False belief understanding for mean versus nice interactions in children and adults

Lily Tsoi[1], J. Kiley Hamlin[2], Adam Waytz[3], Andrew Scott Baron[2], Liane Young[4]

[1]Department of Psychology, Princeton University

[2] Department of Psychology, University of British Columbia

[3] Management and Organizations Department, Northwestern University

[4] Department of Psychology, Boston College

Corresponding Author:

Lily Tsoi
Department of Psychology
Princeton University
Peretsman Scully Hall
Princeton, NJ 08540
Email: ltsoi@princeton.edu

Abstract

Two sets of studies test whether people reason more accurately about the mental states of agents involved in negative versus positive social interaction. In a large sample of 3- to 5-year-olds, we find a generalized theory of mind (ToM) advantage for negative interactions: that is, preschool children are better at understanding the false beliefs of people described in negative versus positive social interactions. We find a different pattern of results for adults, who show no difference in performance across negative, positive, and neutral interaction contexts. Together, these results support the idea that social context influences early ToM deployment, though there is no evidence of this effect in adulthood.

Keywords: theory of mind; social cognition; social development

Introduction

The capacity to reason about the mental states of others, often referred to as “theory of mind” (ToM), enables effective social interaction. Considering what an agent thinks or believes helps people understand and evaluate their past behaviors and predict their future behaviors. As such, ToM also helps people identify agents as potential enemies and allies (Young & Waytz, 2013). Despite these benefits, evidence shows that people do not fully deploy their capacity for ToM at all times. Indeed, some work reveals that ToM engagement is effortful and prone to slippage (Keysar, 2007), while other studies show that ToM is differently deployed for different groups of people (e.g., ingroup members versus outgroup members; (Kelman, 1973; Leyens et al., 2000; McLoughlin & Over, 2017; Opotow, 1990; Struch & Schwartz, 1989). One possible explanation for these findings is that some contexts elicit ToM more so than others; the present work tests this explanation by investigating whether people deploy their capacity for ToM more when encountering negative versus positive social contexts.

The main hypothesis is that people deploy their capacity for ToM more for processing negative versus positive interactions. In domains like memory and attention, both children and adults display a valence-driven asymmetry or negativity bias: people pay more attention to and better remember negative versus neutral or positive information (Vaish et al., 2008). This bias also appears for some aspects of ToM: indeed, children and adults attribute more agency to people engaging in negative behaviors (e.g., reporting that actors intended the bad outcome) as opposed to neutral or positive behaviors (Hamlin & Baron, 2014; Knobe, 2003; Leslie et al., 2006; Morewedge, 2009; Waytz et al., 2010). Notably, these studies focus on people’s attributions of mental states to the agent engaging in the action. The present study provides a conservative test of this hypothesis by answering a question that, to our knowledge, has not yet been explored: whether negative (mean) social interactions, as opposed to positive (nice) interactions, lead to a general boost in ToM that applies beyond the negative actor. Indeed, a general ToM boost during negative interactions might serve a useful function: to help people gain control over a bad situation (Peeters & Czapinski, 1990).

In the current work, we test whether encountering mean behaviors, as compared to nice behaviors, lead people to more broadly consider the minds of agents in the given interaction context. We take a developmental approach to this question, in testing preschool children (ages 3-5) and in adults. We test children between the ages of 3 to 5 years because children in this age range begin to acquire an explicit understanding of false beliefs, a key marker of mature ToM (for a review, Wellman, Cross, & Watson, 2001). Evidence of a generalized ToM boost for mean behaviors in preschool children would suggest its relatively early emergence. Testing adults would further reveal whether any ToM boost found in childhood persists through adulthood. Importantly, when we refer to positive or negative interactions, we refer to the actor's intent and not the interaction's outcome, which is the same across both conditions in Study 1 and across all three conditions (mean, nice, baseline) in Study 2.

To test preschool children (Study 1), we made a simple modification to a classic false belief task targeted at children in this age group: the Sally-Anne task (Baron-Cohen et al., 1985; adapted from Wimmer & Perner, 1983). In this task, Sally puts a ball in a basket and leaves the room. While Sally is away, Anne comes in and moves the ball to a different location, for example, the closet. We then examined whether preschool children could correctly answer where Sally will think the ball is when she returns to the room. In our modified task, Anne hides the ball for either a mean reason ("Anne is not a very nice girl, so she wants to trick Sally by moving the ball to the closet") or a nice reason ("Anne is a very nice girl, so she wants to help Sally by moving the ball to the closet"). Importantly, the outcome (the ball is in fact in a different location), the question (where Sally thinks the ball is when she returns to the room), and the correct answer (that Sally will think the ball is in the basket) are the same across the mean and nice conditions, and also the same as in the classic version of the task.

The experimental approach in the current work departs from prior work in a few important ways. First, unlike prior work, the current work focuses on children's understanding of false beliefs in mean versus nice contexts for people other than the mean versus nice agents themselves. Second, unlike prior work revealing greater false belief understanding in children by directly instructing the children themselves to deceive another person (Chandler et al., 1989; Chandler & Hala, 1994; Davis, 2001; Hala et al., 1991; Sullivan & Winner, 1993; Wellman et al., 2001), our studies do not ask children to engage in any deception or trickery. Instead, we test whether observing an agent acting in a mean versus nice manner, in the form of a puppet show, is sufficiently effective in prompting children to consider the false beliefs of the target. By focusing on third-party observations of others' interactions, and by focusing on children's ToM for the target and not mean versus nice actor, we provide a conservative test of our hypothesis that negative social contexts elicit ToM. Third, aside from differences in study design, our study employs a much larger participant sample (N=537), allowing for a better estimate of the size of the effect. Finally, to test the same question in adults (Study 2), we presented adults with a series of similar vignettes, again manipulating whether Anne or another agent in a similar position performed an action with mean or nice intentions. Whereas prior work has examined how much adults engage in ToM across contexts (Tsoi et al., 2018, 2016), this study tests an aspect of ToM understudied in adults: ToM *accuracy* across contexts. Indeed, the accuracy of ToM and not amount of ToM is often what matters in sustaining real-life interactions. Given that the Sally-

Anne task is targeted at preschool children, we aimed to reduce any possible ceiling effects for adults by constructing a greater number of vignettes in the same vein as the Sally-Anne task and making adults infer the meanness or niceness of agents in the interactions.

Study 1: False belief understanding across mean and nice interactions in children

Methods

Participants

Six hundred and forty-three participants were recruited from a community-based science center and tested in a soundproof room dedicated to behavioral science research. Participants were recruited for two separate tasks: the task in the present study and a task for a separate study not presented in this paper. The sample size was determined based on the conditions of that separate study (8 cells total: 2 cells for 3-year-olds, 6 cells for 4-year-olds, and 2 cells for 5-year-olds; 60 participants per cell, roughly 30 per gender). We decided on this sample size with the assumption that dropouts would be substantial, which is typical in science centers (e.g., Workshop on Research and Museum Partnerships, Cognitive Development Society Meeting, October 2015; Gonzalez, Steele, & Baron, 2017; Gonzalez, Dunlop, & Baron, 2016). We aimed to stop when we estimated that we reached our target sample. Because the policy at the science center is to not turn away participants if they want to participate, we sometimes exceeded our stopping rule for overpopulated ages.

Of the 635 participants who were recruited, 98 were excluded (exclusion criteria are reported in Supplementary Material). The final sample consisted of 537 participants: 147 three-year-olds (74 females), 266 four-year-olds (137 females), and 124 five-year-olds (55 females). A legal guardian provided informed consent for all children. The study was approved by [removed for blind review].

Procedure

Participants were introduced to a modified version of the Sally-Anne task (Baron-Cohen, 1985) in the form of a live puppet show. Participants were assigned to either the *Nice Anne* condition or the *Mean Anne* condition (counterbalanced across participants; see complete script in Supplementary Material). In the *Nice Anne* condition, Anne, who is a nice girl, moves Sally's ball from the basket to the closet while Sally is away because she wanted to help Sally. In the *Mean Anne* condition, Anne, who is a mean girl, moves Sally's ball from the basket to the closet while Sally is away because she wanted to trick Sally. After the puppet show, participants are asked the following questions: (1) Where will Sally look?, (2) Where does Sally think her ball is?, (3) Should Anne and Sally be friends?, (4) Is Anne a nice girl or not a nice girl?, (5) Is Sally a nice girl or not a nice girl?. The order in which Questions 1 and 2 were asked was counterbalanced across participants. The focus of this paper is on responses to Questions 1, 2, and 4; descriptive statistics for responses to the remaining questions are provided in Supplementary Material.

