False belief understanding for negative versus positive 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

Declarations of interest: none

Abstract

The capacity to reason about the mental states of others ("theory of mind"; ToM) facilitates effective social interaction. For instance, ToM helps people identify agents as potential enemies or allies. Despite clear benefits, evidence shows that people do not fully deploy their capacity for ToM at all times or evenly across people. Two sets of studies test whether people reason more accurately about the mental states of agents affected by negative versus positive social behaviors. In Study 1, we test a large sample of three- to five-year-olds ($N = 537$) and find a generalized valenced-driven asymmetry for ToM: that is, preschool children are better at understanding the false beliefs of people described in negative versus positive vignettes. In Study 2 (2a: $N = 88$; 2b: $N = 91$), we find a different pattern of results for adults, who show no difference in performance across negative, positive, and neutral social interactions. Together, these findings support the idea that social context influences early ToM deployment in a way that is not observed in adulthood.

Keywords: theory of mind; social cognition; social development

False belief understanding for negative versus positive interactions in children and adults

The capacity to reason about the mental states of others, also known as mentalizing or "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 behaviors.

In the field of psychology, one well-studied phenomenon is the tendency for negative information to dominate positive information; this valence-driven asymmetry or negativity bias is found in domains such as memory, attention, decision-making, and social cognition (for reviews, see Baumeister et al., 2001; Peeters & Czapinski, 1990; Rozin & Royzman, 2001; Taylor, 1991; Vaish et al., 2008). Just to name a few examples: people spend more time looking at negative versus positive stimuli (Fiske, 1980), learn faster from negative reinforcement versus positive reinforcement (for a review, Öhman & Mineka, 2001), perceive negative information to contain greater informational value than positive information (Peeters & Czapinski, 1990), weigh negative aspects of events more than positive events when making judgments or decisions (Kahneman & Tversky, 1984), use negative information more than positive information to arrive

at an impression of a person (Kanouse & Hanson Jr., 1987), and similarly, require less information to make trait inferences about others when the information is negative versus positive (Aloise, 1993).

With regard to social cognition, a negativity bias emerges early in life: infants discriminate between antisocial actors (who hinder others) and neutral actors, whereas they do not discriminate between prosocial actors and neutral actors (Hamlin et al., 2010), and attribute agency to a mechanical claw that causes a bad outcome but not to a claw that causes a good outcome (Hamlin & Baron, 2014). Preschool children are more likely to judge side effects of an action as intentional when the side effect is morally bad versus good (Leslie et al., 2006). Similar pattern are found in adults: for example, adults are more willing to blame an agent for morally bad side effects than to praise an agent for morally good side effects (Knobe, 2003); adults are more likely to attribute events to external agents (e.g., more likely to think their game partner was human versus a computer) when events are negative rather than neutral or positive (Morewedge, 2009); and adults are more likely to anthropomorphize or perceive their computers to have minds, beliefs, and desires, in the case of computer malfunction (Waytz et al., 2010). This negativity bias is proposed to serve a useful function: to help people gain control over a bad situation (Peeters & Czapinski, 1990).

We hypothesize that people deploy their capacity for ToM more when processing negative versus positive interactions. The present work provides a conservative test of this hypothesis by answering a question that, to our knowledge, has not yet been explored: whether negative social interactions, as opposed to positive interactions, lead people to deploy ToM for people beyond just the negative actor. Notably, much of prior work has focused on people's mental state attributions to agents causing good or bad actions. We hypothesize that when people

encounter someone engaging in behaviors that affect another person (e.g., Person A put items that were on the floor on Person B's desk) because of negative or 'mean' reasons (e.g., they wanted to disrupt Person's B work) versus positive or 'nice' reasons (e.g., they wanted to prevent Person B from tripping over items left on the floor), they will be more likely to consider not just the mental state of the acting agent (Person A), but also the mental states of the agent affected by the negative or positive behavior (Person B). We note that positive or negative interactions in this context refer to the actor's intent and not the interaction's outcome, which is the same across conditions in our two studies.

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 difference between ToM deployment for mean versus nice behaviors in preschool children would suggest its relatively early emergence. 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 negative versus positive 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 large participant sample (N=537), allowing for a better estimate of the size of the effect.

We also tested the same question in adults to further reveal whether any generalized ToM deployment for negative versus positive behaviors found in childhood persists through adulthood. In 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., 2016, 2018), this study tests an aspect of ToM understudied in adults (Zaki & Ochsner, 2011): 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

We note that we did not pre-register Study 1, as the study was conducted before pre-registration was a common step in research pipelines. Study 2 was pre-registered.

