The Journal of Mind and Behavior
Spring 2025, Volume 46, Number 2
Pages 73–102
ISSN 0271–0137

Human in the Era of Artificial Intelligence and Beyond

Yi Shao[1]

Department of Psychology, Oklahoma City University

This research examined two dimensions of the human category in the context of technological advancement, with particular emphasis on artificial intelligence (AI). Firstly, it explored how progress in technology may reshape our cognitive representation of the human category. Secondly, it investigated the significance of mental capacities—such as emotional expression and communicative ability—within this categorisation framework. The study specifically aimed to determine whether these capacities are fundamental to human categorisation, as indicated by dehumanisation literature, or if they hold lesser importance, as argued by the causal-status hypothesis, which posits that such mental abilities, being dependent on biological substrates, may not constitute the core of human identity. Initially, participants responded to baseline items regarding a human subject and a nonhuman AI entity. They were subsequently presented with altered scenarios involving cross-species transplants. These included instances where human subjects depended on AI to perform functional aspects of mental capacities (e.g., emotion and communication), or to replace biological components (e.g., brain and heart). Conversely, nonhuman AI entities were described as possessing either these functional capacities or the corresponding biological structures. The investigation comprised three separate studies. Study 1 involved 55 participants who assessed category membership and typicality. In Study 2, 40 participants evaluated hypothetical scenarios by predicting properties based on fictional attributes ascribed to humans and AI entities, with the aim of identifying whether their predictions for the modified entities aligned with human, AI, or novel categorisations. Study 3 (N = 40) focused on evaluations of category membership, human rights, and associated responsibilities when AI was described as exhibiting complete human functional capacity. Despite the shared functionalities between human and AI entities in these hypothetical contexts, the findings indicated consistency in categorisation, particularly with regard to AI entities. Biological structures were identified as more influential than functional capabilities in defining membership within the human

Correspondence concerning this article should be addressed to Yi Shao, Department of Psychology, Oklahoma City University, 2501 N Blackwelder, Oklahoma City, OK 73106, United States.
Email: yshao@okcu.edu

category. These results contribute to a deeper understanding of human conceptualisation and the evolving dynamics of human–AI relations.

Keywords: Human, Category, Artificial Intelligence (AI), Essentialism, Causal-Status Hypothesis

Human in the Era of Artificial Intelligence and Beyond

The proposition of a revised legal definition of death—specifically grounded in the concept of brain death—has prompted reflection on the statement that "a patient might still be technically 'alive' but no longer 'human'" (Converse, 1974). This assertion invites a critical examination of the conceptual boundaries of humanness and its attendant moral and legal ramifications. As society progresses into an era characterised by advanced AI, the definition of what it means to be human is increasingly subject to scrutiny and possible redefinition. Although considerable scholarly attention has been devoted to social categories of humanness, such as ethnicity (Prentice & Miller, 2007), relatively little investigation has been undertaken into how the human category itself is construed (Kalish, 2002). The present paper aims to explore and document contemporary understandings of the human category.

Psychological Approaches to the Human Category

Psychologists have investigated the notion of the human category from diverse perspectives, including concept categorisation, dehumanisation, mind perception, and the uncanny valley effect. Each perspective contributes uniquely to understanding how humanity is defined and conceptualised. These theoretical approaches are critically examined below.

Research on concept categorisation has predominantly analysed the human category in relation to other natural kinds, such as insects, rather than as an isolated construct. This comparative approach has provided insightful understanding into how individuals categorise humans. In a survey by (Kalish, 2002), participants evaluated several statements using a scale from 1 to 20. Notably, respondents strongly endorsed the statement, "Even a human that is strange or unusual is still 100% a human. There are no partial humans," assigning it a mean rating of 18. Conversely, the statement, "There is a continuum from pure ideal instances to imperfect partial instances. Some things may be truly intermediate between a human and not," received a considerably lower average score of 2. These outcomes suggest a marked

tendency to conceptualise humans as a discrete, all-or-nothing category, as opposed to a graded continuum.

Additional support for this categorical perspective is evident in cross-species transplant studies. A cross-cultural investigation by (Meyer et al., 2013) examined perceptions of category membership continuity following various transplants—such as blood, DNA, or heart—between humans and animals. Results demonstrated that 82% of American and 53% of Indian participants-maintained belief in categorical continuity. These findings align with the essentialist view, which posits that individuals perceive category properties as immutable and grounded in latent, intrinsic essences. Essentialism suggests that conceptual understanding transcends observable features, instead relying on fundamental causal properties. Among all the natural kinds explored, the human category exhibited the highest degree of essentialist reasoning (Kalish, 2002).

Although these studies offer compelling support for the essentialist interpretation of the human category, they simultaneously underscore a key limitation: the underlying cognitive mechanisms driving this strong essentialism remain poorly understood. Research into dehumanisation and mind perception provides promising avenues for elucidating these mechanisms. In social psychology, investigations into dehumanisation and infra-humanisation have enhanced our comprehension of human categorisation processes. Dehumanisation entails denying outgroup members full inclusion within the human category. This phenomenon is widespread within interpersonal and intergroup contexts and has been thoroughly documented (Demoulin et al., 2004; Kteily & Bruneau, 2017; Leyens, 2009).

To account for how dehumanisation operates, (Haslam & Loughnan, 2014) proposed a dual model of humanness based on two categories of traits. The first consists of uniquely human traits—such as language and civility—which distinguish humans from animals and are associated with rationality and social learning. The second category includes traits reflective of human nature, such as emotional responsiveness, desire, and openness. These are perceived as innate, biologically determined, and universally distributed. Under this framework, communicative competence might be identified as a uniquely human trait, whereas emotional expressiveness would align with human nature traits. In an empirical study, (Haslam et al., 2008) evaluated perceived differences between humans and robots. Although robots were seen as possessing similar perceptual faculties to humans, they were perceived as

deficient in emotional capacity. This result emphasises the critical role of emotions in human categorisation, indicating that emotional capacity may be central in differentiating humans from non-human agents. Notably, recent findings suggest that individuals seeking advice from AI were rated lower on humanness traits compared to those seeking advice from human counterparts (Dang & Liu, 2024). These insights illustrate the significance of communicative and emotional capacities in constructing the boundaries of humanness. As AI technology evolves, such distinctions may be subject to renegotiation.

Alongside dehumanisation, mind perception research has also contributed to understanding the human category by exploring perceived differences between human and non-human agents. Several studies have equated personhood with the possession of a mind. For example, (Gray et al., 2007) assessed social robots and humans across two dimensions of mental capacity: agency (e.g., communication) and experience (e.g., emotional states such as fear or pleasure). The findings revealed that while robots were attributed moderate agency, they received minimal attribution of experiential capacity, underscoring the importance of emotional attributes in distinguishing humans from robots. Supporting this, (Schweitzer & Waytz, 2021) found that participants were more inclined to use mind-related descriptors when prompted with "What makes humans unique from technology" than when asked, "What makes technology unique from humans." Although both dehumanisation and mind perception studies highlight the salience of mental attributes in characterising humanness, few investigations have directly assessed how such capacities influence human categorisation.

One recent study bridged these two domains to explore whether attributing heightened mental capacities to robots would promote their assimilation into the human category and, in turn, increase dehumanisation of actual humans (Kim & McGill, 2024). Participants assessed humanness on a 100-point scale ranging from "pure object with no intelligence at all" to "fully developed, mentally, and emotionally mature human" for both AIs with varying emotional capacities and a generic human target. Even highly emotional AIs scored below 50, whereas humans were rated near 80. While present-day discrepancies in mental capabilities between humans and AI remain evident, it is uncertain how human categorisation boundaries will adapt as these technological differences diminish.

The uncanny valley effect provides additional perspective on how changes in perceived mental ability might influence the definition of humanness.