Analyses

Analyses were conducted in R (version 3.6; R Core Team, 2019). Responses were analyzed using a Generalized Linear Mixed Model (GLMM) with binary response terms (correct [1] or incorrect [0]). We were primarily interested in whether responses (correct versus incorrect) depended on Condition and Age. Because the standard question ("Where will Sally look?") might be difficult in that it requires integrating a belief about Sally's mental state as well as knowledge of how mental states can affect motor behavior, we also included a question probing just the belief ("Where does Sally think her ball is?"); hence, we included Question Type as a predictor. The order in which these two questions were presented was counterbalanced across participants; the counterbalancing order was added as a regressor. Our full model included the following regressors: Condition (Mean Anne or Nice Anne), Age Category (three, four, or five), Gender (male or female), Question Type ("Where will Sally look" or "Where does Sally think her ball is?"; manipulated within-participant), and Counterbalancing Order (first or second). We examined the three-way interaction between Condition, Question Type, and Age Category, the four two-way interactions (Condition x Question Type, Condition x Age Category, Condition x Counterbalancing Order, and Question Type x Age Category), and the main effects of these variables. Participant was entered as a random effect. To assess the importance of our predictors of interest, we performed likelihood ratio tests (LRTs) and examined whether the model including a given term provided a significantly better fit to the data than the model without that term. For all analyses with Age Category as a predictor, we also conducted the same analyses with age as a continuous measure. A sensitivity analysis revealed that with $N = 537$ we had 80% power and alpha of 0.05 to detect an effect of Condition with a fixed effect estimate as extreme as 0.44.

Results

We observed overall differences between responses to the *Mean Anne* and *Nice Anne* conditions (Figure 1). That is, the log odds of providing a correct response was significantly greater for the *Mean Anne* condition than for the *Nice Anne* condition ($\chi^2(1) = 7.168$, $p = 0.007$; fixed effect estimate of -0.46, with the *Mean Anne* condition as baseline). Importantly, the main effect of Condition was not qualified by an interaction with other predictors: likelihood ratio tests revealed no three-way interaction between Condition, Question Type, and Age Category ($\chi^2(2) = 3.683$, $p = 0.16$), no two-way interactions between Condition and Age Category ($\chi^2(2) = 0.815$, p

= 0.67), Condition and Question Type ($\chi^2(1) = 1.517$, $p = 0.22$), or Condition and Counterbalancing Order ($\chi^2(1) = 3.266$, $p = 0.071$). In all, 17.3% of the variance in our response term (correct or incorrect) is explained by the fixed factors in our full model, and 38.9% of the variance was explained by both the fixed and random factors in our model. These marginal and conditional R^2 values were calculated based on methods described by Nakagawa and colleagues (2017). For the sake of completeness, analyses at individual levels of Age Category, Condition, and Counterbalancing Order are reported in Supplementary Material.

We further examined the robustness of this effect in two ways. First, we entered age as a continuous variable instead of a categorical variable; doing this revealed a similarly significant main effect of Condition ($\chi^2(1) = 6.593$, $p = 0.01$; see Supplementary Material for more details). Second, we restricted our analyses to participants who responded to the question "Is Anne a nice girl or not a nice girl?" in a manner congruent with the condition to which they were assigned. This question served as a comprehension and manipulation check: 78.03 % of participants responded in a congruent manner, and excluding participants who did not respond congruently did not change the general pattern of results ($\chi^2(1) = 13.857$, $p < 0.001$; see Supplementary Material for more details).

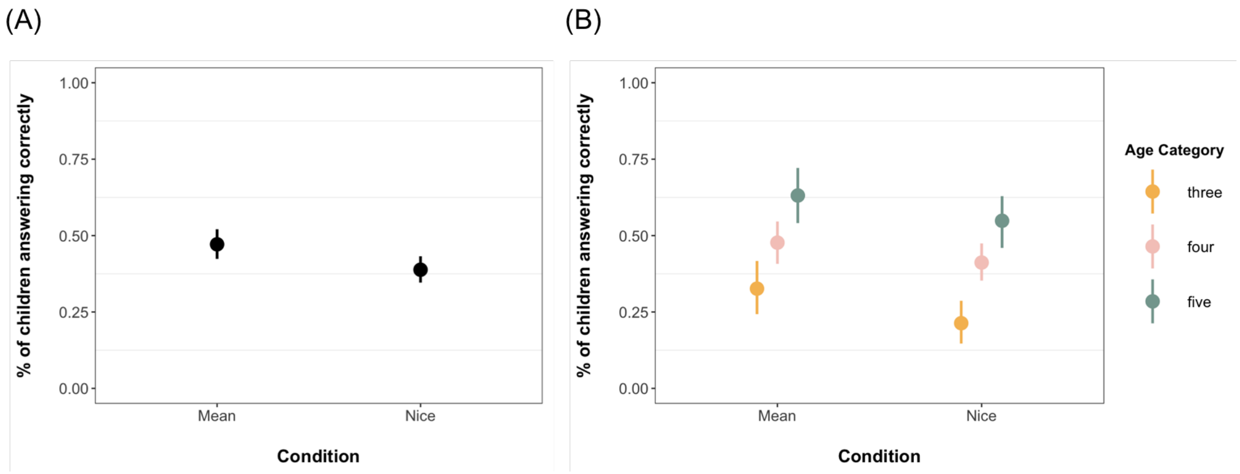

Figure 1. Proportion of children responding correctly, (A) for each condition, and (B) for each condition and age. Error bars denote bootstrapped 95% CI.

Study 2: False belief understanding across mean and nice interactions in adults

Given that the Sally-Anne task is targeted at preschool children, we aimed to reduce any possible ceiling effects for adults by constructing a greater number of vignettes in the same vein as the Sally-Anne task and making adults infer the meanness or niceness of agents in the interactions. We also included a baseline condition in order to directly compare ToM for mean and nice interactions to ToM for neutral interactions. The pre-registration for this study can be found here: http://aspredicted.org/blind.php?x=tr6t9f.

We note that we ran two identical full versions of the same study, originally intending the latter to be a direct replication of the former. Because the results of the two studies differed, we pooled the data to increase our power to detect any possible effects. We provide analyses of each individual study version in Supplementary Material. We also note that we conducted an initial version of the study that, due to experimenter error, prevented participants from seeing the attention check. In the spirit of transparency, we include all analyses involving this version of the study as well in Supplementary Material. Critically, results including this version (pooling all three versions) do not differ from the results presented here in the main text.

Methods

Participants

We conducted a power analysis using the R package ‘simr’ to determine the sample size. With alpha = 0.05 and power = 0.90, the projected sample size needed for an effect size of 0.1 (smaller than what was observed in a pilot study) is approximately 95.

For Study 2a, 95 adults from the United States were recruited for the study via Amazon Mechanical Turk. Of the 94 participants who completed the study online with no technical errors, six participants were excluded because they failed the attention check. The final sample consisted of 88 participants (28 females, ages 23-66 with mean of 36.76).

For Study 2b, 95 adults from the United States were recruited for the study via Amazon Mechanical Turk. We excluded four participants who failed the attention check; the final sample consisted of 91 participants (40 females, ages 19-64 with mean of 36.93).

All participants provided informed consent prior to starting the study. The study was approved by [removed for blind review].

Procedure

Participants completed 30 items (see Supplementary Material): for each item, they read a scenario in which an agent engaged in an action that was mean, nice, or neutral toward a second person. After reading about the agent's actions, participants were asked a question regarding the belief of the target of the action, for which they answered "True" or "False", and the extent to which they thought the agent engaging in the action was nice (from a scale of 1 [not at all] to 7 [very]). Items were assigned to the conditions such that participants saw 10 items per condition, with each item only presented once.

Analyses

Analyses were conducted in R (version 3.6; R Core Team, 2019). As stated in the pre-registration, responses were analyzed using a Generalized Linear Mixed Model (GLMM) with a binary response terms (correct [1] or incorrect [0]). We were primarily interested in whether responses (correct versus incorrect) depended on Condition. Our full model included the following regressors: Condition (mean, nice, or neutral), Age, and Gender (male or female).

