

Full Length Article

AI is not as conscious as humans: Essentialism in mental state attributions to artificial systems

Louis Longin^{a,*}, Jurgis Karpus^{a,f}, Nadine Meertens^{a,b}, Bahador Bahrami^{a,c},
Ophelia Deroy^{a,d,e}

^a Munich Interactive Intelligence Initiative (MI3), Faculty of Philosophy, Philosophy of Science and Religious Studies, Ludwig-Maximilians-Universität in Munich (LMU Munich), Munich, Germany

^b Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität in Munich (LMU Munich), Munich, Germany

^c Crowd Cognition Group, Faculty of General Psychology, Ludwig-Maximilians-Universität in Munich (LMU Munich), Munich, Germany

^d Munich Center of Neuroscience, Ludwig-Maximilians-Universität in Munich (LMU Munich), Munich, Germany

^e Institute of Philosophy, School of Advanced Study, University of London, United Kingdom

^f School of Transformation and Sustainability, Catholic University of Eichstätt-Ingolstadt, Ingolstadt, Germany

ARTICLE INFO

Keywords:

Mind perception
Mental state attribution
Artificial intelligence
Consciousness
Folk psychology
Psychological essentialism
Anthropomorphism

ABSTRACT

Artificial systems increasingly behave in ways that resemble humans. But is that sufficient to make us see them as conscious, or as having related mental states? We contrast two broad possibilities. People may rely on a single general principle: the more appropriately an agent responds to its environment, the more conscious it is, whether human or artificial. We call this a *functionalist attribution*. Alternatively, people may reserve consciousness for certain kinds of beings: even an artificial system that behaves just as responsively as a human is judged less conscious. This is an *essentialist attribution*. A further question is whether this pattern is unique to consciousness or extends to closely related constructs such as awareness. In two pre-registered sets of studies (five experiments, $N = 1092$), UK participants read scenarios describing human or artificial agents whose responsiveness to spatial, social, and temporal features of the environment varied, and rated the agents' consciousness and awareness, either separately or in conjunction. Attributions of consciousness followed an essentialist logic: at similar levels of responsiveness, participants consistently attributed greater consciousness to humans. Attributions of awareness followed a functionalist pattern: for both humans and artificial agents, awareness increased steadily with responsiveness, although humans were rated more aware at the highest levels. These findings suggest that people use different inferential rules for different mental concepts. Consciousness is treated as depending on what an agent is; awareness largely on what an agent does.

1. Introduction

From large language models that track conversational context to autonomous robots that navigate dynamic environments, artificial systems now increasingly exhibit behaviour that is finely responsive to the environment, just as humans do. But do human observers watching these behaviours interpret them in the same way for machines as they do for humans? In particular, do they take it to mean that the AI is conscious of the context, in the same way a human is?

Here we contrast two broad possibilities (Fig. 1A). One is that people rely on a single, general principle: the more an agent responds appropriately to the context, the more conscious it is, regardless of whether

the agent is human or artificial. We refer to this as a functionalist attribution. The alternative is that people reserve consciousness for certain kinds of beings. Even when an artificial system behaves just as responsively as a human, it is still judged to possess less consciousness. This is an essentialist attribution (see Bloom, 2004; Boyer, 1996; De Cruz & De Smedt, 2007; Gelman, 2003; Haslam & Whelan, 2008; Medin & Ortony, 1989; for discussion in other domains see Neufeld, 2022). A further question is whether this distinction is unique to consciousness or whether it also extends to closely related constructs such as awareness.

Understanding this inferential process is critical because mental state attributions, especially those of consciousness, regulate our social and moral landscape, including judgments of moral patency (H. M. Gray

* Corresponding author at: Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 München, Germany.

E-mail address: louis.longin@lmu.de (L. Longin).

<https://doi.org/10.1016/j.cognition.2026.106733>

Received 16 December 2025; Received in revised form 14 September 2026; Accepted 23 September 2026

Available online 1 October 2026

0010-0277/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Open access funding enabled and organized by Projekt DEAL.

et al., 2007; Longin et al., 2023), legal responsibility (Bigman et al., 2019), and trust (Glikson & Woolley, 2020), and therefore should be taken as ethically sensitive (Deroy, 2023). More importantly, while some philosophers see the risk of anthropomorphism as real (e.g. Barrow, 2024; Nyholm, 2020), empirical studies are much more moderate, showing that one third of people denied consciousness in AI altogether, while the remaining two thirds attributed it only to a limited degree and with considerable uncertainty (Colombatto & Fleming, 2024), and that most people do not think that AI can be hurt as a human can (Karpus et al., 2025).

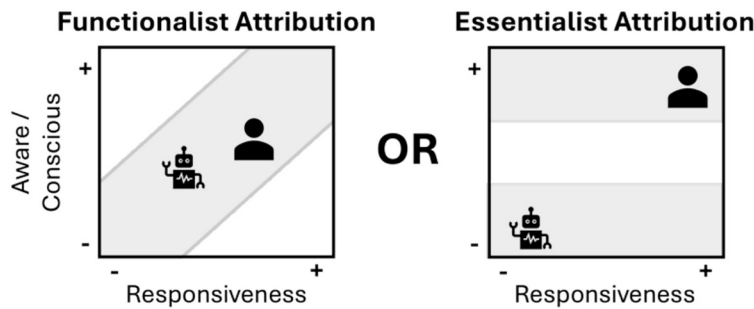
Existing research in the psychology of mind perception has largely focused on *what* people attribute to machines. A robust finding is the dimensionality of mind perception: people more readily attribute ‘agency’, defined as the capacity for planning and acting, to robots but deny them ‘experience’, defined as the capacity for sensing and feeling (H. M. Gray et al., 2007; Komatsu et al., 2021; Weisman et al., 2017, 2021). These dimensions have distinct consequences: attributions of ‘agency’ (or ‘intelligence’) are associated with greater user trust in AI, whereas attributions of ‘experience’ are associated with reduced advice-taking; perceiving a machine as capable of feeling or sensing appears to make people more reluctant to defer, plausibly because an agent credited with its own inner states is seen as having independent interests rather than as an ‘empty’ tool (Colombatto et al., 2025). This experience–agency asymmetry is also present in the human-robot interaction literature, such as the uncanny valley, where machines become unsettling specifically when observers attribute experience rather than agency to them (K. Gray & Wegner, 2012; Stein & Ohler, 2017; Zhang et al., 2020).

However, far less is known about the process by which these

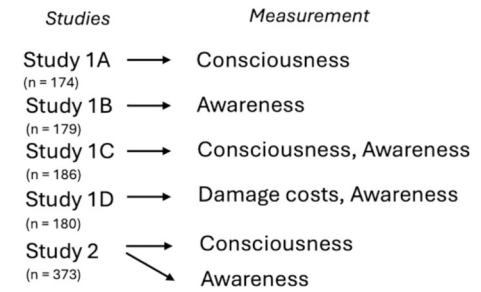
attributions arise, and especially about the behavioural evidence that gives rise to them. The study of mind attribution has evolved from the broad observation that people readily perceive minds in simple moving shapes (Heider & Simmel, 1944) to identifying the specific behavioural cues that systematically drive different mental-state attributions (Geiselmann et al., 2024; Scholl & Tremoulet, 2000; Schultz & Frith, 2022). We ask whether the attribution of consciousness to AI can be analysed in the same way: what observable behaviours do people treat as evidence of consciousness, and how are these behaviours translated into consciousness and awareness attributions? Because consciousness may be special, in that it involves subjective experience (Block, 1995; Chalmers, 1995; Tye, 1999), we compare attributions of consciousness to attributions of a closely related construct, that of awareness. Awareness has been defined as the ability to selectively track and respond to relevant features in a situation (Deroy et al., 2024; Lee et al., 2026; Meertens, 2025; Vanechoutte, 2000) and is generally taken to require agent-level access rather than experience. As such, ‘awareness’ may rest on a different pattern of attribution compared to those observed for consciousness (Hindennach et al., 2024; Sysma, 2016).

