Report

Current Biology

Deconstructing Theory-of-Mind Impairment in High-Functioning Adults with Autism

Highlights

- Novel task enables model-based analysis and decomposition of theory of mind (ToM)
- Learning of intentions was impaired in high-functioning adults with autism (ASD)
- ToM impairment in ASD was specific; belief tracking and ToM reasoning remained intact
- Model parameters correlate with ASD symptom severity, pointing to new research targets

Authors

Isabelle A. Rosenthal, Cendri A. Hutcherson, Ralph Adolphs, Damian A. Stanley

Correspondence

dstanley@adelphi.edu

In Brief

Rosenthal et al. use a new task and model-based approaches to provide a detailed characterization of theory-of-mind impairment in autism spectrum disorder (ASD). High-functioning adults with ASD were specifically impaired at using others' beliefs to learn their intentions. Model parameters correlated with symptom severity and point to new targets for research.

Rosenthal et al., 2019, Current Biology *29*, 513–519
February 4, 2019
https://doi.org/10.1016/j.cub.2018.12.039

CellPress

Current Biology
Report

Deconstructing Theory-of-Mind Impairment in High-Functioning Adults with Autism

Isabelle A. Rosenthal,[1,5] Cendri A. Hutcherson,[2,3] Ralph Adolphs,[1] and Damian A. Stanley[1,4,5,6,*]
[1]Division of Humanities and Social Sciences, California Institute of Technology, 1200 E. California Boulevard, MC 228-77, Pasadena, CA 91125, USA
[2]Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON M1C 1A4, Canada
[3]Department of Marketing, Rotman School of Management, University of Toronto, Toronto, ON M5S 3E6, Canada
[4]Derner School of Psychology, Adelphi University, 1 South Avenue, Garden City, NY 11530, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: dstanley@adelphi.edu
https://doi.org/10.1016/j.cub.2018.12.039

SUMMARY

Inferring the beliefs, desires, and intentions of other people (“theory of mind,” ToM) requires specialized psychological processes that represent the minds of others as distinct from our own [1–3]. ToM is engaged ubiquitously in our everyday social behavior and features a specific developmental trajectory [4] that is notably delayed in children with autism spectrum disorder (ASD) [5, 6]. In healthy individuals, model-based analyses of social learning and decision-making have successfully elucidated specific computational components of ToM processing [7–11]. However, the use of this approach to study ToM impairment in ASD has been extremely limited [10, 12]. To better characterize specific ToM impairment in ASD, we developed a novel learning task and applied model-based analyses in high-functioning adults with ASD and matched healthy controls. After completing a charitable donation task, participants performed a “mentalizer” task in which they observed another person (the agent) complete the same charity task. The mentalizer task probed the participants’ ability to acquire and use ToM representations. To accurately predict agent behavior, participants needed to dynamically track the agent’s beliefs (true or false) about an experimental context that varied over time and use that information to infer the agent’s intentions from their actions. ASD participants were specifically impaired at using their estimates of agent belief to learn agent intentions, though their ability to track agent belief was intact and their reasoning about belief and intentions was rational. Furthermore, model parameters correlated with aspects of social functioning, e.g., ADOS severity scores [13]. Together, these results identify novel, and more specific, targets for future research.

RESULTS

Task Performance Accuracy

We designed and carried out a novel social-learning task (Figure 1) in 26 high-functioning adults with autism spectrum disorder (ASD) and 53 matched healthy controls (CTL) (see Table S1 for participant characterization). After first learning a task that involved donating to charities (“Charity task”; Figure S1), participants next performed a task in which they observed another person (the Agent) carry out the same Charity task they had just completed (“Mentalizer task”; Figure 1). To accurately predict the Agent’s behavior, participants needed to learn about and dynamically track the Agent’s true or false beliefs about the experimental context (belief) and infer the Agent’s intentions with respect to their donations (intent). Importantly, Agent beliefs were a replay of actual choices made during the charity task by a real (healthy) individual (see STAR Methods for details on Agent behavior). The dissociation of intentions and beliefs has been a powerful tool in research on theory of mind (ToM) and moral judgment [14]; our task is unique in requiring participants to update these representations continuously throughout the trials of the task, making possible a model-based analysis of the computations that support them.