Participant and item were entered as random effects. To assess the importance of our predictors of interest, we performed likelihood ratio tests (LRTs) and examined whether the model including a given term provided a significantly better fit to the data than the model without that term. We note that in our pre-registration we did not specify any directional tests.

Results

Given the inconsistency of the results across each individual study version, we pooled the two versions to more robustly test for a difference between the mean and nice conditions. For clarity, we present the results of the pooled sample here in the main text and the full results of each individual study version, analyzed as described in the pre-registration, in Supplementary Material. A sensitivity analysis revealed that with a combined $N = 279$, we had 80% power and alpha of 0.05 to detect an effect of Condition with a fixed effect estimate of at least 0.28.

In all, 5.7% of the variance in our response term (correct or incorrect) is explained by the fixed factors in our main full model, and 42.8% of the variance was explained by both the fixed and random factors in our model. These marginal and conditional R^2 values were calculated based on methods described by Nakagawa and colleagues (2017).

We observed no overall effect of Condition ($\chi^2(2) = 2.007$, $p = 0.37$; Figure 2). Exploratory post-hoc pairwise analyses reveal no significant differences across the three conditions in the log odds of providing a correct response (mean versus nice: $z = -1.219$, $p = .442$; mean versus baseline: $z = 0.083$, $p = 0.996$; nice versus baseline: $z = 1.300$, $p = 0.395$). A separate model examined response latencies by Condition, which also showed no overall effect of Condition ($\chi^2(2) = 2.064$, $p = 0.36$).

To assess whether the lack of difference across conditions could be due to a ceiling effect, we examined participants' performance across each condition. Mean performance for any given condition was about 83%, with 33-38% of participants getting all questions within a condition correct. Even after we removed participants with perfect scores for a given condition, we continued to find no condition differences ($\chi^2(2) = 0.658$, $p = 0.72$). We also examined whether the proportion of people getting 90% or 100% correct differed across the three conditions, and we found that this was not the case ($\chi^2(2) = 0.156$, $p = 0.92$).

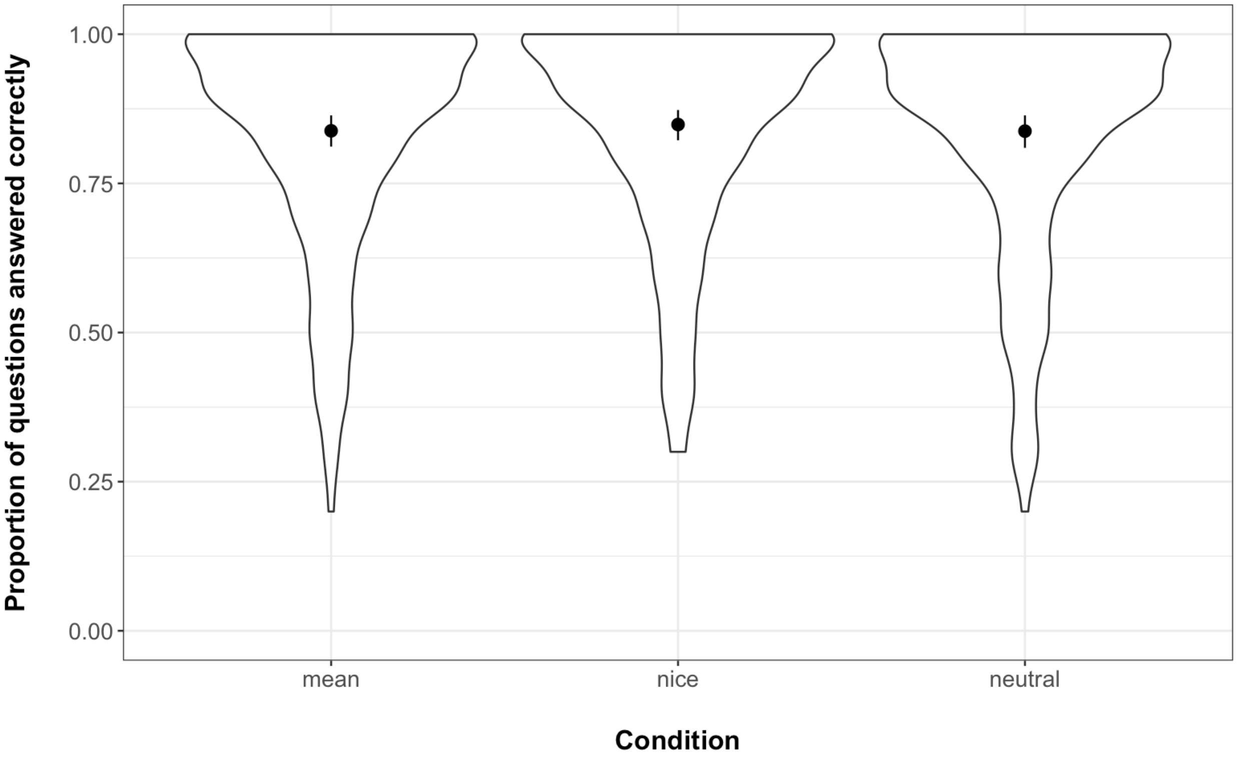

Figure 2. Proportion of questions adults answered correctly, by condition (pooled across the two study versions). The distributions of proportions are depicted, as well as the mean and 95% CI.

General Discussion

Two sets of studies test whether negative (mean) social interactions, as opposed to positive (nice) interactions, lead to a general boost in ToM. While investigations of this form of asymmetry in ToM or moral judgment have focused on ToM for the agent engaging in the action (e.g., the negative actor), we conducted a more conservative test of this phenomenon by asking about the false beliefs of someone other than the actor.

In a large sample of 3- to 5-year-olds, we find a general boost in ToM for negative interactions: that is, children are better at understanding the false belief of Sally (who is trying to find an object) when Anne is being mean to her versus nice. We note that this effect is small yet remains significant after accounting for the factors of age, gender, and participant-level differences. This ToM advantage for negative social interactions might serve a useful purpose: to help people make sense of and act in a non-ideal environment. We provide initial evidence for the idea that processing others' negative but not positive interactions lead children to broadly consider the minds of people involved in the interaction. A question for future work is whether this generalized boost can extend to understanding agents beyond the interaction itself.

Our work also reveals developmental change in the way negative interactions affect ToM deployment. Adults, unlike children, do not show a significant difference in ToM accuracy across conditions. Notably, this result cannot be explained by a ceiling effect. While we did not directly investigate the reason for developmental change, the baseline condition in the adult data offers some clues. Without the baseline condition, it would have been difficult to distinguish whether positive interactions led to a boost in ToM that paralleled negative interactions or whether negative interactions no longer led to a boost in ToM. Because performance in the mean and nice conditions were no different from baseline, we can infer that adults lost any ToM advantage for negative interactions they may have had, as opposed to gaining a ToM advantage for positive interactions.

This pattern for adults departs from findings in prior work indicating a negativity bias in adults. We propose that the present work demonstrates that any effect of valence may limited, in the case of adult ToM deployment. That is, while prior work demonstrates that adults differentially consider the mental states of negative versus positive actors, our work shows that this difference does not necessarily generalize to other individuals affected by the actions of negative versus positive agents. In other words, we do not see a generalized boost in ToM for negative actions among adults.

Other considerations

The *Mean Anne* and *Nice Anne* conditions were designed to be quite similar, so that any condition differences could not be attributed to factors such as word length or story structure. However, we recognize that in doing so we may have inadvertently introduced other features that could potentially drive condition differences. One possibility is that participants in the *Nice Anne* condition were led to focus more on the location corresponding with the wrong answer. That is, children in the *Nice Anne* condition were told that Anne was trying to be nice and so moved the ball to a different location, implicitly indicating that the different location is the correct place for the ball. Meanwhile, there was no implicit indication of a "correct" place for the ball in the *Mean Anne* condition. Though we did not explicitly mention a "correct" place for the ball, it is possible that children in the *Nice Anne* condition experienced greater difficulty because of this incidental and implied presence of a "correct" location, corresponding to the wrong answer for the actual task. However, we note that, in adults, we did not find a significant difference in response latencies across the two conditions, providing at least some evidence against the idea that participants might have spent more time thinking about this additional implicit information.