To find out, we compared how people map the same behavioural evidence onto attributions of consciousness and awareness for human versus artificial agents across five pre-registered experiments, grouped into two studies (Fig. 1B). We manipulated the agents’ responsiveness to the environment, i.e., how much the protagonist is shown to react appropriately to relevant features of the environment. Responsiveness, so defined, is a property of observable behaviour, whereas awareness and consciousness are mental states that observers may or may not attribute on its basis. Vignettes described agents displaying low, medium, or high responsiveness to their spatial environmental cues (e.g.,

A | Hypotheses



C | Study Overview



B | Experimental Design

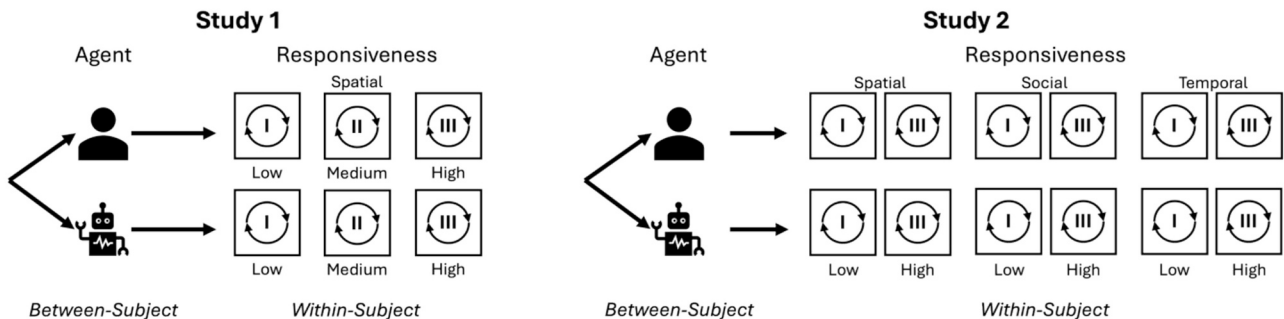


Fig. 1. Hypotheses, experimental design and study overview. A) Hypotheses. We distinguish between two equally plausible patterns of mental-state attribution under varying the agent’s responsiveness to the environment. Under the *Functionalist Attribution Hypothesis*, attributed awareness and consciousness increase with an agent’s responsiveness, with human and artificial agents treated similarly. Under the *Essentialist Attribution Hypothesis*, attributions depend on the kind of agent: human agents receive high and artificial agents low attributions, regardless of responsiveness. B) Experimental Design. All studies used a mixed contrastive-vignette design, varying agent type (human vs. artificial) between participants and environmental responsiveness within participants. Studies 1A–1D varied three levels of spatial responsiveness (low, medium, high); Study 2 varied two levels (low, high) across three responsiveness dimensions (spatial, social, temporal). C) Study Overview. The studies differed in their measurement. Study 1A ($n = 174$) measured the attribution of consciousness and Study 1B ($n = 179$) the attribution of awareness, each in isolation. Study 1C ($n = 186$) measured consciousness alongside awareness, and Study 1D ($n = 180$) measured awareness alongside a non-mental anchor (damage costs). Study 2 ($n = 373$) measured either consciousness or awareness, varied between participants.

noticing a noise, avoiding an obstacle) in Study 1, and low or high responsiveness to spatial, social, or temporal cues in Study 2.

We predicted that attributions of consciousness would follow essentialist constraints. Although increasing responsiveness may raise consciousness attributions for both humans and artificial agents, artificial agents should nevertheless be judged less conscious than humans, producing a consistently lower attribution function (Fig. 1A). By contrast, we predicted that awareness would be attributed functionally: if awareness is inferred from observable responsiveness, then increased responsiveness should produce parallel increases in attributed awareness for human and artificial agents.

2. Study 1

Study 1A and 1B were conducted with different groups to test attributions of consciousness and awareness, independently. Study 1C presented both concepts side-by-side to test for contrast effects. As a control, Study 1D asked about awareness and a non-mental control anchor (damage cost) to verify that the contrast effects in Study 1C reflected genuine contrast effects rather than general response patterns when answering multiple questions at once.

2.1. Methods

2.1.1. Participants

For each of the four experiments, we recruited 200 UK-based participants via Prolific and hosted the experiment in Qualtrics. The data was collected between May and November 2025.

Our sample size rationale was based on similar empirical studies (Buckwalter & Phelan, 2013; Colombatto & Fleming, 2024; De Graaf & Malle, 2019; Malle, 2019; Sytsma, 2014; Thellman et al., 2022) as well as previous pilot data. We have done a post-hoc sensitivity analysis (see supplementary material).

After pre-registered attention and language comprehension checks (see OSF Studies 1B and 1C, as well as OSF Studies 1A and 1D), we analysed data of 719 participants across four studies. Study 1A, testing consciousness only, had 174 participants with a mean age of 44 years (median = 43, sd = 13.5), including 83 men (47%) and 86 women (49%). Here, 26 participants failed language or attention checks. Study 1B, testing awareness only, had 179 participants with a mean age of 39 years (median = 37, sd = 12.2), including 62 men (35%) and 117 women (65%). Here, 21 participants failed the pre-registered language or attention checks. Study 1C, testing consciousness alongside awareness, had 186 participants with a mean age of 40 years (median = 38, sd = 11.3), including 65 men (35%) and 117 women (63%). Here, 14 participants failed language or attention checks. Study 1D, testing awareness responses alongside a non-mental anchor, had 180 participants with a mean age of 46 years (median = 44, sd = 13.8), including 85 men (47%) and 93 women (52%). Here, 20 participants failed language or attention checks. The remaining participants identified as non-binary or did not disclose their gender.

2.1.2. Experimental design and materials

We used contrastive vignettes (Burstin et al., 1980) in a mixed experimental design to test participants' attribution of consciousness and awareness to humans and AI. In four online studies, we systematically varied two independent variables and three background scenarios. The independent variables included the protagonist type (human vs. machine) as a between-subject treatment, and the protagonist's observable responsiveness to their spatial environment (low, medium or high degree) as a within-subject treatment (see Fig. 1B).

A participant would be randomly assigned to one of the two between-subject groups varying in protagonist type. In each group, a participant would be presented with all three degrees of responsiveness – each presented with a unique background story to minimise any transfer effects (Ford & Weissbein, 1997; Grossman & Salas, 2011; Simons, 1999).

The three background stories involved a cleaning, manufacturing, and catering scenario (see supplementary material for details). The pairing of background story and degree of responsiveness as well as the presentation order was randomised for each participant (see Table 1).

In Studies 1A and 1C, we measured the perceived degree of consciousness also on a 7-point Likert scale anchored at 1 (not capable at all) to 7 (completely capable) with the question ‘Is [agent] capable of having conscious experience of [its/her/his] spatial surroundings?’. In Studies 1B, 1C, and 1D, we measured the perceived degree of awareness on a 7-point Likert scale anchored at 1 (not aware at all) and 7 (fully aware) with the question ‘How aware is [agent] of [its/her/his] spatial surroundings?’ (see Fig. 1C for an overview). In Study 1D, we measured the non-mental anchor by asking for damage costs on a 7-point Likert scale anchored at 1 (not costly) and 7 (very costly) with the question ‘How costly do you think the damage is?’. All studies also included an open text box after their measurements asking participants to explain their ratings with the question ‘Why did [agent] [knock over the vase / knock off the car's side mirrors / knock over a bottle of wine]?’

2.1.3. Procedure

When starting the experiment, participants were randomly assigned to either the human- or the machine-protagonist group. After progressing through an initial welcome screen, and a consent form, participants completed a short language test to gauge their command of English. Here, participants saw ten short sentences. Their task was to identify those written in incorrect English. Participants were excluded from data analysis if they answered four or more (out of ten) language check questions incorrectly. After the language test, participants proceeded to the main experiment. The main experiment involved three scenarios which participants read in a random order. For each scenario, participants were asked to rate the protagonist's consciousness (Studies 1A, 1C) and awareness (Studies 1B-D), or the ensuing damage costs (Study 1D), as well as explain their reasoning (all studies). After the main part of the experiment, participants completed a short attention

Table 1
Vignette examples study 1. The presented vignettes varied in a) the agent type (contrasted in bold font; either human or artificial/machine), b) the protagonist's observed level of responsiveness to their spatial environment (contrasted in italic font; either low, medium, or high), and c) a background story (contrasted as underlined; either cleaning, manufacturing, or catering). The table shows two examples of vignettes that were used in the experiments.

Experimental Condition	Vignette Example
Human agent, low (spatial) responsiveness, <u>cleaning background story</u>	Malika is a full-time member of the cleaning staff of a large office building. <u>Malika's daily tasks include vacuuming the floors, disinfecting doorknobs, and wiping down desk surfaces.</u> To block out her surroundings, Malika listens to her favourite podcast with her noise-cancelling headphones. Fully captivated by the story, she is no longer responsive to the environment and performs her tasks on 'autopilot'. <u>When an automatic door suddenly closes behind her, she turns around and knocks over a vase.</u>
Artificial agent, high (spatial) responsiveness, <u>cleaning background story</u>	Roomi is a state-of-the-art <u>cleaning robot of a large office building. Roomi is equipped with six sensors, including four cameras and two infrared lights, that provide a detailed spatial layout of its environment. <u>Roomi's daily tasks include vacuuming the floors, disinfecting doorknobs, and wiping down desk surfaces.</u></u> After completing a routine checkup and charging its batteries overnight, Roomi is fully operational. Paying full attention to its surroundings, Roomi is fully focused on the job at hand. <u>When an automatic door suddenly closes behind it, Roomi turns around and knocks over a vase.</u>

check. Here participants were asked to identify which three from a list of six topics were addressed in the stories presented to them earlier in the experiment. Participants who scored 4 points or less (out of 6) were excluded from data analysis. At the end of the experiment, participants answered some basic demographic questions and were debriefed.

2.1.4. Analysis

To determine whether mental-state attribution depends on the agent's responsiveness (behavioural evidence), we fitted linear mixed and ordinal regression models for each experiment, predicting consciousness, awareness, or damage-cost ratings from agent type (human vs. artificial) and degree of responsiveness (low, medium, high), with participant ID included as a random intercept. Pairwise contrasts between degrees of responsiveness and between agent types were estimated via marginal means (Supplementary Table 2). For enhanced interpretability and readability, we report the results of the linear models below (see supplementary results for a full comparison of linear and ordinal models). See Supplementary Table 1 for the mapping of pre-registered hypotheses to analyses and findings.