Participants

Six hundred and forty-three participants were recruited from a community-based science center and tested in a soundproof testing room dedicated to behavioral science research. Only the participant and the researcher were in the testing room, with family in an adjacent waiting room. 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, as children are transitioning from a high-energy environment to a testing environment (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 the Boston College Institutional Review Board.

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?. Questions 1 and 2 tapped into children's false belief understanding. 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 included a question probing just the belief ("Where does Sally think her ball is?"). The order in which Questions 1 and 2 were asked was counterbalanced across participants. In addition to reporting results for Questions 1 and 2, we report results for Question 4, which served as our comprehension and manipulation check; 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. 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 (Supplementary Material). A sensitivity analysis revealed that with $N = 537$ we had 80% power and alpha of 0.05 to detect an effect of Condition with an unstandardized regression coefficient (odds ratio) as extreme as 0.44.

Results and Discussion

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 ($\chi^2(1) = 7.168$, $p = 0.007$). 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). Effects from the full model are reported in Table 1. 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).

We provide initial evidence for the idea that processing others' negative behaviors lead children to broadly consider the minds of the people involved more so than when they process positive behaviors. That is, in a large sample of 3- to 5-year-olds, we find that 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. Improved performance for negative versus positive social interactions might serve a useful purpose: to help people make sense of and act in a non-ideal social environment.

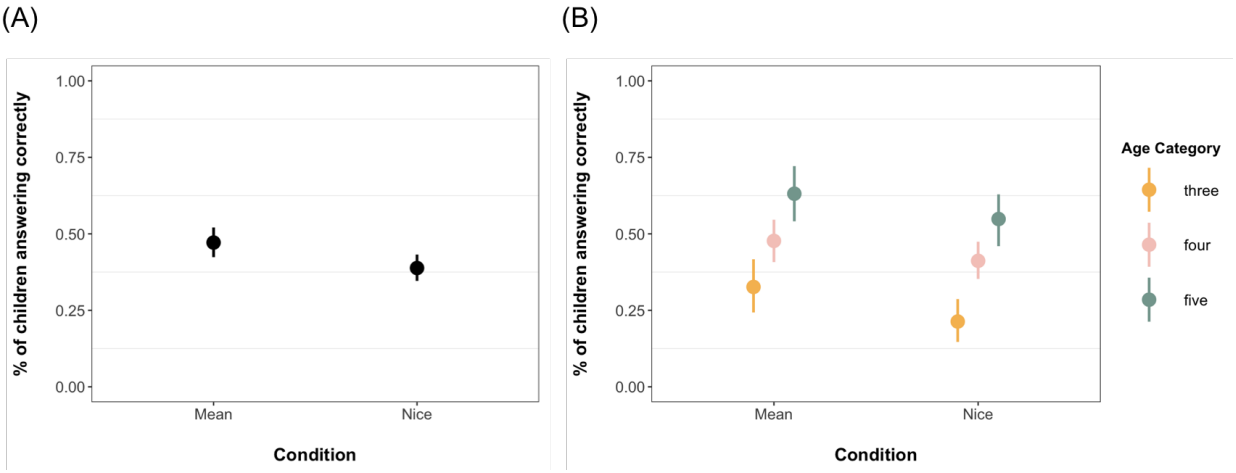

Figure 1. Proportion of children responding correctly, averaging across false belief questions, and separated (A) by condition, and (B) by condition and age. Error bars denote bootstrapped 95% CI.