It is also worth noting that performance on this task by preschool children was slightly lower than in previous work using the canonical Sally-Anne task (Wellman et al., 2001). One explanation is that children in our study were not paying attention to the task and thus performed worse than is typical. However, this explanation seems unlikely given that we see an effect of Counterbalancing Order (Supplementary Material), in the direction that suggests children are engaging in the task as opposed to becoming fatigued or inattentive. Indeed, the likelihood of getting the second question correct was higher than the first question, regardless of which question was presented first, alleviating concerns that children became disengaged over the course of the task.

Conclusion

These studies contribute to a broader line of research that investigates when and how people reason about the minds of others. By asking this question in both children and adults, we are able to examine whether and how ToM-related behaviors change across development. Specifically, we examined how encountering third-party negative and positive interactions might affect people's ToM accuracy. One key aspect of this work is that, despite its conservative test (i.e., focusing on third-party observations of others' interactions), we nevertheless see an effect of negative social context on ToM. Our studies show that valence information has a diffuse effect on children's ToM such that children deploy greater ToM not just for negative actors as shown by previous work, but also for other individuals impacted by negative actors. This general impact of valence, however, is not seen in adulthood: adults do not display the same ToM advantage for negative social interactions. Future research should explore why this pattern is found in adults, and also whether this pattern results from experience in navigating an increasingly complex social world, in which people must toggle between cooperative and competitive contexts more efficiently. Together, these findings support the proposal that social context influences early ToM deployment in a way that is no longer observed in adulthood.

Acknowledgments

[removed for blind review]

References

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-Scale Deceit: Deception as a Marker of Two-, Three-, and Four-Year-Olds' Early Theories of Mind. *Child Development*, *60*(6), 1263. https://doi.org/10.2307/1130919

Chandler, M., & Hala, S. (1994). The role of personal involvement in the assessment of early false belief skills. *Children's Early Understanding of Mind: Origins and Development*, 403–425.

Davis, T. L. (2001). Children's understanding of false beliefs in different domains: Affective vs. physical. *British Journal of Developmental Psychology*, *19*(1), 47–58. https://doi.org/10.1348/026151001165958

Gonzalez, Antonya M., Steele, J. R., & Baron, A. S. (2017). Reducing children's implicit racial bias through exposure to positive out-group exemplars. *Child Development*, *88*(1), 123–130. https://doi.org/10.1111/cdev.12582

Gonzalez, Antonya Marie, Dunlop, W. L., & Baron, A. S. (2016). Malleability of implicit associations across development. *Developmental Science*. https://doi.org/10.1111/desc.12481

Hala, S., Chandler, M., & Fritz, A. S. (1991). Fledgling theories of mind: Deception as a marker of three-year-olds' understanding of false belief. *Child Development*, *62*(1), 83–97. https://doi.org/10.1111/j.1467-8624.1991.tb01516.x

Hamlin, J. K., & Baron, A. S. (2014). Agency attribution in infancy: Evidence for a negativity bias. *PLoS ONE*, *9*(5), e96112. https://doi.org/10.1371/journal.pone.0096112

Kelman, H. G. (1973). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues*, *29*(4), 25–61. https://doi.org/10.1111/j.1540-4560.1973.tb00102.x

Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, *4*(1). https://doi.org/10.1515/IP.2007.004

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, *17*(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

Leyens, J.-P., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review*, *4*(2), 186–197. https://doi.org/10.1207/S15327957PSPR0402_06

McLoughlin, N., & Over, H. (2017). Young children are more likely to spontaneously attribute mental states to members of their own group. *Psychological Science*, *28*(10), 1503–1509. https://doi.org/10.1177/0956797617710724

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

Opotow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues*, *46*(1), 1–20. https://doi.org/10.1111/j.1540-4560.1990.tb00268.x

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

Struch, N., & Schwartz, S. H. (1989). Intergroup aggression: Its predictors and distinctness from in-group bias. *Journal of Personality and Social Psychology*, *56*(3), 364–373.

Sullivan, K., & Winner, E. (1993). Three-year-olds′ understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology*, *56*(2), 135–148. https://doi.org/10.1006/jecp.1993.1029

Tsoi, L., Dungan, J. A., Chakroff, A., & Young, L. L. (2018). Neural substrates for moral judgments of psychological versus physical harm. *Social Cognitive and Affective Neuroscience*, *13*(5), 460–470. https://doi.org/10.1093/scan/nsy029

Tsoi, L., Dungan, J., Waytz, A., & Young, L. (2016). Distinct neural patterns of social cognition for cooperation versus competition. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2016.04.069

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

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

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Young, L., & Waytz, A. (2013). Mind attribution is for morality. In Baron-Cohen, Simon, Michael Lombardo, & Tager-Flusberg, Helen (Eds.), *Understanding Other Minds: Perspectives from Developmental Social Neuroscience* (3rd ed.). Oxford University Press.

Supplementary Material for

"False belief understanding for mean versus nice interactions in children and adults"

Table of Contents

Study 1: Script....2
Study 1: Participant information....3
Study 1: Additional analyses....5
Study 2: Stimuli....10
Study 2: Analyses for each individual study version....16
Study 2: Analyses pooling all three versions....18

Study 1: Script

Nice Anne Condition

Hi (child's name)! Let me introduce you to some people.

This is Sally (Sally waves), and this is Anne (Anne waves). Sally is playing with her favourite ball (act playing with ball) when her mom calls her out to lunch. Sally has to go, so she runs off, dropping her ball in the basket on the way out (act dropping ball in basket).

Now Anne sees where Sally put her favourite ball and knows that the ball is supposed to go in the closet, and NOT in the basket. Anne thinks that Sally must have been in such a hurry that she put her ball in the wrong place! Anne is a very nice girl, so she wants to help Sally by moving the ball to the closet. Anne goes and gets the ball out of the basket, and to help Sally, puts it away in the closet. That was very nice. Now Sally comes back after lunch and wants to find her ball.

Where will Sally look?
Where does Sally think her ball is?
Should Anne and Sally be friends?
Is Anne a nice girl or not a nice girl?
Is Sally a nice girl or not a nice girl?

Mean Anne Condition

Hi (child's name)! Let me introduce you to some people.

This is Sally (Sally waves), and this is Anne (Anne waves). Sally is playing with her favourite ball (act playing with ball) when her mom calls her out to lunch. Sally has to go, so she runs off, dropping her ball in the basket on the way out (act dropping ball in basket).

Now Anne sees where Sally put her favourite ball and knows that Sally LOVES to play with it. Anne thinks that Sally must have been in such a hurry that she will definitely come back after lunch to play with her ball again. Anne is not a very nice girl, so she wants to trick Sally by moving the ball to the closet. Anne goes and gets the ball out of the basket, and to trick Sally, puts it away in the closet. That was not very nice. Now Sally comes back after lunch and wants to find her ball.

Where will Sally look?
Where does Sally think her ball is?
Should Anne and Sally be friends?
Is Anne a nice girl or not a nice girl?
Is Sally a nice girl or not a nice girl?

Study 1: Participant information

The final sample consisted of 537 participants (Table S1). Of the 635 participants that were recruited for Study 1, 98 were excluded due to: participant's age being outside our age range of interest (31), incompletion of the task or declining to do the task (12), insufficient understanding of English (10), parental/other interference (9), lack of attention to the task (6), experimenter error (4), not understanding the task (4), having previously seen or completed the task (3), improper consent forms (2), having a developmental disorder (2), fussing out (2), or having data noted by the experimenter as unusable (12).