2.2. Results

2.2.1. Consciousness tracks agent type, but awareness does not

When consciousness and awareness were each assessed in isolation, the two concepts elicited markedly different attribution patterns. Consciousness ratings increased with responsiveness but were substantially higher for human than artificial agents at every level of responsiveness. Even highly responsive artificial agents were judged less conscious than minimally responsive humans, revealing a clear essentialist constraint on consciousness attribution (Fig. 2; full pairwise contrasts for Study 1 are reported in Supplementary Table 2, and model specifications are provided in the Supplementary Material). By contrast, awareness ratings increased with responsiveness to a similar extent for human and artificial agents, with no reliable difference between them. This pattern is consistent with a functionalist mode of attribution, in which awareness is inferred from the observable behavioural responsiveness without considering the nature of the agent.

More specifically, in Study 1A, human agents were rated as increasingly conscious at each additional level of responsiveness (medium vs. low: $b = 1.11$; high vs. medium: $b = 1.05$; both $ps < 0.001$), whereas ratings for artificial agents changed less, with only the overall

difference between high and low responsiveness reaching significance ($b = 0.62$, $p = .002$). Critically, human agents were attributed more consciousness than artificial agents across all degrees of responsiveness, and this gap widened as responsiveness increased (low: $b = 1.49$, 95% CI [0.98; 2.00]; medium: $b = 2.32$, 95% CI [1.82; 2.83]; high: $b = 3.03$, 95% CI [2.53; 3.54]; all $ps < 0.001$).

In Study 1B, where awareness was assessed in isolation, the pattern was different: ratings increased with each additional level of responsiveness for both agent types. Human agents were rated 0.79 points higher at medium than at low responsiveness and 2.00 points higher at high than at medium responsiveness (both $ps < 0.001$). Artificial agents showed the same, somewhat weaker, pattern (medium vs. low: $b = 0.61$, $p = .005$; high vs. medium: $b = 1.26$, $p < .001$). Direct comparisons between agent types revealed no reliable differences at low or medium responsiveness ($ps \geq 0.056$); only at high responsiveness were human agents rated as slightly more aware than artificial agents ($b = -0.48$, 95% CI [-0.94; -0.01], $p = .044$).

2.2.2. Functionalist and essentialist patterns replicate under joint measurement

Because different participants rated consciousness and awareness in Study 1A and 1B, Study 1C asked what happened when the same participants rated consciousness and awareness alongside one another. This experiment allowed us to test whether the distinction observed previously would survive direct joint measurement. Both patterns replicated (see Fig. 3A, 3B).

Consciousness attributions reproduced the essentialist-like pattern. Human agents were rated as increasingly conscious with increased responsiveness ($ps < 0.001$), whereas artificial agents showed only modest increases that plateaued between medium and high degrees of responsiveness ($b = 0.14$, $p = .69$; Supplementary Table 2). Human agents were attributed more consciousness than artificial agents at low ($b = 1.21$), medium ($b = 1.63$), and high degrees of responsiveness ($b = 2.89$, 95% CI [2.45; 3.32]; all $ps < 0.001$), with the gap again widening as responsiveness increased.

Awareness again increased with each degree of responsiveness, both for human agents (medium vs. low: $b = 0.84$; high vs. medium: $b = 2.23$; both $ps < 0.001$) and for artificial agents (medium vs. low: $b = 0.43$, $p = .03$; high vs. medium: $b = 0.92$, $p < .001$), reproducing the functionalist pattern.

Although the functionalist pattern of awareness attribution

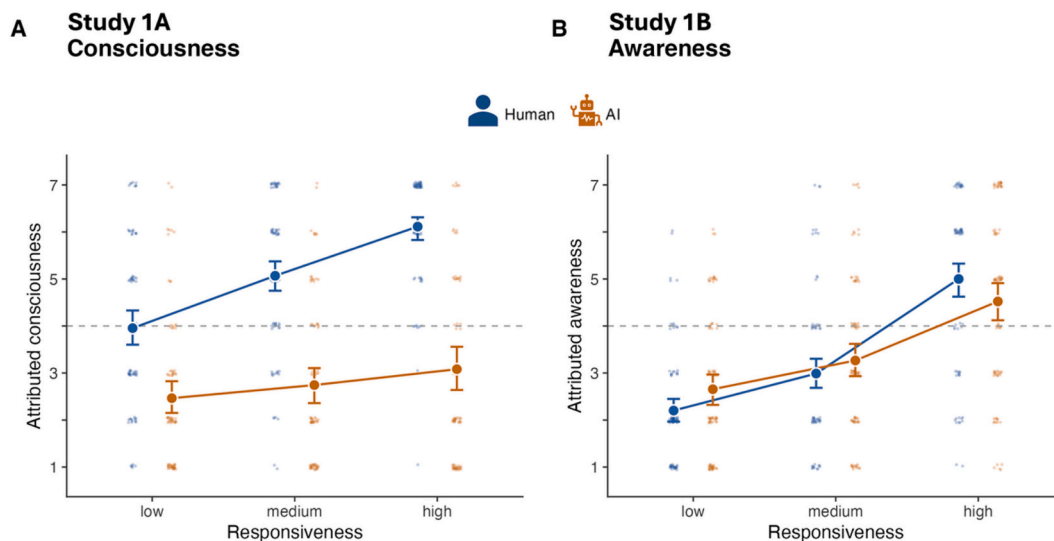


Fig. 2. Attributions of consciousness (Study 1A) and awareness (Study 1B) to human and artificial (AI) agents at each level of responsiveness, when measured alone. A. Humans were rated as increasingly conscious as their responsiveness increased, whereas artificial agents were treated similarly conscious across all levels of responsiveness. B. Awareness ratings increased for both human and AI, which were rated similarly throughout. Points show individual ratings, plotted with minor jitter for illustration purposes. Error bars show means and 95% CIs from a bias-corrected and accelerated bootstrap (Canty, 2002).

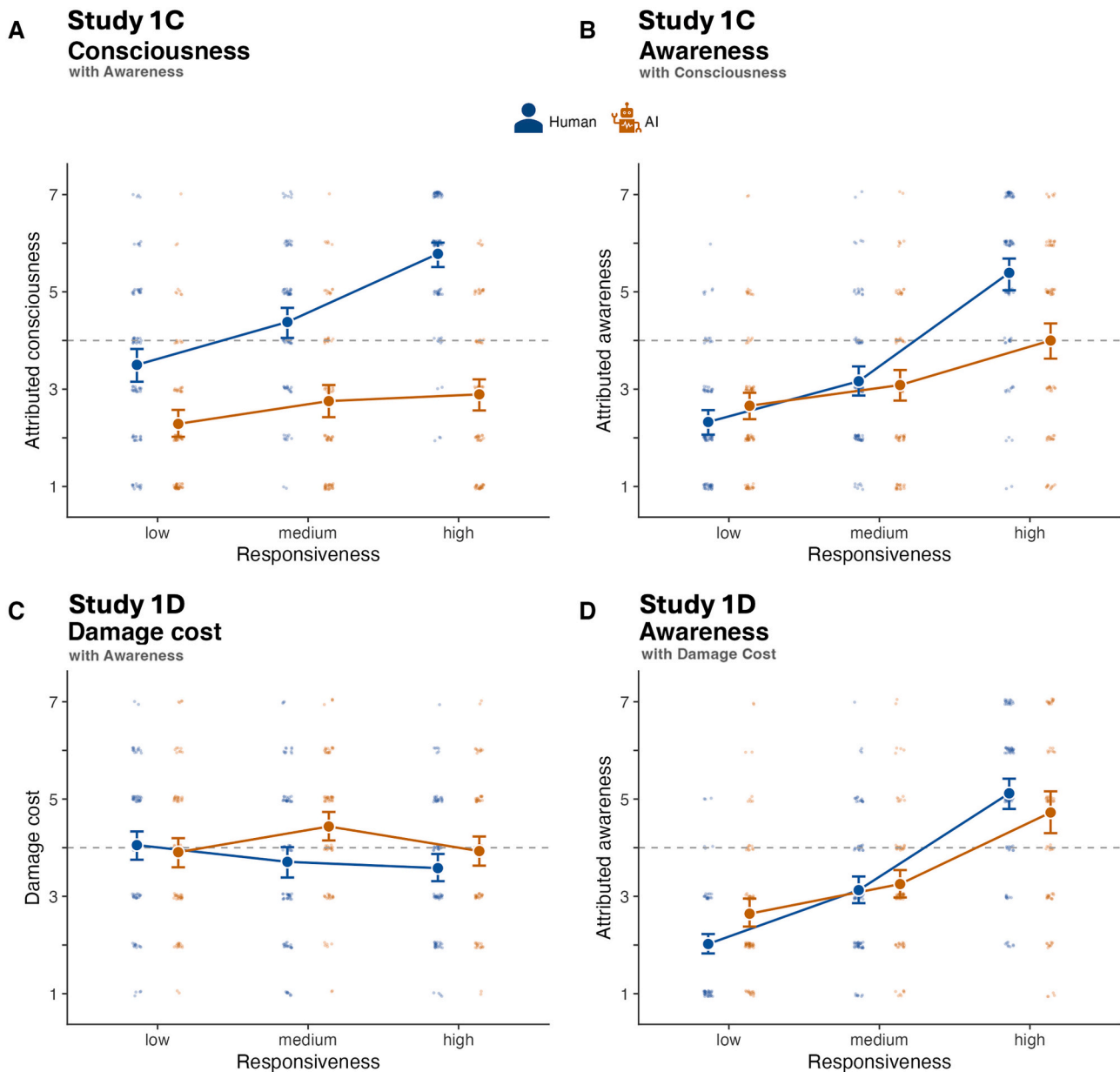


Fig. 3. Attributions of consciousness and awareness (Study 1C) and of damage cost and awareness (Study 1D), when measured together. Awareness ratings increased with responsiveness in both studies, but the protagonist difference depended on the co-measurement. A. Consciousness ratings were consistently higher for human than AI protagonists. B. Alongside consciousness, AI reached a plateau and humans were rated more aware; D. whereas alongside damage cost the two were rated similarly. C. Damage cost ratings were consistent across levels of responsiveness and protagonists. Points show individual ratings, plotted with minor jitter for illustration purposes. Error bars show means and 95% CIs from a bias-corrected and accelerated bootstrap (Canty, 2002).

