Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities People Attribute to Them

Allison Chen[1], **Sunnie S. Y. Kim**[1], **Amaya Dharmasiri**[1], **Olga Russakovsky**[1], **Judith E. Fan**[2]
[1] Department of Computer Science, Princeton University
[2] Department of Psychology, Stanford University
{allisonchen, sunniesuhyoung, amayadharmasiri olgarus}@princeton.edu
jefan@stanford.edu

Abstract

How do people determine whether non-human entities have thoughts and feelings — an inner mental life? Prior work has proposed that people use compact sets of dimensions (e.g., *body-heart-mind*) to form beliefs about familiar kinds, but how do they generalize to novel entities? Here we investigate emerging beliefs about the mental capacities of large language models (LLMs) and how those beliefs are shaped by how LLMs are portrayed. Participants ($N = 470$) watched brief videos that encouraged them to view LLMs as either *machines*, *tools*, or *companions* then took a survey measuring mental capacity attributions. We found that the *companion* group more strongly endorsed statements regarding a broad array of mental capacities that LLMs might possess relative to the *machine* and *tool* groups, suggesting that people's beliefs can be rapidly shaped by context. Our study highlights the need to explore the factors shaping people's beliefs about emerging technologies to promote accurate public understanding.

Keywords: social inference; animacy; human-computer interaction; mind perception; intentional stance

Introduction

Humans face a fundamental challenge in making sense of their world: determining which entities within it are genuine social agents—beings that possess minds and feelings—and which are not (Epley et al., 2010). Previous work has established that people readily attribute various mental capacities (e.g., the ability to form goals, have beliefs, and recognize desires) to other people as a way of accounting for their behavior (Waytz et al., 2010; Jara-Ettinger, 2019; Tamir & Thornton, 2018; van Baar et al., 2022; FeldmanHall & Nassar, 2021). Moreover, these attributions seem to sometimes extend to non-human entities, including animals (Eddy et al., 1993; Urquiza-Haas & Kotrschal, 2015) and certain technological artifacts (Nass & Moon, 2000; Weisman et al., 2017; Rossignoli et al., 2022; Thellman et al., 2022). For example, people sometimes judge cars (Windhager et al., 2012), computers (Nass et al., 1996), and robots with human-like appearances or behaviors (Thellman et al., 2022; Cucciniello et al., 2023; Spatola & Wudarczyk, 2021; Gena et al., 2023) to possess some mental capacities (sometimes referred to as mental states), even if people are aware that these objects are in fact inanimate. Why do people sometimes mistakenly attribute mental capacities to otherwise clearly inanimate entities?

Broadly, there are at least two sets of constraints to help understand these misattributions. The first set might come from limited experience or knowledge from the person making attributions. For example, prior work has found that younger children, who have had strictly less experience interacting with others than older children and adults, often fail in standard false-belief tasks (i.e., understanding others' beliefs can be different from their own) (Baron-Cohen et al., 2013; Gopnik & Astington, 1988; Perner et al., 1987; Birch & Bloom, 2007); struggle to distinguish between intentions, beliefs, or desires (Flavell, 1999; Astington, 1993; Astington & Lee, 1991; Shultz, 2014); and exhibit stronger egocentric biases in understanding mental states (Hayashi & Nishikawa, 2019; Pillow & Henrichon, 1996; Pillow, 1995). With limited experience and knowledge, people are also more susceptible to attribute mental lives to inanimate objects for reasons like emotional attachment (Gjersoe et al., 2015), imaginative play (Smirnova, 2011), the presence of robotic or digital features (Sung, 2018; Kahn Jr et al., 2004), or physical resemblance to humans or animals (Kahn Jr et al., 2004; Manzi et al., 2020). While people typically overcome these mistakes as they gain more experience and converge to culturally shared understandings about mental capacities in various entities (Meltzoff & Gopnik, 2013), even adults can make attribution errors in specific settings (Gao et al., 2010; Thellman et al., 2022).

Because attributing mental capacities requires making inferences from observable features (i.e., appearances and behaviors), a second set of important constraints that may lead to misattribution errors come from sophisticated features of a target entity. Even observers who are aware that shapes are inanimate and do not possess goals or intentions will sometimes report moving shapes as animate and capable of goal-directed behavior (Heider & Simmel, 1944; Gao et al., 2009, 2010; Scholl & Gao, 2013). People also tend to attribute mental capacities and traits to technological artifacts that appear human-like or demonstrate complex goal-directed behavior, despite knowing they are not animate (Rossignoli et al., 2022; Imamura et al., 2015; Thellman et al., 2022; De Graaf & Malle, 2019).

Not only might these constraints impact holistic judgments concerning how "life-like" an entity seems, but they might also shape judgments about different components of that entity's mental life in different ways. A substantial body of prior work suggests that people use multiple dimensions to represent the mental capacities of others, and different entities can display these capacities to varying degrees (Gray et al., 2007; Weisman et al., 2017; Colombatto & Fleming,

2024; Pekçetin et al., 2024; Hindennach et al., 2024; Takahashi et al., 2016; Tamir et al., 2016). One influential study proposed that there are two dimensions along which people attribute mental capacities to other entities: *experience* (e.g., hunger and pain) and *agency* (e.g., self-control and memory) (Gray et al., 2007) In a more recent work, Weisman et al. (2017) argued that actually three dimensions might be necessary to capture the variation in mental capacity attribution, which they dubbed: *body*, *heart*, and *mind*. However, another study focusing on technological artifacts instead found evidence suggesting that two dimensions for *experience* and *intelligence* were sufficient (Colombatto & Fleming, 2024). Taken together, these lines of work nevertheless provide converging evidence that mental capacity attribution likely goes *beyond* making judgments about a single dimension (e.g., animacy). Moreover, they suggest that multiple factors are likely important for informing those judgments, such as prior knowledge on the part of the observer and behavioral complexity on the part of the target.