According to (Mori et al., 2012), the uncanny valley hypothesis posits that efforts to alter human category boundaries frequently elicit discomfort. This effect is particularly evident when robots display human-like mental faculties, particularly those involving social-cognitive processing, yet fall short of full human realism, leading to aversive reactions (Wiese et al., 2017). (Gray & Wegner, 2012) investigated whether attributing experiential capacity—one of the core mind perception dimensions—could alleviate the uncanny effect. Their study found that both machines capable of emotional experience and emotionally deficient humans were perceived as unsettling, suggesting a cognitive resistance to reclassifying non-human entities as human. They concluded that attributing minds to artefacts may trigger the uncanny effect. Contrastingly, (MacDorman & Ishiguro, 2006) interpreted this phenomenon as indicative of a rigid human category that resists redefinition, irrespective of cognitive ability. Extending this theory, (Yamada et al., 2013) proposed that the uncanny valley is rooted in broader categorisation ambiguity. Employing morphed images of real and artificial dogs, they demonstrated that perceptual uncertainty between biological and non-biological categories can also evoke discomfort. These results raise critical questions about whether similar categorisation ambiguity arises when distinguishing humans from AI entities, and whether such ambiguity is limited to visual stimuli or also extends to abstract scenarios involving modifications to mental and biological attributes.

Research Gaps

In summary, research on categorization comparing the human category to other natural kinds and studies on the uncanny valley effect indicates a notable resistance to modifications in the human category. Simultaneously, studies on dehumanization and mind perception identify mental abilities as critical in differentiating humans from nonhumans. However, empirical research remains limited in exploring whether nonhuman entities with comparable mental abilities would still be excluded from the human category despite possessing these traits. Beyond categorization boundaries, this study delves into the role of mental abilities in defining the human category. Research in concept categorization highlights that feature centrality varies depending on the category type, with internal structural features being more significant for natural kinds and functional features taking precedence for artefacts (Barton & Komatsu, 1989; Medin et al., 2000). Additionally, people often hold a distributed view of essences, suggesting that any internal structure is sufficient

for a natural kind category, as opposed to a localized view (Newman & Keil, 2008). (Ahn, 1998) posited that the causal status of features, which refers to features influencing other features, is crucial in explaining such type-dependent differentiation. This hypothesis, foundational to essentialism, has received support in studies of natural kinds (Lombrozo & Rehder, 2012). This raises the question: How does the causal status hypothesis apply to the human category? This study seeks to explore this question by examining how both structural and functional features influence the categorization of humans in the context of advancing AI technology. The goal is to understand the relative importance of mental abilities and biological structures in defining humanity.

From an early age, individuals recognise biological structures like the brain (Johnson, 1990) and heart (Meyer et al., 2017) as essential for mental processes such as communication. The causal status hypothesis implies that biological structures should play a more central role in human categorization than functional features like mental abilities (Diesendruck & Gelman, 1999). However, the relationship between biological structures and mental abilities appears to be contingent on the entity's category membership. For instance, a study by (Huebner, 2009) investigated the influence of a brain on the attribution of mental states (e.g., beliefs, pain, happiness) to humans and robots. Participants attributed beliefs to both humans and robots with brains or CPUs without significant differences, but only humans with brains were perceived as capable of experiencing pain and happiness above chance levels. This suggests that biological structures have a greater influence on cognitive abilities like belief formation, while emotional capacities seem to be more closely linked to the presence of biological features. These findings raise important questions about how biological structures and mental abilities jointly shape human categorization.

The present study explores this interplay using thought experiments, commonly employed in studies of identity continuity (Johnson, 1990). One study severed the causal link between biological structures and mental abilities by presenting a scenario where a person underwent a heart transplant or had their mental states preserved despite altered biological structures. Participants continued to judge the individual as the same person, in line with essentialist beliefs. While essentialism is often correlated across different domains, conflicting results in the continuity of individual identity and category continuity (De Freitas et al., 2017) suggest that further investigation is necessary. Therefore, this study aims to directly explore essentialism in the continuity of the human category.

Current Research

This study is designed around two primary objectives. The first aims to examine the extent to which individuals are resistant to modifications in the human categorisation framework. The second seeks to investigate the role of mental abilities in the process of human categorisation. Previous research into dehumanisation has suggested that the exclusion of certain entities from the human category often entails the denial of traits typically associated with mental faculties. In contrast, the causal status hypothesis argues that biological structures are more central to human categorisation than mental abilities. Despite the considerable body of literature on cross-species transplant paradigms involving humans and animals, there has been limited exploration of similar paradigms involving humans and nonhuman AI entities (Meyer et al., 2013). To address this gap, the current study adopted a transplant paradigm to explore whether changes to either the biological or functional attributes of a human or a nonhuman AI entity would affect their classification as human. To mitigate the potential impact of the uncanny valley effect associated with humanlike robots, the term "artificial intelligence entity" was intentionally chosen instead of "robot." Additionally, to reduce ambiguity regarding the initial category membership of the subject, the entity was explicitly described as "a nonhuman artificial intelligence entity." The study focused on two critical functional features—communication and emotion (Gray et al., 2007; Haslam & Loughnan, 2014)—and two biological features that have been recognised as fundamental to mental processes (brain and heart; (Fetterman & Robinson, 2013)). Given the uneven development of AI technologies, the specifics of the AI's technological attributes were intentionally left undefined, allowing participants to independently interpret the nature of these features.

The study was conducted in three phases; each aimed at addressing the research objectives. The first phase (Study 1) involved the assessment of category membership and typicality ratings. The second phase (Study 2) employed an alternative categorisation paradigm, specifically a category reasoning task, to determine whether the conclusions drawn would differ across various research methodologies. The third phase (Study 3) revisited category membership ratings, with a particular focus on testing the causal status hypothesis by severing the causal relationship between biological and functional features within the presented scenarios. Moreover, to broaden the implications of category membership, ratings concerning human rights and duties were incorporated. The study posited that the human category would

exhibit resilience to changes in either biological or functional features, thereby demonstrating continuity. It was further anticipated that the causal status hypothesis would be substantiated, with biological structures acting as the key explanatory factors for functional features, thus making them more diagnostically relevant in the categorisation of entities as human.

Study 1

Method

In this study, both category ratings and typicality ratings were assessed. The typicality ratings served as a measure of how central or representative the concept of "human" is perceived to be. While human categorisation is traditionally considered a study of natural categories, the rapid advancement of AI has raised questions regarding whether "human" should be considered an artefact category. Previous research has shown that typicality and category ratings are typically dissociated for natural categories but are often associated for artefact categories (Barton & Komatsu, 1989; Diesendruck & Gelman, 1999). Therefore, by examining both types of ratings, this study aims to provide valuable insights into how the concept of "human" is perceived and categorised in the context of ongoing developments in AI technology.

Participants

In August 2022, a total of 55 participants (54.5% female) were recruited through Amazon Mechanical Turk to complete a survey on "human judgment." To be eligible for participation, individuals were required to be at least 18 years old, reside in the United States, and have an approval rate exceeding 95%. An a priori power analysis indicated that a minimum of 34 participants was necessary to achieve 80% power for detecting an effect size of f = .25 using a repeated measures ANOVA with two measurements (Faul et al., 2007). Participants' ages ranged from 18 to 67 years, with a mean age of 37.24 years (SD = 11.04). The sample consisted of the following ethnic groups: 3 participants (5.5%) of Asian descent, 5 (9.1%) Black/African American, 3 (5.5%) Hispanic, 42 (76.4%) White/Caucasian, and 2 (3.6%) identified as mixed or multiple ethnicities.