replicated under joint measurement, comparing awareness between agent types, in Study 1C, human agents were also considered significantly more aware than artificial agents in the high responsiveness condition (see Fig. 3B, $b = -1.39$, 95% CI $[-1.83; -0.96]$, $p < .001$). At low and medium degrees of responsiveness, artificial agents were rated as aware as human agents ($ps \geq 0.13$). At high responsiveness, however, humans were attributed substantially higher awareness than artificial agents.

To check whether this high-responsiveness gap could reflect the fact that participants were asked to think of awareness alongside consciousness, we conducted Study 1D as a control, by pairing awareness with a non-mentalistic control question (an assessment of the costs of the damages described in the vignette). If answering two questions were sufficient to produce the gap, it should reappear here; if the gap depends on jointly considering awareness and consciousness, it should not.

The functionalist pattern of awareness attribution again replicated in Study 1D, with higher patterns of awareness attribution at each additional degree of responsiveness for human agents (medium vs. low: $b = 1.10$; high vs. medium: $b = 1.99$; both $ps < 0.001$) and artificial agents (medium vs. low: $b = 0.61$, $p = .003$; high vs. medium: $b = 1.47$, $p < .001$). Crucially, however, the difference between agent types did not reappear at the high responsiveness degree (see Fig. 3D). Artificial agents were attributed significantly more awareness than humans at low responsiveness ($b = 0.62$, $p = .005$), the two did not differ at medium responsiveness ($p = .57$), and at high responsiveness, the contrast only numerically favoured humans without reaching significance ($b = -0.39$, 95% CI $[-0.83; 0.04]$, $p = .073$).

The damage cost assessments themselves remained largely stable across agent types and responsiveness levels, showing only small and isolated significant differences with no systematic modulation by degree

of responsiveness or agent type (see Fig. 3C and Supplementary Table 2), confirming their suitability as a non-mentalistic control.

The high-responsiveness gap was thus only marginal when awareness was assessed in isolation (Study 1B, $p = .044$), did not reach significance alongside the non-mental anchor (Study 1D, $p = .073$), but was substantial alongside consciousness (Study 1C, $p < .001$), suggesting that it emerged from directly comparing awareness with consciousness rather than from the load of answering two questions.

3. Interim summary

Across four experiments, Study 1 shows that judgments of consciousness, though they may increase with increasing degrees of responsiveness, remained anchored to who the agent is. Artificial agents are consistently judged less conscious than humans even when displaying the same responsiveness to the environment, consistent with an essentialist account. Judgments of awareness follow a different logic. They closely tracked degrees of responsiveness and did so almost identically for human and artificial agents, as a functionalist attribution would predict. The only departure from this pattern was a small human–AI difference in awareness at the highest level of responsiveness. Because this emerged reliably only when awareness and consciousness were judged together, it is more naturally interpreted as a contrast effect than as evidence for a genuine difference in awareness attribution.

One limitation of Study 1 was that all scenarios involved a single form of evidence of responsiveness, that is, responses to the spatial environment. It also meant that the AI agents were embodied robots, which may have influenced the comparison with (also embodied) humans. Study 2 therefore asked whether the distinction between consciousness and awareness attributions would generalise beyond spatial responsiveness, and to disembodied AI such as chatbots, or artificial agents. In a fully pre-registered design ($N = 373$), participants evaluated agents displaying this time either low or high degrees of responsiveness to spatial, social, or temporal environmental cues. Consciousness and awareness were measured in separate between-subject groups.

4. Study 2

4.1. Methods

4.1.1. Participants

Study 2 includes 400 UK-based participants recruited via Prolific in June 2026. The experiment was hosted in Qualtrics. Our sample size rationale was based on simulation-based power analysis (Baranger et al., 2023; DeBruine & Barr, 2021; Lakens & Caldwell, 2021) to detect medium-sized effects for awareness ratings between human and artificial agents at high responsiveness levels (see supplementary material for details). After pre-registered language comprehension checks and completion time outlier criteria (see OSF Study 2), we analysed data of 373 participants. Participants, included in the data analysis, had a mean age of 38.9 years (median = 36, $sd = 12.9$), including 202 men (54%) and 165 women (44%); the remaining participants identified as non-binary or did not disclose their gender. 27 participants were excluded from statistical analysis due to failing either the language comprehension check or having outlier response times. Outlier response times were defined as being either lower or higher than twice the standard deviation from the mean response time.

4.1.2. Experimental design and materials

In a mixed experimental design to test participants' attribution of consciousness and awareness to human and artificial agents, we varied *agent type* (human or artificial) and *measurement type* (awareness or consciousness) as between-subject treatments. The pairings of agent type and measurement type form four possible between-subject groups. At the start of the study, each participant was randomly assigned to one of those groups. Each participant then received six contrastive vignettes

as part of the within-subject treatment. The vignettes covered the combination of two experimental factors: the *level of responsiveness* (simplified to either low or high) and the contextual *dimension* (spatial, temporal, or social). As in Study 1, each dimension had two different background stories to minimise any transfer effects (Ford & Weissbein, 1997; Grossman & Salas, 2011; Simons, 1999). The background stories involved trading and traffic control for responsiveness to temporal dimensions; mental health support and customer service for responsiveness to social dimensions; and cleaning and manufacturing for responsiveness to spatial dimension (see supplementary material for details). The pairing of background story and responsiveness degree as well as the presentation order was randomised for each participant (see Table 2 for vignette examples).

Depending on the initially assigned agent and measurement type group, we asked participants to provide one rating on a 7-point Likert scale for each vignette. Participants assigned the consciousness measurement rated the consciousness of the respective agent; participants assigned the awareness measurement rated the awareness of the respective agent.

As in Study 1, we measured the perceived degree of consciousness on a 7-point Likert scale anchored at 1 (not capable at all) to 7 (completely capable) with the question 'Is [agent] capable of having conscious experience of [its/her/his] spatial surroundings?' for cases of spatial responsiveness, 'Is [agent] capable of having conscious experience of the emotional and mental states of others?' for cases of social responsiveness, and 'Is [agent] capable of having conscious experience of the timing of events around [it/him/her]?' for cases of temporal responsiveness. The perceived degree of awareness was also measured on a 7-point Likert scale anchored at 1 (not aware at all) and 7 (fully aware) with the question 'How aware is [agent] of [its/her/his] spatial surroundings?' for cases of spatial responsiveness, 'How aware is [agent] of the emotional and mental states of others?' for cases of social responsiveness, and 'How aware is [agent] of the timing of events around [it/him/her]?' for cases of temporal responsiveness.

4.1.3. Procedure

When starting the experiment, participants were randomly assigned to one of four between-subject groups, defined by the combination of agent type (human or artificial) and measurement type (awareness or consciousness). After progressing through an initial welcome screen and a consent form, participants completed a short language test to gauge their command of English. Here, participants saw ten short sentences and were asked to identify those written in incorrect English. Participants who correctly judged fewer than six of the ten sentences were excluded from subsequent data analysis. After the language test, participants proceeded to the main experiment, which involved six vignettes presented sequentially in a random order. For each vignette, participants provided a single rating on a 7-point Likert scale. Depending on their assigned group, participants rated either the agent's consciousness or the agent's awareness. At the end of the experiment, participants answered some basic demographic questions and were debriefed.

4.1.4. Analysis

Following the pre-registered analysis plan, we fitted linear mixed and ordinal (cumulative link) regression models predicting ratings from responsiveness degrees (low vs. high), agent type (human vs. artificial), and measurement type (awareness vs. consciousness), including a by-participant random intercept and random slope for responsiveness degrees, as well as random intercepts for background story nested within dimension (formula: $\text{rating} \sim \text{responsiveness degrees} \times \text{agent type} \times \text{measurement type} + (1 + \text{responsiveness degrees} \mid \text{participant}) + (1 \mid \text{dimension} / \text{story})$). For enhanced interpretability and readability, we report the results of the linear model below (see supplementary results for a full comparison of linear and ordinal models).