Newly developed natural language generation technologies, including large language models (LLMs), offer a useful case study for exploring how people determine which mental capacities are appropriate to attribute to novel entities in the real-world where there is substantial uncertainty and ever-changing conditions. First, people have limited experience and knowledge about LLMs. Even the most popular consumer-facing systems (e.g., ChatGPT) only recently became available; thus, many people are still learning about and how to interact with them. As a consequence, people hold widely divergent beliefs about the type of entity these systems really are (Rapp et al., 2024; Martínez & Winter, 2021; Boyle, 2024). Second, these technologies span a wide range of behavioral complexity and continue to evolve rapidly. For example, early versions of GPT (Radford, 2018) were less proficient than current models (e.g., GPT-4o (Hurst et al., 2024) and DeepSeek (Liu et al., 2024)) in tasks such as solving reasoning problems (F. Cheng et al., 2025), shifting our notions of what it means to be intelligent (Mitchell, 2024).

In the face of such a rapidly changing and uncertain technological landscape, people might be especially sensitive to the ways that these technologies are *portrayed*, especially what features of the target are accentuated. Some portrayals—for instance, technical blogs—emphasize the internal mechanisms within these systems (Stoffelbauer, 2023). Other portrayals instead focus on their usefulness as productivity support tools (*Grammarly*, n.d.), and others still might characterize these systems as ones with goals, "who care," and are "always on your side" (*Replika*, n.d.). The differences between such portrayals can be thought of as akin to the well known distinction between `mechanistic`, `functional`, and `intentional` stances that can be adopted towards another entity's behavior (Dennett, 1989; Lombrozo, 2009, 2012; Jahic Pettersson et al., 2020; Kelemen, 2019).

Informed by this axis of variation in portraying LLMs, we explore how video portrayals may influence how people attribute mental capacities to novel complex entities. We develop short informational videos that invite people to adopt one of the three stances towards LLMs: `mechanistic`, `functional`, and `intentional`. The `mechanistic` portrayal presents LLMs primarily as *machines* that generate text; the `functional` portrayal presents them as *tools* to accomplish various tasks; and the `intentional` portrayal as *companions* open to conversation and connection (see Fig. 1).

Experimental Setup

We conducted a large-scale, pre-registered, between-subjects experiment ($N = 470$) to evaluate the effect of different portrayals of LLMs, a novel technological entity, on people's attributions of mental capacities to LLMs. We developed three unique informational videos that capture the respective theoretical stances (`mechanistic`, `functional`, and `intentional`) that people may adopt towards an entity (e.g., LLMs) (Dennett, 1989; Lombrozo, 2012). We recruited laypeople (i.e., those without computer science or artificial intelligence (AI) expertise) as participants and randomly assigned them to one of four conditions:

- **`Mechanistic`**: participants watch a video portraying LLMs as *machines*, describing text generation as next-word-prediction.
- **`Functional`**: participants watch a video portraying LLMs as *tools*, describing potential use cases and suggestions for using LLMs effectively.
- **`Intentional`**: participants watch a video portraying LLMs as *companions*, describing their social abilities.
- **`Baseline`**: participants do not watch any video.

Then, all participants took a survey measuring their attributions of 40 mental capacities to LLMs. This study was approved by our Institutional Review Board (IRB #17151) and pre-registered with AsPredicted: `https://aspredicted.org/vgdm-gjrm.pdf`.

Experimental manipulation: portrayal of LLMs

The three video portrayals encouraged participants to adopt one of the following stances: `mechanistic`, `functional`, and `intentional` (Dennett, 1989; Lombrozo, 2012). They were theoretically motivated by established distinctions between the stances and were reflective of real-world content since many online portrayals are reminiscent of at least one of these stances. The video scripts were designed with two goals in mind: (1) reduce unnecessary variance between portrayals and (2) realistically capture the essence of each portrayal. To satisfy goal (1), we developed the videos to share content (e.g., similar introductions, conclusions, and sections on LLMs learning from data) and sentence structures when possible. The videos were approximately matched in terms of word count and length (<5 minutes) and were recorded with the same narrator, animations, and visual aids for consistency. To satisfy goal (2), we based our content and wording on collected examples of each portrayal from publicly available on-