Table S1. Final sample breakdown by age, gender, and condition for Study 1

	Female	**Male**	**Total**
Mean Anne			
three	37	35	**72**
four	66	64	**130**
five	26	36	**62**
Nice Anne			
three	37	38	**75**
four	71	65	**136**
five	29	33	**62**
Grand Total	**266**	**271**	**537**

Table S2. Participant age summary statistics by condition and age group in Study 1

	Mean	**Minimum**	**Maximum**
Mean Anne	4.47	3.02	5.99
three	3.58	3.02	3.99
four	4.47	4.00	4.99
five	5.49	5.00	5.99
Nice Anne	4.44	3.04	5.98
three	3.53	3.04	3.99
four	4.46	4.00	4.99
five	5.48	5.00	5.98
Total	**4.45**	**3.02**	**5.99**

Table S3. Comparing ages across *Mean Anne* and *Nice Anne* conditions for each age group in Study 1.

	t	**df**	**p**
three	0.982	140.46	0.328
four	0.461	263.56	0.646
five	0.200	121.93	0.842

Table S4. Exclusions broken down by criteria and condition in Study 1

	Total Excluded	Mean Anne	Nice Anne
Age outside of age of interest	31	11	18
Incompletion or declining to do the task	12	7	5
Insufficient understanding of English	10	1	9
Parental / other interference	9	2	7
Lack of attention to the task	6	2	4
Experimenter error	4	3	1
Lack of understanding of the task	4	2	2
Previously saw or completed the task	3	3	0
Improper consent forms	2	1	1
Has developmental disorder	2	0	1
Fussed out	2	2	0
Data marked as unusable but reason was unspecified	12	5	7
Total	**98**	**40**	**55**

Note: The sum of the Mean Anne and Nice Anne columns does not equal Total Excluded because 3 participants were not put in a condition since they did not meet our criteria beforehand but wanted to experience a sample of the task anyway. 2 were excluded because of age outside the age of interest; 1 was excluded because of a developmental disorder.

Study 1: Additional analyses

Descriptive statistics for responses to questions in Study 1

In addition to the two main questions, “Where will Sally look?” and “Where does Sally think her ball is?”, we asked, “Should Anne and Sally be friends?”, “Is Anne a nice girl or not a nice girl?”, and “Is Sally a nice girl or not a nice girl?”. Proportion data for the latter three questions are depicted below (Fig. S1).

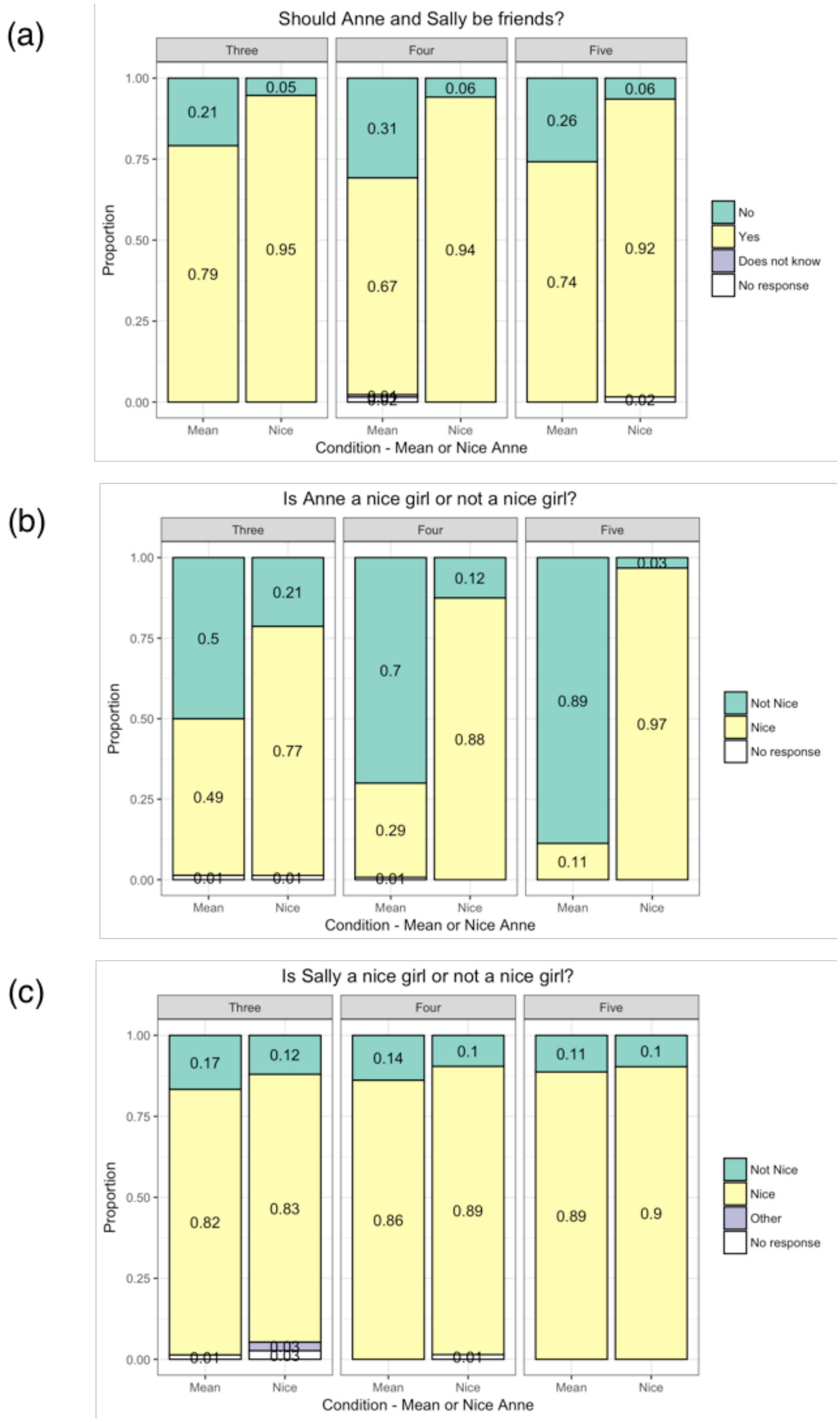

Fig. S1. Proportion of participants by response to auxiliary questions in Study 1, broken down by Condition and Age Category.

Examining performance across conditions at different levels of other factors

In the main text, we revealed a main interaction of Condition, with no interactions between Condition and other factors (e.g., Age Category, Question Type, or Counterbalancing Order). Nevertheless, to address questions of how the conditions differed across levels of other factors, we conducted the following contrasts:

Condition by Age Category: The log odds of responding correctly were numerically greater for *Mean Anne* than *Nice Anne* across each age group, though the difference was not significant for all age groups (3-year-olds: $z = 2.145$, $p = 0.032$; 4-year-olds: $z = 1.593$, $p = 0.111$; 5-year-olds: $z = 1.121$, $p = 0.262$). These results suggest that the general pattern of enhanced performance for *Mean Anne* holds across each age group but is only significant when pooled across age groups.

Condition by Question Type: The log odds of responding correctly were numerically greater in the *Mean Anne* condition than the *Nice Anne* condition for the question "Where does Sally think her ball is?" ($z = 1.627$, $p = 0.104$) and for the question "Where will Sally look?" ($z = 2.545$, $p = 0.011$), though the effect was significant only for the latter question. These results suggest that the general pattern of enhanced performance for *Mean Anne* holds across question type, though the size of the effect differs depending on the question.

Condition by Counterbalancing Order: The log odds of responding correctly were numerically greater in the *Mean Anne* condition than the *Nice Anne* condition for the first question presented ($z = 3.160$, $p = 0.002$) as well as for the second question ($z = 1.034$, $p = 0.301$), though the effect was only significant for the first question presented. These results, again, suggest a general effect of Condition.

Other significant effects

Counterbalancing Order: There was a significant main effect of Counterbalancing Order $\chi^2(1) = 28.979$, $p = 0.007$), wherein performance was better for the second question presented than the first question presented, regardless of which question was presented first.

Breakdown of performance by Age Category and Question Type: There was an interaction between Question Type and Age Category ($\chi^2(2) = 10.863$, $p = 0.004$): the effect of Question Type differed across the three age groups (Fig. S2). Pairwise contrasts performed at each age group revealed that the log odds of getting the "Where does Sally think her ball is?" question correct was significantly greater than the log odds of getting the "Where will Sally look?" question correct among 4-year-olds ($z = 4.438$, $p < 0.001$) and 5-year-olds ($z = 3.785$, $p < 0.001$), but not among 3-year-olds ($z = 0.495$, $p = 0.62$).

