Social Neuroscience

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/psns20

Exploring how harming and helping behaviors drive prediction and explanation during anthropomorphism

Lasana T. Harris , Noor van Etten & Tamara Gimenez-Fernandez

To cite this article: Lasana T. Harris , Noor van Etten & Tamara Gimenez-Fernandez (2020): Exploring how harming and helping behaviors drive prediction and explanation during anthropomorphism, Social Neuroscience, DOI: 10.1080/17470919.2020.1799859

To link to this article: https://doi.org/10.1080/17470919.2020.1799859

Published online: 16 Aug 2020.

SOCIAL NEUROSCIENCE
https://doi.org/10.1080/17470919.2020.1799859

Exploring how harming and helping behaviors drive prediction and explanation during anthropomorphism

Lasana T. Harris[a], Noor van Etten[b] and Tamara Gimenez-Fernandez[c]

[a]Department of Experimental Psychology, University College London, London, UK; [b]Department of Social and Organizational Psychology, Leiden University, Leiden, Netherlands; [c]Department of Experimental Psychology, Autonomous University of Madrid, Madrid, Spain

ABSTRACT

Cacioppo and colleagues advanced the study of anthropomorphism by positing three motives that moderated the occurrence of this phenomenon; belonging, effectance, and explanation. Here, we further this literature by exploring the extent to which the valence of a target's behavior influences its anthropomorphism when perceivers attempt to explain and predict that target's behavior, and the involvement of brain regions associated with explanation and prediction in such anthropomorphism. Participants viewed videos of varying visually complex agents - geometric shapes, computer generated (CG) faces, and greebles - in nonrandom motion performing harming and helping behaviors. Across two studies, participants reported a narrative that explained the observed behavior (both studies) while we recorded brain activity (study one), and participants predicted future behavior of the protagonist shapes (study two). Brain regions implicated in prediction error (striatum), not language generation (inferior frontal gyrus; IFG) engaged more to harming than helping behaviors during the anthropomorphism of such stimuli. Behaviorally, we found greater anthropomorphism in explanations of harming rather than helping behaviors, but the opposite pattern when participants predicted the agents' behavior. Together, these studies build upon the anthropomorphism literature by exploring how the valence of behavior drives explanation and prediction.

ARTICLE HISTORY
Received 21 March 2019
Revised 19 June 2020
Published online 15 August 2020

KEYWORDS
Anthropomorphism; social cognition; belonging; narrative; prediction; explanation

Anthropomorphism – engaging social cognition to nonhuman entities – demonstrates the pervasiveness of social perception, allowing human beings to imbue nonhuman entities with mental states (Epley, Akalis et al., 2008a; Epley, Waytz et al., 2008b; Epley et al., 2007). An agent – an entity that originates its own behavior – that is not human can trigger anthropomorphism since people engage social cognition by default (Fiske & Taylor, 1991, 2010). Such magical thinking satisfies the fundamental human motives to belong (Baumeister & Leary, 1995), to make sense of or explain the world around us (Harris, 2017; Waytz, Morewedge et al., 2010b), and to exert a degree of control over our environment (Burger & Cooper, 1979; Rothbaum, Weisz, & Snyder, 1982; Rotter, 1966). Concurrent with the rise of anthropomorphism research, social neuroscience emerged as a methodological tool to facilitate exploring the black box of social cognition (Harmon-Jones & Devine, 2003; Ochsner & Lieberman, 2001). After pioneering such psychophysiological techniques since the 1970's, Cacioppo advised that social psychological theory should play a central role in guiding brain imaging studies (Cacioppo et al., 2003). Here, we present research at the convergence of these two contributions, exploring how social psychology theory about social cognition during person perception can be applied to the study of anthropomorphism and the brain. Specifically, we examine the impact of the valence of the agents' behavior on brain mechanisms implicated in explanation and prediction during anthropomorphism to provide converging evidence for behavioral data.

Motives for anthropomorphism

Cacioppo and colleagues found evidence for three distinct motives for anthropomorphism; effectance, explanation, and belonging (Epley et al., 2007). Effectance motivation describes a perceiver's desire to have mastery over their environment (White, 1959). Therefore, anthropomorphism is more likely in cases where nonhuman agents do not function the way they are intended (Johnson & Barrett, 2003; Waytz et al., 2010b); by anthropomorphizing the agent, the perceiver retains control over the agent, attributing the dysfunction to the mind of the agent. Importantly, anthropomorphism in this instance also allows the perceiver to *predict* the future behavior of the agent, further regaining control. This latter function of the effectance motive is implicit in

CONTACT Lasana T. Harris lasana.harris@ucl.ac.uk 26 Bedford Way, WC1H 0AP, London, UK.

the literature, but consistent with the function of social cognition in the person perception literature (see Andrews, 2012).

Also consistent with the person perception literature is the second motive for anthropomorphism: explanation. Imbuing an object with a mental life allows the perceiver to *explain* its behavior because the agent's mental states are responsible for driving its behavior (Dennett, 1989). This satisfies the fundamental human need for understanding (Baumeister & Newman, 1994), providing reasons why the agent engaged in a particular behavior. This social motive is also consistent with the effectance motive given that explanation and prediction are usually discussed as co-occurring during social cognition (see Fiske & Taylor, 1991, 2010). However, the explanation motive stands apart since it is described usually as accessibility and applicability of explanations, rather than a strict focus of the engagement of explanatory processes for their own psychological benefit.

The final motive of belonging argues that a perceiver will see human beings in agents that are not human in a bid to form social connections when the perceiver is socially isolated (Eyssel & Reich, 2013). This social motive to belong (Baumeister & Leary, 1995) is consistent with hyper-social behavior common amongst human beings (Hawkes, 2014; Hrdy, 2009; Tomasello & Gonzalez-Cabrera, 2017; Tomasello et al., 2012). Thus, anthropomorphic perceptions should increase when the fundamental need to belong is threatened.

The three motives for anthropomorphism are a subset of a broader set of motives driving social cognition during person perception (for one account, see Fiske, 2003). In addition to effectance (control), belonging, and explanation (understanding), theorists argue that people engage social cognition to other people as a way of self-enhancing, and of trusting other people (Fiske, 2003). Self-enhancement is satisfied through impression management concerns (Kowalski & Leary, 1990), and human beings have a fundamental need to trust other people (Bowlby, 1969). Both of these motives depend on the knowledge that a social target has a mind, and that said mind is capable of both forming an evaluation of the perceiver and of having good or bad intentions that would promote helping or harming behaviors toward the perceiver. Therefore, these latter two motives are irrelevant in the case of anthropomorphism if the perceiver preserves the belief that the non-human agent does not indeed actually have a mind, therefore is not forming an impression of the perceiver, and is not capable of harboring good or ill intentions toward the perceiver. This belief keeps the anthropomorphized agent beyond the boundaries of moral protection, highlighting a fundamental difference between social cognition to human and non-human targets.

Dissociating social cognition to people and objects

Engaging social cognition to people raises impression or reputation management concerns (Fiske & Taylor, 1991, 2010) and makes morality salient (Bandura, 1989; Dennett, 1989). As mentioned above, anthropomorphism does not trigger such processing since people may preserve the belief that an anthropomorphized agent is not a human being. Therefore, it is useful to think about social cognition to agents on a continuum, with perception of full human beings at one end where morality and reputation concerns reside, and anthropomorphized agent perception at the other. Importantly, this continuum metaphor does not suggest that we consider anthropomorphism to be the opposite of full human perception. Stated differently, we did not put these two concepts on a continuum to posit them as opposites, but rather to suggest that though anthropomorphism is not the same as fully human perception, it is also not a discrete category. Therefore, our continuum is not a comprehensive continuum that describes the perception of all entities, including non-anthropomorphic simple objects. Rather, it is a continuum of social cognition (specifically mental state attribution or mind perception), with anthropomorphism anchoring one end, and fully human perception at the other.

At this point, it is also necessary to differentiate anthropomorphism from related psychological concepts. In the literature, researchers sometimes conflate the terms animacy, agency, and anthropomorphism. There are many living entities that are not animate (e.g., a tree), and there are many animate living entities that are not human (e.g., a squirrel). Therefore, all living entities are not animate, and an animate entity need not be alive. Nonetheless, animacy can be defined as the attribution of life to something. This is not the same as anthropomorphism, which describes attributing a mind (not a life) to a non-human entity. Our definition of anthropomorphism also differs from agency since agency is sufficient for anthropomorphism, but not necessary. For instance, a doll can be anthropomorphized even though it does not originate its own behavior. Thus, though anthropomorphism and agency are terms often used interchangeably (Waytz et al., 2010a), anthropomorphism goes beyond merely attributing life to an inanimate object or describing observable behavior; the differences lies in qualities that people think of as distinctly human (possessing a mental life or mind). Perceivers use such a distinction to determine when

agent perception results anthropomorphism (see Epley et al., 2008a; Waytz et al., 2010a). Moreover, although primary emotions like anger or happiness are not distinctly human, they are still part of the experience of being human (Demoulin et al., 2004).

Consistent with the distinction in motives driving social cognition to humans and non-human agents, separate but overlapping brain networks engage during social cognition to human and to non-human targets (Harris & Fiske, 2008; Harris et al., 2005). Specifically, areas of medial prefrontal cortex (MPFC), superior temporal sulcus (STS), temporo-parietal junction (TPJ), anterior temporal pole (ATP) precuneus, and posterior cingulate cortex (PCC) engage during social cognition (Amodio & Frith, 2006; Frith & Frith, 2001; Gallagher & Frith, 2003; Mars et al., 2012; Van Overwalle, 2009). However, studies of anthropomorphism, while relying on the STS for biological motion detection (Puce & Perrett, 2003; Servos et al., 2002; Vaina et al., 2001) and the perception of some objects such as greebles (Gauthier et al., 2004), including the fusiform face area (Moran et al., 2012; Schultz et al., 2003), often depends on the amygdala (Harris & Fiske, 2008; Heberlein & Adolphs, 2004) instead of MPFC, TPJ, and precuneus. Such dissociation may allow preservation of the belief that anthropomorphized agents are not human beings. However, such differences may also be due to the different visual complexity between geometric shapes and human beings, independent of morality.