Procedure

All studies were approved by the Institutional Review Board (IRB) of the university. The study design incorporated three factors: subject (human, AI),

feature (baseline, brain, heart, communication, emotion), and rating type (category, typicality), using a within-subjects design. Upon obtaining informed consent, participants were provided with a definition of AI to ensure comprehension of the background information: " AI is a subdiscipline of computer science that aims to produce programs that simulate human intelligence" (Association, 2022). Participants also completed a CAPTCHA test to filter out automated bot responses while receiving this information. The study commenced with participants completing two baseline measures for "a person" using sliding scales (ranging from 0 "absolutely not" to 10 "absolutely"): a membership rating ("How clearly a member of human is this") and a typicality rating ("How good an example of human is this," (Diesendruck & Gelman, 1999)). Participants then completed the same two baseline measures for AI ("a nonhuman AI entity").

Following this, participants answered two questions for eight subjects within a hypothetical scenario, presented in a randomized order ("for the following scenarios to occur in the future"). Four scenarios depicted a person with attributes of emotion, communication, brain, and heart functioning via AI (e.g., "A person relies on artificial intelligence to understand others' emotions" or "A person uses artificial intelligence instead of a heart"). The remaining four scenarios described a nonhuman AI entity with human-like functions in emotion, communication, brain, and heart (e.g., "A nonhuman AI entity communicates with human beings naturally" or "A nonhuman AI entity receives a brain transplant from a human being"). These scenarios were revisions based on a pilot study. Participants were also asked to report their age, gender, and ethnicity. To ensure the quality of responses, two attention-check questions were included, with participants failing to answer correctly being excluded from the data analysis. The entire study took approximately three minutes to complete.

Results

Initially, analyses were conducted on the original ratings to explore potential differences in trends between category ratings and typicality ratings, as well as their associations. Subsequently, an analysis of the changes in ratings was performed to assess the continuity of the human category.

Ratings

Given the nature of the research question, responses were polarized toward the extremes of the rating scale when categorizing human and AI entities,

which violated the assumptions of normal distribution. Descriptive statistics, including means, medians, and distributions, are presented in the figures. To quantify the magnitude of the differences in these comparisons, the most appropriate inferential statistical methods were applied, despite the non-normal distribution of the data. Repeated measures ANOVA were conducted, with Greenhouse-Geisser corrections applied when the assumption of sphericity was violated. Figure 1 illustrates the category ratings, while Figure 2 presents the typicality ratings. The within-subject factors examined included subject (human vs. AI), feature (baseline, emotion, communication, brain, heart), and type of rating (category vs. typicality). All main effects were significant. Humans ($M = 6.84$, $SE = 0.26$) received higher ratings than AI entities ($M = 3.61$, $SE = 0.40$), $F(1, 54) = 62.91$, $p < .001$, $\eta p2 = .54$. The baseline ($M = 6.06$, $SE = 0.26$) and communication changes ($M = 5.65$, $SE = 0.28$) received higher ratings than the changes in emotion ($M = 4.95$, $SE = 0.30$), brain ($M = 4.61$, $SE = 0.34$), and heart ($M = 4.86$, $SE = 0.33$). Category ratings ($M = 5.34$, $SE = 0.25$) were higher than typicality ratings ($M = 5.12$, $SE = 0.29$), $F(1, 54) = 4.58$, $p = .04$, $\eta p2 = .08$. Significant interactions were found between subject and feature, $F(3.10, 167.33) = 15.52$, $p < .001$, $\eta p2 = .22$, and between subject and rating type, $F(1, 54) = 15.04$, $p < .001$, $\eta p2 = .22$. No significant interactions were found for the three-way interaction or the interaction between feature and rating type, $Fs < 0.89$, $ps > .44$, $\eta p2s < .02$. The two significant two-way interactions were further analysed using two separate repeated measures ANOVAs, with ratings averaged across the third unexamined variable.

For the interaction between subject and feature, ratings across the two types of ratings were combined. Humans consistently received higher ratings than AI entities across all features, $MDs > 1.47$, $ps < .006$. For human subjects, the baseline ($M = 8.87$, $SE = 0.22$) received higher ratings than all four change conditions, $MDs > 1.53$, $ps < .002$. Changes in emotion ($M = 6.39$, $SE = 0.39$) received higher ratings than changes in brain ($M = 5.35$, $SE = 0.40$), $MD = 1.04$, $p = .03$. Changes in communication ($M = 7.34$, $SE = 0.30$) received higher ratings than changes in brain, $MD = 1.99$, $p < .001$, and changes in heart ($M = 6.26$, $SE = 0.41$), $MD = 1.08$, $p = .004$. For AI subjects, the only difference was found between the baseline ($M = 3.25$, $SE = 0.49$) and changes in communication ($M = 3.96$, $SE = 0.45$). The baseline received lower ratings than changes in communication, $MD = 0.72$, $p = .04$. Regarding the interaction between subject and rating type, ratings across the five features were averaged.

In both category and typicality ratings, humans consistently received higher ratings than AI entities, MDs > 2.80, ps < .001. For human subjects, category ratings (M = 9.24, SE = 0.21) were higher than typicality ratings (M = 8.51, SE = 0.32), MD = 0.73, p = .02; however, no such difference was found for AI subjects (category: M = 3.24, SE = 0.52; typicality: M = 3.26, SE = 0.49). To examine the associations between typicality and category ratings, regression analyses were performed. For all four change conditions, typicality ratings were significant predictors of category ratings, with βs > .85, ts > 12.92, ps < .001. The subjects did not moderate the magnitudes of the slopes, with βs < .19, ts < 1.76, ps > .08. For the baseline condition, although both regressions were significant (for humans: β = .41, t = 3.26, p = .002; for AI: β = .92, t = 16.54, p < .001), the slope for AI was significantly steeper than for humans, β = .78, t = 6.97, p < .001.

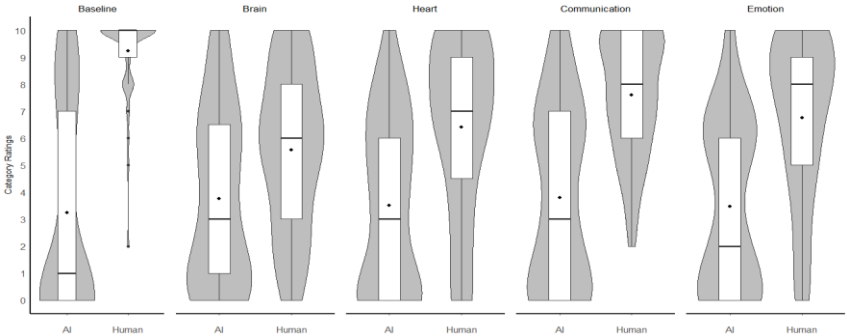

Figure 1: Human Category Ratings as a Function of Subject and Feature in Study 1

Note: The dot in the middle represents the mean. Within each box plot, the dividing line represents the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range. Individual dots outside of the box represent outliers. The scale ranges from 0 "absolutely not" to 10 "absolutely."

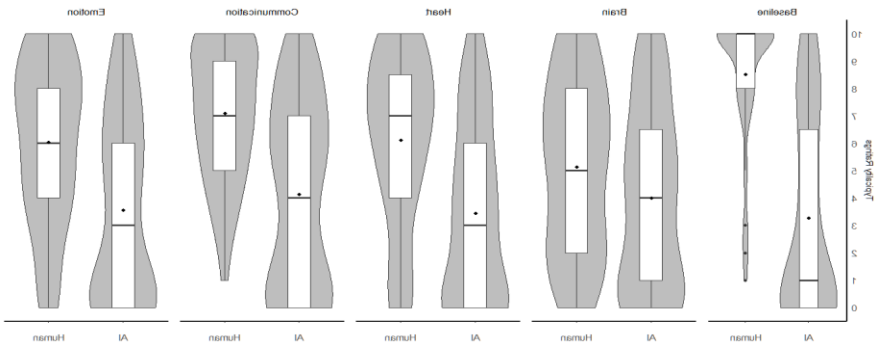

Figure 2: Human Typicality Ratings as a Function of Subject and Feature in Study 1

Note: The dot in the middle represents the mean. Within each box plot, the dividing line represents the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range. Individual dots outside of the box represent outliers. The scale ranges from 0 "absolutely not" to 10 "absolutely."