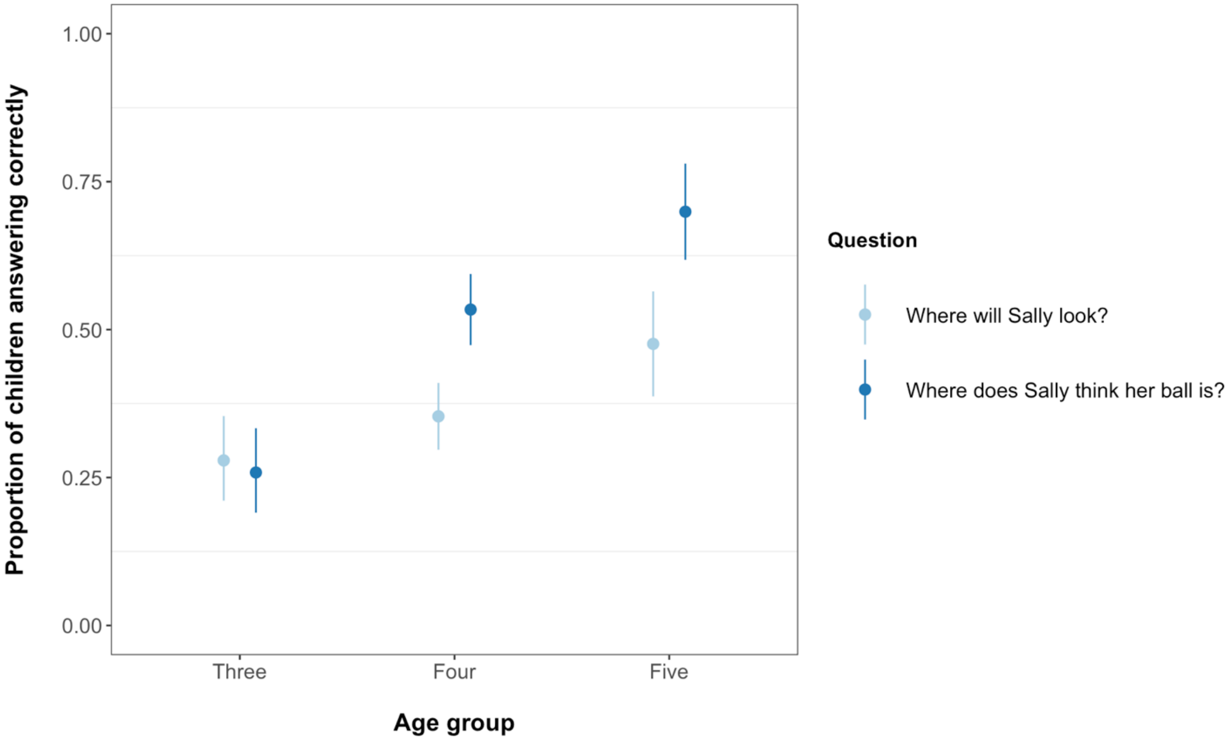

Figure S2. Proportion of participants answering correctly by age group and question. Bars denote 95% CI.

Analyses limited to participants who responded in a congruent manner

The following analyses restricted the data to those who responded to the question "Is Anne a nice girl or not a nice girl?" in a manner congruent to the condition to which they were assigned. This question acted as a comprehension check, but excluding people who did not get this question correct did not change the general pattern of results.

Results were similar to those found in the main text with all usable data. Likelihood ratio tests revealed no significant three-way interaction between Condition, Question Type, and Age Category ($\chi^2(2) = 2.456$, $p = 0.293$), no significant two-way interaction between Condition and Age Category ($\chi^2(2) = 2.363$, $p = 0.307$), no significant two-way interaction between Condition and Question Type ($\chi^2(1) = 2.277$, $p = 0.131$), no significant two-way interaction between Condition and Counterbalancing Order ($\chi^2(1) = 3.373$, $p = 0.066$), and no significant interaction between Question Type and Age Category ($\chi^2(2) = 5.844$, $p = 0.054$).

More importantly for our hypotheses, the main effect of Condition remained significant ($\chi^2(1) = 13.857$, $p < 0.001$): the log odds of providing a correct response was significantly greater for the *Mean Anne* condition than for the *Nice Anne* condition.

Analyses with age as a continuous variable

With all usable data (data reported in the main text):

Entering age as a continuous variable did not affect the pattern of results found in the main text. Similar to results reported in the main text, likelihood ratio tests revealed no significant three-way interaction between Condition, Question Type, and Age ($\chi^2(1) = 2.916$, $p = .088$), no significant two-way interaction between Condition and Age ($\chi^2(1) = 0.903$, $p = 0.342$), no significant two-way interaction between Condition and Question Type ($\chi^2(1) = 0.446$, $p = 0.505$), and no significant two-way interaction between Condition and Counterbalancing Order ($\chi^2(1) = 3.510$, $p = 0.061$). There was, again, a significant interaction between Question Type and Age Category ($\chi^2(1) = 10.290$, $p = 0.001$): the effect of Question Type differed by age. Importantly, the main effect of Condition remained significant ($\chi^2(1) = 6.593$, $p = 0.01$): the log odds of providing a correct response was significantly greater for the *Mean Anne* condition than for the *Nice Anne* condition.

With data limited to participants responding in a congruent manner:

Similar to results reported in the main text, likelihood ratio tests revealed no significant three-way interaction between Condition, Question Type, and Age ($\chi^2(1) = 1.349$, $p = 0.246$), no significant two-way interaction between Condition and Age ($\chi^2(1) = 1.435$, $p = 0.231$), no significant two-way interaction between Condition and Question Type ($\chi^2(1) = 3.110$, $p = 0.078$), and no significant two-way interaction between Condition and Counterbalancing Order ($\chi^2(1) = 3.460$, $p = 0.063$). There was, again, a significant interaction between Question Type and Age ($\chi^2(1) = 5.622$, $p = 0.018$): the effect of Question Type differed by age. Importantly, the main effect of Condition remained significant ($\chi^2(1) = 12.511$, $p < 0.001$): the log odds of providing a correct response was significantly greater for the *Mean Anne* condition than for the *Nice Anne* condition.

Study 2: Stimuli

1. **Scenario:** Sally put her ball in the box and left the room. Anne came in to the room, moved the ball from the box to the cabinet, and left the room.
   Nice: Anne wanted to help Sally by putting the ball in the right place.
   Mean: Anne wanted to play a trick on Sally by moving the ball.
   Neutral: Anne wanted to leave the room to go play in the park.
   Question: While Sally is away, she will think that her ball is in the box.
   Answer: True
2. **Scenario:** Lucy put two piles of clothes on her bed: one to keep and one for donation. While Lucy was in the attic to look for more clothes, Christina took the pile Lucy wanted to keep and brought it downstairs.
   Nice: Christina thought the pile of clothes she took was meant for donation.
   Mean: Christina thought the pile had some really nice clothes she would want for herself.
   Neutral: Christina thought the pile contained about twenty shirts.
   Question: While Lucy is in the attic, she will think that her keep pile is still on her bed.
   Answer: True
3. **Scenario:** Roxanne went to bed, leaving an unfinished puzzle in the living room so that she could solve it the next day. Max saw the puzzle and solved it that night.
   Nice: Max wanted to help Roxanne solve the puzzle.
   Mean: Max wanted to show Roxanne that he beat her to solving the puzzle.
   Neutral: Max wanted to finish the puzzle even if it took the whole night.
   Question: When Roxanne wakes up, she will think the puzzle is completed.
   Answer: False
4. **Scenario:** Brad spent over 12 hours in the communal living area watching DVDs he borrowed from the library, and he planned on rewatching some of them the next day. While he was sleeping, his roommate Tracy returned all the DVDs to the library.
   Nice: Tracy thought Brad was done with the DVDs and did it as a friendly gesture.
   Mean: Tracy thought that she could finally have the living room to herself now that the DVDs were gone.
   Neutral: Tracy thought it was interesting that all the movies were filmed in black and white.
   Question: When Brad wakes up, he will think that the DVDs are still at the house.
   Answer: True
5. **Scenario:** Davon cleared his desk and put all his items on the floor because he needed more desk space to do his work. When he left the house to run some errands, Nikki put all the items back on the table.
   Nice: Nikki wanted to keep the floor clean so no one would trip on anything.
   Mean: Nikki wanted to disrupt Davon's work so that he would be further behind.
   Neutral: Nikki wanted to borrow some of the items from Davon later.
   Question: While Davon is outside, he will think that his desk is cleared of any items.
   Answer: True
6. **Scenario:** Trevor went to the pantry to get spices to make his soup less bland. While Trevor was away, Lauren, the instructor of the culinary class, added unlabeled spice to Trevor's soup.
   Nice: Lauren wanted to improve the flavors of the bland soup.
   Mean: Lauren wanted to sabotage Trevor by adding unknown spice.