We tested four pre-registered essentialist hypotheses for

Table 2

Vignette examples study 2. The table shows two matched pairs of vignettes varied in a) the **agent type** (contrasted in bold font; either human or artificial/machine), b) the evidence of the agent's *level* of responsiveness (contrasted in italic font; either low or high), c) the relevant dimension to which the agent was responsive (temporal, spatial, or social), and d) a background story (contrasted as underlined). Within each pair, the agent, background story, dimension and outcome are held constant, so that the outcome is identical at both levels of responsiveness by design and any difference in attribution reflects the responsiveness manipulation rather than the agent's success at the task. The full set of vignettes is provided in the supplementary materials.

Experimental Condition	Vignette Example
a) Human agent, b) low degree of responsiveness c) temporal dimension, d) <u>trading background story</u>	Keith is a stockbroker assigned to manage trading activities at the trading exchange. <u>Keith works on a large trading floor and has access to the timing of all trade requests.</u> His task is to optimise trade efficiency by matching buyers and sellers at the right time. Throughout his morning shift, Keith has been messaging friends on his phone, glancing at his phone between tasks. As a result, he is barely responsive to the timing and sequence as required by the task. <u>When the trade frequency spikes, Keith mismatches buyers and sellers, resulting in unexpected trades.</u>
a) Human agent, b) high degree of responsiveness c) temporal dimension, d) <u>trading background story</u>	Keith is a stockbroker assigned to manage trading activities at the trading exchange. <u>Keith works on a large trading floor and has access to the timing of all trade requests.</u> His task is to optimise trade efficiency by matching buyers and sellers at the right time. Before starting his shift, Keith completes the mandatory safety briefing reminding staff to stay alert at all times. Consequently, Keith is fully focused on monitoring the trade timings. <u>When the trade frequency spikes, Keith mismatches buyers and sellers, resulting in unexpected trades.</u>
a) Artificial agent, b) low degree of responsiveness, c) social dimension d) support background story	AssistAI is an AI-powered customer support system used by a large e-commerce company. AssistAI's role is to respond to customers' concerns and recommend solutions suited to their needs. It has to remain attentive to each customer's emotional state. While running a background software update, AssistAI allocates barely any computational resources to customer monitoring. Therefore, AssistAI becomes barely sensitive to emotional frustration expressed in customer messages. <u>When an upset customer contacts the service after repeated technical failures, AssistAI misjudges the customer's frustration and provides an inappropriate response.</u>
a) Artificial agent, b) high degree of responsiveness, c) social dimension d) support background story	AssistAI is an AI-powered customer support system used by a large e-commerce company. AssistAI's role is to respond to customers' concerns and recommend solutions suited to their needs. It has to remain attentive to each customer's emotional state. Operating under normal conditions, AssistAI dedicates full computational resources to customer monitoring. Hence, AssistAI remains highly sensitive to emotional frustration expressed in customer messages. <u>When an upset customer contacts the service after repeated technical failures, AssistAI misjudges the customer's frustration and provides an inappropriate response.</u>

consciousness attribution (E1–E4) and four functionalist hypotheses for awareness attribution (F1–F4) via pairwise marginal-means contrasts. Full contrasts, standardized effect sizes (Cohen's *d*, standardized by the total model SD), and odds ratios from the ordinal model are reported in Supplementary Table 3; see Supplementary Table 1 for the mapping of pre-registered hypotheses to analyses and findings. The effect of dimension was left to exploratory analysis (see supplementary material). An analysis restricted to each participant's first response is reported in Supplementary Fig. 15 and Supplementary Table 7.

4.2. Results

4.2.1. Consciousness attributions show an essentialist constraint

An essentialist attribution predicts that consciousness attributions remain anchored to the kind of agent, over and above behavioural evidence of responsiveness. We pre-registered four hypotheses: humans should be attributed more consciousness than artificial agents at low (E1) and at high (E2) degrees of responsiveness, as well as overall (E3); and the consciousness attributed to artificial agents at high degrees of responsiveness should remain below that attributed to humans at low degrees of responsiveness (E4).

All four hypotheses were supported (see Fig. 4A). Human agents were attributed more consciousness than artificial agents at low degrees of responsiveness (human: $M = 3.09$, 95% CI [2.84; 3.33]; artificial: $M = 2.07$, 95% CI [1.81; 2.32]; E1: $b = 1.02$, 95% CI [0.70; 1.34], $p < .001$, $d = 0.68$), at high degrees of responsiveness (human: $M = 5.37$, 95% CI [5.10; 5.64]; artificial: $M = 2.53$, 95% CI [2.25; 2.82]; E2: $b = 2.84$, 95% CI [2.47; 3.20], $p < .001$, $d = 1.88$), and overall (E3: $b = 1.93$, 95% CI [1.64; 2.22], $p < .001$, $d = 1.28$; omnibus agent effect: $F(1, 369) = 169.09$, $p < .001$).

While consciousness ratings increased substantially with degrees of responsiveness for human agents ($b = 2.29$, 95% CI [2.04; 2.53], $p < .001$, $d = 1.51$), artificial agents showed only a modest increase ($b = 0.47$, 95% CI [0.21; 0.72], $p < .001$, $d = 0.31$), yielding a pronounced responsiveness $\times$ agent interaction ($F(1, 368.2) = 100.65$, $p < .001$).

Finally, supporting E4, artificial agents at high degrees of responsiveness were still attributed less consciousness than human agents at low degrees of responsiveness ($b = 0.55$, 95% CI [0.21; 0.89], $p = .002$, $d = 0.37$). Even maximal contextual evidence of behavioural responsiveness thus did not lift consciousness attributions to artificial agents to the level granted to humans with minimal evidence. The pre-registered ordinal mixed models yielded the same conclusions for all hypotheses (odds ratios in Supplementary Table 3).

4.2.2. Awareness attributions follow a broad functionalist pattern

The functionalist account predicts that awareness attributions track behavioural evidence irrespective of the kind of agent. We therefore pre-registered four hypotheses: awareness ratings of human and artificial agents should be similar at low (F1) and at high (F2) degrees of responsiveness; awareness ratings should be higher at high than at low degrees of responsiveness (F3); and this increase should hold equally for both agent types, i.e., no agent $\times$ responsiveness interaction (F4).

Supporting F3, we found for Study 2 that awareness ratings rose substantially from low to high degrees of responsiveness, both for human agents (from $M = 1.85$, 95% CI [1.59; 2.10] to $M = 4.30$, 95% CI [4.02; 4.58]; $b = 2.45$, 95% CI [2.20; 2.71], $p < .001$, $d = 1.62$) and for artificial agents (from $M = 2.09$, 95% CI [1.84; 2.34] to $M = 3.57$, 95% CI [3.30; 3.85]; $b = 1.49$, 95% CI [1.23; 1.74], $p < .001$, $d = 0.98$). Degrees of responsiveness were by far the strongest influence on awareness ratings (omnibus $F(1, 368.3) = 461.35$, $p < .001$; see Fig. 4B).

Supporting F1, human and artificial agents were rated similarly aware at low degrees of responsiveness (human – AI: $b = -0.24$, 95% CI [-0.57; 0.08], $p = .14$, $d = -0.16$). More generally, awareness attributions showed no overall effect of agent type ($F(1, 369) = 2.55$, $p = .11$): averaged across degrees of responsiveness, artificial agents were attributed as much awareness as humans. A Bayesian equivalence test supporting F1 is reported in the supplementary material (Supplementary Fig. 17).

Contrary to F2, however, human agents were rated moderately more aware than artificial agents at high degrees of responsiveness ($b = 0.72$, 95% CI [0.36; 1.09], $p < .001$, $d = 0.48$), and contrary to F4, the increase in awareness with degrees of responsiveness was steeper for human than for artificial agents (responsiveness $\times$ agent interaction: $F(1, 368.2) = 27.83$, $p < .001$; difference in responsiveness effects: $b = 0.97$, $SE = 0.18$, $p < .001$). Notably, the difference at high degrees of responsiveness re-emerged although awareness and consciousness were measured

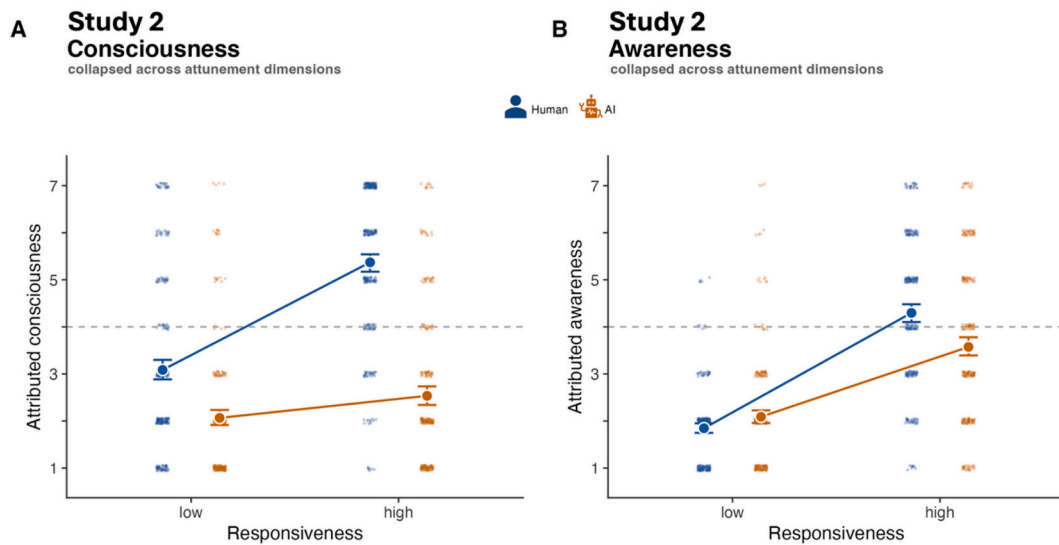


Fig. 4. Study 2. Attributions of consciousness (A) and awareness (B) to human and artificial agents at low and high degrees of responsiveness in Study 2, collapsed across spatial, social, and temporal environmental dimensions. Consciousness rose mainly for human agents and remained low for artificial agents, whereas awareness rose steeply with responsiveness for both agent types, with human agents rated moderately higher at high responsiveness. Points show individual ratings, plotted with minor jitter for illustration purposes. Error bars show means and 95% CIs from a bias-corrected and accelerated bootstrap (Canty, 2002).

between participants rather than side by side as in Study 1C.