Explanation versus prediction

Explanation, effectance, and belonging overlap between social cognition to human and non-humans. Explanation and prediction (effectance) in particular are relevant for social cognition to humans. Social psychological theory argues that the primary function of social cognition to humans is to explain and predict behavior (Fiske & Taylor, 1991, 2010). Specifically, knowing something about a person's mind makes salient their intentions, goals, emotional states, and personality traits; information that can be used to understand or explain why they engaged in past and current behavior, and predict what behaviors they may engage in the future. Such mental state information not only satisfies the core human need for understanding, but also offers a degree of control over the perceiver's outcomes regarding that person. For instance, if Sally thinks that Anne harbors negative intentions toward her, Sally can adjust her behavior and future interactions with Anne to minimize the likelihood that Anne can act on those ill intentions. Therefore, there are survival benefits related to engaging social cognition to any human being.

However, explanation and prediction are not opposite sides of the same coin. For instance, one might consider explanation as relying on inductive reasoning, while prediction depends on deductive reasoning. However, it is possible to make predictions without deductive reasoning (Andrews, 2009, 2012); infants without this advanced cognitive ability can still predict the behavior of agents based on normative inferences (Phillips et al., 2002; Trevarthen, 1979). Moreover, studies with adults also question the role of social cognition in generating behavioral predictions. Personality traits are notoriously poor predictors of behavior (Paunonen & Jackson, 1985; Pervin, 1985), and at least one study that does not require trait generation demonstrates that norms better predict people's behavior than traits (Harris et al., 2016). The question remains whether such distinctions between explanation and prediction are present during anthropomorphism.

Valence of behavior

There is a burgeoning literature on the valence of the behavior of the anthropomorphic agent, and its impact on the extent to which the perceiver anthropomorphizes the agent. For instance, agents who commit harms are attributed less agency than non-harmful agents (Khamitov et al., 2016). Agents that are harmed, however, are anthropomorphized more than non-harmed agents (Swiderska & Küster, 2018; Ward et al., 2013). Agents that are helped are also anthropomorphized more, but only when the perceiver takers the perspective of the helper (Tanibe et al., 2017). However, in all these studies, the anthropomorphized agent was a robot, avatar, or corporation; a much more visually (and conceptually) complex entity than simple geometric shapes.

The results above run counter to a negativity bias described in the person perception literature, such that harmful behaviors tend to better capture the attention, are better remembered, and lead to more dispositional attributions than helpful behaviors (Cacioppo & Gardner, 1999; Carretié et al., 2001; Mogg & Bradley, 1998; Mogg et al., 2000; Peeters & Czapinski, 1990; Taylor, 1991). In addition, negative events are more likely to be attributed to external sources, such as an anthropomorphized agent (Morewedge, 2009). Perhaps the difference between these two literatures hinges on the relevance of morality. If an agent is human, then that person's mind motivated their behavior, and they can be held accountable for harmful behaviors; thus possessing a mind is necessary for such accountability. However, if an agent is not human, then moral rules need not apply, so thinking about their mind is superfluous. Given the

low prevalence of witnessing geometric shapes compared to robots, avatars, and corporations as agents, agentic shapes engaged in harmful behaviors may be emotionally salient. Together, this literature suggests additional criteria may moderate anthropomorphism besides the three motives identified above, extending the work of Cacioppo and colleagues.

Here, we test the extent to which the valence of behavior (harming versus helping) interacts with motives for anthropomorphism, specifically explanation and prediction. Specifically, in the first study, we explore brain mechanisms that underlie language generation (explanation) and decision-making (prediction) across harming and helping behaviors of different types of agents. We also vary the visual complexity of the agents to explore whether this variable interacts with valence. Hence, we aim to determine whether visual complexity (manipulated by the type of agent) or morality (manipulated by the valence of behavior) contributed to differential processing in brain regions associated with prediction and explanation during anthropomorphism. In the second study, we focus on the anthropomorphism of geometric shapes, and explore the extent to which prediction and explanation are influenced by the motive to belong. In both studies, we hypothesize that there may be differential impacts of the valence of behavior on explanation and prediction. Moreover, consistent with the literature, we hypothesize that feeling socially isolated may impact both explanation and prediction.

Study one

We constructed a brain imaging study to directly assess the impact of anthropomorphism on brain regions implicated in explanation and prediction. We conceptually replicated the classic Heider and Simmel (1944) paradigm of geometric shapes in nonrandom and random motion, adding two other agent categories with increased visual complexity as stimuli; greebles and computer-generated (CG) faces. These latter two agents matched the movement trajectories of the shapes, allowing us to determine whether the visual complexity of the agent mattered for both explanation and prediction. Greebles are objects that drive activity in the STS and fusiform gyrus (Puce & Perrett, 2003; Servos et al., 2002; Vaina et al., 2001), providing us an agent that was more visually complex than a geometric shape, but not as complex as humans. CG faces are much closer to human faces on the human-object spectrum, but are not actual humans' faces. Nonetheless, they are substantially more visually complex than both greebles and geometric shapes. We compare activity to these different agents performing helping and harming behaviors in a brain region associated with language generation (IFG) and one associated with prediction error (striatum). We expect that if certain agents or valenced behaviors drive more predictive or explanatory processes, we should detect differences in the respective brain regions.

Method

Participants

Twenty participants completed the brain imaging paradigm. We lost four participants to data recording errors, resulting in a final sample of 16 participants recruited from an American University population. All participants gave informed consent before beginning the experiment, and the study received ethical approval from the University Institutional Review Board.

Materials

We created 180 videos, varying the colors, families, or identities of the agents (geometric shapes, greebles, or CG faces) across two stimuli sets (see Figure 1). Each video lasted 20 seconds. The geometric shapes were circles, squares, and pentagons, and were either red, yellow, blue, purple, orange, or green. The greebles came from a database of such objects (see Gauthier & Tarr, 1997), and equal numbers were selected from the same two "families". CG faces came from a face database

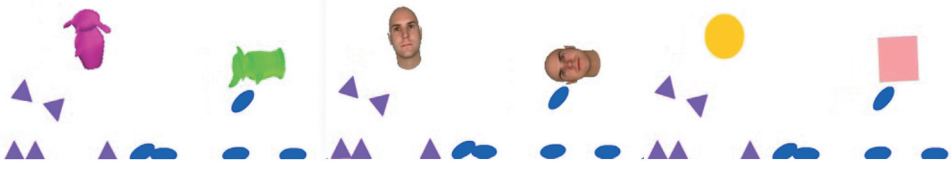

Figure 1. *Stimuli of Agents in Motion*. Screenshots of the three types of agents used in study one. Here, the protagonist agent (left) was engaged in sorting behavior, separating the purple triangles from the blue ovals.

(Oosterhof & Todorov, 2008), and each face selected was at the midpoint of dominance and trustworthiness dimensions. Within each stimulus set, there were 30 videos for each agent category, 10 displaying helping behavior, 10 displaying harmful behavior, and 10 engaged in random motion. The videos depicted different kinds of behaviors, including a protagonist agent trapped in structures, going up inclines and steps, sorting and arranging objects, or avoiding other agents. Each video contained two primary agents (the protagonist and a supporting agent) that both belonged to a category (geometric shapes, greebles, CG faces), except for the avoiding video where small dots served as the agents to be avoided in addition to the two primary agents. To manipulate the identities of the primary agents, we substituted the geometric shapes with either greebles or CG faces. Therefore, the movement pattern of the primary agents was identical across these three conditions. Harming and helping behaviors were differentiated by the goals of the supporting agent, such that harming agents hindered the protagonist agent with the task, while helping agents assisted.

Procedure

We counterbalanced the two stimuli sets across participants, such that half of the participants viewed videos from one set, and the other half of participants from the second set. Participants viewed 90 videos of agents in random and nonrandom motion. Participants were instructed to tell stories silently in their heads about the action they observed in the videos. Each video was randomly presented, followed by a two to eight second jittered fixation cross. After scanning, participants observed the 20 shape videos displaying harming and helping behaviors, and wrote down the story they had told themselves in the scanner about the action. Participants were paid $20 USD for their participation, fully debriefed, and thanked.

fMRI acquisition and data analysis

We used a 3.0 Tesla GE Signa Excite head-dedicated scanner to collect structural images (T1-weighted MPRAGE: 256 × 256 matrix; FOV = 256 mm; 116 1-mm sagittal slices) followed by functional images (EPI sequence: TR = 2000 ms; TE = 25 ms; FOV = 192 cm; flip angle = 75°; echo spacing = 0.29 ms; 39 slices; voxel size: 3 x 3 x 3 mm3</sup>). A computer presented the stimuli projected to a screen mounted at the rear of the scanner bore. Stimuli were reflected through a filter and a mirror, which participants viewed while supine.

BOLD data preprocessing

Both image preprocessing and statistical analysis used Brain Voyager QX (http://www.brainvoyager.de). Before statistical analysis, image preprocessing consisted of: 1) slice acquisition order correction; 2) 3D rigid-body motion correction; 3) voxelwise linear detrending across time; 4) temporal bandpass filtering to remove low and high frequency (scanner and physiology related) noise. We corrected distortions of EPI images with a simple affine transformation. We registered functional images to the structural images and interpolated to cubic voxels. After coregistering participants' structural images to a standard image using a 12-parameter spatial transformation, we similarly transformed their functional data, along with a standard moderate degree of spatial smoothing (Gaussian 8 mm FWHM).

BOLD data analysis strategy

Data analysis used the general linear model available on the Brain Voyager QX software package. We conducted a random-effects general linear model (GLM) analysis on blood oxygen level dependent (BOLD) signal with predictors during the video displays. We also added predictors for motion correction to the model. We convolved the predictors with a standard canonical hemodynamic response function. We transformed structural and functional data of each participant to standard Talairach stereotaxic space (Talairach & Tournoux, 1988).