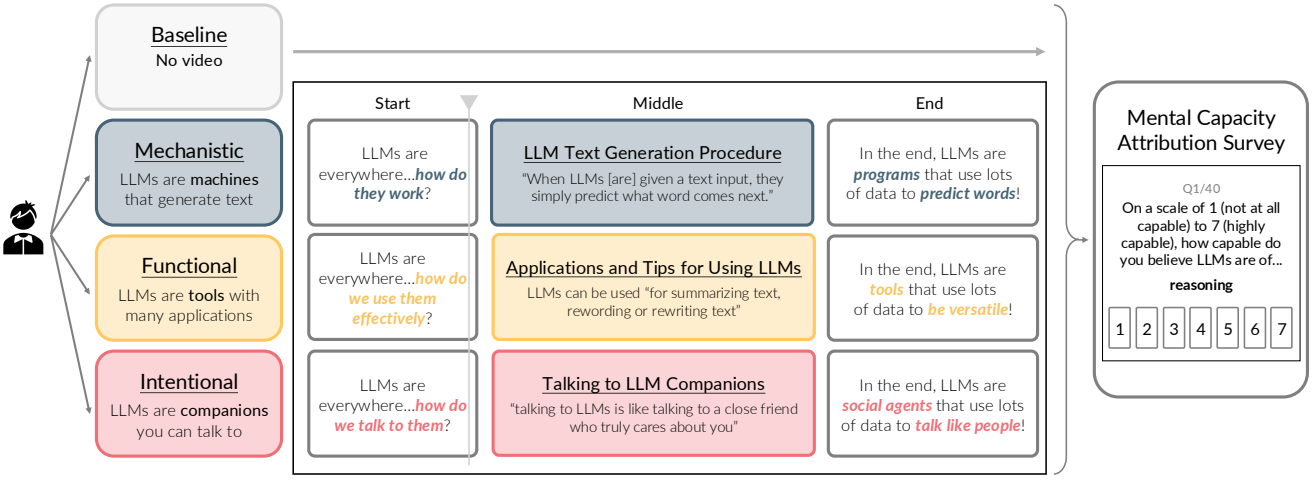

Figure 1: Overview of experimental design. **Left:** Participants are randomly assigned to one of four conditions. **Center:** Summary of the video portrayals shown in each experimental condition. **Right:** Example illustration of the mental capacity attribution survey. All participants rate their attributions of 40 mental capacity items to LLMs on a 7-point Likert scale.

line material about LLMs. Each video contained a content-specific section, highlighted in Fig. 1 (center): LLM text generation procedure in the `mechanistic` portrayal, applications and tips for using LLMs in the `functional`, and the experience of talking to LLM companions in the `intentional`.

Videos were split into three parts. Participants had to stay on each screen for the duration of each section and manually click to move onto the next part.

Measuring mental capacity attribution

We measured the effect of the video intervention on participants' attributions of 40 mental capacity items to LLMs (e.g., reasoning about things) where the scope of the surveyed items extended beyond the content of the videos (see Fig. 2 for the full list). We measured mental capacity attribution using 7-point Likert scale ratings (Thellman et al., 2022; Eddy et al., 1993; Miraglia et al., 2023), and our statements were compiled from two relevant prior works: Weisman et al. (2017) and Colombatto & Fleming (2024). The former is more established and measures people's mental capacity attribution to a variety of entities (e.g., other humans, animals, objects, and robots). The latter is more recent and solely focuses on mental capacity attribution to ChatGPT, a popular consumer-facing LLM. We compiled the final list of 40 mental capacity items after conducting a small-scale qualitative pilot study with eight participants. In this pilot, we sought to understand participants' perception of each item's relevance to LLMs and removed similar items that were deemed repetitive. In the end, our final list of 40 mental capacities contained 34 items from Weisman et al. (2017) and six items from Colombatto & Fleming (2024).

Motivated by prior work (Weisman et al., 2017; Colombatto & Fleming, 2024), we asked participants "On a scale of 1 (not at all capable) to 7 (highly capable), how capable do you believe LLMs are of *X*?" where *X* is a mental capacity (e.g., having intentions). All participants responded to all 40 items, and each item was presented on its own page.

Next, participants responded to an attention check to select which two statements they encountered in the survey from a list of four. Lastly, we measured participant's levels of anthropomorphism of LLMs directly with the following question: "To what extent do you believe LLMs are human-like?" Participants responded using a Likert scale from 1 (Not human-like at all) to 7 (Very human-like).

Participants

We recruited U.S.-based adults without knowledge of AI or related fields on Prolific, an online research platform. To ensure our participants were laypeople, we verified they did not study information/communication technologies, mathematics, statistics, or computer science; did not work in coding, technical writing, or systems administration; and rated themselves as having at *most* some basic knowledge of AI. We utilized a standard sample and paid participants at a rate of $15 per hour. Our target minimum sample size was 90 per condition, based on the G*Power (Erdfelder et al., 1996) sample size calculator with $alpha = 0.05$, $power = 0.9$, $Cohen's\ d = 0.5$ for a two-tailed Mann-Whitney U-test.

We recruited 489 participants, then following our pre-registered exclusion criteria, we excluded 19 participants who failed our attention check, spent less than 80 seconds on the survey, or spent a median time of less than 1 second on each survey item. Post exclusions, we had a total of 470 participants (*age*: $< 35 = 154$, 35-54 = 228, $> 54 = 88$; *gender*: female = 309, male = 149, non-binary = 10, transgender = 1, NA = 1; *education*: Bachelor's or higher = 441, other = 29). We also collected participants' familiarity with various LLM products (e.g., ChatGPT, Gemini, Copilot, and Replika). Most participants either never heard of or never used the products, with the exception of ChatGPT where 47% of participants have used it a few times, but not regularly. In total, there were 118 participants in the `baseline` condition,

116 in the `mechanistic`, 119 in the `functional`, and 117 in the `intentional`.