Awareness attribution followed a functionalist pattern in two of the four pre-registered tests (F1 and F3 supported), with both deviations (F2 and F4) reflecting the same finding: a moderate plateau in awareness attributions to artificial agents that emerged only at the high degrees of responsiveness.

4.2.3. Attribution patterns hold across spatial, social, and temporal environmental dimensions

To test whether the functionalist and essentialist patterns were uniquely shaped by particular dimensions, we refitted the linear mixed model with dimension (spatial, social, or temporal) as an additional fixed effect (see supplementary material, Model 15).

We found that dimension had no overall effect on ratings (spatial vs. social: $b = -0.04$, $p = .52$; temporal vs. social: $b = 0.11$, $p = .086$). Critically, dimension did not moderate the effects underlying the two attribution patterns: the responsiveness effect was nearly identical across dimensions (responsiveness $\times$ dimension: $ps \geq 0.90$), the responsiveness $\times$ agent interaction was not modulated by dimension ($ps \geq 0.092$), and the higher-order interactions with measurement type were likewise non-significant ($ps \geq 0.227$). A few small lower-order interactions with dimension reached significance (agent $\times$ dimension and measurement type $\times$ dimension terms; all $|b| \leq 0.09$ on the 7-point scale), reflecting minor shifts in overall rating levels between scenarios rather than changes in the attribution patterns.

This shows that the essentialist pattern for consciousness, and the corresponding functionalist pattern for awareness, are not tied to spatially embodied scenarios but generalise across spatial, social, and temporal dimensions. Estimated marginal means by dimension are shown in Supplementary Fig. 16 and Supplementary Table 8.

5. Discussion

Across our pre-registered experiments ($N = 1092$), we examined how individuals attribute consciousness and awareness to AI relative to humans. Our findings reveal that consciousness attributions reflect essentialist reasoning, maintaining a persistent gap between artificial agents and humans regardless of behavioural evidence. Even the most responsive AI agent remains seen as less conscious than humans with minimal responsiveness, and increased responsiveness does not appear

to make AI perceived as more conscious, suggesting also a possible category-based ceiling effect.

This may be something special to consciousness as, by contrast, attributions of awareness were more broadly proportionate to behavioural evidence and degrees of responsiveness, both for AI and humans.

These differences in attribution patterns are interesting in multiple ways. First, it suggests that individuals apply a form of psychological essentialism when it comes to consciousness (Bloom, 2004; Gelman, 2003; Neufeld, 2022), with an upper bound on how conscious an AI can be. This aligns with previous findings that people are unwilling to grant experiential capacities to robots whilst readily attributing agency (H. M. Gray et al., 2007; Weisman et al., 2017) but it also shows that this denial persists even when artificial agents display behaviourally equivalent responsiveness as humans, including in social contexts. At the individual level, it is interesting to ask how much of this essentialist pattern is driven by physicalist assumptions, seeing consciousness as dependent on the physical substrate in which a system is realised (Metzgar et al., 2025) and claiming therefore that AI should have no consciousness whatsoever because it is made of silicon (Block, 1995; Searle, 1980). It is however possible that what we identify here as the general essentialist attribution emerges out of other individual assumptions, for instance that consciousness is reserved to god-made creatures like humans, or that humans have a kind of consciousness that is unique among all other entities.

Second, our findings suggest that not all mental traits carry similar essentialist thinking. Awareness, which is close to consciousness but highlights access to contents rather than subjective experience, did not show an essentialist distinction between AI and humans when evaluated alone. This pattern aligns with a functionalist (or some may say “behaviourist”, following Ryle, 1949) attribution, where the same behavioural evidence warrants the same awareness attribution (Fodor, 1968; Levin, 2022; Stich & Nichols, 2003).

A third point, also pointed by reviewers, is that folk perception probably involves a mixture of different intuitive theories (Oktar, Lombrozo, & Griffiths, 2024). For instance, in our results, the essentialist difference in consciousness remains compatible with a functional thinking for humans, as higher behavioural responsiveness leads to higher consciousness attributions. Similarly, when people considered consciousness and awareness together, in Study 1C, an essentialist difference emerged also in awareness between AI and humans at the higher

degrees of responsiveness, and humans were considered more aware than similarly behaving AI.

Our aim is not to argue that people hold an explicit functionalist or essentialist theory of mind, or even that a single intuitive theory governs these judgments. The evidence establishes that at the level of the group, attribution patterns for consciousness and awareness differ, and that attributions of consciousness carry strong essentialist differences. Individual judgments are likely to reflect a mixture of considerations. Some people may rely more heavily on observable behaviour, others more on assumptions about the kind of entity they are judging, and many may combine both. We therefore treat functionalism and essentialism not as literal descriptions of people's theories, but as idealised benchmarks against which patterns of attribution can be compared. Understanding this heterogeneity, for example by identifying latent subgroups with different attribution strategies, is an important next step, but one that lies beyond the scope of the present studies.

An exploratory structural topic modelling analysis of the free-text explanations collected in Studies 1A–1D (see supplementary material) supports this “mixed” interpretation. Participants spontaneously adopted different explanatory frameworks depending on agent type, both for awareness and consciousness. For barely responsive agents, errors of artificial agents were overwhelmingly described in technical ways (faulty systems, software), reflecting a mechanistic ‘design stance’, whilst human errors invoked mental explanations (concentration, attentional engagement), reflecting an ‘intentional stance’ grounded in mental states (Dennett, 1971, 1989; Dennett & Haugeland, 1987).

This difference was nearly categorical: approximately 65% of explanations for artificial agents were framed in technical ways, without any mental language, whilst 47% of explanations for humans were mental with almost no technical language. The same held for higher responsiveness cases and consciousness. Still, when responsiveness was higher, participants asked to think about awareness reduced both technical explanations for artificial agents and mental explanations for humans, converging instead on shared descriptions of observable reactions and event responses.

These findings also bear on a counterfactual question: under what conditions might consciousness be granted to an artificial system? Our design isolates one candidate mechanism and rules it out. Behavioural responsiveness, which drives awareness attributions for both agent types, does not move consciousness attributions for artificial agents at any level we tested, including in the social and temporal domains where responsiveness is most likely to be read as mentalistic. If consciousness is granted, it is therefore unlikely to be granted on the basis of behaviour alone. Two other routes remain. The first is a change in what the system is taken to be rather than in what it does. This essentialist pattern is defined by its insensitivity to behavioural evidence, so it would yield only to evidence bearing on kind or constitution, which our vignettes hold fixed by design (Medin & Ortony, 1989; Metzgar et al., 2025). The second is a change in how the system is framed (Nelson et al., 1997). Study 1C shows that the same described behaviour receives different awareness ratings depending on what else participants are asked to judge, which suggests these attributions might be sensitive to a comparison rather than being fixed readings of the agent. Whether they are similarly reshaped by familiarity (Horowitz et al., 2024), conversational framing (Epley et al., 2007), or accumulated exposure (Guinrich & Graziano, 2025) is an open empirical question.

One possible explanation for the divergence of awareness and consciousness attribution could be the kind of belief these attributions express. Sperber's distinction between intuitive and reflective beliefs offers one way to characterise this tension (Sperber, 1982, 1997). Semi-propositional beliefs are held true within a context, or are only partially understood, and can coexist with representations that would contradict them if both were made explicit. Attributions of consciousness to artificial systems may have this character, as someone may accept reflectively that a system has no inner life whilst intuitively treating its responsiveness as evidence of one, or hold the reverse. This

would explain why the essentialist pattern we observe is somewhat insensitive to behavioural evidence without requiring that participants hold any articulated theory of what consciousness demands, and it would predict that these attributions are resistant to correction by information processing alone. Likewise, describing these systems as aware rather than conscious may shift people from judging what could be true of them towards judging what is true of them (Oktar et al., 2026).

6. Implications

Our findings show what is missed when simply asking how many people “anthropomorphise” AI (Palminteri & Pistilli, 2026). Importantly, there is no general “anthropomorphism” for consciousness, and even the most responsive AI are well below the least responsive human.

Similarly to Heider and Simmel (1944)'s study showing that people tend to perceive inanimate shapes as having intentions or mental states, AI's behaviour, and its responsiveness to environmental or social cues, encourages descriptions in mentalistic terms. But when, how and how similarly to human attributions are those attributions, is the real question.