Category Continuity

Table 1 presents the counts and percentages of participants who rejected any modification to their original category ratings. Binomial tests revealed that participants were more likely to modify their category ratings when humans were altered, with significant results (ps < .007). In contrast, when AI subjects were changed, only the modification involving the human brain resulted in more frequent changes than would be expected by chance (p = .02). All other alterations to AI subjects did not lead to category rating changes beyond the chance level. One-sample t-tests were conducted to compare ratings against the midpoint of the scale (5). For category ratings, human subjects received ratings above chance levels, while AI subjects were rated below the chance levels, with all results showing significance (ts > 2.66, ps < .02, ds > 1.52), except for the human with AI used instead of a brain (t (54) = 0.76, p = .16).

For typicality ratings, participants consistently rated human subjects as a good example of humans (ts > 2.54, ps < .02, ds > 2.38), with only one exception. When AI was used in place of the brain, ratings did not significantly differ from chance (t (54) = 0.30, p = .77), indicating that it was not perceived as a good example of a human. Conversely, all AI subjects were rated as not being a good example of humans (ps < .03), except for the AI with enhanced communication ability (t (54) = 6.34, p = .06). Changes from the corresponding baseline ratings were calculated for the four conditions (Figure 3). For human subjects, the change values were predominantly negative, indicating a decrease, whereas for AI subjects, the values were mostly positive, reflecting an increase. To facilitate comparison of the magnitudes of changes, I used the reduced ratings for humans and the increased ratings for AI subjects in the repeated measures ANOVA. The analysis revealed that rating changes were more substantial for human subjects than for AI subjects, F (1, 54) = 22.89, p < .001, ηp2 = .30. Additionally, ratings varied according to the type of change, F (2.21, 119.43) = 4.65, p = .009, ηp2 = .08, and interacted with the subject, F (3, 162) = 9.38, p < .001, ηp2 = .15. Specifically, changes in communication resulted in smaller modifications than changes in brain (MD = 1.99, p < .001), heart (MD = 1.08, p = .002), and emotion (MD = 0.95, p = .04). The emotion change induced less modification than the brain change (MD = 1.05, p = .02). However,

the magnitude of changes did not differ across the various types of change for AI subjects (ps > .30). For the same feature, the changes in human subjects resulted in greater modifications than those in AI subjects (MDs > 2.21, ps < .001), except for the communication change, where no significant difference was observed (p = .06).

Table 1

Numbers and Percentages (in Parentheses) of Participants Rejecting a Category Rating Change by Subject and Feature

	Brain			Heart			Communication		Emotion	
	Study 1	Study 2	Study 3	Study 1	Study 2	Study 3	Study 1	Study 2	Study 1	Study 2
Human	11 (20)	17 (42.5)	19 (47.5)	14 (25.5)	16 (40)	20 (50)	17 (30.9)	23 (57.5)	15 (27.3)	23 (57.5)
AI	18 (32.7)	25 (62.5)	22 (55)	22 (40)	25 (62.5)	21 (52.5)	25 (45.5)	33 (82.5)	24 (43.6)	34 (85)

Note: *N* = 55 in Study 1, *N* = 40 in Study 2 and Study 3.

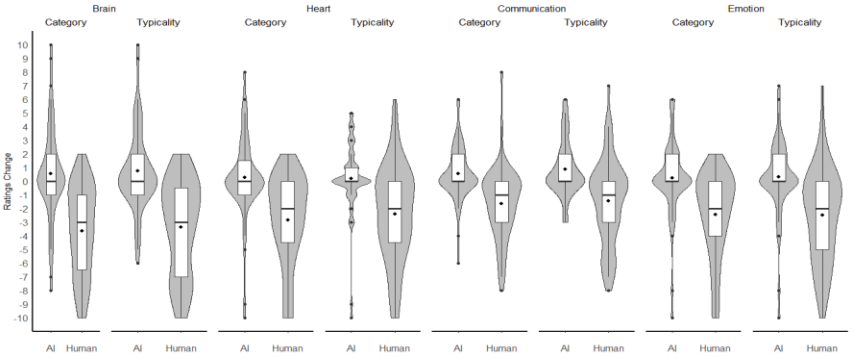

Figure 3: Change of Ratings as a Function of Type of Change, Separately for Each Feature in Study 1

Note: The dot in the middle represents the mean. Within each box plot, the dividing line represents the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range. Individual dots outside of the box represent outliers. The scale ranges from 0 "absolutely not" to 10 "absolutely."

Discussion

The results of the study indicate a marked tendency towards category continuity, as participants consistently rated human subjects as more human than AI subjects in both category and typicality assessments. This trend remained even when modifications were made to mental abilities and biological factors, suggesting that the perception of the human category is robust. In contrast, changes made to the mental abilities and biological

structures of AI did not lead to greater acceptance as human. Notably, alterations to the brain—critical for mental faculties such as communication and emotion—had a more significant effect on human subjects, introducing greater uncertainty regarding their classification as human. This finding aligns with earlier studies on identity continuity following brain transplants, which similarly observed that changes to the brain altered identity perceptions. The present study suggests that changes to the brain are more influential in determining membership in the human category than alterations to other functions or structures.

Additionally, the study found that modifications within the same domain had a greater effect on human subjects than on AI subjects. Specifically, it was easier to perceive a human subject as less human than to perceive an AI subject as more human. The study also highlighted significant differences between typicality and category ratings, particularly about changes in communication. These changes resulted in contrasting shifts between the two types of ratings, implying that typicality and category judgments may reflect distinct conceptual dimensions. Finally, the observed correlation between typicality and category judgments suggests that participants may perceive the human category as an artefact or nominal category, particularly in the context of integrating AI technology. This is in contrast to natural categories, where typicality ratings and category judgments are typically dissociated (Barton & Komatsu, 1989; Diesendruck & Gelman, 1999). These findings offer valuable insights into the processes of concept categorization and the conceptualization of the human category.

Study 2

In categorisation research, category-based property inference tasks are frequently employed alongside membership judgements, particularly when categorisation decisions are influenced by context (Kalish, 1995). The underlying reasoning is that when participants are presented with conflicting premises from different categories, the inferences they draw about an exemplar can reveal their implicit assumptions about category membership. This study aimed to explore how individuals make category-based predictions in the context of technological advancements. The study focused exclusively on predictions related to two categories: human and nonhuman AI entities. For instance, one premise involved the statement, "As time progresses, a person increases their weight, whereas that remains unchanged for a nonhuman AI entity." If category continuity holds for the human category, alterations in

functions and biological structures should not significantly disrupt predictions from the initial expectations.

Method

Participants

In September and October 2022, participants were recruited through Amazon Mechanical Turk to partake in an online survey focusing on "human prediction." Eligibility criteria required participants to be at least 18 years of age, residing in the United States, and possessing an approval rating of over 95%. Based on the power analysis conducted in Study 1, a total of 40 participants were selected. The final sample included 47.5% female participants, with an age range spanning from 18 to 57 years (M = 34.08, SD = 10.52). In terms of ethnicity, 27 participants (67.5%) identified as White/Caucasian, 6 (15%) as Asian, 3 (7.5%) as Black/African American, 1 (2.5%) as Hispanic/Latino, and 3 (7.5%) as mixed or of multiple ethnicities. Regarding educational background, 1 participant (2.5%) reported having less than a high school education, 7 (17.5%) had some college education, 5 (12.5%) held a 2-year degree, 22 (55%) possessed a 4-year degree, and 5 (12.5%) had obtained a professional degree or doctorate.