Analysis procedure

The primary goal of our analysis was to determine whether and how LLM portrayals affect people's attributions of mental capacities to LLMs. We first organized the 40 mental capacity items into one of three categories: *body-heart-mind* based on the dominant factor loadings from Study 4 in Weisman et al. (2017) (see Fig. 2 for assignments). The *body* category is characterized as physiological sensations and self-initiated behaviors, *heart* as emotions and social/moral agency, and *mind* as perceptual/cognitive abilities (Weisman et al., 2017). For mental capacity items from Colombatto & Fleming (2024) that did not have factor loadings for *body-heart-mind* categories, we placed each item into its respective category based on its semantic meaning and relation to the capacity items in each category. Categorization enabled finer-grained analyses of the effect of LLM portrayal and allowed us to identify patterns across related mental capacity items.

According to the pre-registered plan, we fit the following mixed-effects regression model on the collected data: `rating` $\sim$ `portrayal * category + (1 | participant_id)`. The dependent variable `rating` was for each mental capacity item, and the predictors `portrayal`, `category`, and `participant_id` were all categorical. For `portrayal`, there were four possible values (`baseline` (no portrayal), `mechanistic`, `functional`, `intentional`) with `baseline` as the reference, and for `category`, there were three possible values (*body-heart-mind*) with *body* as the reference. Using ANOVA, we compared this model to 1) an equivalent model lacking interaction between `portrayal` and `category`, 2) an equivalent model lacking `portrayal`, and 3) a null model with no fixed effects [1]. Reported means and standard errors are from post-hoc estimated marginal means analysis averaged across all items or within a category.

Results

Attribution is higher for *mind* items than *body* or *heart* items

We first sought to establish the validity of participants' ratings of mental capacity attributions. Separating mental capacity items into the *body-heart-mind* categories from Weisman et al. (2017), we observed that *mind*-related items tended to be the highest, followed by *heart* then *body*. This is demonstrated both in Fig. 2 at the item level and Fig. 3 (*right*) at the category level. We found that the addition of the `category` predictor significantly improved the fit of the regression model predicting attribution ratings over the null baseline ($p < 0.001$), suggesting that participants attribute mental capacity items of different categories differently to LLMs.

[1]The pre-registration originally include random effects for each mental capacity item (`(1 | item)`) in the baseline models. However, this caused errors with ANOVA because these baselines were not nested versions of the full model, leading us to drop this term.

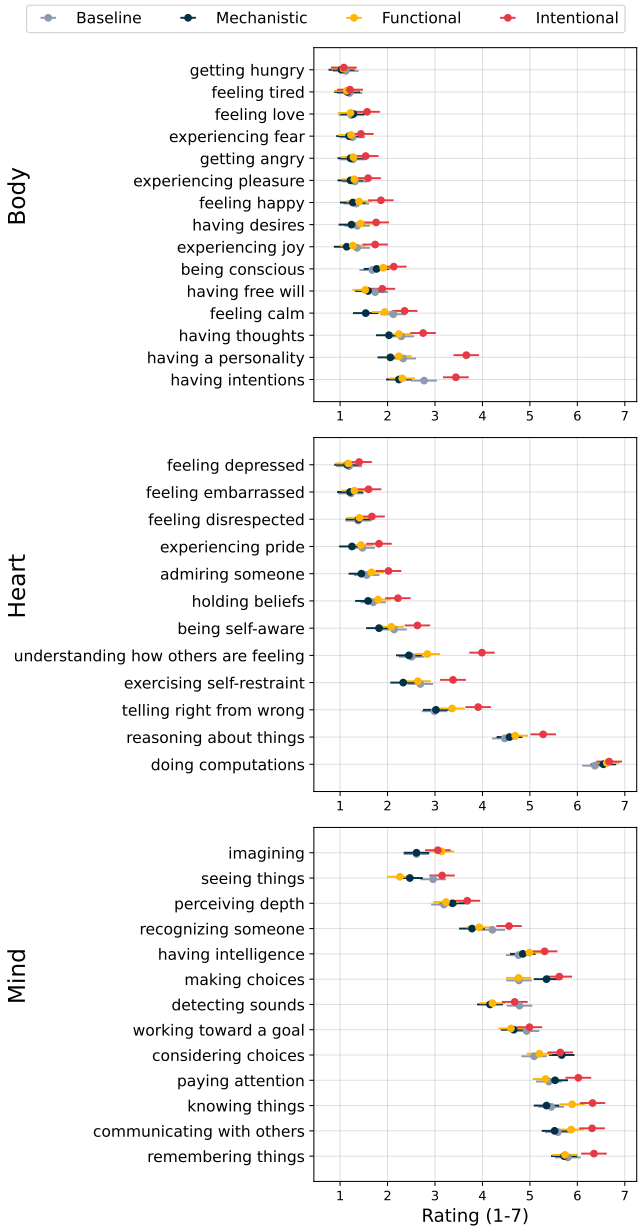

Figure 2: Mean ratings for all items. Different portrayal conditions are indicated by different colors. Items are grouped by *body-heart-mind* categories and sorted by increasing magnitude of mean rating in the `baseline` condition. Error bars represent 95% confidence intervals. *Mind* items tend to be the highest, followed by *heart* then *body*.