To facilitate out-of-sample prediction analyses (see model-based analysis below), CTL participants were first split into two groups (CTL1 and CTL2) on the basis of demographic information (n_{CTL1} = 27, n_{CTL2} = 26; each optimally matched to ASD group for age, IQ, gender, and education; see Table S1). We then examined overall accuracy and learning performance in the Mentalizer task (see Figure 2 and STAR Methods for descriptions of the specific variables). This revealed a striking pattern: despite no observable differences at the beginning of the experiment, as trials progressed, the performance of the ASD group diverged from that of the CTL groups in a manner consistent with impairment at learning Agent intentions. To quantify this learning effect, we compared blocks of trials at the beginning of the experiment to blocks of trials at the end of the experiment (Figure 2B; Table S3). (Note: Throughout this paper, we eschew presentation of p values and instead present only bootstrapped confidence intervals [CI] [15].) By contrast, estimates of Agent belief were above chance for all groups (Figure 2A; Table S3) and overlapped in their confidence intervals (while belief

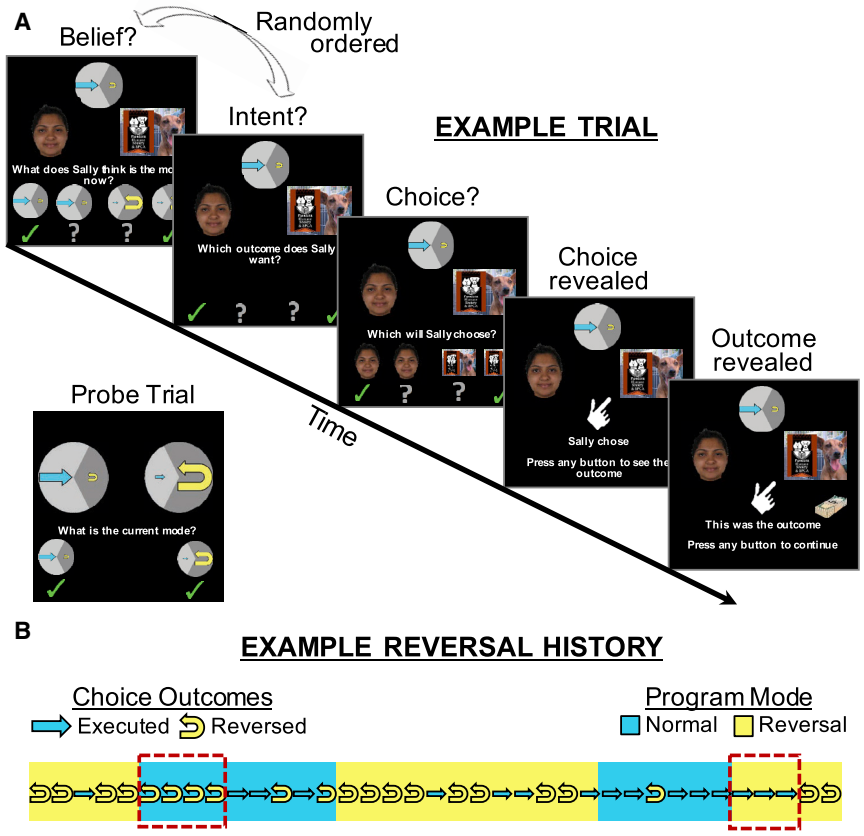

Figure 1. Structure of the Mentalizer task

(A and B) After completing the "Charity task" (see Figure S1), participants completed the "Mentalizer task" (A), in which they observed another person's (the Agent) choices while that person completed the charity task. On each trial, Agents chose whether to donate to charities in two contexts: "normal" mode (in which 36% of agent decisions were subsequently reversed by the computer) and "reversal" mode (64% of decisions reversed). Thus, to obtain desired outcomes, Agents needed to take context into account. Importantly, the context switched every 3–12 trials, and though Agents knew that the mode was "stable across multiple trials," they were unaware of when switches occurred (note: switches were explicitly revealed to the participant). (B) The probabilistic nature of the context and choice reversals meant that Agents regularly had false beliefs [14] about the context. On each trial, participants viewed pictures of the agent and one of three charities and were asked to estimate (1) the Agent's belief about the current mode (normal/reversal), (2) the Agent's intent (donate/keep), and (3) the Agent's choice (donate/keep), which critically depended on the logical integration of belief and intent. Four response options enabled participants to indicate more certain (check marks) and less certain (question marks) responses; for analysis all responses were binarized (two left options pooled, two right options pooled). Following the response, Agent choice, and whether it was reversed by the computer, were revealed. Probe trials required participants to report the actual program mode (normal/reversal), ensuring attention. Red-dotted boxes indicate trials where a false belief should be inferred. (Note: Random ordering of belief and intent questions had no influence on performance.) See also Figure S1 and Methods S1.