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

Table 1

Effects from the Full Model in Study 1, with Age as a Categorical Variable

Predictors	Odds Ratios	Std. Beta	CI	Standardized CI	*z*	*p*
(Intercept)	0.62	-0.47	0.42 – 0.93	-0.88 – -0.07	-2.29	0.02*
Condition [reference: Mean]	0.41	-0.90	0.24 – 0.70	-1.44 – -0.36	-3.27	<0.001*
Question Type [reference: Look]	1.41	0.34	0.90 – 2.21	-0.11 – 0.79	1.48	0.14
Age Category [linear]	2.29	0.83	1.22 – 4.29	0.20 – 1.46	2.57	0.01*
Age Category [quadratic]	0.96	-0.05	0.58 – 1.59	-0.55 – 0.46	-0.18	0.86
Counterbalancing Order [reference: first]	1.79	0.58	1.16 – 2.75	0.15 – 1.01	2.64	0.01*
Gender [reference: Female]	0.81	-0.22	0.57 – 1.14	-0.56 – 0.13	-1.23	0.22
Condition x Question Type	1.24	0.21	0.65 – 2.35	-0.43 – 0.85	0.64	0.52
Condition x Age Category [linear]	0.93	-0.07	0.38 – 2.28	-0.96 – 0.83	-0.15	0.88
Condition x Age Category [quadratic]	1.25	0.23	0.60 – 2.61	-0.51 – 0.96	0.60	0.55
Condition x Counterbalancing Order	1.73	0.55	0.94 – 3.17	-0.06 – 1.16	1.76	0.08
Question Type x Age Category [linear]	1.93	0.66	0.85 – 4.39	-0.16 – 1.48	1.57	0.12
Question Type x Age Category [quadratic]	1.10	0.10	0.57 – 2.13	-0.56 – 0.75	0.29	0.77
Condition x Question Type x Age Category [linear]	1.89	0.64	0.58 – 6.23	-0.55 – 1.83	1.05	0.29
Condition x Question Type x Age Category [quadratic]	0.46	-0.77	0.18 – 1.20	-1.72 – 0.18	-1.58	0.11
Random Effects						
σ^2	3.29					
$\tau_{00\ Subject}$	1.17					
ICC	0.26					
$N_{Subject}$	537					
Observations	1073					
Marginal R^2 / Conditional R^2	0.173 / 0.389					

Note. Reference levels are provided above. Polynomial functions (linear, quadratic) were fit to the levels of the Age Category variable because Age Category was an ordered factor (Three < Four < Five). CI = 95% confidence interval. *$p < 0.05$, two-tailed.

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 possiblethat 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 possible 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 with 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.

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

In this study, we tested whether the contextual effect (generalized ToM deployment for negative versus positive behaviors) found in childhood persists through adulthood. Similar to Study 1, we presented adults with a series of vignettes, again manipulating whether Anne or another agent in a similar position performed an action with mean or nice intentions. 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 (Studies 2a and 2b), originally intending the latter to be a direct replication of the former. Because the results of the two studies differed, we pooled the data from Studies 2a and 2b to increase our power to detect any possible effects; the results reported here in the main text used this pooled dataset. In Supplementary Material, we provide (1) analyses of each individual study version and an initial version of the study prior to these two versions that, due to experimenter error, prevented participants from seeing the attention check (we include the results in the spirit of transparency), and (2) results pooling Study 2a, Study 2b, and the initial flawed version (Study 2-pre), which 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, *M* = 36.76, *SD* = 10.57).

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, *M* = 36.93, *SD* = 10.73).

All participants provided informed consent prior to starting the study. The study was approved by the Boston College Institutional Review Board.

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 and Discussion

Given the inconsistency of the results across each individual study version, we pooled the data from Studies 2a and 2b 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 an unstandardized regression coefficient (odds ratio) as extreme as 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).

Effects from the full model are reported in Table 2. 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 ($\chi^2(2) = 2.064$, $p = 0.36$; Table S16).

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 ($\chi^2(2) = 0.658$, $p = 0.72$; Table S17). 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 ($\chi^2(2) = 0.156$, $p = 0.92$). The analyses above treated the response variable as categorical (correct versus incorrect). We also performed an alternative analysis that compared participants' scores across conditions using transformed scores (using the Tukey's Ladder of Powers transformation) to reduce the skewness of participants' scores; the results, again, revealed no effect of condition ($F(2, 562) = 0.20$, $p = 0.81$). While a ceiling effect could potentially explain this pattern of results, our analyses that account for it still reveal no differences across conditions.

We note that prior piloting of these stimuli for a neuroimaging study (Tsoi and Young, in prep) revealed no differences in response time across the three conditions ($\chi^2(2) = 0.84$, $p = 0.99$), providing some evidence from an independent sample that suggests that people were unlikely to find one condition more difficult or complex than the other. Moreover, that study also found no differences across conditions in overall levels of neural responses in brain regions implicated in social cognition, providing a different but converging line of evidence that adults may not deploy ToM differently when processing positive versus negative interactions.

However, we note that 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 may not necessarily generalize to other individuals affected by the actions of those negative versus positive agents.

In summary, this study 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.