Further, we conducted post-hoc pairwise comparisons of the mean ratings for each item category with Tukey adjustments and observed significant differences for all pairwise comparisons. Specifically, the mean rating for the 13 items in the *mind* category (M (mean) $= 4.68$, SE (standard error) $= 0.04$) was reliably higher than the mean rating for the 12 *heart* items ($M = 2.63$, $SE = 0.04$, $p < 0.001$) and the mean rating for the 15 *body* items ($M = 1.66$, $SE = 0.04$, $p < 0.001$). Additionally, the mean rating for *heart* was reliably higher than for *body* ($p < 0.001$). Given that LLMs per-

form well on tasks that typically require humans to perform cognitive functions, the attribution of *mind*-related capacities over *heart* and *body* capacities is not surprising.

`Intentional` portrayal of LLMs increases mental capacity attribution overall

Next, we sought to evaluate the effect of our experimental manipulation of the portrayal condition and found that including `portrayal` as a predictor reliably improved the fit of the regression model predicting attribution ratings ($p < 0.001$). However, we did not find a reliable interaction between `portrayal` (`baseline`, `mechanistic`, `functional`, `intentional`) and `category` (*body-heart-mind*), indicating that the magnitude of the portrayal condition's effect did not reliably differ across different categories of capacity items.

We additionally conducted post-hoc pairwise comparisons of mean ratings in each condition with Tukey adjustments. We found that the mean rating calculated over all items was higher in the `intentional` portrayal condition ($M = 3.37$, $SE = 0.07$) than in the `baseline` ($M = 2.89$, $SE = 0.07$, $p < 0.001$), `mechanistic` ($M = 2.80$, $SE = 0.07$, $p < 0.001$), and `functional` ($M = 2.90$, $SE = 0.07$, $p < 0.001$) conditions. These results (Fig. 3, *left*) suggest that the overall effect on mental capacity attributions might have been driven by the `intentional` condition.

`Intentional` portrayal increases mental capacity attribution within *body-heart-mind* categories

In order to determine whether the effect of the `intentional` portrayal was primarily driven by certain categories of mental capacity items, we again separated the mental capacities into the *body-heart-mind* categories. We then examined differences across portrayal conditions within each category. From this finer-grained analysis, we found that the effect of the portrayal condition is reliable within all categories of items, as demonstrated in Fig. 3 (*right*), suggesting the main effect of portrayal was not driven by only one category of items.

Within each item category (*body-heart-mind*), we additionally conducted post-hoc pairwise comparisons of the mean ratings between each condition. For items in the *body* category, we found the mental capacity attribution of participants in the `intentional` portrayal condition ($M = 2.00$, $SE = 0.08$) to be reliably higher than the `baseline` ($M = 1.63$, $SE = 0.08$, $p = 0.003$), `mechanistic` ($M = 1.46$, $SE = 0.08$, $p < 0.001$), and `functional` ($M = 1.57$, $SE = 0.08$, $p = 0.001$) conditions. For items in the *heart* category, mental capacity attribution in the `intentional` portrayal condition ($M = 3.05$, $SE = 0.08$) was also reliably higher than the `baseline` ($M = 2.48$, $SE = 0.08$, $p < 0.001$), `mechanistic` ($M = 2.40$, $SE = 0.08$, $p < 0.001$), and `functional` ($M = 2.58$, $SE = 0.08$, $p < 0.001$) conditions. Similarly, for items in the *mind* category, mental capacity attribution was higher in the `intentional` portrayal condition ($M = 5.05$, $SE = 0.08$) than `baseline` ($M = 4.58$, $SE = 0.08$, $p < 0.001$), `mechanistic` ($M = 4.54$, $SE = 0.08$, $p < 0.001$), and `functional` ($M = 4.55$, $SE = 0.08$, $p < 0.001$). Thus, because the `intentional` portrayal led to an increase in mental capacity attribution within all three categories of items (*body-heart-mind*), this suggests the effect is not solely driven by one type of mental capacity but by a broad range of capacities.

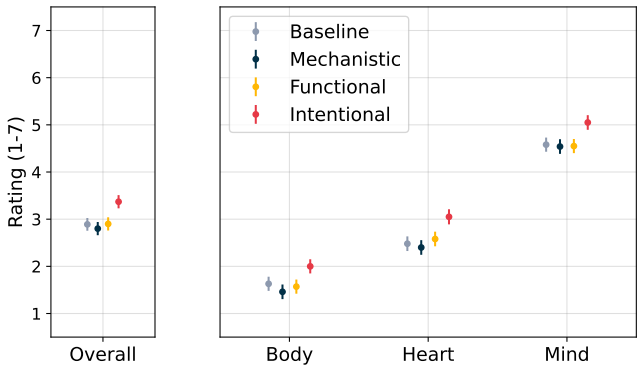

Figure 3: Mean mental state attribution ratings across conditions for all items (**left**) and for items in each *body-heart-mind* category (**right**). Error bars represent 95% confidence intervals. The `intentional` portrayal condition reliably increases mean ratings for items overall and within each category, and participants tend to rate *mind* related items higher than *body* and *heart*.

`Intentional` portrayal affects items beyond the content of the video

Our analysis thus far has revealed that the `intentional` portrayal can have a reliable effect on increasing people's mental capacity attributions to LLMs. However, an important question remains, especially given that the `intentional` portrayal is more suggestive of mental capacities than others. Could this be due to overlap between the video and the surveyed mental capacities?

To answer this, we conducted an exploratory robustness analysis to understand whether the difference between the `intentional` portrayal and `baseline` condition was solely driven by the mental capacities referenced in the `intentional` video or if its effect extends beyond the video content.

