. L. (2011). Psychological essentialism and its association with stigmatization. *Personality and Individual Differences, 50,* 95–100. http://dx.doi.org/10.1016/j.paid.2010.09.006

Ishiguro, H. (2006). Android science: Conscious and subconscious recognition. *Connection Science, 18,* 319–332. http://dx.doi.org/10.1080/09540090600873953

Ishiguro, H. (2007). Android science. In S. Thrun, R. Brooks, & H. Durrant-Whyte (Eds.), *Robotics research* (pp. 118–127). Berlin Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-3-540-48113-3_11

Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology, 36,* 660–676. http://dx.doi.org/10.1006/jesp.2000.1430

Jentsch, E. (1906). Zur Psychologie des Unheimlichen [On the psychology of the uncanny]. *Psychiatrisch-Neurologische Wochenschrift, 8,* 195–198.

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition, 30,* 340–352. http://dx.doi.org/10.3758/BF03194935

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6,* 390. http://dx.doi.org/10.3389/fpsyg.2015.00390

Kruglanski, A. W. (2013). *Lay epistemics and human knowledge: Cognitive and motivational bases*. Berlin, Germany: Springer Science & Business Media.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random and fixed effects for linear mixed effect models* (lmer objects of lme4 package). R package version 2.0–11. Retrieved from http://CRAN.R-project.org/package=lmerTest

Li, A. X., Florendo, M., Miller, L. E., Ishiguro, H., & Saygin, A. P. (2015, March). Robot form and motion influences social attention. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 43–50). New York, NY: ACM.

MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25,* 695–710. http://dx.doi.org/10.1016/j.chb.2008.12.026

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems, 7,* 297–337. http://dx.doi.org/10.1075/is.7.3.03mac

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition, 146,* 22–32. http://dx.doi.org/10.1016/j.cognition.2015.09.008

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosnaidou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, United Kingdom: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511529863.009

Mitchell, W. J., Szerszen, K. A., Sr., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & Macdorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception, 2,* 10–12. http://dx.doi.org/10.1068/i0415

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

Owen, H. E., Halberstadt, J., Carr, E. W., & Winkielman, P. (2016). Johnny Depp, reconsidered: How category-relative processing fluency determines the appeal of gender ambiguity. *PLoS ONE, 11,* e0146328. http://dx.doi.org/10.1371/journal.pone.0146328

Pollick, F. E. (2010). In search of the uncanny valley. In P. Daras & O. M. Ibarra (Eds.), *User centric media* (pp. 69–78). Berlin Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-12630-7_8

Powers, K. E., Worsham, A. L., Freeman, J. B., Wheatley, T., & Heatherton, T. F. (2014). Social connection modulates perceptions of animacy. *Psychological Science, 25,* 1943–1948. http://dx.doi.org/10.1177/0956797614547706

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science, 16,* 202–206. http://dx.doi.org/10.1111/j.1467-8721.2007.00504.x

Proulx, T., & Inzlicht, M. (2012). The five "A" s of meaning maintenance: Finding meaning in the theories of sense-making. *Psychological Inquiry, 23,* 317–335. http://dx.doi.org/10.1080/1047840X.2012.702372

Rajaram, S., & Geraci, L. (2000). Conceptual fluency selectively influences knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1070–1074. http://dx.doi.org/10.1037/0278-7393.26.4.1070

Raymond, J. (2009). Interactions of attention, emotion and motivation. *Progress in Brain Research, 176,* 293–308. http://dx.doi.org/10.1016/S0079-6123(09)17617-3

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8,* 338–342. http://dx.doi.org/10.1006/ccog.1999.0386

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9,* 45–48. http://dx.doi.org/10.1111/1467-9280.00008

Roets, A., Kruglanski, A. W., Kossowska, M., Pierro, A., & Hong, Y. Y. (2015). The motivated gatekeeper of our minds: New directions in need for closure theory and research. *Advances in Experimental Social Psychology, 52,* 221–283. http://dx.doi.org/10.1016/bs.aesp.2015.01.001

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7,* 413–422. http://dx.doi.org/10.1093/scan/nsr025

Saygin, A. P., & Stadler, W. (2012). The role of appearance and motion in action prediction. *Psychological Research, 76,* 388–394. http://dx.doi.org/10.1007/s00426-012-0426-z

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review, 2,* 87–99. http://dx.doi.org/10.1207/s15327957pspr0202_2

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (Eds.), *The mind in context* (pp. 105–125). New York, NY: Guilford Press.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry, 14,* 296–303. http://dx.doi.org/10.1080/1047840X.2003.9682896

Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence, 16,* 337–351. http://dx.doi.org/10.1162/pres.16.4.337

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2013). *Mediation: R package for causal mediation analysis*. R package version 4. 4. Retrieved from http://CRAN.R-project.org/package=mediation

Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior, 27,* 741–749. http://dx.doi.org/10.1016/j.chb.2010.10.018

Viola, V., Tosoni, A., Brizi, A., Salvato, I., Kruglanski, A. W., Galati, G., & Mannetti, L. (2015). Need for cognitive closure modulates how perceptual decisions are affected by task difficulty and outcome relevance. *PLoS ONE, 10,* e0146002. http://dx.doi.org/10.1371/journal.pone.0146002

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14,* 383–388. http://dx.doi.org/10.1016/j.tics.2010.05.006

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/b17198

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1235–1253. http://dx.doi.org/10.1037/0278-7393.19.6.1235

Wiese, H., Schweinberger, S. R., & Neumann, M. F. (2008). Perceiving age and gender in unfamiliar faces: Brain potential evidence for implicit and explicit person categorization. *Psychophysiology, 45,* 957–969. http://dx.doi.org/10.1111/j.1469-8986.2008.00707.x

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology, 81,* 989–1000. http://dx.doi.org/10.1037/0022-3514.81.6.989

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17,* 799–806. http://dx.doi.org/10.1111/j.1467-9280.2006.01785.x

Winkielman, P., Olszanowski, M., & Gola, M. (2015). Faces in-between: Evaluative responses to faces reflect the interplay of features and task-dependent fluency. *Emotion, 15,* 232–242. http://dx.doi.org/10.1037/emo0000036

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Erlbaum.

Złotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2015). Persistence of the uncanny valley: The influence of repeated interactions and a robot's attitude on its perception. *Frontiers in Psychology, 6,* 883. http://dx.doi.org/10.3389/fpsyg.2015.00883

Received April 26, 2016
Revision received July 22, 2016
Accepted July 25, 2016 ■