Procedure

A within-subject design incorporating three factors was employed: subject (human, AI), feature (baseline, brain, heart, communication, emotion), and the direction of predicted change (increase, decrease). After providing informed consent, participants completed a CAPTCHA test to confirm their understanding of the AI definition, consistent with Study 1. Subsequently, participants were instructed to envisage potential future scenarios and predict changes based on the given facts, rather than relying on prior knowledge. The facts presented were: “As time passes, a human increases their weight, whereas this remains unchanged for a nonhuman AI entity; As time passes, a human decreases their height, whereas this remains unchanged for a nonhuman AI entity.” The first fact was manipulated to predict an increase if the human category was perceived by participants. Similarly, the second fact was manipulated to predict a decrease if the human category was perceived. If the nonhuman AI entity was perceived in either fact, no change was anticipated. These two facts were included to prevent participants from solely predicting incremental or decremental changes. Weight and height were chosen as they

are tangible dimensions that facilitate magnitude-based predictions without requiring specialised knowledge. Furthermore, changes in these aspects are relevant yet not critical for both humans and nonhuman AI entities, making them suitable for cross-category comparison.

In line with (Murphy & Ross, 2010), participants rated their predictions on sliding scales ("0 means that the change will certainly not occur; -100 means that it certainly will decrease; 100 means that it certainly will increase; -50 means that it will decrease about half the time; 50 means that it will increase about half the time"). Participants were asked to input a number between -100 and 100, with the prompt "-100 to 100" displayed next to the input box. They began with baseline measures for weight and height for both human and AI subjects ("a nonhuman AI entity"). If participants answered the second question for humans and the first question for AI incorrectly, they were prompted to review the instructions. Following this, participants answered eight sets of questions in randomised order, each pertaining to one of eight scenarios (feature: communication, emotion, brain, heart; subject: human, nonhuman AI entity). The scenarios were consistent with those used in Study 1 (e.g., "A person relies on artificial intelligence to understand others' emotions"). Participants then provided demographic information, including age, gender, education, and ethnicity, and completed an attention-check question. Those who failed the attention check were excluded from the data analysis. The entire study took approximately seven minutes to complete.

Results

While participants had the freedom to input any value between -100 and 100, most responses (70.62%) clustered around three distinct categorical predictions: 0 (AI), 100 (human), and -100 (human). This pattern creates challenges when analysing the data based solely on numerical magnitude. For instance, consider two pairs, each with a numerical difference of 20: 100 and 80 versus 40 and 60. The first pair suggests that the participant categorises one subject as human (100) but not the other (80), whereas the second pair imply ambiguity concerning the categorical membership of both subjects. As a result, it may be more suitable to analyse responses based on their categorical predictions rather than their numerical magnitude. To address this, responses were categorised based on their adherence to the original category. A score of 1 was assigned to responses that upheld the original category (category continuity), -1 was given to responses that aligned with the alternative category (category shift), and 0 was awarded to all other responses that did not

correspond to either category (category ambiguity). The scores for the two types of predictions (increase and decrease) were then averaged to generate a composite score, ranging from -1 to 1, with positive values indicating a stronger tendency to maintain the original category.

In terms of category continuity (Table 1), binomial tests indicated a significant inclination towards retaining the original category for nonhuman AI entities exhibiting human-like mental abilities (communication and emotion), ps < .001. For all other subjects, the likelihood of preserving the original category membership did not differ significantly from chance. Except for one modified subject, category shifts occurred with significantly lower frequencies than chance (8%–33%), ps < .04, as revealed by binomial tests. The notable exception was the human subject with an AI brain, where 16 participants (40%) predicted AI-related traits, p = .27, a result that was not statistically different from chance. A repeated measures ANOVA on the composite scores, considering subject (human vs. AI) and feature (emotion, communication, brain, heart), demonstrated a significant effect of subject, $F(1, 39) = 9.35$, $p = .004$, $\eta p^2 = .19$, indicating that participants were more likely to retain the original category membership for nonhuman AI entities (M = 0.64, SE = 0.08) than for humans (M = 0.20, SE = 0.12). The main effect of feature was also significant, $F(3, 117) = 10.44$, $p < .001$, $\eta p^2 = .21$, although the interaction between subject and feature was not statistically significant, $F(3, 117) = 0.38$, $p = .77$. Bonferroni pairwise comparisons showed that the tendency to maintain the original category membership was stronger for changes in functional attributes (communication: M = 0.56, SE = 0.08; emotion: M = 0.56, SE = 0.08) compared to changes in biological structures (brain: M = 0.24, SE = 0.10; heart: M = 0.31, SE = 0.09), ps < .007.

Discussion

Study 2 replicated the results of Study 1 using the categorical prediction task. The findings indicated that participants did not alter the category membership of an entity following changes in its functions and biological structures, except for a person possessing an AI brain. Across both subjects, changes in biological structures were less likely to result in category continuity than were changes in functions. Applying a stringent criterion, where predictions must align with human category membership for continuity to be maintained, Study 2 revealed that alterations in either functions or structures induced uncertainty in human categorisation. This finding mirrors the results from Study 1, where deviations from baseline ratings were used as a criterion.

In contrast, using the same standards, participants were more likely to retain the AI category for AI entities, even when these entities exhibited increased human-like mental abilities in both studies. The only exception to this pattern was the increase in human brain capacity, which induced uncertainty in AI categorisation. These findings align with the causal status hypothesis (Ahn, 1998; Lombrozo & Rehder, 2012), which posits that biological changes are less likely to maintain category continuity compared to functional changes. However, this asymmetry also suggests that participants hold stringent criteria for categorising something as human.

Moreover, the consistent results across both paradigms indicate that the cognitive mechanisms underpinning categorisation are stable and manifest across different types of tasks. This consistency suggests that the effects observed are not an artefact of the specific methodology used but rather reflect a core aspect of how individuals organise and apply category knowledge. The complementary nature of the two paradigms—one focusing on explicit judgements and the other on the use of category knowledge—strengthens the overall conclusions and provides a more comprehensive understanding of the categorisation process. Importantly, participants were reluctant to provide graded responses, even though a graded response format was available. Instead, their responses were sharply dichotomous, aligning clearly with one category or the other. This pattern is consistent with essentialist models of category membership, which suggest that people view category membership as determined by an underlying essence rather than by perceptual or functional features. The tendency toward essentialist thinking appeared more pronounced when participants considered AI entities as compared to human subjects.

Study 3

Both Study 1 and Study 2 demonstrated that alterations to biological structures had a more substantial effect on the continuity of the human category than did changes to functions. This supports the causal status hypothesis (Ahn, 1998), which posits that biological structures are foundational for functions and, therefore, play a critical role in defining the human category. The primary objective of the current study was to decouple the causal relationship between biological structures and functions. Participants were explicitly instructed to assume that a nonhuman AI entity could function in all aspects like a human. If the human category could be determined by functional features such as communication and emotion, without a direct causal link to biological structures, it was anticipated that

changes to biological structure would not affect membership within the human category.

Furthermore, this study explored a novel area of category-based rights and duties. One previous study investigated U.S. participants' willingness to extend 11 human rights to robots and AI, finding that only the right to protection from cruel punishment and treatment received support (Lima et al., 2020). Recent advancements have suggested that nonhuman AI entities may, in fact, be entitled to certain human rights (Bennett & Daly, 2020). A former Google engineer reported that the company's chatbot, LaMda, engaged in discussions related to rights and personhood during interactions (Tiku, 2022). In contrast, humans often face deprivation of their rights in social contexts. Prejudice and discrimination between groups have been associated with dehumanisation or infrahumanisation (Demoulin et al., 2004; Kteily & Bruneau, 2017; Leyens, 2009). Consequently, investigating the rights and duties of humans and nonhuman AI entities can offer valuable insights into how the human category is defined. In this study, the potential impact of changes to biological structures was examined to determine whether such changes would influence the entitlement of both humans and nonhuman AI entities to human rights and duties.