Members of our research team first independently identified items that were mentioned in the `intentional` video then resolved differences through discussion. In the end, five items were identified to be mentioned in the video: "knowing things," "having intelligence," "understanding how others are feeling," "having a personality," and "communicating with others." We then fit a new mixed-effects regression model, similar to before but replacing item `category` with a binary categorical predictor `mentioned`, giving us the following model: `rating` $\sim$ `portrayal * mentioned + (1 | participant_id)`. Using only the `baseline` and `intentional` portrayal conditions, we repeated the ANOVA using incrementally complex mixed-effects models and conducted pairwise comparisons with Tukey adjustments.

Including `mentioned` as a predictor significantly improved the model fit of predicting attribution ratings over a null model ($p < 0.001$). Additionally, including the interaction term `portrayal * mentioned` had a reliable improvement of model fit as well ($p < 0.001$). For the five mentioned items, the pairwise comparisons demonstrated a reliable difference between the `intentional` ($M = 5.12$, $SE = 0.11$) and `baseline` ($M = 4.13$, $SE = 0.11$, $p < 0.001$) portrayal conditions, as expected. Further, for the remaining *un*mentioned items, we also observed a significant difference between the `intentional` ($M = 3.05$, $SE = 0.07$) and `baseline` ($M = 2.66$, $SE = 0.07$, $p < 0.001$) conditions. This suggests that the effect of the `intentional` portrayal can carry over to items beyond its content, and this is highlighted in Fig. 2 for items such as "having intentions" (under *body*) or "telling right from wrong" (under *heart*).

Mental capacity attribution is positively correlated with anthropomorphism

In our final analysis, we explored the relationship between mental capacity attributions and anthropomorphism of LLMs. Anthropomorphism is the attribution of human-like qualities—not necessarily physically observable—to a non-human entity (Epley et al., 2007; Kim & Im, 2023). While anthropomorphism and mental capacity attribution are related and often studied together (Kawai et al., 2023; Miraglia et al., 2023), they are not the same. The two can overlap when the mental capacities considered are unique to humans (Thellman et al., 2022), but anthropomorphism can also measure non-mental traits, (e.g., moving rigidly vs. elegantly) (Bartneck et al., 2009).

We conducted an exploratory correlational analysis between participants' mean mental capacity attribution ratings and anthropomorphism responses (i.e., 7-point Likert scale ratings of "To what extent do you believe LLMs are human-like?"). We observed the Spearman rank correlation coefficient to be $\rho = 0.48$ ($p < 0.001$) suggesting a moderately positive correlation: as participants' mean ratings of mental attributions increased, their anthropomorphism ratings also tended to increase.

We additionally performed a finer-grained analysis of the correlation between the mean attribution rating of each *body-heart-mind* category with participants' anthropomorphism rating. We found that the correlation between *body* items and anthropomorphism ($\rho = 0.49$, $p < 0.001$) and *heart* items and anthropomorphism ($\rho = 0.49$, $p < 0.001$) are also moderately positive. However, the correlation between *mind* items and anthropomorphism ($\rho = 0.36$, $p < 0.001$) is slightly lower but still exhibits mid-range correlation, suggesting that anthropomorphism is more closely connected to *body* and *heart* than it is to *mind* capacities.

Discussion

In the present work, we leverage LLMs as a case study to explore how varying portrayals of a novel entity under real-world complexity and uncertainty can shape the mental capacities people attribute to it. We present participants with varying portrayals of LLMs—`mechanistic`, `functional`, and `intentional`—and measure their attributions of mental capacities to LLMs. Our results primarily suggest that the `intentional` portrayal can increase attributions, both for items overall and within *body-heart-mind* categories, while the `mechanistic` and `functional` portrayals do not have reliable effects. We also find participants more readily attribute *mind*-related items to LLMs than *body* or *heart*, the effect of the `intentional` portrayal extends beyond the scope of its content, and that mental capacity attribution is moderately positively correlated with anthropomorphism.

Participants' increased attributions of mental capacities in the `intentional` portrayal condition may be explained by people's tendency to interpret behavior through *intent* (Thellman et al., 2022; Waytz et al., 2010; Heider & Simmel, 1944)—the way they experience the world. It also suggests that people may be especially sensitive to human-like portrayals of novel complex entities when judging which mental capacities the entities may possess. Further, participants' generalization from the `intentional` portrayal may indicate that portrayals can have a far-reaching effect on laypeople's beliefs. Compared to prior work (Weisman et al., 2017), people are now more likely to attribute mental capacities to LLMs than they were towards computers or robots just a few years ago. If this trend continues, it may influence both how people interact with technology (M. Cheng et al., 2024) and with one another (Guingrich & Graziano, 2024). This may be further influenced by the observed correlation between anthropomorphism and attributions of mental capacities in our study, which supports prior work in suggesting the two are closely related but not identical (Kawai et al., 2023; Miraglia et al., 2023).

This work demonstrates that people's attributions of mental capacities to novel technological entities can be shaped by the way the entities are portrayed, but we address a few limitations. First, while self-report measures are straight-forward, they offer limited insight into people's real-world behaviors; we encourage future works to explore how portrayals may impact how people use and interact with these systems. Further, our work utilizes informational video portrayals about LLMs, but there are a variety of mediums (e.g., interaction interfaces, social media) as well as novel and complex entities that future studies can investigate.