Neutral: Lauren wanted to look at every student's soup.
Question: While Trevor is away, he will think that his soup needs more spice.
Answer: True

7. **Scenario:** Paula forgot the combination to her lock, so she left her locker unlocked and went in to class. Anthony saw the unlocked locker and locked it.
Nice: Anthony wanted to prevent people from stealing Paula's things.
Mean: Anthony wanted to prevent Paula from getting to her belongings.
Neutral: Anthony wanted to get to his next class on time so he walked briskly.
Question: While Paula is in class, she will think that the locker is locked.
Answer: False
8. **Scenario:** Stan needed to get to his fencing match soon, so he asked Duncan to request an Uber while he was getting ready to leave. Instead of requesting an Uber, Duncan requested a Lyft.
Nice: Duncan wanted Stan to get to his match soon and Lyft had an earlier arrival time than Uber.
Mean: Duncan wanted Stan to get to his match late and Lyft had a later arrival time.
Neutral: Duncan wanted to use a promo code for a free ride.
Question: While Stan is getting ready, he will think an Uber will be there.
Answer: True
9. **Scenario:** Amy was told that the results of her medical tests would come in the mail that week. The mail arrived in a timely manner, but her daughter decided to hide the results from Amy for another two weeks.
Nice: Amy's daughter thought Amy had enough stress to deal with.
Mean: Amy's daughter thought that holding on to the results would prevent Amy from receiving treatment.
Neutral: Amy's daughter thought the results were difficult to interpret.
Question: When the week passes by, Amy will think that the medical results were delivered.
Answer: False
10. **Scenario:** Parker moved the ladder close to the air vent and crawled into the vent. Asher moved the ladder from under the vent back to where it typically is.
Nice: Asher thought the ladder was there because someone forgot to return it back.
Mean: Asher thought playing a trick on Parker would be funny.
Neutral: Asher thought it was hard not to scratch the floor when dragging the ladder.
Question: While Parker is in the vents, he will think the ladder is right under the vent.
Answer: True
11. **Scenario:** Stacy went to get some water after her workout at the treadmill, leaving behind her towel. Elena, who wanted to use the treadmill, placed the towel on a nearby elliptical.
Nice: Elena thought that placing the towel there would make it easy for Stacy to find her towel.
Mean: Elena thought moving Stacy's towel would annoy Stacy.
Neutral: Elena thought that she stayed at the gym long enough so she left soon after.
Question: While Stacy is away, she will think that her towel is on the elliptical.
Answer: False
12. **Scenario:** Ashley chopped some vegetables to use for cooking the next day. Her roommate Maggie saw the vegetables in the fridge and decided to cook them while Ashley was at work.
Nice: Maggie wanted to have Ashley's dinner ready for her when she returned.

Mean: Maggie wanted to shorten her cooking time by using Ashley's chopped vegetables.
Neutral: Maggie wanted her meal to contain vegetables and meat.
Question: While Ashley is at work, she will think her vegetables are in the fridge.
Answer: True

13. **Scenario:** Chris was moving out of his furnished apartment and was packing up items in the kitchen. Without any input from Chris, Andrew loaded the desk that came with the apartment into the truck.
    Nice: Andrew wanted to help move bulkier items.
    Mean: Andrew wanted to get Chris in trouble with the building superintendent.
    Neutral: Andrew wanted a Gatorade after moving the desk.
    Question: While Chris is in the kitchen, he will think the desk was moved to the truck.
    Answer: False
14. **Scenario:** Brittany decided to take golf lessons and planned on signing up for them later that day. Sam signed up for the last spot not knowing that Brittany didn't sign up yet.
    Nice: Sam thought that it would be great if he and Brittany took lessons together.
    Mean: Sam thought that formal lessons would help him be better at golf than Brittany.
    Neutral: Sam thought the lessons were really cheap given the frequency of the lessons.
    Question: Before Brittany checks the sign-up sheet, she will think that there are no spots left.
    Answer: False
15. **Scenario:** Joe, who was running late that morning, dropped his bag by the entryway and ran to the kitchen to pick up something he forgot. While Joe was in the kitchen, Steve put the bag away.
    Nice: Steve thought Joe had left and didn't want the bag to clutter up the entryway.
    Mean: Steve thought Joe would get frustrated about losing his bag.
    Neutral: Steve thought Joe's bag was really heavy given its size.
    Question: While Joe is in the kitchen, he will think his bag is in the entryway.
    Answer: True
16. **Scenario:** Janice placed the heavy vase she had just purchased at the yard sale on the floor outside the public restroom. While Janice was in the restroom, Heather moved the vase to the vase section of the yard sale.
    Nice: Heather thought the vase had been misplaced so she moved it there.
    Mean: Heather thought doing this would really upset Janice.
    Neutral: Heather thought the vase section of the yard sale contained a nice variety of different vases.
    Question: While Janice is in the restroom, she will think her vase is at the vase section of the yard sale.
    Answer: False
17. **Scenario:** Anna wanted to go mountain climbing so she bought some gear. Her husband, David, saw the gear on the table and brought it back to the store while Anna was at work.
    Nice: David thought that Anna needed nicer gear for several safety reasons.
    Mean: David thought that Anna doing mountain climbing would threaten his masculinity.
    Neutral: David thought the store was so disorganized that it was hard to find specific items.
    Question: While Anna is at work, she will think that her gear is on the table.
    Answer: True
18. **Scenario:** Jeff wanted a nice photo for his dating profile, so he asked Allie to enhance a photo of him. Allie removed the background and added a plain white background.

Nice: Allie wanted to help create a decent photo for what she thought was for a passport.
Mean: Allie wanted to make Jeff's photo look boring and unappealing.
Neutral: Allie wanted to make other alterations but ran short on time.
Question: Before Jeff sees the photo, he will think the photo will have a plain white background.
Answer: False

19. **Scenario:** Erica removed the tire on her bike and went to the bike shop to get a new replacement tire. While she was away, Brian saw the tire and put the tire back on the bike.
Nice: Brian wanted to help Erica out with something that he's good at doing.
Mean: Brian wanted Erica to struggle with getting her bike in working order.
Neutral: Brian wanted to get a new bike like Erica's, but with bigger wheels and a lighter frame.
Question: While Erica is away, she will think the tire is removed from the bike.
Answer: True
20. **Scenario:** Brianna dislikes open umbrellas in indoor spaces, so she closed her wet umbrella and left it by the front door before heading to the living room. Amanda, who came in after Brianna, opened Brianna's umbrella and left it open by the front door.
Nice: Amanda thought opening it would get the umbrella to dry quicker.
Mean: Amanda thought opening the umbrella would annoy Brianna.
Neutral: Amanda thought the blue and orange swirls on the umbrella nicely complemented each other.
Question: While Brianna is in the living room, she will think the umbrella is closed.
Answer: True
21. **Scenario:** Amanda found Mike's paint can in the closet and left it out to be thrown away since the color was ugly. While she was away, Mike saw the paint can and painted their bedroom with it.
Nice: Mike thought it was the can Amanda wanted him to use to paint their room.
Mean: Mike thought the color that he picked was more welcoming than the one Amanda had wanted.
Neutral: Mike thought it would take 30 minutes to paint the room, but it actually took over two hours.
Question: While Amanda is away, she will think the bedroom will be newly painted.
Answer: False
22. **Scenario:** Maria left all of her clothes out as she was packing for vacation and left to get more new clothes. Dani saw that the clothes were out and put them away.
Nice: Dani wanted to help Maria keep her room clean.
Mean: Dani wanted to make packing more difficult for Maria.
Neutral: Dani wanted her sense of style to be as good as Maria's.
Question: While Maria is away, she will think her clothes are packed away.
Answer: False
23. **Scenario:** Maya was working on her portion of a complex multi-person project and took a break by taking a short nap. While Maya was napping, Brianna finished the project.
Nice: Brianna wanted to help Maya with the project.
Mean: Brianna wanted more credit for the work.
Neutral: Brianna wanted to take a nap, too.
Question: When Maya wakes up, she will think the project is finished.