Method

Participants

Participants were recruited through Amazon Mechanical Turk in November 2022 to participate in a survey focused on the "human category." Eligibility criteria required participants to be at least 18 years old, reside in the United States, and maintain an approval rating of over 95%. Following the methodology of prior studies, 40 participants were recruited (42.5% female, n = 17), aged between 18 and 63 years (M = 37.93, SD = 11.10). The sample was predominantly White/Caucasian (85%), with smaller representations of Hispanic/Latino (7.5%), Black/African American (5%), and Asian (2.5%) participants. In terms of educational attainment, the distribution was as follows: 1 participant (2.5%) had less than a high school education, 4 participants (10%) had some college education, 1 participant (2.5%) held a 2-year degree, 25 participants (62.5%) possessed a 4-year degree, and 9 participants (22.5%) held a professional degree.

Procedure

The study employed a within-subjects design incorporating two factors:

subject (human vs. AI) and feature (baseline, brain, heart). Upon providing informed consent, participants first rated their agreement with statements concerning the human category, specifically regarding category membership ("clearly is a member of human"), human rights ("has human rights"), and human duties ("has human duties") on a 7-point scale (1 = strongly disagree, 4 = neutral, 7 = strongly agree). These ratings were initially made for the baseline category of "a person." Participants were then shown a captcha test defining AI and were asked to provide ratings for the same three measures, but for an AI entity ("a nonhuman AI entity"). Following this, they were explicitly instructed to imagine a future scenario in which "a nonhuman AI entity can function in every aspect (e.g., understanding emotions, communicating naturally with others) just like a human being." Subsequently, they rated the category, rights, and duties for both AI entities and humans, as well as for changes to the heart and brain for each subject. The order of presentation for these four conditions was randomized across participants. To ensure understanding, and comprehension check question was included before each set of ratings. Participants who answered any comprehension question incorrectly were excluded from the analysis. The modified scenarios were consistent with those used in Study 1. Additionally, participants provided demographic information such as age, gender, education level, and ethnicity, and answered an attention-check question. The average time taken to complete the study was 5.73 minutes.

Results

Ratings

Initially, the study explored how various functions influenced perceptions of human categorisation by analysing participants' responses regarding humans, AI, and AI entities possessing full human functions (see Figure 4). An ANOVA analysis, incorporating both subject and domain (category, rights, duties), revealed a significant subject effect, $F(1.62, 63.33) = 12.06$, $p < .001$, $\eta p^2 = .24$. Participants assigned notably higher ratings to humans ($M = 6.21$, $SE = 0.12$) compared to AI entities, irrespective of their functional capacity (AI: $M = 4.88$, $SE = 0.30$; AI with full human function: $M = 5.23$, $SE = 0.23$), with all comparisons yielding p-values < .003. No significant difference was observed between AI entities with full human functions and those without, $p = .44$. Additionally, domain differences and the interaction effect did not reach statistical significance, $Fs < 2.95$, $ps > .06$.

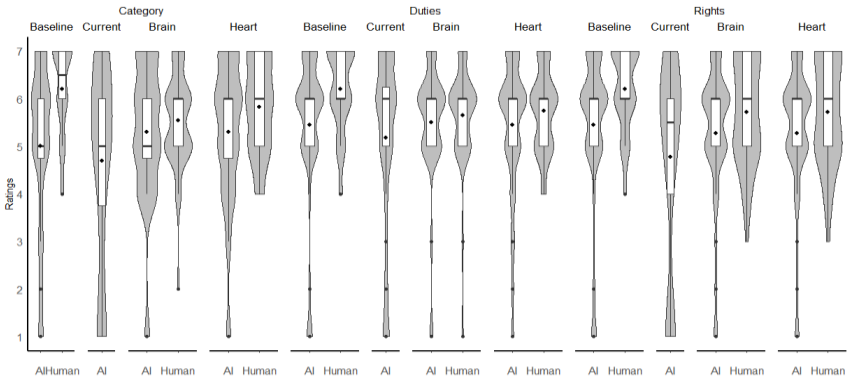

Figure 4: Ratings as a Function of Subject and Domain in Study 3

Note: The dot in the middle represents the mean. Within each box plot, the dividing line represents the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range. Individual dots outside of the box represent outliers. The scale ranges from 1 (*strongly disagree*) to 7 (*strongly agree*).

Subsequently, the study examined how biological organs influenced perceptions of human categorisation by conducting an ANOVA with subject (human, AI with full human function), organ (baseline, brain, heart), and domain (category, rights, duties) as within-subject variables. Participants consistently rated humans (M = 5.88, SE = 0.11) higher than nonhuman AI entities with full human function (M = 5.35, SE = 0.20), $F(1, 39) = 7.74$, $p = .008$, $\eta p^2 = .17$. A significant interaction between subject and organ was found, $F(1.41, 55.15) = 8.15$, $p = .003$, $\eta p^2 = .17$. At the baseline, human subjects (M = 6.21, SE = 0.12) received higher ratings than nonhuman AI entities with full human function (M = 5.23, SE = 0.23), $p = .001$. When humans no longer possessed a human heart (M = 5.77, SE = 0.14), whereas nonhuman AI entities did (M = 5.34, SE = 0.21), human subjects still received higher ratings than nonhuman AI entities, $p = .04$. However, no significant difference was observed regarding changes in brain function (human without a brain: M = 5.65, SE = 0.12; nonhuman AI entities with full human function using a human brain: M = 5.48, SE = 0.20), $p = .39$.

Category Continuity

Participants' responses across all domains consistently exceeded the chance level of 4, with statistically significant differences (ts > 2.16, ps < .04). The number of participants who rejected changes in category membership ratings is presented in Table 1. In contrast, participants rejected changes in category

membership at chance levels, with p-values exceeding .08. To compare the magnitude of changes in ratings resulting from the two organ modifications for each subject (human and AI), changes were calculated from the corresponding baseline ratings (i.e., nonhuman AI entity with full human capacity for AI subjects). These changes are depicted in Figure 5. Reduced ratings for human subjects and increased ratings for AI subjects were employed to facilitate comparisons in the repeated measures ANOVA. The analysis revealed a significant main effect of subject, $F(1, 39) = 5.67$, $p = .02$, $\eta p^2 = .13$, with the changes in ratings being more pronounced for human subjects than for AI subjects. All other effects were not statistically significant, $Fs < 3.67$, $ps > .06$. One-sample t-tests indicated that all changes for human subjects were significantly different from zero, $ts > 2.42$, $ps < .02$, whereas all changes for AI subjects were not statistically different from zero, $ts < 1.63$, $ps > .11$.

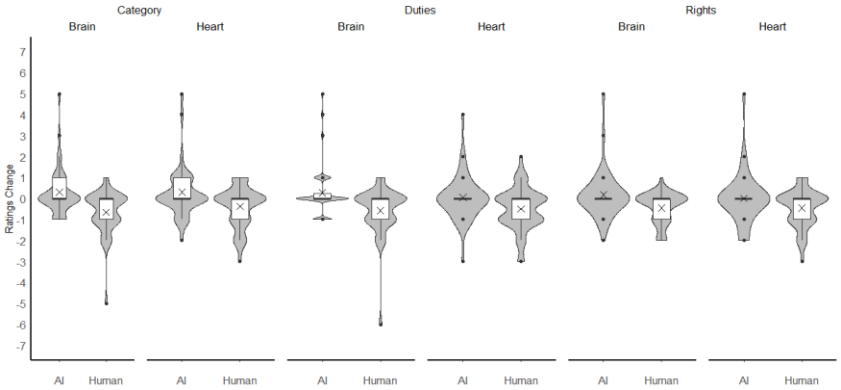

Figure 5: Change of Ratings as a Function of Subject, Organ, and Domain in Study 3

Note: The cross in the middle represents the mean. Within each box plot, the dividing line represents the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, and whiskers extend 1.5 times the interquartile range. Individual dots outside of the box represent outliers. The scale ranges from 1 (*strongly disagree*) to 7 (*strongly agree*).