We first conducted region of interest (ROI) analyzes on brain regions associated with explanation and prediction; we identified brain regions using the Neurosynth database by searching for the terms "prediction" (bi-lateral putamen in the striatum: $x = (-)14$, $y = 10$, $z = -10$; see Figure 2(a)) and "language network" (bi-lateral inferior frontal gyrus; BA 45: $x = (-)47$, $y = 24$, $z = 16$ IFG). We then drew 10 X 10 × 10 voxel cubes around the target voxel for each brain region, before extracting the average data for each of our predictors. We ran 3 *agent* X 3 *behavior* repeated measures analysis of variance (ANOVA) on each ROI. We followed up significant main effects and interactions with simple effect contrasts, Bonferroni corrected for multiple comparisons, $\alpha = 2.78E\text{-}3$. We only reported below marginal or significant differences if the confidence intervals (CI) for the simple effect contrast did not include zero.

We then performed whole brain contrasts on the data, focusing on harming versus helping behavior, nonrandom versus random motion, and deviant cell contrasts for each of the three agents (we reported these results in Table 1–5). In addition, we performed a weighted analysis such that we computed the average number of mental state words (e.g., want) and human words (e.g., friend) generated across all participants independently for each type of behavior (see

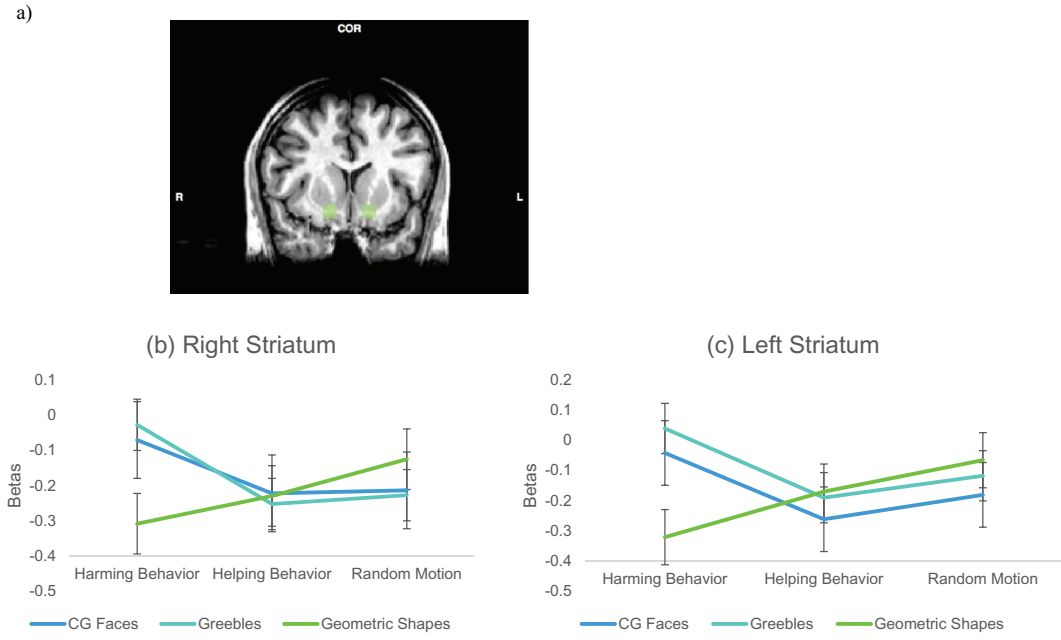

Figure 2. *Striatum ROI Brain Activity*. (a) The location of left and right striatum ROIs centered on the putamen at x = (-)14, y = 10, z = −10. (b) Extracted betas from the right putamen. Error bars show standard error of the mean. (c) Extracted betas from the left putamen.

Table 1. Nonrandom versus random movement.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Right Precentral Gyrus (BA 4)	*46, −12, 43*	29	5.21	1.23E-04
Right Parietal Lobe (BA 40)	*37, −44, 34*	96	5.16	1.29E-04

Note: Data represents the results of a whole brain contrast between non-random and random movement of agents in the videos, collapsed across the other independent variables. Voxels counts are taken at 3 x 3 x 3 mm3</sup> resolution. All statistics are uncorrected.

Table 2. Harming behavior versus helping behavior.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Right Anterior Cingulate Cortex (BA 33)	*4, 19, 20*	21	5.44	7.10E-05
Right Anterior Cingulate Cortex (BA 33)	*7, 19, 16*	14	5.36	8.00E-05
Right Anterior Cingulate Cortex (BA 24)	*3, 30, 8*	11	5.55	6.00E-05
Left Cerebellar Tonsil	*−23, −55, −36*	86	5.62	5.40E-05
Left Uvula	*−7, −63, −27*	12	5.37	7.80E-05

Note: Data represents the results of a whole brain contrast between harming and helping behaviors of agents in the videos, collapsed across the other independent variables. Voxels counts are taken at 3 x 3 x 3 mm3</sup> resolution. All statistics are uncorrected.

Table 3. CG faces versus greebles and geometric shapes deviant cell contrast analysis.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Superior Temporal Gyrus (BA 10)	*−4, 66, 23*	11	4.29	6.72E-04

Note: Data represents the results of a whole brain contrast between CG faces and the other two types of agents in the videos, collapsed across the other independent variables. Voxels counts are taken at 3 × 3 × 3 mm3</sup> resolution. All statistics are uncorrected.

Table 4. Greebles versus CG faces and geometric shapes deviant cell contrasts analysis.

Brain Region	Talairarch Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Fusiform Gyrus (BA 20)	*−48, −4, −24*	74	−4.5	4.61E-04
Left Inferior Temporal Gyrus (BA 20)	*−40, −9, −32*	28	−4.46	4.91E-04
Left Superior Temporal Gyrus (BA 38)	*−31, 10, −31*	58	−4.38	5.53E-04
Left Uncus (BA 20)	*−29, −5, −35*	42	−4.32	6.11E-04

Note: Data represents the results of a whole brain contrast between greebles and the other two types of agents in the videos, collapsed across the other independent variables. Voxels counts are taken at 3 × 3 × 3 mm3</sup> resolution. All statistics are uncorrected.

Table 6 for means) from the participants' narratives post-scanning. We then used these means as weights for each type of behavior during a whole brain GLM ANOVA of the brain data, independently for each type of agent, and collapsed across agents. Because of the low prevalence of human word use (all means below 1), we only ran this analysis for mental state words usage (we report these results in Table 7–10).

Table 5. Geometric shapes versus CG faces and greebles deviant cell contrast analysis.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Right Rectal Gyrus (BA11)	*4, 36, −20*	13	3.96	1.32E-03
Right Tonsil	*46, −36, −35*	89	4.21	9.22E-04

Note: Data represents the results of a whole brain contrast between geometric shapes and the other two types of agents in the videos, collapsed across the other independent variables. Voxels counts are taken at 3 × 3 × 3 mm3</sup> resolution. All statistics are uncorrected.

Table 6. Mean word usage.

Behavior	Mental State Word Mean	Human Word Mean
Harming Behavior	5.02	4.17E-06
Helping Behavior	3.48	2.51E-01
Random Motion	0.08	1.33E-04

Table 7. Weighted analyses for CG faces.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Superior Parietal Lobule (BA7)	*−25, −64, 52*	5868	4.39	7.82E-04
Right Precuneus (BA 7)	*21, −70, 53*	2868	4.29	8.53E-04
Left Inferior Occipital Gyrus (BA 18)	*−33, −83, −14*	13,365	4.57	6.47E-04
Right Middle Occipital Gyrus (BA 18)	*35, −79, −8*	12,940	4.46	7.49E-04
Left Precentral Gyrus (BA 6)	*−25, −14, 53*	3080	4.53	6.91E-04
Right Precentral Gyrus (BA 6)	*23, −15, 53*	908	4.29	8.72E-04
Left Inferior Parietal Lobule (BA 40)	*−57, −27, 32*	2175	4.59	6.68E-04
Left Temporal Lobe (BA 21)	*−42, −9, −9*	4328	−4.70	6.60E-04
Right Middle Temporal Gyrus (BA 39)	*48, −76, 30*	804	−4.10	1.25E-03
Right Insula (BA 13)	*43, −14, 1*	5556	−4.37	8.02E-04
Left Anterior Cingulate Cortex (BA 31)	*−1, −40, 41*	234	−3.86	1.59E-03
Left Cerebellum, Culmen	*−18, −45, −16*	821	−3.93	1.41E-03

Note: Data represents the results of a whole brain contrast using weighted means just for CG faces. Brain regions with *-ve* statistical values are inversely correlated. Voxels counts are taken at 1 x 1 x 1 mm resolution. All statistics are uncorrected.

Results and discussion

Striatum ROIs

We computed a 3 *agent* (CG face, greeble, shape) X 3 *behavior* (helping, harming, random) ANOVA on the left striatum ROI. We did not find significant *agent* or *behavior* main effects, but we did find a significant *agent* X *behavior* interaction, F (4, 60) = 4.45, p =.003, *partial η2</sup>* = 0.23, Ω = 0.92 (see Figure 2(c)).