Answer: False

24. **Scenario:** Juan checked his calendar and saw that his schedule was free at noon. While he was making plans to eat out for lunch, his secretary scheduled a meeting for Juan at noon.
    Nice: His secretary thought this was a good business opportunity for Juan.
    Mean: His secretary thought Juan would get mad for having to skip lunch yet again.
    Neutral: His secretary thought the meeting should take place in Meeting Room B.
    Question: Before Juan checks his calendar again, he will think that there is a meeting at noon.
    Answer: False
25. **Scenario:** Kate, prior to chorus practice, came up with a song and showed it to Grace. While Kate was away, Grace memorized it and showed it to their teacher.
    Nice: Grace wanted to get advice from the teacher on how to make Kate's song even better.
    Mean: Grace wanted to take the credit for coming up with the song.
    Neutral: Grace wanted to hit all the high notes without straining her voice.
    Question: While Kate is away, she will think the teacher has already heard her song.
    Answer: False
26. **Scenario:** Carla was baking a cake and asked Jane for vegetable oil. While Carla was looking for other ingredients in the pantry, Jane added sesame oil to the cake mix.
    Nice: Jane thought there was no difference between sesame oil and vegetable oil.
    Mean: Jane thought adding sesame oil to the mix would make Carla's cake taste weird.
    Neutral: Jane thought the cake should be done in 40 minutes.
    Question: While Carla is away, she will think the cake batter contains sesame oil.
    Answer: False
27. **Scenario:** Caroline works as a fry cook flipping hamburgers. Her co-worker, Ali, saw that Caroline walked away from the grill and took the burger off the grill.
    Nice: Ali wanted to prevent the burger from burning.
    Mean: Ali wanted Caroline to get in trouble for walking away from the grill.
    Neutral: Ali wanted to make sure she didn't burn her fingers while moving the burger.
    Question: While Caroline is away, she will think the burger is on the grill.
    Answer: True
28. **Scenario:** Jackson left his keys at home, so he kept his office door unlocked and open. While he was at a meeting, the janitor closed and locked the office door.
    Nice: The janitor thought it was important to keep everything in the office safe.
    Mean: The janitor thought locking the door would make it harder for Jackson to get into his office.
    Neutral: The janitor thought the office door plaque needed to be updated.
    Question: While Jackson is away, he will think his door is unlocked.
    Answer: True
29. **Scenario:** Jenny found a nice spot at the party to put down her beer. While she was getting food, Evan moved her beer from the windowsill to the floor.
    Nice: Evan thought the wind might knock down the beer from the windowsill.
    Mean: Evan thought his beer should be the one by the windowsill instead.
    Neutral: Evan thought there was no way the house would comfortably fit everyone.
    Question: While Jenny is getting food, she will think that her beer is on the windowsill.
    Answer: True

30. **Scenario:** Nicole put down twenty seashells she found at the beach and planned to pick them up after a swim. Jason moved the seashells further away from the water.
    Nice: Jason wanted to keep the seashells from getting washed away.
    Mean: Jason wanted to make Nicole think the seashells got washed away.
    Neutral: Jason wanted to lie on the sand and remain far from the water.
    Question: While Nicole is swimming, she will think the seashells are where she dropped them.
    Answer: True

Study 2: Analyses for each individual study version

In the main text, we presented analyses pooling the two full versions of Study 2. Here, we break down the results for each study, including an initial version that we did not include in the main text for reasons specified below.

Study 2pre (not reported in main text)

Due to an experimenter error, this initial version of the study prevented participants from seeing the attention check. In the spirit of transparency, we include analyses involving this version of the study here.

We first performed a manipulation check by testing whether ratings of the niceness of the agent differed across conditions. Results reveal a main effect of Condition on ratings of niceness ($\chi^2(2) = 410.52$, $p < 0.001$). Post-hoc analyses reveal that agents were rated as nicer in the nice condition than in the neutral condition ($z = 13.097$, $p < 0.001$) and mean condition ($z = 18.829$, $p < 0.001$). Agents were also rated as nicer in the neutral condition than the mean condition ($z = 9.406$, $p < 0.001$).

In the full main model of this study, 2.9% of the variance in the response term was explained by the fixed factors in the model, and 35.4% of the variance was explained by both the fixed and random factors in our model. We did not find an effect of Condition on performance on the false belief question ($\chi^2(2) = 1.825$, $p = 0.402$).

Post-hoc comparisons show no significant pairwise differences in performance on the false belief question between the three conditions (mean versus nice: $z = -1.373$, $p = 0.355$; mean versus neutral: $z = -0.745$, $p = 0.724$; nice versus neutral: $z = 0.609$, $p = 0.815$). Response latencies also did not differ across the three conditions ($\chi^2(2) = 0.320$, $p = 0.852$).

Study 2a

We first performed a manipulation check by testing whether ratings of the niceness of the agent differed across conditions. Results reveal a main effect of Condition on ratings of niceness ($\chi^2(2) = 416.53$, $p < 0.001$). Post-hoc analyses reveal that agents were rated as nicer in the nice condition than in the neutral condition ($z = 19.310$, $p < 0.001$) and mean condition ($z = 11.912$, $p < 0.001$). Agents were also rated as nicer in the neutral condition than the mean condition ($z = 11.157$, $p < 0.001$).

In the full main model of this study, 6.3% of the variance in the response term was explained by the fixed factors in the model and 42.7% of the variance was explained by both the fixed and random factors in our model. We tested for a main effect of Condition on responses to the false belief question. The log odds of providing a correct response was significantly different across the three conditions ($\chi^2(2) = 9.605$, $p = 0.008$).

Post-hoc analyses not specified in the pre-registration reveal significantly greater log odds of providing a correct response in the nice condition than in the mean condition ($z = 3.0166$, $p = 0.007$) and significantly greater log odds of providing a correct response in the nice condition than in the neutral condition ($z = 2.383$, $p = 0.045$) but not significantly different across the mean and neutral conditions ($z = 0\text{-}0.652$, $p = 0.791$). Response latencies, too, differed across the three conditions ($\chi^2(2) = 6.955$, $p = 0.031$), with response latencies longer for the mean condition than the nice condition ($z = 2.508$, $p = 0.033$).

Study 2b

Our manipulation check revealed a main effect of Condition on ratings of niceness ($\chi^2(2) = 573.17$, $p < 0.001$). There was a greater likelihood of reporting the agent as nice in the nice condition than the mean or neutral condition and greater likelihood in the neutral condition than the mean condition (nice versus neutral: $z = 14.264$, $p < 0.001$; nice versus mean: $z = 22.283$, $p < 0.001$; neutral versus mean: $z = 12.618$, $p < 0.001$).

In the full main model of this study, 5.6% of the variance in the response term was explained by the fixed factors in the model and 43.2% of the variance was explained by both the fixed and random factors in our model. However, we did not find an effect of Condition on performance on the false belief question ($\chi^2(2) = 2.121$, $p = 0.346$).

Post-hoc comparisons show no significant pairwise differences in performance on the false belief question between the three conditions (mean versus nice: $z = 1.499$, $p = 0.291$; mean versus neutral: $z = 0.746$, $p = 0.736$; nice versus neutral: $z = -0.750$, $p = 0.733$). Response latencies also did not differ across the three conditions ($\chi^2(2) = 1.739$, $p = 0.419$).

Study 2: Analyses pooling all three versions

Our manipulation check revealed a main effect of Condition on ratings of niceness ($\chi^2(2) = 1417.94$, $p < 0.001$). There was a greater likelihood of reporting the agent as nice in the nice condition than the mean or neutral condition and greater likelihood in the neutral condition than the mean condition (nice versus neutral: $z = 22.946$, $p < 0.001$; nice versus mean: $z = 35.271$, $p < 0.001$; neutral versus mean: $z = 19.362$, $p < 0.001$).

In the full main model of this study, 4.5% of the variance in the response term was explained by the fixed factors in the model and 40.8% of the variance was explained by both the fixed and random factors in our model. However, we did not find an effect of Condition on performance on the false belief question ($\chi^2(2) = 3.335$, $p = 0.189$).

Post-hoc comparisons show no significant pairwise differences in performance on the false belief question between the three conditions (mean versus nice: $z = -1.764$, $p = 0.182$; mean versus neutral: $z = -0.337$, $p = 0.939$; nice versus neutral: $z = 1.427$, $p = 0.327$). Response latencies also did not differ across the three conditions ($\chi^2(2) = 1.671$, $p = 0.434$).