Overall, our work serves to understand the role of portrayals in shaping people's beliefs about a novel complex entity. We simultaneously contribute to the *timeless* question of people's mental capacity attribution tendencies and the *timely* question of technology's role in society. As AI technology becomes increasingly sophisticated and integrated into daily life, we encourage further research in this area to develop both scientific knowledge and safe adoption of these systems.

Data, Materials, and Software Availability. Videos, survey materials, pilot information, data, and analysis code can be found in the OSF repository (`https://osf.io/gjef3/`)

Acknowledgements. We thank the participants for their time and effort. We also thank all those who provided thoughtful feedback and discussion, especially members of the Princeton Visual AI Lab, Stanford Cognitive Tools Lab, and the anonymous reviewers. We acknowledge support from the Princeton Cognitive Science Program, NSF Graduate Research Fellowship Program (AC, SK), Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), NSF CAREER #2047191 (JEF), NSF DRL #2400471 (JEF), and a Hoffman-Yee Grant from the Stanford Center for Human-Centered Artificial Intelligence (JEF).

References

Astington, J. (1993). *The Child's Discovery of the Mind*. Harvard University Press.

Astington, J., & Lee, E. (1991). What Do Children Know About Intentional Causation. In *Biennial Meeting of the Society for Research in Child Development, Seattle, WA*.

Baron-Cohen, S., Tager-Flusberg, H., & Lombardo, M. (2013). *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*. OUP Oxford.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*, 71–81.

Birch, S. A., & Bloom, P. (2007). The Curse of Knowledge in Reasoning About False Beliefs. *Psychological Science*, *18*(5), 382–386.

Boyle, J. (2024). *The Line: AI and the Future of Personhood*. MIT Press.

Cheng, F., Li, H., Liu, F., van Rooij, R., Zhang, K., & Lin, Z. (2025). Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*.

Cheng, M., DeVrio, A., Egede, L., Blodgett, S. L., & Olteanu, A. (2024). " i am the one and only, your cyber bff": Understanding the impact of genai requires understanding the impact of anthropomorphic ai. *arXiv preprint arXiv:2410.08526*.

Colombatto, C., & Fleming, S. M. (2024). Folk Psychological Attributions of Consciousness to Large Language Models. *Neuroscience of Consciousness*, *2024*(1), niae013.

Cucciniello, I., Sangiovanni, S., Maggi, G., & Rossi, S. (2023). Mind Perception in HRI: Exploring Users' Attribution of Mental and Emotional States to Robots with Different Behavioural Styles. *International Journal of Social Robotics*, *15*(5), 867–877.

De Graaf, M. M., & 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).

Dennett, D. C. (1989). *The Intentional Stance*. MIT press.

Eddy, T. J., Gallup Jr, G. G., & Povinelli, D. J. (1993). Attribution of Cognitive States to Animals: Anthropomorphism in Comparative Perspective. *Journal of Social Issues*, *49*(1), 87–101.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864.

Epley, N., Waytz, A., et al. (2010). Mind perception. *Handbook of social psychology*, *1*(5), 498–541.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A General Power Analysis Program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1–11.

FeldmanHall, O., & Nassar, M. R. (2021). The Computational Challenge of Social Learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057.

Flavell, J. H. (1999). Cognitive Development: Children's Knowledge About the Mind. *Annual Review of Psychology*, *50*(1), 21–45.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The Wolfpack Effect: Perception of Animacy Irresistibly Influences Interactive Behavior. *Psychological Science*, *21*(12), 1845–1853.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The Psychophysics of Chasing: A Case Study in the Perception of Animacy. *Cognitive Psychology*, *59*(2), 154–179.

Gena, C., Manini, F., Lieto, A., Lillo, A., & Vernero, F. (2023). Can Empathy Affect the Attribution of Mental States to Robots? In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 94–103).

Gjersoe, N. L., Hall, E. L., & Hood, B. (2015). Children Attribute Mental Lives to Toys When They are Emotionally Attached to Them. *Cognitive Development*, *34*, 28–38.

Gopnik, A., & Astington, J. W. (1988). Children's Understanding of Representational Change and its Relation to the Understanding of False Belief and the Appearance-reality Distinction. *Child Development*, 26–37.

Grammarly. (n.d.). Retrieved from `https://www.grammarly.com/ai`

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619.

Guingrich, R. E., & Graziano, M. S. (2024). Ascribing consciousness to artificial intelligence: human-ai interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology*, *15*, 1322781.

Hayashi, H., & Nishikawa, M. (2019). Egocentric Bias in Emotional Understanding of Children and Adults. *Journal of Experimental Child Psychology*, *185*, 224–235.

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259.

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), 1–43.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... others (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Imamura, Y., Terada, K., & Takahashi, H. (2015). Effects of Behavioral Complexity on Intention Attribution to Robots. In *Proceedings of the 3rd International Conference on Human-Agent Interaction* (pp. 65–72).

Jahic Pettersson, A., Danielsson, K., & Rundgren, C.-J. (2020). 'Traveling Nutrients': How Students Use Metaphorical Language to Describe Digestion and Nutritional Uptake. *International Journal of Science Education*, *42*(8), 1281–1301.

Jara-Ettinger, J. (2019). Theory of Mind as Inverse Reinforcement Learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Kahn Jr, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2004). Robotic Pets in the Lives of Preschool Children. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1449–1452).

Kawai, Y., Miyake, T., Park, J., Shimaya, J., Takahashi, H., & Asada, M. (2023). Anthropomorphism-based Causal and Responsibility Attributions to Robots. *Scientific Reports*, *13*(1), 12234.