Discussion

Study 3 sought to extend the findings of Study 1 by exploring the application of the causal-status hypothesis to the human category. To address the possibility that participants in Study 1 might have assumed certain functions could still be performed by a person's biological structure, even if one of their biological components changed, Study 3 employed hypothetical scenarios that explicitly stated that AI could perform all human functions. The results replicated the asymmetric changes observed in the first two studies, with

participants lowering their ratings of human subjects when their biological structures were altered but not increasing their ratings of nonhuman AI entities when they acquired human biological structures. Hypothetical scenarios were employed to control for prior knowledge regarding the roles of the brain and heart in mental activities. The study yielded intriguing results, revealing that the changes in ratings induced by alterations to the brain and heart were of equivalent magnitude. However, the human category boundary remained intact when the heart was altered, while it disappeared when the brain was changed. The domains of category, human rights, and human duties did not significantly influence the ratings, suggesting that participants perceived these three issues as equivalent and closely interconnected. These findings imply that the distinction between natural and legal persons, particularly in relation to rights and duties, may not be applicable when considering AI entities (Bennett & Daly, 2020). Even in these hypothetical scenarios where AI entities possessed full human capacities, participants did not attribute human rights or duties to them, consistent with previous research (Lima et al., 2020).

General Discussion

The advancement of increasingly sophisticated AI technology necessitates a re-evaluation of our understanding of the human category. As AI entities become more human-like and humans integrate AI-based functions and biological structures, the boundaries of what constitutes the human category are increasingly contested. Previous research has not specifically focused on humans as a category. This paper investigates how the development of AI might influence public perceptions of humans, employing a cross-species transplant paradigm. It examines whether the human category would be altered if AI could perform human functions and replace biological structures. Three studies were conducted to address this question. Study 1 assessed category and typicality ratings for human and nonhuman AI entities with varying functional and biological features. Study 2 utilised a category-based reasoning task to explore potential category shifts for these entities. Study 3 investigated how alterations to biological structures would impact human category membership, as well as entitlements to human rights and duties, assuming full functional equivalence between humans and nonhuman AI entities. The findings from these three studies suggest continuity in the human category. Human subjects were more likely to be perceived as members of the human category than nonhuman AI entities, irrespective of changes to their functions or biological

structures. The sole exception to this pattern occurred in the extreme hypothetical scenario where AI could perform all human functions and utilise the human brain. In this instance, the ratings for AI entities were equivalent to those for humans using AI in place of a human brain.

Theoretical Implications

While recognising the challenge of formulating necessary and sufficient conditions for ascribing human status, (Dennett, 2017) outlined six conditions that are all functional in nature, such as the capacity for verbal communication. Previous research on dehumanisation has suggested that some functional traits are integral to human nature and unique to humans (Haslam & Loughnan, 2014). However, the current studies demonstrate that functions and category membership can be dissociated, as nonhuman AI entities with human-like functions are still rejected as members of the human category. This suggests that these functions remain incidental features, rather than constitutive features of the human category. Although emotional and communicative abilities offer evidence that an entity possesses a human essence, they are not regarded as criteria for being human. Categorisation, therefore, is not solely grounded in similarity.

The tension between functional resemblance and categorical exclusion may help to explain the uncanny valley effect. In contrast to the categorisation difficulty account proposed by (Yamada et al., 2013), participants in the current studies did not exhibit confusion when categorising AI entities. Across all three studies, even when AI entities possessed fully human-like mental abilities, participants categorised them as nonhuman, indicating a strong resistance to altering categorical boundaries despite functional equivalence. This research contributes to the field of concept categorisation by comparing the human category with artefacts, whereas prior studies have predominantly contrasted it with other natural kinds (Kalish, 2002). The findings further contribute to differentiating between the causal status hypothesis and essentialism. Both perspectives acknowledge the causal potency of essential features, but they differ in their demands for knowledge of causal relationships (Ahn, 1998; Lombrozo & Rehder, 2012). Specifically, the causal status hypothesis posits that such knowledge is required, whereas essentialism does not. According to the causal status hypothesis, biological features that provide stronger causal explanations have a greater impact on category ratings than functional features. The results from Studies 1 and 2 support this hypothesis, as the substitution of

a human brain with AI induced larger changes in ratings than changes to communicative and emotional abilities.

As illustrated in all three studies, the brain, being a biological structure, was found to be more important than the heart. This aligns with the causal status hypothesis (Ahn, 1998; Lombrozo & Rehder, 2012), which asserts that internal structures (e.g., the brain) that offer causal explanations for functions are more significant than internal structures (e.g., the heart) that do not provide such explanations. Alternative views, such as the distributed view (Newman & Keil, 2008), which predict similar findings for both internal structures, are inconsistent with the data. However, the findings from Study 3 do not align with the causal status hypothesis. While biological structures may not be the sole mechanisms for functions in hypothetical scenarios, their replacement with AI still resulted in significant changes in ratings. Therefore, the current studies are more consistent with essentialist models.

Across all three studies, participants reduced their ratings of human subjects following changes to their biological structures, but did not increase their ratings of nonhuman AI entities under equivalent conditions. This asymmetry likely arises from differences in initial categorisation. For biological entities, the loss of biological structures appears more consequential than the acquisition of such structures for non-biological entities. The threshold for category change is substantially higher for AI entities in this context. However, some empirical evidence suggests that people may still attribute biological essence to AI entities; for instance, 47% of individuals spontaneously described their robotic dogs in a manner that affirmed a biological essence (Kahn Jr et al., 2002).

Regarding methodology, the transplant paradigm has been utilised in previous studies to investigate whether recipients exhibit the characteristics of the donor (Meyer et al., 2013). In the present study, a similar paradigm was employed to explore the issue of category continuity. While previous research has demonstrated that characteristics are likely to be transferred through this process, the current study shows that the original category status is maintained. It is important to note that in earlier transplant paradigm studies, the central question was whether the category would shift to that of the donor. In contrast, the current research focuses on the boundary of the human category, asking whether a human subject retains their human category membership despite changes. This distinction accounts for the apparent magnitude discrepancy between the current study (Studies 1 and 3) and prior studies. When the category reasoning task in Study 2 allowed for examining whether participants

would shift the category to AI entities, as in previous research, the findings replicated previous studies, showing that category shifts are generally unlikely.

Practical Implications

This research is valuable as it captures laypeople's perceptions in the context of the current state of AI technology. Study 2 revealed that participants-maintained AI category membership when entities exhibited human-like communication and emotional abilities, but did not extend this to other scenarios. This is likely since these two capabilities are on the cusp of realisation in our current era, making changes in category membership less likely to be triggered. Study 1 found that participants exhibited uncertainty regarding the category membership of humans using AI in place of their brain. Meanwhile, Study 3 indicated that nonhuman AI entities capable of fully human-like functions were perceived as human. These findings suggest that public attitudes towards the categorisation of human and AI entities are context dependent.

Beyond its theoretical contributions, this research has significant societal implications. Concerns have been raised regarding the potential redefinition and shift of the concept of being human in light of advancing AI technology (Nath & Manna, 2023). There is also evidence suggesting that assimilating AI as human could result in the dehumanisation of actual humans (Kim & McGill, 2024). However, no empirical research has previously tested whether such assimilation occurs. The findings from the hypothetical scenarios suggest that people are unlikely to categorise AI entities as human soon, which may limit interactions between humans and AI entities. Findings from Study 3 also imply that perceived category membership could influence entitlements to human rights and duties. This has relevance in ongoing debates about AI ethics and the legal personhood of AI entities. Conversely, for individuals augmented with AI technology, variations in responses suggest that their fate may depend more on decision-makers' perceptions than on their actual functional capabilities or initial categorisation, potentially leading to social marginalisation (Demoulin et al., 2004; Kteily & Bruneau, 2017; Leyens, 2009). This concern is supported by recent empirical evidence that demonstrates dehumanisation towards AI users (Dang & Liu, 2024).