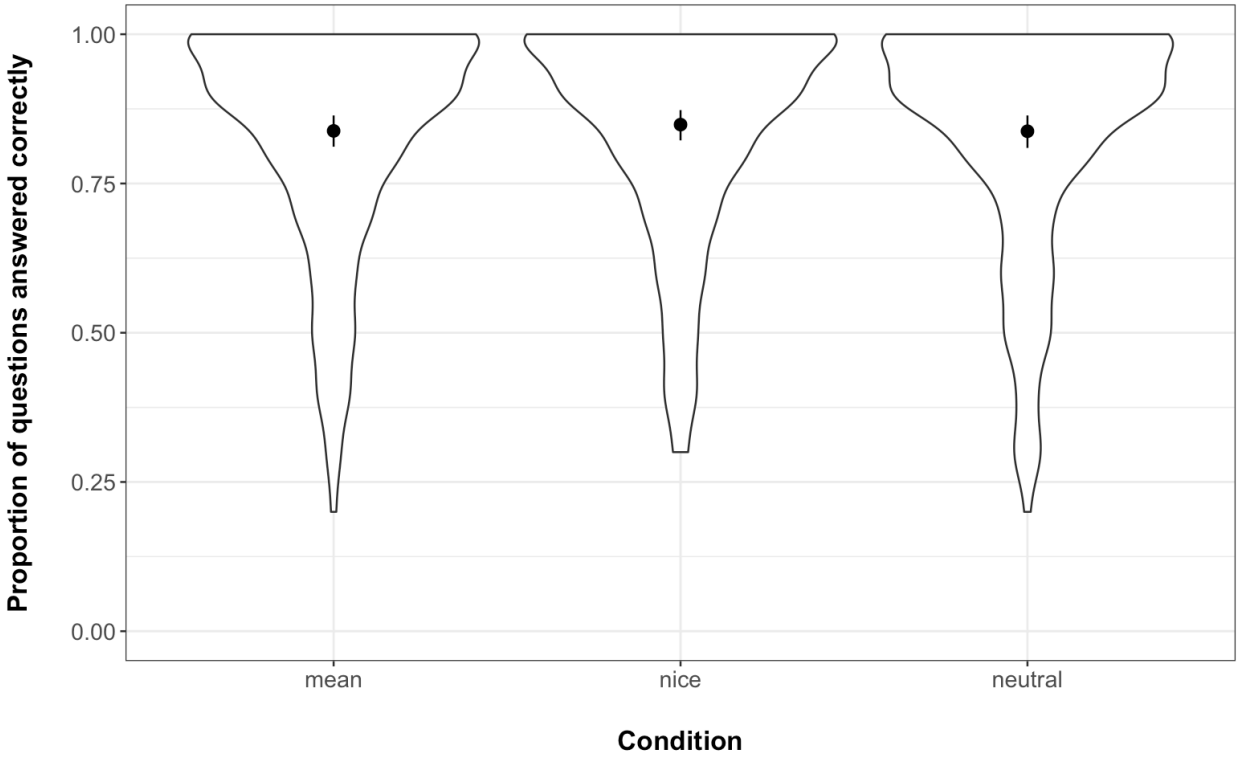

Figure 2. Proportion of questions adults answered correctly, by condition (pooled across the two study versions). The distributions of proportions are depicted, as well as the mean and 95% CI.

Table 2

Effects from the Full Model in Combined Study 2a and Study 2b: Response Term is Response on False Belief Question

Predictors	Odds Ratios	Std. Beta	CI	Standardized CI	z	p
(Intercept)	1.47	2.34	0.62 – 3.51	1.90 – 2.79	10.31	< 0.001*
Condition - Nice [reference: Mean]	1.13	0.12	0.93 – 1.38	-0.08 – 0.32	1.22	0.22
Condition - Neutral [reference: Mean]	0.99	-0.01	0.81 – 1.21	-0.20 – 0.19	-0.08	0.93
Gender [reference: Female]	0.96	-0.04	0.63 – 1.46	-0.46 – 0.38	-0.19	0.85
Age (scaled)	1.05	0.56	1.03 – 1.08	0.35 – 0.78	5.26	<0.001*
Random Effects						
σ^2	3.29					
$\tau_{00\ Subject}$	1.55					
$\tau_{00\ Item}$	0.59					
ICC	0.39					
$N_{Subject}$	189					
N_{Item}	30					
Observations	5670					
Marginal R^2 / Conditional R^2	0.057 / 0.428					

Note. Reference levels are provided above. * $p < 0.05$, two-tailed

General Discussion

Two sets of studies test whether negative (mean) social interactions, as opposed to positive (nice) interactions, lead to differential deployment of 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 (here, the person impacted by the actor's actions). 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 versus positive 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.