For CG faces, we found a marginal difference between harming and helping behavior, t (15) = 2.51, p = .024, 95% CIs [0.03, 0.41], such that harming behavior engaged the brain region more than helping behavior. There were no differences between harming or helping behavior and random behavior. However, greebles showed a marginal difference between harming and random behavior, t (15) = 2.23, p = .042, 95% CIs [0.01, 0.31], with more brain activity for harming relative to random behavior. There was

Table 8. Weighted analyses for Greebles.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Fusiform Gyrus (BA 19)	*−35, −80, −13*	10,170	4.29	8.89E-04
Right Inferior Occipital Gyrus (BA 17)	*23, −92, −9*	4471	4.15	1.02E-03
Right Inferior Occipital Gyrus (BA 19)	*44, −70, −5*	5910	4.44	8.14E-04
Left Precentral Gyrus (BA 6)	*−25, −14, 54*	2729	4.95	5.89E-04
Right Middle Frontal Gyrus (BA 6)	*24, −15, 54*	1241	4.50	7.58E-04
Left Precuneus (BA 7)	*−24, −67, 51*	7070	4.42	7.84E-04
Right Precuneus (BA 7)	*19, −74, 54*	4532	4.11	1.06E-03
Left Inferior Parietal Lobule (BA 40)	*−57, −26, 34*	1962	4.33	9.02E-04
Left Superior Frontal Gyrus (BA 9)	*−31, 49, 34*	2428	−4.14	1.08E-03
Right Middle Frontal Gyrus (BA 9)	*28, 39, 36*	3108	−4.20	9.97E-04
Right Anterior Cingulate Cortex (BA 32)	*3, 16, 31*	1916	−3.90	1.50E-03
Right Anterior Cingulate Cortex (BA 24)	*3, −1, 40*	271	−3.83	1.68E-03
Left Temporal Lobe (BA 21)	*−42, −8, −10*	3965	−4.33	9.10E-04
Right Superior Temporal Gyrus (BA 38)	*36, 6, −17*	480	−4.06	1.18E-03
Right Angular Gyrus (BA 39)	*50, −64, 35*	729	−3.85	1.61E-03

Note: Data represents the results of a whole brain contrast using weighted means just for greebles. Brain regions with *-ve* statistical values are inversely correlated. Voxels counts are taken at 1 x 1 x 1 mm resolution. All statistics are uncorrected.

Table 9. Weighted analysis for geometric shapes.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Superior Parietal Lobule (BA 7)	*−23, −71, 56*	5822	3.82	2.95E-03
Right Precuneus (BA 7)	*19, −76, 54*	5106	3.57	3.60E-03
Left Precentral Gyrus (BA 6)	*−26, −15, 52*	1912	4.19	2.42E-03
Right Precentral Gyrus (BA 6)	*24, −17, 53*	571	3.39	4.71E-03
Left Inferior Occipital Gyrus (BA 19)	*−45, −71, −5*	2920	3.74	3.27E-03
Right Inferior Occipital Gyrus (BA 19)	*44, −70, −4*	5237	3.79	3.29E-03
Left Inferior Parietal Lobule (BA 40)	*−55, −29, 33*	1556	3.73	3.32E-03
Left Inferior Parietal Lobule (BA 40)	*−31, −49, 42*	533	3.66	3.65E-03
Left Middle Temporal Gyrus (BA 21)	*−65, −28, −10*	443	−3.14	6.94E-03

Note: Data represents the results of a whole brain contrast using weighted means just for geometric shapes. Brain regions with *-ve* statistical values are inversely correlated. Voxels counts are taken at 1 × 1 × 1 mm resolution. All statistics are uncorrected.

Table 10. Weighted analysis for all agents.

Brain Region	Talairach Coordinates (*x, y, z*)	Voxels	Z-value	P-value
Left Precentral Gyrus (BA 6)	*−26, −14, 53*	4054	4.31	1.90E-03
Right Middle Frontal Gyrus (BA 6)	*24, −15, 53*	1739	3.97	2.30E-03
Left Precuneus (BA 7)	*−24, −67, 51*	10,084	4.09	2.00E-03
Right Precuneus (BA 7)	*19, −74, 54*	6883	3.84	2.42E-03
Left Inferior Parietal Lobule (BA 40)	*−57, −25, 33*	3942	4.00	2.46E-03
Right Postcentral Gyrus (BA 5)	*22, −40, 68*	2534	−3.42	4.19E-03
Left Postcentral Gyrus (BA 5)	*−23, −42, 65*	1819	−3.45	4.20E-03
Right Angular Gyrus (BA 39)	*50, −69, 32*	6647	−3.67	3.15E-03
Left Superior Temporal Gyrus (BA 39)	*−62, −61, 26*	2367	−3.58	3.50E-03
Left Middle Frontal Gyrus	*−26, 37, 37*	5404	−3.34	4.75E-03

Note: Data represents the results of a whole brain contrast using weighted means for all agents. Brain regions with *-ve* statistical values are inversely correlated. Voxels counts are taken at 1 × 1 × 1 mm resolution. All statistics are uncorrected.

no such difference between helping and random behavior, but we did find marginally more engagement during harming compared to helping behavior, t (15) = 2.34, p = .034, 95% CIs [0.02, 0.44]. Lastly, shapes only showed a marginal difference between harming and random behavior, t (15) = 1.92, p = .027, 95% CIs [0.04, 0.48], with more brain activity to random rather than harming behavior. No other difference was significant for shapes.

For harming behavior, CG faces engaged the region marginally more than geometric shapes, t (15) = 2.59, p = .021, 95% CIs [0.05, 0.51], and greebles engaged more than geometric shapes, t (15) = 3.58, p = .003, 95% CIs [0.09, 0.58], but CG faces and greebles did not differ on harming behavior. None of the agents differed for helping behavior, or for random motion.

We found very similar effects in the right striatum. We did not find a significant *agent* main effect, but we did find a marginally significant *behavior* main effect, F (2, 30) = 2.75, p = .080, *partial η2</sup>* = 0.16, Ω = 0.51. This main effect was qualified by a significant *agent* X *behavior* interaction, F (4, 60) = 3.66, p = .010, *partial η2</sup>* = 0.20, Ω = 0.85 (see Figure 2(b)). For CG faces, there was a marginal difference between harming and random behavior, t (15) = 2.55, p = .022, 95% CIs [0.02, 0.26], such that harming behavior engaged the brain region more than random motion. We found no such difference between helping behavior and random motion, or between harming and helping behavior. Greebles showed a similar pattern to faces; a marginal difference between harming and random behavior, t (15) = 2.83, p = .013, 95% CIs [0.05, 0.35], with more brain activity for harming relative to random behavior. However, there was marginally more engagement during harming compared to helping behavior, t (15) = 2.99, p = .009, 95% CIs [0.06, 0.39]. Finally, geometric shapes only showed a marginal difference between harming and random behavior, t (15) = 2.43, p = .028, 95% CIs [0.02, 0.34], with more brain activity to random rather than harming behavior.

For harming behavior, greebles engaged more than geometric shapes, t (15) = 3.51, p = .003, 95% CIs [0.11, 0.45], but CG faces and greebles did not differ on harming behavior, nor did CG faces and geometric shapes differ. None of the agents differed for helping behavior, or for random motion.

Together, the pattern of results for brain regions implicated in prediction suggest valence of behavior and agent complexity mattered. Specifically, harming behaviors from more visually complex agents engaged these regions more than helping behaviors or random motion.

IFG ROIs

We computed a 3 *agent* (CG face, greeble, geometric shape) X 3 *behavior* (helping, harming, random) ANOVA on the left IFG ROI. We did not find any significant main effects or interactions. We computed a similar analysis on right IFG, and also found no significant main effects or interactions. This suggests that a brain region implicated in language generation did not differentiate valence of behavior or type of agent in our paradigm.

Study two

The first study provided partial support for the notion that harming behaviors drove brain mechanisms implicated in prediction during anthropomorphism more than helping behaviors. We did not find a similar effect for brain regions implicated in language production, which is associated with explanation. However, there are a number of limitations with the first study that affect the causal inferences that can be drawn from the data. Firstly, we relied on ROI analyses of a relatively small area of cortex, when activation patterns for both constructs typically engage more than the limited number of voxels we explored. Secondly, the search terms "prediction" and "language network" are analogues of the kinds of processes we expect to be active during anthropomorphism, not necessarily the specific processes themselves. Thirdly, the repeated measures design meant that participants saw the different types of agents performing exactly the same actions with different intentions (harming versus helping), which could have led to spillover effects from one agent to another, and from one valence to the other. Finally, the effects and sample sizes are rather small.

In the second study, we attempted to replicate the finding that the valence of behavior influenced the extent to which participants would explain and predict the anthropomorphic agent's behavior in an online study. We used only videos of geometric shapes in nonrandom motion in a conservative attempt to replicate the previous effects, ignoring agent complexity. Moreover, we explicitly assessed both explanation and prediction rather than simply relying on reverse inferences from brain activity. Specifically, participants explained the current behavior of the main protagonist shape by telling a narrative that fit the behavior, and predicted the behavior of the protagonist shape in future shape-to-shape interactions that were either relevant or irrelevant to the observed behavior. We included the relevant and irrelevant behavior during the prediction task to add precision to our prediction measure. Specifically, participants should predict from past behavior only for relevant behavior; prediction to irrelevant behavior suggest a process separate from prediction (the observed behavior does not inform the irrelevant prediction) such as generalization or some other psychological construct. Thus, the irrelevant behavior condition served as a control to allow us to better interpret the prediction results. Finally, given the central role of belonging as a motive for anthropomorphism, we manipulated the extent to which participants believed their future would be filled with social relationships or isolation to assess the impact of this motive on explanation and prediction.

Method

Participants

Ninety-two participants participated in the study, (27 males, 65 females), ranging in age from 16 to 42 (M = 21.86, SD = 3.18). The majority of participants were Dutch natives at University (n = 85) and had attained a high educational level (n = 61 graduates on VWO-level, n = 22 university bachelor or master graduates, n = 9 graduates on havo- or HBO-level). All participants who began the study completed it, thus there was no attrition or exclusion of participants. The study received ethical approval from the Psychology Department's Ethical Review Board.

Measures

We programmed the experiment using Qualtrics. To account for the participants' need for belonging prior to the experiment, we used the Need to Belong Scale; a ten item measure about the individual's urge for belonging, including items like: "I do not like being alone" and "My feelings are easily hurt when I feel that others do not accept me" (Leary et al., 2013). We found no differences on this measure and do not discuss it further.

Moreover, we used the first thirty items of the Eysenck Personality Questionnaire translated in Dutch (Sanderman et al., 1991) to manipulate loneliness. Participants either agreed or disagreed to several personality-related questions. For the manipulation check, participants rated their feelings on six emotions (sad, happy, lonely, at ease, tense, and satisfied) on a five-point scale (1 = absolutely disagree through 5 = absolutely agree).