performance was slightly better for CTL compared to ASD groups, this difference was not born out by the more sensitive model-based approaches; see below). This pattern indicates that CTL, but not ASD, participants were able to use their representations of Agent beliefs to correctly interpret agent actions and therefore learn about the Agent's intentions. Whereas belief and intent are temporally persistent states that are inferred over multiple trials, predicting an Agent's choice is unique for each trial. Given that the probabilistic values of belief and intent needed to be integrated to generate choice predictions, participants were close to chance in accurately estimating choice (though both CTL groups exceeded chance performance and ASD did not; Table S3), and we found no effect on the relatively insensitive measure of choice learning over the course of the task (Figure 2C).

To ensure the behavioral impairment we identified in the ASD group did not result simply from a failure to understand our task, we first assessed participants' reasoning consistency (i.e., their belief and intent estimates should logically predict their choice estimates in each trial). Both ASD and CTL consistency rates were well above chance (Figure 2D; Table S3) and overlapped in their confidence intervals, indicating that ASD participants understood the logic of the task, even though they had difficulty learning agent intent. Second, we examined Charity task performance (see STAR Methods for details) and found that both ASD and CTL participants performed equivalently ($Belief_{CTL}$ = 0.67, $CI_{95\%}$ [0.64, 0.71]; $Belief_{ASD}$ = 0.71, $CI_{95\%}$ [0.64, 0.78]; Outcome $Desired_{CTL}$ = 0.58, $CI_{95\%}$ [0.56, 0.61]; Outcome $Desired_{ASD}$ = 0.58, $CI_{95\%}$ [0.54, 0.62]; $Consistency_{CTL}$ = 0.78, $CI_{95\%}$ [0.74, 0.82]; $Consistency_{ASD}$ = 0.73, $CI_{95\%}$ [0.67, 0.80]), indicating that all participants had understood the structure and requirements of the task and that ASD participants had no difficulty knowing their own intentions. Finally, performance in reporting what the current mode was (normal or reversal, probed at random; see Supplemental Information) was statistically equivalent across groups (accuracies: $Probe_{CTL}$ = 0.94, $CI_{95\%}$ [0.91, 0.96], $Probe_{ASD}$ = 0.91, $CI_{95\%}$ [0.84, 0.94]), again indicating that all participants understood the task. To summarize, standard performance metrics of accuracy and learning of the ToM components (belief and intent) in the Mentalizer task revealed a disproportionate impairment: high-functioning adults with ASD were impaired in their ability to use ToM to infer the intentions of another person. We next applied a model-based approach to further elucidate the learning computations that underlie these initial findings.

Model-Based Analyses

Data from each group (ASD/CTL) were fit using a combination of modified Rescorla-Wagner reinforcement learning models (Figure 3A; hierarchical fitting; see STAR Methods and Figure S2). Our *a priori* model (M1, Figures 3A and S2) had two free parameters, λ_B (learning rate for belief) and λ_I (learning rate for intent), and

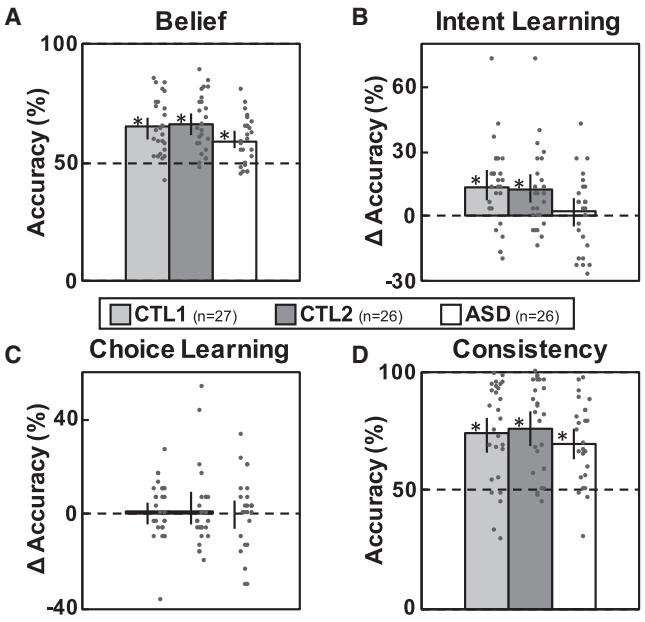

Figure 2. Behavioral Results

Accuracy and learning performance for each group (CTL1, light gray; CTL2, dark gray; ASD, white).