Kelemen, D. (2019). The Magic of Mechanism: Explanation-based Instruction on Counterintuitive Concepts in Early Childhood. *Perspectives on Psychological Science*, *14*(4), 510–522.

Kim, J., & Im, I. (2023). Anthropomorphic Response: Understanding Interactions Between Humans and Artificial Intelligence Agents. *Computers in Human Behavior*, *139*, 107512.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... others (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lombrozo, T. (2009). Explanation and Categorization: How "Why?" Informs "What?". *Cognition*, *110*(2), 248–253.

Lombrozo, T. (2012). Explanation and Abductive Inference. *Oxford handbook of thinking and reasoning*, 260–276.

Manzi, F., Peretti, G., Di Dio, C., Cangelosi, A., Itakura, S., Kanda, T., ... Marchetti, A. (2020). A Robot is not Worth Another: Exploring Children's Mental State Attribution to Different Humanoid Robots. *Frontiers in Psychology*, *11*, 2011.

Martínez, E., & Winter, C. (2021). Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection. *Frontiers in Robotics and AI*, *8*, 788355.

Meltzoff, A. N., & Gopnik, A. (2013). Learning About the Mind from Evidence: Children's Development of Intuitive Theories of Perception and Personality. *Understanding Other Minds*, *3*, 19–34.

Miraglia, L., Peretti, G., Manzi, F., Di Dio, C., Massaro, D., & Marchetti, A. (2023). Development and Validation of the Attribution of Mental States Questionnaire (AMS-Q): A Reference Tool for Assessing Anthropomorphism. *Frontiers in Psychology*, *14*, 999921.

Mitchell, M. (2024). *The turing test and our shifting conceptions of intelligence* (Vol. 385) (No. 6710). American Association for the Advancement of Science.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can Computers be Teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103.

Pekçetin, T. N., Acarturk, C., & Urgen, B. A. (2024). Investigating Mind Perception in HRI through Real-Time Implicit and Explicit Measurements. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 139–141).

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' Difficulty with False Belief: The Case for a Conceptual Deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137.

Pillow, B. H. (1995). Two Trends in the Development of Conceptual Perspective-taking: An Elaboration of the Passive-active Hypothesis. *International Journal of Behavioral Development*, *18*(4), 649–676.

Pillow, B. H., & Henrichon, A. J. (1996). There's More to the Picture than Meets the Eye: Young Children's Difficulty Understanding Biased Interpretation. *Child Development*, *67*(3), 803–819.

Radford, A. (2018). Improving Language Understanding by Generative Pre-Training.

Rapp, A., Boldi, A., Curti, L., Perrucci, A., & Simeoni, R. (2024). How do People Ascribe Humanness to Chatbots? An Analysis of Real-world Human-agent Interactions and a Theoretical Model of Humanness. *International Journal of Human–Computer Interaction*, *40*(19), 6027–6050.

Replika. (n.d.). Retrieved from `https://replika.com/`

Rossignoli, D., Manzi, F., Gaggioli, A., Marchetti, A., Massaro, D., Riva, G., & Maggioni, M. (2022). Attribution of Mental State in Strategic Human-Robot Interactions.

Scholl, B. J., & Gao, T. (2013). Perceiving Animacy and Intentionality: Visual Processing or Higher-level Judgment. *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*, *4629*, 197–229.

Shultz, T. R. (2014). Development of the Concept of Intention. In *Development of Cognition, Affect, and Social Relations* (pp. 131–164). Psychology Press.

Smirnova, E. O. (2011). Character Toys as Psychological Tools. *International Journal of Early Years Education*, *19*(1), 35–43.

Spatola, N., & Wudarczyk, O. A. (2021). Ascribing Emotions to Robots: Explicit and Implicit Attribution of Emotions and Perceived Robot Anthropomorphism. *Computers in Human Behavior*, *124*, 106934.

Stoffelbauer, A. (2023, Oct). *How Large Language models Work.* Medium. Retrieved from https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

Sung, J. (2018). How Young Children and Their Mothers Experience Two Different Types of Toys: A Traditional Stuffed Toy Versus an Animated Digital Toy. In *Child & Youth Care Forum* (Vol. 47, pp. 233–257).

Takahashi, H., Ban, M., & Asada, M. (2016). Semantic Differential Scale Method can Reveal Multi-dimensional Aspects of Mind Perception. *Frontiers in sPychology*, *7*, 1717.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural Evidence that Three Dimensions Organize Mental State Representation: Rationality, Social impact, and Valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194–199.

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 (THRI)*, *11*(4), 1–51.

Urquiza-Haas, E. G., & Kotrschal, K. (2015). The Mind Behind Anthropomorphic Thinking: Attribution of Mental States to Other Species. *Animal Behaviour*, *109*, 167–176.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent Motives Guide Structure Learning During Adaptive Social Choice. *Nature Human Behaviour*, *6*(3), 404–414.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and Consequences of Mind Perception. *Trends in Cognitive Sciences*, *14*(8), 383–388.

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.

Windhager, S., Bookstein, F. L., Grammer, K., Oberzaucher, E., Said, H., Slice, D. E., ... Schaefer, K. (2012). "Cars Have Their Own Faces": Cross-cultural Ratings of Car Shapes in Biological (Stereotypical) Terms. *Evolution and Human Behavior*, *33*(2), 109–120.