Finally, participants had to make a prediction about the future behavior of the geometric shapes by answering eight questions immediately following each video. Four questions asked about helping behavior showed in the videos: the likelihood of the geometric shape helping the other shape to clear items away, climb a staircase, escape a closed space, and sorting elements in the future. The exact same questions addressed harming behaviors by asking about the likelihood of one geometric shape preventing another geometric shape from carrying out these four actions. Participants answered these questions on a five-point scale (1 = highly unlikely through 5 = highly likely) each time after viewing a movie.

Procedure

We recruited participants using a snow-ball technique by putting advertisements on social media platforms. We did not inform participants of the true nature of the experiment, but told them that it was about the individual perceptions of certain geometric shapes. The only selection criteria were that participants must be Dutch-speakers and 18 years of age or older. Once participants contacted the experimenter, we sent them an internet link to participate in the study. Participants were asked to read the instructions carefully and eliminate distractions. Participation took place anonymously and on a voluntary basis. As a reward, participants received €3. The experiment took approximately thirty minutes.

First, participants reported their demographic information: gender, age, nationality, religious background, and educational level. Next, we asked them to answer all questions as honestly as possible without thinking too long about their answers. They were told that they had to fill out a questionnaire about their personality before viewing videos involving geometrical shapes in motion. We used no anthropomorphic terms to describe what

would be showed to prevent participants from starting to anthropomorphize before they had seen any of the videos.

Subsequently, we manipulated loneliness using an approach in the literature (Twenge et al., 2001). Specifically, we asked participants to complete the Eysenck Personality Questionnaire. When finished, we told participants that their results were being processed and would add up to a short personality description. We then randomly assigned participants to one of our two between-subject conditions. In the experimental condition, participants read that their personality type was an indicator of becoming lonely later in life, specifically we told them: "You are the type of person who ends up lonely later in life. Even though you have friends and relationships now, they are not likely to hold in the future. Chances are that you will end up being alone more and more." In the control condition, we told participants that they would end up having lots of successful relationships and never be lonely, specifically: "You are the type of person who has rewarding relationships throughout life. The friends and relationships that you have now are likely to hold in the future. Chances are that you will always have friends and people around who care about you." The loneliness manipulation was immediately followed by the manipulation check.

We then showed participants fifteen short videos of moving geometrical shapes; 5 each displaying harming, helping, and random behaviors. The videos always involved two shapes at the time. A square was present in every single video. The other two shapes were a circle always engaging in "helping" behavior and a pentagon always portraying "harming" behavior. In the random videos, the presence of either the pentagon or the circle was varied. The videos also showed inanimate objects used by the shapes, like a staircase or triangles that were put away in a box. The helping behavior consisted of the circle helping the square escape from a closed box, helping escape from a closed circle, sorting objects, packing objects away, and climbing a staircase. The harming behavior entailed the pentagon disallowing the square to perform these actions. In the random videos, both geometric shapes moved around the screen in a random motion. All participants described as accurately as possible what they had witnessed immediately after viewing each of the fifteen videos in a blank box on the screen without a time limit. Participants then made predictions about the behavior of the geometric shapes by answering questions: "How likely is this shape to help another shape to escape a trap?; How likely is this shape to help another shape climb a staircase?; How likely is this shape to help another shape put objects away?; How likely is this shape to help another shape sort objects?" We also asked questions about harming behavior where we replaced the phrase "to help" with "to prevent" in the four questions above. Finally, participants filled out the Need to Belong Scale. We then debriefed participants and solicited information for payment.

Data analysis strategy

We computed two valenced scales for the manipulation check questions, averaging the three positive emotions (happy, at ease, satisfied) and the three negative emotions (sad, lonely, tense). We then conducted reliability analysis for both scales, revealing good reliability for the positive (Cronbach's $\alpha = 0.88$) and negative scales (Cronbach's $\alpha = 0.79$).

We defined words as anthropomorphic if they consisted of attributing personality traits or a humanlike mind to the shapes such as emotions, intentions, and conscious awareness. We did not count words as anthropomorphic if they only described the shapes in terms of being alive and able to move by themselves. We counted both primary and secondary emotions as anthropomorphic. We also included verbs related to human actions (e.g., celebrating, giving a high five, crying when hindered), mental state verbs (e.g., wish, want, hope) or the inference of a human relationship between shapes (e.g., siblings, friends, enemies, or parent and child). We counted the number of anthropomorphic inferences in the text descriptions by listing all such verbs and nouns. If participants used the same word more than once, it was counted individually each time. We divided all words into two groups: anthropomorphic and non-anthropomorphic. We computed an anthropomorphic ratio for helping, harming, and random behavior by dividing the number of anthropomorphic words for a particular behavior by the total number of words (anthropomorphic and non-anthropomorphic). We then ran an ANOVA on these ratios, and followed up significant main effects and interactions with simple effect contrasts, Bonferroni corrected for multiple comparisons with an $\alpha = 8.33\text{E-}2$. To provide further evidence for the anthropomorphism of the geometric shapes, we submitted the ratios, collapsed across the isolation conditions, to a one-sample *t*-test against zero, Bonferroni corrected for multiple comparisons with an $\alpha = 1.67\text{E-}2$.

We computed dependent variables for relevant and irrelevant prediction behavior by averaging across the appropriate questions across the videos. This resulted in averaged responses to questions about predicted valenced behavior (helping or harming) that was either relevant or irrelevant to the geometrics shape behavior depicted in the video after observing helping or harming

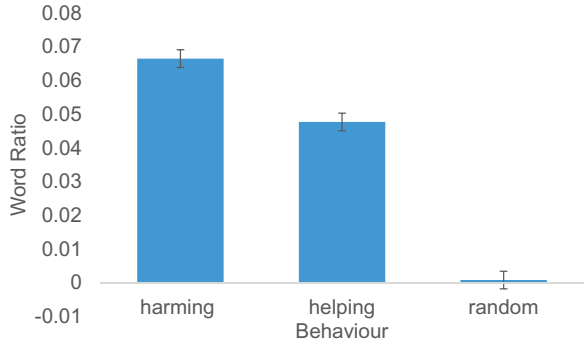

Figure 3. *Anthropomorphic Word Use During Narrations*. Means capture the amount of anthropomorphic words relative to the total number of words used when participants were relaying a narrative that explained the observed behavior of the agents. Error bars represent the standard error of the mean.

shape behaviors. We then ran an ANOVA on these prediction likelihood ratings, and followed up significant main effects and interactions with simple effect contrasts, Bonferroni corrected for multiple comparisons with an α = 2.08E-3. We excluded the random videos from this analysis since there was no relevant behavior.

Results and discussion

Manipulation check

The loneliness manipulation significantly influenced responses on the positive affect (Levene's test for Equality of Variances significant, F = 21.92, p = 1.00E-05, therefore, corrected statistics) t (70.14) = –4.17, p = 8.60E-05, 95% CI [–1.18, –0.42]; participants in the control condition reported significantly higher positive emotions (M = 4.06, SD = 0.67) than participants in the lonely condition (M = 3.26, SD = 1.09). However, the manipulation only marginally influenced negative affect, t (90) = 1.78, p = .078, 95% CI [–0.05, 0.84]; participants in the lonely condition (M = 2.58, SD = 1.15) reported feeling slightly more negative emotions than participants in the control condition (M = 2.18, SD = 0.98). Therefore, we concluded that the manipulation was partially effective.

Explanations

We ran a 2 *isolation* (lonely, not lonely) X 3 *word type* (helping ratio, harming ratio, random ratio) mixed ANOVA on word usage to determine differences in the amount of anthropomorphism in the explanation of the behavior of the shapes. We found a significant main effect of word type, F (2, 180) = 142.70, p = 7.43E-38, *partial η2</ sup>* = 0.61, Ω = 1.00 (see Figure 3). We found a significant difference between helping ratio and harming ratio, t (91) = 4.33, p = 3.10E-5, 95% CIs [0.01, 0.03], such that the harming ratio was greater than the helping ratio (M_{diff} = 0.02, SD_{diff} = 0.04). This suggests participants anthropomorphized harming behavior more than helping behavior. We also found significant differences between both the harming ratio and the random ratio, and the helping ratio and the random ratio, respectively t (91) = 14.89, p = 3.99E-26, 95% CIs [0.06, 0.07], such that the harming ratio was greater than the random ratio, (M_{diff} = 0.07, SD_{diff} = 0.04), and t (91) = 15.04, p = 2.04E-26, 95% CIs [0.04, 0.05], such that the helping ratio was greater than the random ratio, (M_{diff} = 0.05, SD_{diff} = 0.03). Together, these results suggest that participants did anthropomorphize harming and helping behavior more than random movement, but anthropomorphized harming behavior the most, consistent with study one and our hypotheses. Moreover, the isolation main effect was not significant, F (1, 90) = 0.01, p = .939, and it did not interact with the word type main effect, F (2, 180) = 0.15, p = .853. This suggests that feeling isolated or not did not affect the extent to which participants anthropomorphized the movement of the geometric shapes.

For tests against zero, we found that only the harming ratio, t (91) = 15.00, p = 2.87E-26, 95% CIs [0.06, 0.08], and helping ratio, t (91) = 15.78, p = 8.72E-28, 95% CIs [0.04, 0.05], were significantly different from zero, while the random ratio was not, t (91) = 1.92, p = .059, 95% CIs [–3.0E-5, 1.6E-3]. This suggests that participants only anthropomorphized during helping and harming, but not random movement of the geometric shapes.

Together, these results support the notion that harming behavior generates more anthropomorphism than helping behavior, though both types of behavior generated anthropomorphism. However, since word generation is also a measure of explanation, we can also conclude that harming behavior increased the motive to explain the behavior.