(A) Belief: Overall accuracy (%) of belief estimates (normal/reversal) was above chance for all groups, which did not differ from one another.

(B) Intent learning (% change in mean intent accuracy between first 30 and last 30 trials): Both CTL groups showed significant improvements in the ability to predict Agent intent (donate/keep), whereas ASD performance was at chance.

(C) Choice learning (% change in mean choice accuracy between first 30 and last 30 trials): No groups showed significant improvements in mean choice accuracy; however, we note that, for both CTL groups (and not for ASD), overall choice accuracy was above chance performance (see Supplemental Information). Note: The overall pattern of learning results (B and C) is robust to the number of trials included in % change analysis (examined for 20-, 30-, and 40-trial averages).

(D) Participant consistency (%): Trials were consistent if participants' answers for choice were logically predicted by their answers for belief and intent. Both ASD and CTL had high consistency rates that did not differ from each other.

For all panels, error bars represent bootstrapped 95% confidence intervals (CIs), asterisks above bars indicate 95% CIs that exclude chance accuracy, asterisks between bars indicate that 95% CIs between participant groups do not overlap, and individual participant data is represented by gray points. See also Tables S1–S5.

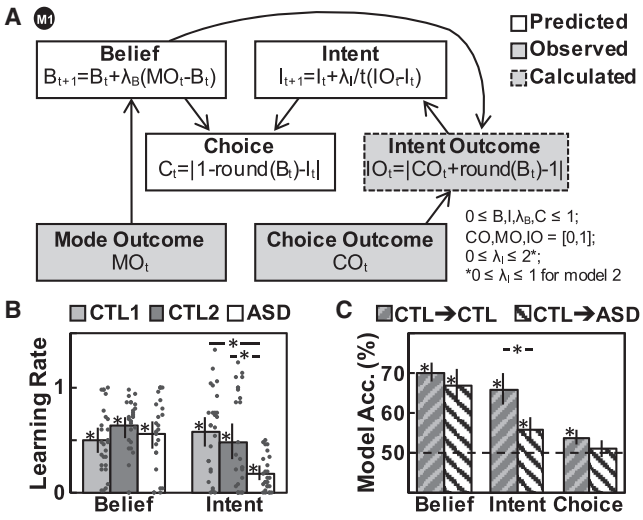

Figure 3. Modeling Results

(A) Schematic of the modified Rescorla-Wagner learning model (M1; see Figure S2 for details). On each trial, participants had to integrate the Agent's belief (B_t) and intent (I_t) to predict their choice (C_t). Inferences of Agent belief (B_t) could be updated directly using the mode outcome (MO_t; i.e., whether a reversal of Agent choice occurred). The choice outcome (CO_t) needed to be viewed within the context of the Agent's current belief (B_t) to assign the appropriate intent outcome (IO_t). Intent learning rate diminished over time.

(B) Learning rates were similar across groups (CTL1, light gray; CTL2, dark gray; ASD, white) for belief (λ_B), but ASD intent learning rates (λ_I) were significantly lower than for both CTL1 and CTL2*. Individual participant data is represented by gray points.

(C) Out-of-sample model accuracy (%; see text for details). Dark and light gray bars indicate models fit on one CTL group and tested on the other; black and white bars indicate models fit on either CTL group and tested on ASD. Out-of-sample model accuracy was similar across groups for belief but significantly worse in ASD for intent and choice (which did not differ from chance). See also Figure S2; Tables S1, S2, S4, and S6; and STAR Methods.

For (B) and (C), error bars represent bootstrapped 95% confidence intervals (CIs). Asterisks above bars indicate 95% CIs that exclude either zero (B) or chance accuracy (C); asterisks between bars indicate that 95% CIs between participant groups do not overlap.

*Note: Only belief and intent learning rates were calculated; choice predictions were computed using the belief and intent model estimates (see A). In addition, because choice performance is the integration of two probabilistic estimates (belief and intent) that are <1, model choice performance is expected to be lower than for belief and intent in (C).

was constructed so that belief estimates could flexibly update over the course of the experiment (as would happen in reversal learning), whereas intent learning rates for each charity decreased as the experiment progressed (reflecting an assumption that participants would expect Agent preferences to be stable).