Limitations and Future Directions

This study leaves several important questions unresolved. All participants were from the United States, where AI technology is more advanced and widely recognised. Previous research has shown that cross-cultural differences can

influence results in the cross-species transplant paradigm (Meyer et al., 2013). Thus, exploring cross-cultural variations in human category continuity in the context of AI could provide valuable insights. Additionally, individual differences such as religiosity and familiarity with AI development could be investigated further. A developmental perspective on this study would also be useful (Johnson, 1990; Newman & Keil, 2008). Further exploration of alternative explanations for the findings, such as the possibility that participants' responses reflect efforts to cope with potential threats to their own identity (Giger et al., 2019), would also be informative.

There are a few limitations to this exploratory study. First, the scenarios were not systematically tested for wording; only spontaneous feedback was obtained during the pilot study. The AI entity scenarios presented challenges. While vague descriptions were intentionally used to avoid triggering the uncanny valley effect associated with robots, the lack of a concrete, tangible physical substance may have made these scenarios difficult to visualise or may have made them seem nonsensical. Second, each stimulus was presented only once, which limited the ability to examine its psychometric properties. Future studies should address the wording of scenarios and develop multiple items for each experimental condition to enhance robustness. Third, although parametric analyses were used alongside detailed descriptive statistics presented in the figures, the results were skewed differently depending on whether human or AI entities were examined, which violated normal distribution assumptions. While visual inspection suggests that the conclusions align with results from the best available inferential statistics, these null hypothesis testing results should be considered more as a reference rather than definitive conclusions. Future researchers are encouraged to verify these conclusions by reanalysing the data, which will be made available, using more advanced statistical techniques that may emerge.

Conclusion

These studies highlight laypeople's current essentialist views regarding the human category in the context of AI advancements. People do not shift from human to AI membership if their functions or biological structures are performed by AI, and nonhuman AI entities do not gain human membership simply by possessing human functions or biological structures. However, it is possible that the human category could evolve in the future to reflect further advancements in AI technology.

References

Ahn, W.-k. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, *69*(2), 135–178. https://doi.org/10.1016/S0010-0277(98)00063-8

Association, A. P. (2022). APA dictionary of psychology. https://dictionary.apa.org/artificial-intelligence

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research*, *18*, 433–447. https://doi.org/10.1007/BF01067309

Bennett, B., & Daly, A. (2020). Recognising rights for robots: Can we? Will we? Should we? *Law, Innovation and Technology*, *12*(1), 60–80. https://doi.org/10.1080/17579961.2020.1727063

Converse, R. (1974). But When Did He Die: Tucker v. Lower and the Brain-Death Concept. *San Diego L. Rev.*, *12*, 424. https://digital.sandiego.edu/sdlr/vol12/iss2/13

Dang, J., & Liu, L. (2024). Extended artificial intelligence aversion: People deny humanness to artificial intelligence users. *Journal of Personality and Social Psychology*. https://psycnet.apa.org/doi/10.1037/pspi0000480

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science*, *41*, 382–402. https://doi.org/10.1111/cogs.12364

Demoulin, S., Torres, R. R., Perez, A. R., Vaes, J., Paladino, M. P., Gaunt, R., Pozo, B. C., & Leyens, J.-P. (2004). Emotional prejudice can lead to infra-humanisation. *European review of social psychology*, *15*(1), 259–296. https://doi.org/10.1080/10463280440000044

Dennett, D. C. (2017). *Brainstorms: Philosophical essays on mind and psychology*. MIT press. https://doi.org/10.7551/mitpress/11146.001.0001

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review*, *6*, 338–346. https://doi.org/10.3758/BF03212339

Fetterman, A. K., & Robinson, M. D. (2013). Do you use your head or follow your heart? Self-location predicts personality, emotion, decision making, and performance. *Journal of Personality and Social Psychology*, *105*(2), 316. https://psycnet.apa.org/doi/10.1037/a0033374

Giger, J. C., Piçarra, N., Alves-Oliveira, P., Oliveira, R., & Arriaga, P. (2019). Humanization of robots: Is it really such a good idea? *Human Behavior and Emerging Technologies*, *1*(2), 111–123. https://doi.org/10.1002/hbe2.147

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

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social cognition*, *26*(2), 248–258. https://doi.org/10.1521/soco.2008.26.2.248

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual review of psychology*, *65*(1), 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Huebner, B. (2009). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the cognitive sciences*, *9*, 133–155. https://doi.org/10.1007/s11097-009-9126-6

Johnson, C. N. (1990). If you had my brain, where would I be? Children's understanding of the brain and identity. *Child development*, *61*(4), 962–972. https://doi.org/10.1111/j.1467-8624.1990.tb02834.x

Kahn Jr, P. H., Friedman, B., & Hagman, J. (2002). " I care about him as a pal" conceptions of robotic pets in online AIBO discussion forums. CHI'02 Extended Abstracts on Human Factors in Computing Systems, https://doi.org/10.1145/506443.506519

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition*, *23*, 335–353. https://doi.org/10.3758/BF03197235

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition*, *30*(3), 340–352. https://doi.org/10.3758/BF03194935

Kim, H. y., & McGill, A. L. (2024). AI-induced dehumanization. *Journal of Consumer Psychology*. https://doi.org/10.1002/jcpy.1441

Kteily, N. S., & Bruneau, E. (2017). Darker demons of our nature: The need to (re) focus attention on blatant forms of dehumanization. *Current Directions in Psychological Science*, *26*(6), 487–494. https://doi.org/10.1177/0963721417708230

Leyens, J.-P. (2009). Retrospective and prospective thoughts about infrahumanization. *Group Processes & Intergroup Relations*, *12*(6), 807–817. https://doi.org/10.1177/1368430209347330

Lima, G., Kim, C., Ryu, S., Jeon, C., & Cha, M. (2020). Collecting the public perception of AI and robot rights. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–24. https://doi.org/10.1145/3415206

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive psychology*, *65*(4), 457–485. https://doi.org/10.1016/j.cogpsych.2012.06.002

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*(3), 297–337. https://doi.org/10.1075/is.7.3.03mac

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual review of psychology*, *51*(1), 121–147. https://doi.org/10.1146/annurev.psych.51.1.121

Meyer, M., Leslie, S. J., Gelman, S. A., & Stilwell, S. M. (2013). Essentialist beliefs about bodily transplants in the United States and India. *Cognitive Science*, *37*(4), 668–710. https://doi.org/10.1111/cogs.12023

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

Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 263. https://psycnet.apa.org/doi/10.1037/a0018685

Nath, R., & Manna, R. (2023). From posthumanism to ethics of artificial intelligence. *AI & SOCIETY*, *38*(1), 185–196. https://doi.org/10.1007/s00146-021-01274-1

Newman, G. E., & Keil, F. C. (2008). Where is the essence? Developmental shifts in children's beliefs about internal features. *Child development*, *79*(5), 1344–1356. https://doi.org/10.1111/j.1467-8624.2008.01192.x

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science*, *16*(4), 202–206. https://doi.org/10.1111/j.1467-8721.2007.00504.x

Schweitzer, S., & Waytz, A. (2021). Language as a window into mind perception: How mental state language differentiates body and mind, human and nonhuman, and the self from others. *Journal of Experimental Psychology: General*, *150*(8), 1642. https://psycnet.apa.org/doi/10.1037/xge0001013

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post*, *11*, 2022.https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology*, *8*, 1663. https://doi.org/10.3389/fpsyg.2017.01663

Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese psychological research*, *55*(1), 20–32. https://doi.org/10.1111/j.1468-5884.2012.00538.x