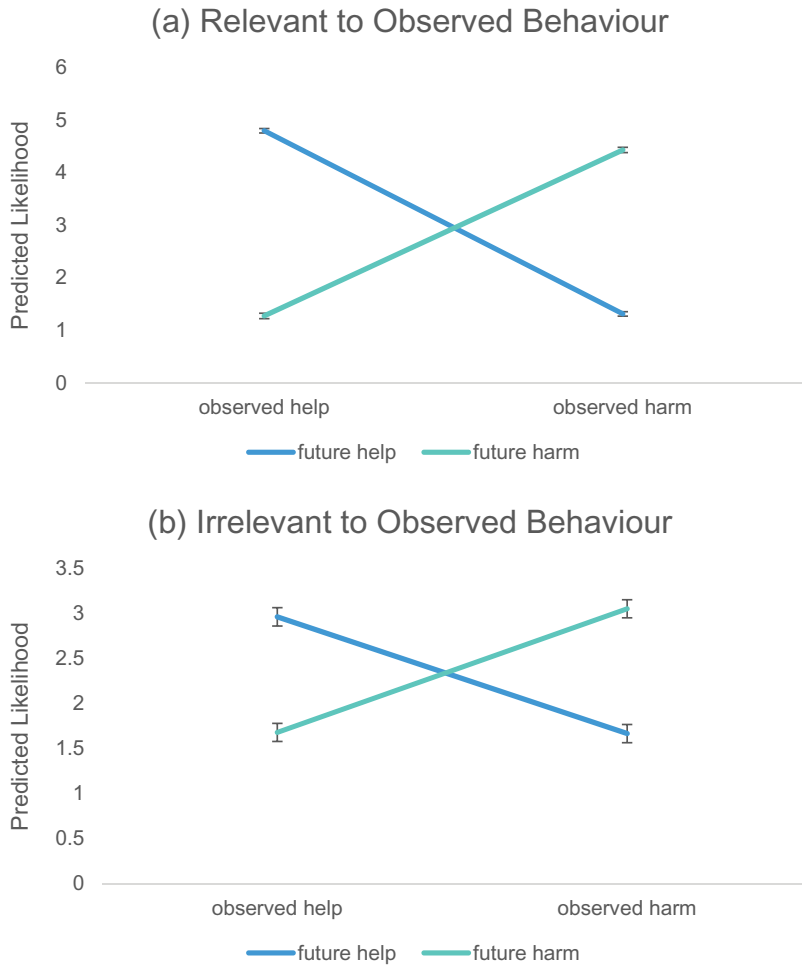

Figure 4. *Predicted Likelihood of Agent Future Behavior*. Means capture the likelihood of an agent engaging in similar behavior in the future for (a) relevant and (b) irrelevant behaviors. Error bars represent the standard error of the mean.

Predictions

We ran a 2 *predicted behavior* (helping, harming) X 2 *relevance of the question to the depicted shape behavior* (relevant, irrelevant) X 2 *observed shape behavior* (helping harming) X 2 *isolation* (lonely, not lonely) mixed ANOVA on behavioral predictions. We found a significant main effect of *observed shape behavior, F* (1, 90) = 9.49, *p* = .003, *partial η2</sup>* = 0.10, *Ω* = 0.86, such that predictions based on observing helping behaviors (*M* = 2.68, *SD* = 0.71) were rated as more likely than predictions based on observing harming behaviors (*M* = 2.62, *SD* = 0.60). We found a main effect of *relevance, F* (1, 90) = 75.16, *p* = 1.67E-13, *partial η2</sup>* = 0.46, *Ω* = 1.00, such that responses to questions relevant to the depicted behavior (*M* = 3.00, *SD* = 0.72) were higher than responses to questions irrelevant to the predicted behavior (*M* = 2.34, *SD* = 0.97). We found a main effect of *predicted behavior, F* (1, 90) = 7.52, *p* = .007, *partial η2</sup>* = 0.08, *Ω* = 0.77, such that future helping behavior (*M* = 2.69, *SD* = 0.69) was predicted as more likely than future harming behavior (*M* = 2.61, *SD* = 0.72). These main effects were qualified by significant two-way interactions, specifically *observed shape behavior* X *relevance, F* (1, 90) = 39.53, *p* = 1.13E-8, *partial η2</sup>* = 0.31, *Ω* = 1.00, *observed shape behavior* X *predicted behavior, F* (1, 90) = 757.39, *p* = 1.33E-45, *partial η2</sup>* = 0.89, *Ω* = 1.00, and *relevance* X *predicted behavior, F* (1, 90) = 46.48, *p* = 1.03E-9, *partial η2</sup>* = 0.34, *Ω* = 1.00. All two-way interactions and main effects were qualified by a significant three-way interaction between *predicted behavior, relevance*, and *observed shape behavior, F* (1, 90) = 265.17, *p* = 1.43E-28, *partial η2</sup>* = 0.74, *Ω* = 1.00 (see Figure 4).

To unpack the interaction, we first consider simple effects separately for relevant and irrelevant behavior. For relevant behavior, observed helping behavior led to significantly more predicted helping rather than predicted harming behavior, *t* (91) = 45.31, *p* = 3.12E-64, 95% CIs [3.37, 3.68], and observed harming behavior led

to significantly more predicted harming rather than predicted helping behavior, t (91) = 37.90, p = 1.58E-57, 95% CIs [2.96, 3.29]. We found a similar pattern for the irrelevant behavior, such that observing helping behavior led to significantly more predicted helping rather than predicted harming behavior, t (91) = 9.36, p = 5.47E-15, 95% CIs [1.01, 1.56], and observing harming behavior led to significantly more predicted harming rather than predicted helping behavior, t (91) = 11.05, p = 1.68E-18, 95% CIs [1.14, 1.64].

A similar pattern emerged when we consider simple effects based on the predicted behavior. When predicting future helping behavior, observing relevant helping rather than harming behavior led to increased predictions, t (91) = 48.17, p = 1.48E-66, 95% CIs [3.34, 3.63]. A similar effect emerged for observing irrelevant helping rather than harming behavior, t (91) = 9.85, p = 5.23E-16, 95% CIs [1.04, 1.56]. Similarly, when predicting future harming behavior, observing harming rather than helping behavior led to increased predictions, t (91) = 37.43, p = 4.60E-57, 95% CIs [2.99, 3.33], as did predicting irrelevant harming rather than helping behavior, t (91) = 10.86, p = 4.11E-18, 95% CIs [1.12, 1.62].

However, the pattern was different when we considered simple effects separately for observing helping and harming behavior. When participants observed helping behavior, predicting relevant helping behaviors was *more* likely than predicting irrelevant helping behaviors, t (91) = 13.39, p = 3.20E-23, 95% CIs [1.56, 2.11]. However, predicting relevant harming behaviors was *less* likely than predicting irrelevant harming behaviors, t (91) = −6.40, p = 6.64E-09, 95% CIs [−0.53, −0.28]. We again saw a similar pattern when participants observed harming behaviors, such that predicting relevant harming behaviors was *more* likely than predicting irrelevant harming behaviors, t (91) = 12.84, p = 3.93E-22, 95% CIs [1.17,

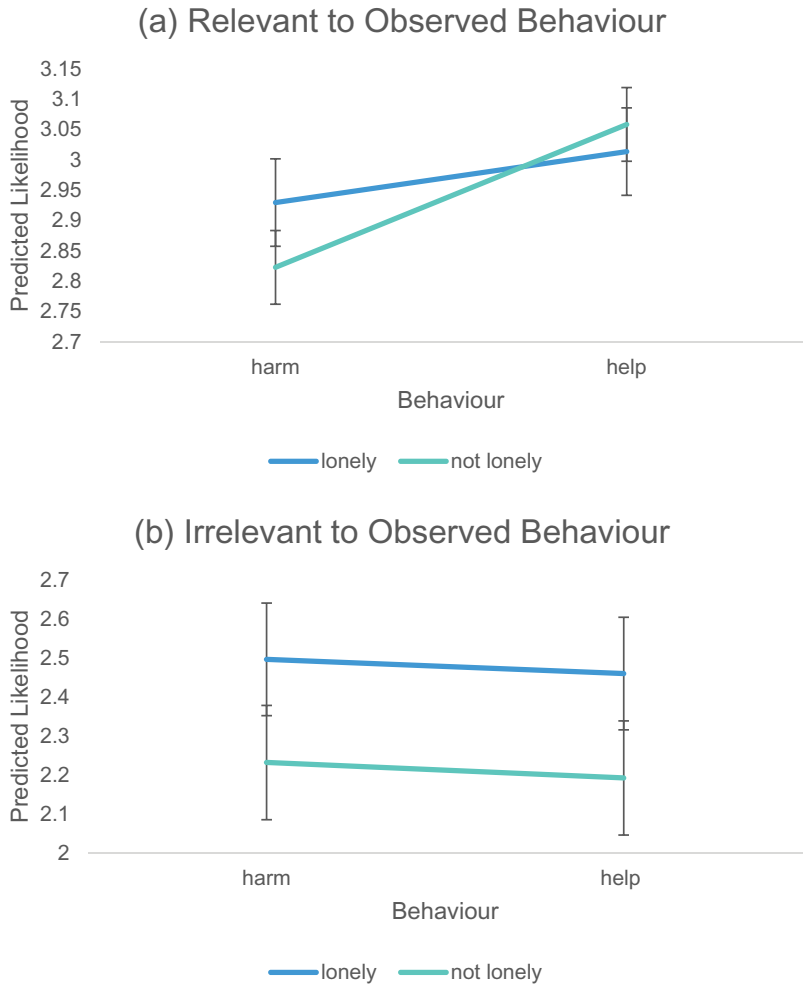

Figure 5. *Predicted Likelihood of Agent Future Behavior*. Means capture the likelihood of an agent engaging in similar behavior in the future for participants in the (a) lonely and (b) not lonely conditions. Error bars represent the standard error of the mean.

1.60]. However, predicting relevant helping behaviors was *less* likely than predicting irrelevant helping behaviors, t (91) = −6.54, p = 3.54E-09, 95% CIs [−0.46, −0.25].