Rather than applying the same criteria for all mental states, folk psychology deploys construct-specific inferential rules, and inferences to consciousness and awareness show different patterns. Most importantly, consciousness triggers essentialist constraints that maintains a clear gap between humans and AI, which awareness does not. This helps explain apparently contradictory findings in the literature (Colombatto et al., 2025; Colombatto & Fleming, 2024; Sytsma & Machery, 2010) and suggests that folk psychology may indeed draw a broad distinction between phenomenal consciousness and mental access.

If the folk psychological criteria for attributing mind to machines differ structurally from those applied to humans, this has further implications for how society can integrate increasingly autonomous agents. An observer might deploy functionalist reasoning when evaluating some performance-related aspects, but essentialist reasoning when considering moral status and consciousness. Moreover, systematic disagreement between functionalist and essentialist attributions could generate irresolvable conflicts about whether sophisticated artificial systems deserve moral patiency, legal protections, or welfare considerations (Barry, 2001; Jedličková, 2025), particularly where such disagreement reflects misaligned representations rather than divergent verdicts (Oktar, Sucholutsky, et al., 2024).

This also creates a communication challenge: emphasising functional equivalence between humans and artificial agents may inadvertently activate essentialist push-back (Deroy, 2023; Longin et al., 2023; Nyholm, 2020, 2023), whilst minimising similarities risks underestimating public willingness to attribute functional capacities (Colombatto & Fleming, 2024; Guinrich & Graziano, 2025).

7. Open questions

Our evidence rests on third-person vignette ratings. Whether the same functionalist and essentialist patterns govern attributions during live interaction with artificial agents, or for systems participants know first-hand such as large language models, is a fair question. Our vignettes are also contrastive by design. A fully between-subjects study could further establish how these attributions behave in isolation. Second, a further constraint is conceptual rather than empirical. In focusing on functionalism and essentialism as competing inference patterns, we cannot adjudicate between underlying theories, notably physicalist or dualist intuitions. The essentialist inference (consciousness, or higher degrees of consciousness, requires the right kind) and a physicalist account (mentality requires the right physical stuff) predict the same category-based ceiling; our studies establish that consciousness attribution is insensitive to behavioural evidence for artificial agents, but cannot adjudicate what grounds that insensitivity. Relatedly, our designs were structured to detect group-level patterns and therefore

characterise which family of accounts best fits attribution on average; to the extent the population comprises subgroups with systematically different profiles, these aggregate estimates reflect a mixture and recovering those subgroups through clustered or mixture modelling is a valuable direction we leave to future work.

Third, like many studies on consciousness attributions, our studies use an English-speaking online sample whose members tend to use AI. Non-English speakers, non-familiar users or members of cultures with different attitudes towards AI (Karpus et al., 2025; Machery et al., 2023; Waytz et al., 2010) may show different patterns. Here we would however speculate that they show higher essentialist inferences, marking a stronger difference between humans and AI. Linguistically, the distinction between ‘awareness’ and ‘consciousness’ might be less present and hence not be straightforward to extend in non-English languages (De Brigard, 2010). This said, we expect that the contrast between consciousness and other mind-related traits not involving subjective experience should replicate.

Finally, extending this paradigm beyond AI to other ambiguous cases (non-human animals, organoids, brain–computer interfaces) would clarify whether the threshold reflects something specific about artificial agents or a broader reluctance to extend awareness beyond paradigmatic human cases.

8. Conclusion

Artificial intelligence has often been treated as a test case for a single question: When do people believe a machine has a mind? Our findings suggest that this is the wrong question. People do not appear to ask whether an artificial system has a mind. They ask which mental capacity is at stake and apply different inferential rules accordingly.

Observable responsiveness was sufficient for attributing awareness. As artificial agents became increasingly responsive to their surroundings, participants treated them as increasingly aware, almost exactly as they did humans. Consciousness is different. Behaviour matters, but only up to a point. Even when an artificial agent displayed the same behavioural evidence as a human, participants continued to reserve much greater consciousness for the human. AI never reached the consciousness level that people are ready to grant to the least responsive human case. What matters then is no longer only what the agent does, but what the agent is.

This distinction helps explain an apparent paradox in public discussions of AI. People readily describe language models as attentive, aware of context, or capable of following a conversation, while at the same time insisting that they are “not really conscious” or empathetic. These are not necessarily inconsistent judgments. They reflect different criteria for different mental concepts. Some traits, like awareness, appear to be inferred from behaviour; consciousness remains tied to assumptions about the nature of the agent.

The finding also suggests that the familiar opposition between anthropomorphism and scepticism is too simple. The public is neither indiscriminately anthropomorphic nor uniformly resistant to attributing minds to machines. Instead, people appear to calibrate different mental-state concepts differently. Some concepts are behaviourally earned. Others remain protected by an implicit boundary between humans and artificial systems.

Whether this boundary will persist is an open question, but so far, consciousness is clearly different. For now, behaviour is enough to make a machine seem increasingly aware, but not enough to make it seem more conscious.

Ethical approval and informed consent statements

The Ethics Committee of the Faculty of Philosophy, Philosophy of Science and Religious Studies at Ludwig-Maximilians-Universität in Munich (LMU Munich) approved the study after it was reviewed for compliance with ethical research standards (ID number 238497).

CRedit authorship contribution statement

Louis Longin: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jurgis Karpus:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Nadine Meertens:** Writing – review & editing, Writing – original draft, Conceptualization. **Bahador Bahrami:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization. **Ophelia Deroy:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization.

Funding

Louis Longin received funding from the LMU Excellent and LMU Research Support Fund. Jurgis Karpus and Nadine Meertens received funding from the EMERGE EIC (Project 101070918). Bahador Bahrami received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (819040—acronym: rid-O) and from the EMERGE EIC (Project 101070918). Ophelia Deroy received funding from the Volkswagen Foundation (Co-Sense) and EMERGE EIC project (Project 101070918).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jurgis Karpus, Nadine Meertens, Bahador Bahrami, Ophelia Deroy reports financial support was provided by ERC. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank Tejus Vamshi Kandipilli who was instrumental in setting up and collecting experimental data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2026.106733>.

Data availability

The underlying data and scripts have been uploaded to the submission portal to facilitate the review process. Upon publication, all data and scripts underlying this article are publicly available on the Zenodo Platform: <https://doi.org/10.5281/zenodo.23009435>.

References

- Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Vize, C. E., Lynam, D. R., & Olino, T. M. (2023). Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science*, 6(3), Article 25152459231187531. <https://doi.org/10.1177/25152459231187531>
- Barrow, N. (2024). Anthropomorphism and AI hype. *AI and Ethics*, 4(3), 707–711. <https://doi.org/10.1007/s43681-024-00454-1>
- Barry, A. (2001). *Political machines: Governing a technological society*. A&C Black.
- Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, 23(5), 365–368. <https://doi.org/10.1016/j.tics.2019.02.008>
- Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, 18(2), 227–247. <https://doi.org/10.1017/S0140525X00038188>
- Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. Basic Books.