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 increased ToM that paralleled negative interactions or whether negative interactions no longer led to an increase in ToM. Because performance in the mean and nice conditions were no different from baseline, we can infer that adults do not show increased ToM deployment for negative interactions, as opposed to also showing increased ToM for positive interactions.

The question investigated in this work opens up exciting new avenues for research. One avenue for future research could explore why the pattern differs for adults and children, and also whether the adult pattern results from experience in navigating an increasingly complex social

world, in which people must toggle between cooperative and competitive contexts more efficiently (and deploy ToM robustly for both).

A second line of research can explore the scope of this socially-dependent asymmetry in ToM. In this set of studies, participants observed others' interpersonal actions. It may be the case that a socially-dependent asymmetry toward negative interactions may be found when observing but not directly engaging in interactions. Indeed, our prior work has found that when preschool children are directly engaging in interactions that require false belief understanding (i.e., require children to plan false beliefs in others' minds), younger preschoolers perform better when they are working toward a cooperative goal versus a competitive goal (Tsoi et al., in press). One potential, albeit difficult to test, theory is that ToM developed primarily to promote cooperative interactions, which could illuminate how human societies are so adept at starting and maintaining large and complex cooperative ties with one another. However, when observing others' interactions, ToM may better serve as a cheater detection mechanism, identifying actors who do not adhere to cooperative standards, especially in situations involving social exchanges (Cosmides, 1989). Characterizing the nuances of socially-dependent asymmetries in ToM will help disentangle the mixed findings in the literature.

A third line of research could examine this contextual effect across different ToM tasks. We show that the contextual effect can be found in a classic explicit ToM task that asks children to deliberately consider agents' beliefs, but it is an open question as to whether this effect can also be found for ToM tasks assessing implicit ToM. A large body of psychological literature with adults indicate that there are cases in which people do not automatically or spontaneously attribute minds to others (e.g., dehumanization; for reviews, see Fiske, 2009; Haslam, 2006). Whether children, too, will rely on contextual information in their spontaneous ToM deployment,

often assessed with spontaneous measures such as eye gaze (Baillargeon et al., 2010; Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017), is a topic for future research. If children also show this contextual effect for implicit ToM tasks, this result would indicate that social context plays an even greater role in the development of sociocognitive processes. Future tests of this question can provide further insight into the limits of the contextual effect with regard to ToM.

Overall, these studies contribute to a broader line of research that investigates when and how people reason about the minds of others. These findings support the proposal that social context influences early ToM deployment in a way that is not observed in adulthood.

Acknowledgments

We thank Sophie Cooper for assistance with data collection. We also thank members of the Boston College Morality Lab and anonymous reviewers for comments on earlier versions of this manuscript. We would also like to thank the staff of the Living Lab and Science World at TELUS World of Science and participating families. This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada (A.S.B.; # 435-2013-0286), a grant from the National Science Foundation (L.Y. and A.W.; grant number 1627157) and a National Science Foundation Graduate Research Fellowship (L.T.; grant number 1258923).

Open Practices

Complete materials (e.g., scripts and scenarios) for all studies are provided in Supplementary Material. Complete data, and data analysis code for all studies can be found here: https://github.com/tsoices/ToM-mean-nice-children-adults.

References

Aloise, P. A. (1993). Trait confirmation and disconfirmation: The development of attribution biases. *Journal of Experimental Child Psychology*, *55*(2), 177–193. https://doi.org/10.1006/jecp.1993.1010

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

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.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

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

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889.

Fiske, S. T. (2009). From dehumanization and objectification to rehumanization: Neuroimaging studies on the building blocks of empathy. *Annals of the New York Academy of Sciences*, *1167*(1), 31–34. https://doi.org/10.1111/j.1749-6632.2009.04544.x

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

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341

Kanouse, D. E., & Hanson Jr., L. R. (1987). Negativity in evaluations. In *Attribution: Perceiving the causes of behavior.* (pp. 47–62). Lawrence Erlbaum Associates, Inc.

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

Kiley Hamlin, J., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations: Social evaluation by 3-month-old infants. *Developmental Science*, *13*(6), 923–929. https://doi.org/10.1111/j.1467-7687.2010.00951.x

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

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522. https://doi.org/10.1037/0033-295X.108.3.483

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.)*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

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.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

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

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

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

Tsoi, L., Hamlin, J. K., Waytz, A., Baron, A. S., & Young, L. L. (in press). A cooperative advantage for theory of mind in children and adults. *Social Cognition*.

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.

Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy Into social cognition research. *Psychological Inquiry*, *22*(3), 159–182. https://doi.org/10.1080/1047840X.2011.551743