The *isolation* main effect was not significant, F (1, 90) = 2.47, p = .120. We did however find a marginally significant *isolation* X *observed shape behavior* interaction, F (1, 90) = 3.51, p = .064, *partial η2</sup>* = 0.04, Ω = 0.46, and a marginally significant *isolation* X *relevance* interaction, F (1, 90) = 2.78, p = .099, *partial η2</sup>* = 0.03, Ω = 0.38. These were qualified by a significant *isolation* X *observed shape behavior* X *relevance* three-way interaction, F (1, 90) = 6.04, p = .016, *partial η2</sup>* = 0.06, Ω = 0.68 (see Figure 5). No other interactions were significant.

To unpack this interaction, we consider simple effect contrasts separately for the lonely and not lonely conditions. For the lonely condition, we found that observing relevant helping rather than relevant harming behavior *did not* lead to increased predictions, t (43) = 2.74, p = .009, 95% CIs [0.02, 0.15]. A similar lack of a significant effect emerged for observing irrelevant helping rather than irrelevant harming behavior, t (43) = −1.49, p = .145, 95% CIs [−0.09, 0.01]. However, we found a different pattern for these contrasts in the not lonely condition, such that observing relevant helping rather than harming behavior *did* lead to increased predictions, t (47) = 5.07, p = 7.00E-6, 95% CIs [0.14, 0.33], but observing irrelevant helping rather than harming behavior *did not* lead to a significant difference in predictions, t (47) = −1.13, p = .263, 95% CIs [−0.11, 0.03].

When examining simple effects within observed shape behavior, for the lonely condition, we found that observing relevant rather than irrelevant helping behavior led to increased predictions, t (43) = 5.29, p = 4.00E-6, 95% CIs [0.34, 0.77], as did observing relevant rather than irrelevant harming behavior, t (43) = 4.92, p = 1.30E-5, 95% CIs [0.26, 0.61]. A similar pattern emerged in the not lonely condition, such that observing relevant rather than irrelevant helping behavior led to increased predictions, t (47) = 7.34, p = 2.54E-9, 95% CIs [0.63, 1.10], as did observing relevant rather than irrelevant harming behavior, t (47) = 6.39, p = 6.96E-8, 95% CIs [0.40, 0.78]. We found no further simple effect contrast differences.

Unlike the results for the explanation dependent variable, the prediction dependent variable shows more prediction of the helping behavior rather than the harming behavior. Specifically, when participants generated explanations, satisfying the motive to understand, the negativity bias remained, but they considered future helping behavior as more likely when predicting the future. Perhaps in the case of explanation, negative behavior perhaps loomed larger, while in the case of predicting, positive behavior seemed more reliable. Further research is necessary to further parse such effects.

General discussion

Across two studies, we demonstrate a difference in the valence of behavior, and between prediction and explanation, during anthropomorphism. Specifically, we find that harming behaviors led to more explanation, while both harming and helping behavior led to more predictions about the future behavior of the anthropomorphized agents. Moreover, we found evidence that brain regions engaged in prediction were sensitive to differences in the visual complexity of the anthropomorphized agents, engaging more to visually complex objects. In addition, the findings of the whole brain contrasts are consistent with the literature, such that greebles and CG faces engage parts of the temporal lobe. Together, these findings further the literature on the motives of anthropomorphism, finding support for the role of effectance, understanding, and belonging motives while also highlighting the impact of visual complexity of the agent and valence of behavior.

Limitations

However, there are a number of limitations with the current pair of studies. Firstly, our measures of explanation double as our measure of anthropomorphism. Therefore, it is not possible to separate this motives for anthropomorphism from actual anthropomorphism. We decided not to explicitly ask participants to make ratings of the extent to which they perceived a mind in the agents in an attempt not to bias them into anthropomorphizing the observed motion. But the question remains whether dissociating the explanation motive from independently measured anthropomorphism would have led to different results.

Secondly, our sample size for the brain imaging study is small, and our ROIs not ideal. Regarding the ROIs, we randomly selected voxels from big swaths of the brain depicted as active for each of our key terms in the Neurosynth database. This approach allowed us unbiased ROIs, but the possibility exists that other ROIs centered around other locations in the database may have shown a different pattern. We cannot rule out such limitations, and do not test additional ROIs because of the concern surrounding multiple comparisons. However, future studies could independently define the ROIs using localizer tasks with a larger sample, reducing this limitation.

Next, our selection of brain regions, though guided by Neurosynth, still relied on reverse inferences. For instance, though studies of silent reading do engage the IFG (Assadollahi et al., 2009; Berl et al., 2005; Chu et al., 2013; Joubert et al., 2004; Stasenko et al., 2020), our task involved the generation of silent narratives, not reading. Similarly, the striatum is sensitive to prediction error, but predictive processes involve other brain regions beyond the striatum, including the orbito-frontal cortex (OFC; Tanaka et al., 2006). In addition, it seems plausible that the present method and approach might simply not have been sensitive enough to detect this seemingly extremely subtle semantic difference via BOLD activation in the IFG.

Finally, our visual complexity manipulation can also be interpreted as a "humanness" manipulation, such that the more visual complex shapes appeared more human. This interpretation does not invalidate our conclusions, and future research can further tease visual complexity from perceived humanness.

Implications

Our results suggest that the valence of behavior matters for anthropomorphism. The persistence of a negativity bias during explanation suggests that as human beings, we prioritize understanding negative behavior. However, given the previous findings in the literature demonstrating that negative behavior leads to less anthropomorphism for more complex entities (cadavers, robots, avatars, and corporations), this priority may not extend to all agents. Further research is necessary to better understand how the visual and conceptual complexity of the agent, and its degree of humanness, influences mental state attribution for negative behavior.

Moreover, our results suggest that all motives for anthropomorphism may not be equal. The motive for understanding may be privileged over belonging and effectance motives during anthropomorphism given that belonging did not affect explanations for the different valenced behaviors, but did interact with effectance motives indicated by the prediction results. This makes evolutionary sense, and suggests that anthropomorphism can be encouraged when interacting with artificial intelligence or animations by having the agents display negative behaviors.

The valence difference, however, was not as consistent for predictions, and showed a bias toward positive behaviors as visual complexity increased, consistent with the finding for more complex entities (Khamitov et al., 2016), but switched to a bias for negative behaviors for less complex agents when people were not socially isolated. This interaction result suggests that valence may be more context specific for the effectance and isolation motives. Moreover, our data suggests that both motives interact; a novel finding in the anthropomorphism literature.

Future directions

These results suggest that effectance and understanding are relevant for anthropomorphism. It may not be the case that people consciously consider these social motives when anthropomorphizing. More likely, these motives are activated by the agent, and anthropomorphism addresses these motives. An interesting future question surrounds whether self-enhancement and trusting motives may also be relevant for anthropomorphism. We argued in the introduction that such motives depend on a human target, and should be irrelevant for anthropomorphism when the perceiver maintains the belief that the agent is not human. However, the preponderance of humanized animals such as pets, artificial intelligence, and robots suggests that human beings may begin to consider these other two motives during anthropomorphism. Future studies can more directly tests these motives using a more precise experimental manipulation akin to Heider and Simmel (1944) animations, particularly if the brain will be explored. Presumably the context will again matter, and it should be possible to produce effects with these motives under circumstances where the distinction between human and not is blurred.

Another possible future direction surrounds dissociating anthropomorphism to the moral agent and the moral patient in valenced behavior situations. Our current analyses did not differentiate anthropomorphism to the two primary agents in our videos, but it is possible that the amount of anthropomorphism to each differed. Such a notion is consistent with the literature on anthropomorphism and moral behavior (Swiderska & Küster, 2018; Tanibe et al., 2017; Ward et al., 2013). Future research can use the word ratio measure that we employ to dissociate anthropomorphism to the moral agent and patient.

In closing, our studies build on the work of Cacioppo and colleagues, demonstrating nuances between social motives and valenced behavior when anthropomorphism nonrandom motion of geometric shapes. We used brain imaging, guided by social psychological theory, to provide converging evidence for behavioral data, consistent with the approach to social neuroscience research advocated by Cacioppo and colleagues, resulting in this additional contribution to the literature. Thus, the legacy of Cacioppo's work lives on in social neuroscience and has implications as anthropomorphism of non-human entities becomes more common-place

in modern, technologically driven societies.

Acknowledgments

We would like to thank Beatrice Capestany and Alyssa Fowers for building the animations, and Beatrice Capestany for collecting and assisting with analyzing the fMRI data. This work was funded by start-up funds to the first author by Duke University, and an award to the second author by Leiden University. The second study comprised the second author's MSc thesis.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Duke University; Universiteit Leiden.

ORCID

Lasana T. Harris http://orcid.org/0000-0002-8503-3201

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268. https://doi.org/10.1038/nrn1884

Andrews, K. (2009). Understanding norms without a theory of mind. *Inquiry*, *52*(5), 433–448. https://doi.org/10.1080/00201740903302584

Andrews, K. (2012). Can Apes Read Minds?. MIT Press: Cambridge MA

Andrews, K. (2012). *Do apes read minds? Toward a new folk psychology*. MIT Press.

Assadollahi, R., Meinzer, M., Flaisch, T., Obleser, J., & Rockstroh, B. (2009). The representation of the verb's argument structure as disclosed by fMRI. *BMC Neuroscience*, *10*(1), 3. https://doi.org/10.1186/1471-2202-10-3

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175-1184. doi:10.1037/0003-066X.44.9.1175

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Baumeister, R. F., & Newman, L. S. (1994). How stories make sense of personal experiences: Motives that shape autobiographical narratives. *Personality and Social Psychology Bulletin*, *20*(6), 676–690. https://doi.org/10.1177/0146167294206006

Berl, M. M., Balsamo, L. M., Xu, B., Moore, E. N., Weinstein, S. L., Conry, J. A., Pearl, P. L., Sachs, B. C., Grandin, C. B., Frattali, C., Ritter, F. J., Sato, S., Theodore, W. H., & Gaillard, W. D. (2005). Seizure focus affects regional language networks assessed by fMRI. *Neurology*, *65*(10), 1604–1611. https://doi.org/10.1212/01.wnl.0000184502.06647.28

Bowlby, J. (1969). *Attachment and loss. Vol. 1: Attachment*. Basic Books.