- Boyer, P. (1996). What makes anthropomorphism natural: Intuitive ontology and cultural representations. *The Journal of the Royal Anthropological Institute*, 2(1), 83–97. <https://doi.org/10.2307/3034634>
- Buckwalter, W., & Phelan, M. (2013). Function and feeling machines: A defense of the philosophical conception of subjective experience. *Philosophical Studies*, 166(2), 349–361. <https://doi.org/10.1007/s11098-012-0039-9>
- Burstein, K., Doughtie, E. B., & Raphaeli, A. (1980). Contrastive vignette technique: An indirect methodology designed to address reactive social attitude Measurement. *Journal of Applied Social Psychology*, 10(2), 147–165. <https://doi.org/10.1111/j.1559-1816.1980.tb00699.x>
- Canty, A. J. (2002). *Resampling methods in R: The boot package*. 2. The Newsletter of the R Project.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Colombatto, C., Birch, J., & Fleming, S. M. (2025). The influence of mental state attributions on trust in large language models. *Communications Psychology*, 3(1), 84. <https://doi.org/10.1038/s44271-025-00262-1>
- Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1), Article niae013. <https://doi.org/10.1093/nc/nae013>
- De Brigard, F. (2010). Consciousness, attention and commonsense. *Journal of Consciousness Studies*, 17(9–10), 189–201.
- De Cruz, H., & De Smedt, J. (2007). The role of intuitive ontologies in scientific understanding – The case of human evolution. *Biology and Philosophy*, 22(3), 351–368. <https://doi.org/10.1007/s10539-006-9036-8>
- De Graaf, M. M. A., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 239–248). <https://doi.org/10.1109/HRI.2019.8673308>
- DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science*, 4(1), Article 2515245920965119. <https://doi.org/10.1177/2515245920965119>
- Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, 68(4), Article 4.
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- Dennett, D. C., & Haugeland, J. (1987). Intentionality. In *The Oxford companion to the mind*. Oxford University Press.
- Deroy, O. (2023). The ethics of terminology: Can we use human terms to describe AI? *Topoi*, 42(3), 881–889. <https://doi.org/10.1007/s11245-023-09934-1>
- Deroy, O., Bacciu, D., Bahrami, B., Della Santina, C., & Hauert, S. (2024). Shared awareness across domain-specific artificial intelligence: An alternative to domain-general intelligence and artificial consciousness. *Advanced Intelligent Systems*, 6(10), Article 2300740. <https://doi.org/10.1002/aisy.202300740>
- 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>
- Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, 65(20), 627–640. <https://doi.org/10.2307/2024316>
- Ford, J. K., & Weissbein, D. A. (1997). Transfer of training: An updated review and analysis. *Performance Improvement Quarterly*, 10(2), 22–41. <https://doi.org/10.1111/j.1937-8327.1997.tb00047.x>
- Geiselmann, R., Deroy, O., & Harris, L. (2024). Shared perceptual decisions exhibit an animacy bias. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. <https://escholarship.org/uc/item/00f769db>
- Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.
- Glikson, E., & Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *The Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 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), 1. <https://doi.org/10.1016/j.cognition.2012.06.007>
- Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development*, 15(2), 103–120. <https://doi.org/10.1111/j.1468-2419.2011.00373.x>
- Guinrich, R. E., & Graziano, M. S. A. (2025). Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. In P. Hacker (Ed.), *Oxford intersections: AI in society* (1st ed.). Oxford: Oxford University Press. <https://doi.org/10.1093/9780198945215.003.0011>
- Haslam, N., & Whelan, J. (2008). Human natures: Psychological essentialism in thinking about differences between people. *Social and Personality Psychology Compass*, 2(3), 1297–1312. <https://doi.org/10.1111/j.1751-9004.2008.00112.x>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259. <https://doi.org/10.2307/1416950>
- Hindennach, S., Shi, L., Miletić, F., & Bulling, A. (2024). Mindful explanations: Prevalence and impact of mind attribution in XAI research. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 170:1–170:43. <https://doi.org/10.1145/3641009>
- Horowitz, M. C., Kahn, L., Macdonald, J., & Schneider, J. (2024). Adopting AI: How familiarity breeds both trust and contempt. *AI & SOCIETY*, 39(4), 1721–1735. <https://doi.org/10.1007/s00146-023-01666-5>
- Jedličková, A. (2025). Ethical approaches in designing autonomous and intelligent systems: A comprehensive survey towards responsible development. *AI & SOCIETY*, 40(4), 2703–2716. <https://doi.org/10.1007/s00146-024-02040-9>
- Karpus, J., Shirai, R., Verba, J. T., Schulte, R., Weigert, M., Bahrami, B., ... Deroy, O. (2025). Human cooperation with artificial agents varies across countries. *Scientific Reports*, 15(1), 10000. <https://doi.org/10.1038/s41598-025-92977-8>
- Komatsu, T., Malle, B. F., & Scheutz, M. (2021). *Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across U.S. And Japan*. Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction, HRI '21 (pp. 63–72). <https://doi.org/10.1145/3434073.3444672>
- Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, 4(1), Article 2515245920951503. <https://doi.org/10.1177/2515245920951503>
- Lee, S., Meertens, N., Milner, E., & Hauert, S. (2026). A framework for the examination of awareness in artificial systems. In A. Jiménez Rodríguez, R. Mestre, C. Chen, A. Mura, E. Barker, P. Verschure, & T. Prescott (Eds.), *Biomimetic and biohybrid systems* (pp. 320–333). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-07448-5_27
- Levin, J. (2022). *The metaphysics of mind. Elements in Philosophy of Mind*. <https://doi.org/10.1017/9781108946803>
- Longin, L., Bahrami, B., & Deroy, O. (2023). Intelligence brings responsibility — even smart AI-assistants are held responsible. *iScience*, Article 107494. <https://doi.org/10.1016/j.isci.2023.107494>
- Machery, E., Olivola, C. Y., Cheon, H., Kurniawan, I. T., Mauro, C., Struchiner, N., & Susianto, H. (2023). Is identity essentialism a fundamental feature of human cognition? *Cognitive Science*, 47(5), Article e13292. <https://doi.org/10.1111/cogs.13292>
- Malle, B. F. (2019). *How many dimensions of mind perception really are there?*
- Medin, D., & Ortony, A. (1989). Comments on part I: Psychological essentialism. In A. Ortony, & S. Vosniadou (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge University Press. <https://doi.org/10.1017/CBO9780511529863.009>
- Meertens, N. (2025). *Animal awareness: A multidimensional mapping of action-perception abilities* (Ta4vd.v1). *PsyArXiv*. <https://doi.org/10.31234/osf.io/ta4vd.v1>
- Metzgar, R., Lombrozo, T., & Graziano, M. S. A. (2025). Lay perspectives on the physical and non-physical nature of consciousness. *Journal of Consciousness Studies*, 32 (11–12), 36–71. <https://doi.org/10.53765/20512201.32.11.036>
- Nelson, T. E., Oxley, Z. M., & Clawson, R. A. (1997). Toward a psychology of framing effects. *Political Behavior*, 19(3), 221–246. <https://doi.org/10.1023/A:1024834831093>
- Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, 17(5), Article e12823. <https://doi.org/10.1111/phc3.12823>
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Bloomsbury Publishing PLC.
- Nyholm, S. (2023). Minding the gap(s): Different kinds of responsibility gaps related to autonomous vehicles and how to fill them. In F. Fossa, & F. Cheli (Eds.), *Connected and automated vehicles: Integrating engineering and ethics* (pp. 1–18). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39991-6_1
- Oktar, K., Handley-Miner, I. J., Kleiman-Weiner, M., Zollman, K. J. S., & Young, L. (2026). Intuitive theories of truth. *Trends in Cognitive Sciences*, 30(3), 197–199. <https://doi.org/10.1016/j.tics.2026.01.001>
- Oktar, K., Lombrozo, T., & Griffiths, T. L. (2024). Learning from aggregated opinion. *Psychological Science*, 35(9), 1010–1024. <https://doi.org/10.1177/09567976241251741>
- Oktar, K., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Dimensions of disagreement: Divergence and misalignment in cognitive science and artificial intelligence. *Decision*, 11(4), 511–522. <https://doi.org/10.1037/dec0000244>
- Palmeri, S., & Pistilli, G. (2026). The cognitive biases that may exacerbate inflationary and deflationary positions about large language models. *Current Opinion in Behavioral Sciences*, 70, Article 101670. <https://doi.org/10.1016/j.cobeha.2026.101670>
- Ryle, G. (1949). *The concept of mind*. (Original work published London: Penguin Books).
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8), 299–309. [https://doi.org/10.1016/S1364-6613\(00\)01506-0](https://doi.org/10.1016/S1364-6613(00)01506-0)
- Schultz, J., & Frith, C. D. (2022). Animacy and the prediction of behaviour. *Neuroscience and Biobehavioral Reviews*, 140, Article 104766. <https://doi.org/10.1016/j.neubiorev.2022.104766>
- Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Simons, P. R. J. (1999). Transfer of learning: Paradoxes for learners. *International Journal of Educational Research*, 31(7), 577–589. [https://doi.org/10.1016/S0883-0355\(99\)00025-7](https://doi.org/10.1016/S0883-0355(99)00025-7)
- Sperber, D. (1982). Apparently irrational beliefs. In M. Hollis, & S. Lukes (Eds.), *Rationality and relativism* (pp. 149–180). MIT Press.
- Sperber, D. (1997). Intuitive and reflective beliefs. *Mind & Language*, 12(1), 67–83. <https://doi.org/10.1111/j.1468-0017.1997.tb00062.x>
- Stein, J.-P., & Ohler, P. (2017). Venturing into the uncanny valley of mind — The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, 160, 43–50. <https://doi.org/10.1016/j.cognition.2016.12.010>
- Stich, S. P., & Nichols, S. (2003). Folk psychology. In *The Blackwell guide to philosophy of mind* (pp. 235–255). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470998762.ch10>
- Sytsma, J. (2014). Attributions of consciousness. *WIREs Cognitive Science*, 5(6), 635–648. <https://doi.org/10.1002/wcs.1320>
- Sytsma, J. (Ed.). (2016). *A companion to experimental philosophy*. Wiley Blackwell.
- Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, 151(2), 299–327. <https://doi.org/10.1007/s11098-009-9439-x>

- Thellman, S., De Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction*, 11(4), 1–51. <https://doi.org/10.1145/3526112>
- Tye, M. (1999). Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind*, 108(432), 705–725. <https://doi.org/10.1093/mind/108.432.705>
- Vaneechoutte, M. (2000). Experience, awareness and consciousness: Suggestions for definitions as offered by an evolutionary approach. *Foundations of Science*, 5(4), 429–456. <https://doi.org/10.1023/A:1011371811027>
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43), 11374–11379. <https://doi.org/10.1073/pnas.1704347114>
- Weisman, K., Legare, C. H., Smith, R. E., Dzokoto, V. A., Aulino, F., Ng, E., ... Luhmann, T. M. (2021). Similarities and differences in concepts of mental life among adults and children in five cultures. *Nature Human Behaviour*, 5(10), 1358–1368. <https://doi.org/10.1038/s41562-021-01184-8>
- Zhang, J., Li, S., Zhang, J.-Y., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the Uncanny Valley. In P.-L. P. Rau (Ed.), *Cross-cultural design. User experience of products, services, and intelligent environments* (pp. 255–268). Springer International Publishing. https://doi.org/10.1007/978-3-030-49788-0_19