Burger, J. M., & Cooper, H. M. (1979). The desirability of control. *Motivation and emotion, 3*(4), 381-393. doi:10.1007/BF00994052

Cacioppo, J. T., Berntson, G. G., Lorig, T. S., Norris, C. J., Rickett, E., & Nusbaum, H. (2003). Just because you're imaging the brain doesn't mean you can stop using your head: A primer and set of first principles. *Journal of Personality and Social Psychology*, *85*(4), 650. https://doi.org/10.1037/0022-3514.85.4.650

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual Review of Psychology*, *50*(1), 191–214. https://doi.org/10.1146/annurev.psych.50.1.191

Carretié, L., Mercado, F., Tapia, M., & Hinojosa, J. A. (2001). Emotion, attention, and the 'negativity bias', studied through event-related potentials. *International Journal of Psychophysiology*, *41*(1), 75–85. https://doi.org/10.1016/S0167-8760(00)00195-1

Chu, Y. H., Lin, F. H., Chou, Y. J., Tsai, K. W. K., Kuo, W. J., & Jääskeläinen, I. P. (2013). Effective cerebral connectivity during silent speech reading revealed by functional magnetic resonance imaging. *PLoS One*, *8*(11), e80265. https://doi.org/10.1371/journal.pone.0080265

Demoulin, S., Leyens, J. P., Paladino, M. P., Rodriguez-Torres, R., Rodriguez-Perez, A., & Dovidio, J. (2004). Dimensions of "uniquely" and "non-uniquely" human emotions. *Cognition and Emotion*, *18*(1), 71–96. https://doi.org/10.1080/02699930244000444

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008a). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. Psychological Science, 19(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008b). When We Need A Human: Motivational Determinants of Anthropomorphism. Social Cognition, 26(2),143–155. doi:doi.https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 121-122). IEEE.

Fiske, S. T. (2003). Five core social motives, plus or minus five. In S. Spencer, S. Fein, M. Zanna, & J. Olson (Eds.), *Motivated social perception: The Ontario symposium* (pp. 233-246). Lawrence Erlbaum.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). McGraw-Hill.

Fiske, S. T., & Taylor, S. E. (2010). *Social cognition: From brains to culture*. McGraw-Hill.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current directions in psychological science, 10*(5), 151-155. doi:10.1111/1467-8721.00137

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7*(2), 77-83. doi:10.1016/S1364-6613(02)00025-6

Gauthier, I., Behrmann, M., & Tarr, M. J. (2004). Are Greebles like faces? Using the neuropsychological exception to test the rule. *Neuropsychologia*, *42*(14), 1961–1970. https://doi.org/10.1016/j.neuropsychologia.2004.04.025

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6

Harmon-Jones, E., & Devine, P. G. (2003). Introduction to the special section on social neuroscience: Promise and caveats. *Journal of Personality and Social Psychology*, *85*(4), 589. https://doi.org/10.1037/0022-3514.85.4.589

Harris, L., Lee, V. K., Thompson, E. H., & Kranton, R. (2016). Exploring the generalization process from past behaviour to predicting future behaviour. *Journal of Behavioural Decision Making*, *29*(4), 419–436. https://doi.org/10.1002/bdm.1889

Harris, L. T. (2017). *Invisible mind: Flexible social cognition and dehumanization*. MIT Press.

Harris, L. T., & Fiske, S. T. (2008). The brooms in Fantasia: Neural correlates of anthropomorphizing objects. *Social Cognition*, *26*(2), 210–223. https://doi.org/10.1521/soco.2008.26.2.210

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage*, *28*(4), 763–769. https://doi.org/10.1016/j.neuroimage.2005.05.021

Hawkes, K. (2014). Primate sociality to human cooperation: Why us and not them? *Human Nature*, *25*(1), 28–48. https://doi.org/10.1007/s12110-013-9184-x

Heberlein, A. S., & Adolphs, R. (2004). Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences*, *101*(19), 7487–7491. https://doi.org/10.1073/pnas.0308220101

Heider, F., & Simmel, M. (1944). An experimental study of apparent behaviour. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Hrdy, S. (2009). *Mothers and others: The evolutionary origins of mutual understanding*. Harvard University Press.

Johnson, A. H., & Barrett, J. (2003). The role of control in attributing intentional agency to inanimate objects. *Journal of Cognition and Culture, 3*(3), 208-217. doi:10.1163/156853703322336634

Joubert, S., Beauregard, M., Walter, N., Bourgouin, P., Beaudoin, G., Leroux, J. M., Karama, S., & Lecours, A. R. (2004). Neural correlates of lexical and sublexical processes in reading. *Brain and Language*, *89*(1), 9–20. https://doi.org/10.1016/S0093-934X(03)00403-6

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition*, *146*, 33–47. https://doi.org/10.1016/j.cognition.2015.09.009

Kowalski, R. M., & Leary, M. R. (1990). Strategic self-presentation and the avoidance of aversive events: Antecedents and consequences of self-enhancement and self-depreciation. *Journal of Experimental Social Psychology*, *26*(4), 322–336. https://doi.org/10.1016/0022-1031(90)90042-K

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the need to belong scale: Mapping the nomological network. *Journal of Personality Assessment*, *95*(6), 610–624. https://doi.org/10.1080/00223891.2013.819511

Mars, R. B., Neubert, F. X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. (2012). On the relationship between the "default mode network" and the "social brain". *Frontiers in Human Neuroscience*, *6*, 189. https://doi.org/10.3389/fnhum.2012.00189

Mogg, K., & Bradley, B. P. (1998). A cognitive-motivational analysis of anxiety. *Behaviour Research and Therapy*, *36*(9), 809–848. https://doi.org/10.1016/S0005-7967(98)00063-1

Mogg, K., McNamara, J., Powys, M., Rawlinson, H., Seiffer, A., & Bradley, B. P. (2000). Selective attention to threat: A test of two cognitive models of anxiety. *Cognition & Emotion*, *14*(3), 375–399. https://doi.org/10.1080/026999300378888

Moran, J. M., Jolly, E., & Mitchell, J. P. (2012). Social-cognitive deficits in normal aging. *Journal of Neuroscience*, *32*(16), 5553–5561. https://doi.org/10.1523/JNEUROSCI.5511-11.2012

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*(4), 535–545. https://doi.org/10.1037/a0016796

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist*, *56*(9), 717. https://doi.org/10.1037/0003-066X.56.9.717

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review*, *92*(4), 486. https://doi.org/10.1037/0033-295X.92.4.486

Peeters, G., & Czapinski, J. (1990). Positive- negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, *1*(1), 33–60. https://doi.org/10.1080/14792779108401856

Pervin, L. A. (1985). Personality: Current controversies, issues, and directions. *Annual Review of Psychology*, *36*(1), 83–114. https://doi.org/10.1146/annurev.ps.36.020185.000503

Phillips, A. T., Wellman, H., & Spelke, E. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, *85*(1), 53–78. https://doi.org/10.1016/S0010-0277(02)00073-2

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 435–445. https://doi.org/10.1098/rstb.2002.1221

Rothbaum, F., Weisz, J. R., & Snyder, S. S. (1982). Changing the world and changing the self: A two-process model of perceived control. *Journal of personality and social psychology, 42*(1), 5-37.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 609.

Sanderman, R., Arrindell, W. A., & Ranchor, A. V. (1991). *Eysenck personality questionnaire (EPQ)*. Noordelijk Centrum voor Gezondheidsvraagstukken.

Schultz, R. T., Grelotti, D. J., Klin, A., Kleinman, J., Van der Gaag, C., Marois, R., & Skudlarski, P. (2003). The role of the fusiform face area in social cognition: Implications for the pathobiology of autism. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1430), 415–427. https://doi.org/10.1098/rstb.2002.1208

Servos, P., Osu, R., Santi, A., & Kawato, M. (2002). The neural substrates of biological motion perception: An fMRI study. *Cerebral Cortex*, *12*(7), 772–782. https://doi.org/10.1093/cercor/12.7.772

Stasenko, A., Hays, C., Wierenga, C. E., & Gollan, T. H. (2020). Cognitive control regions are recruited in bilinguals' silent reading of mixed-language paragraphs. *Brain and Language*, *204*, 104754. https://doi.org/10.1016/j.bandl.2020.104754

Swiderska, A., & Küster, D. (2018). Avatars in pain: Visible harm enhances mind perception in humans and robots. *Perception*, *47*(12), 1139–1152. https://doi.org/10.1177/0301006618809919

Talairach, J., & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. 1988. *Theime, Stuttgart, Germany, 270* (132), 90128-5.

Tanaka, S. C., Samejima, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S., & Doya, K. (2006). Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Networks*, *19*(8), 1233–1241. https://doi.org/10.1016/j.neunet.2006.05.039

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *Plos One*, *12*(7), e0180952. https://doi.org/10.1371/journal.pone.0180952

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events. The mobilisation- minimisation hypothesis. *Psychological Bulletin*, *110*(1), 67–85. https://doi.org/10.1037/0033-2909.110.1.67

Tomasello, M., & Gonzalez-Cabrera, I. (2017). The role of ontogeny in the evolution of human cooperation. *Human Nature*, *28*(3), 274–288. https://doi.org/10.1007/s12110-017-9291-1

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of cooperation: The interdependence hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207

Trevarthen, C. (1979). Communication and co-operation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before Speech* (pp. 321–347). Cambridge University Press.

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: Effects of social exclusion on aggressive behaviour. *Journal of Personality and Social Psychology*, *81*(6), 1058. https://doi.org/10.1037/0022-3514.81.6.1058

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences*, *98*(20), 11656–11661. https://doi.org/10.1073/pnas.191374198

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

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

Waytz, A., Cacioppo, J. T., & Epley, N. (2010a). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3),219–232. doi:doi.https://doi.org/10.1177/1745691610369336

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010b). Making sense by making sentient: Effectance motivation increases anthropomorphism.. Journal of Personality and Social Psychology, 99(3),410–435. doi:doi.https://doi.org/10.1037/a0020240

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333.