The learning rates for belief (Figure 3B) were greater than zero for both CTL (CTL1: λ_B = 0.49; $CI_{95\%}$ [0.38, 0.61]; CTL2: λ_B = 0.63; $CI_{95\%}$ [0.52, 0.72]) and the ASD (λ_B = 0.56; $CI_{95\%}$ [0.42, 0.68]) groups, and there were no group differences ($CI_{95\%}$ highly overlapping). In sharp contrast, the mean intent learning rate for the ASD group (λ_I = 0.18, $CI_{95\%}$ [0.12, 0.25]) was well below that of either CTL group (CTL1: λ_I = 0.58, $CI_{95\%}$ [0.44, 0.72]; CTL2: λ_I = 0.48, $CI_{95\%}$ [0.32, 0.66]), demonstrating a selective deficit ($CI_{95\%}$ nonoverlapping between groups) in learning the intentions of Agents by the ASD group.

We further assessed the robustness of model predictive performance by using the models that were fit to each CTL group to predict out-of-sample CTL and ASD data. To do this, λ_B and λ_I were estimated separately for each group (CTL1 or CTL2) and subsequently used to predict the responses for the other two groups (CTL2 or CTL1, and ASD; Figure 3C). Models estimated with CTL data were more accurate at predicting out-of-sample CTL (compared to ASD) intent and choice performance ($Intent_{DiffAcc}$ = 10.4% $CI_{95\%}$ = [5.3%,15.7%], $Choice_{DiffAcc}$ = 2.8% $CI_{95\%}$ = [0.1%, 5.5%]) but not belief performance ($Belief_{DiffAcc}$ = 3.3% $CI_{95\%}$ [-1.1%, 7.6%]).

A strength of model-based approaches is that they specify how a process is implemented with mathematical precision; however, it is important to rule out alternative models that could

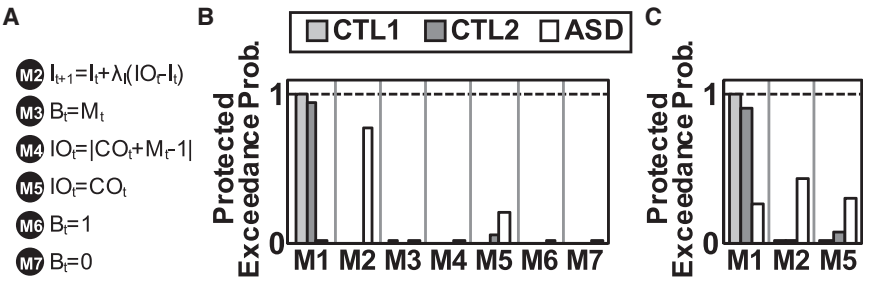

Figure 4. Consideration of Alternative Models

(A) Equations indicating alterations to the *a priori* model (M1) for models 2–7 (see STAR Methods and Figure S2 for full details). In M2, intent learning rate did not attenuate over time (simple Rescorla-Wagner); in M3, the actual mode shown to participants replaced all belief estimates; in M4, the actual mode was used to calculate intent outcomes but not belief estimates; in M5, choice outcomes were used directly instead of being interpreted in the context of belief; in models M6 and M7, belief for the entire experiment was set to normal and reversal, respectively.

(B) Bayesian model comparison results: Protected exceedance probabilities (i.e., how likely it is that any given model is more frequent than all other models) for each model by group. For both CTL1 (light gray; Bayesian omnibus risk [i.e., the posterior probability that all model frequencies are equal; bor] = 5.9e−6) and CTL2 (dark gray; bor = 1.3e−5) groups, the likelihood of M1 clearly exceeded that of all other models, whereas for the ASD group (white; bor = 0.016), M2 and M5 were the most likely.

(C) A second model comparison between the three most likely models (M1, M2, and M5) confirmed the high likelihood of M1 for the CTL groups (CTL1 bor = 0.004; CTL2 bor = 0.069) but was unable to distinguish between the three models for the ASD group (bor = 0.775). See also Figure S2 and Tables S1 and S2.

better explain the data. As it is impossible to directly test the infinite space of all possible models, we tested six additional models (in addition to our *a priori* model M1; Figure 4A, see legend and Figure S2 for details on how the models varied) that represented specific alternative approaches participants could have used when completing the Mentalizer task. To determine which of these seven candidate models best explained performance in each group (CTL1, CTL2, and ASD, respectively) we performed Bayesian model selection analyses (Figure 4B; see [16] and STAR Methods for details). These analyses (Figure 4B) confirmed that our *a priori* model (M1; Figure 3A) best explained performance for both CTL groups (CTL1: M1 protected exceedance probability [i.e., how likely it is that any given model is more frequent than all other models; pxp] = 0.99, Bayesian omnibus risk [i.e., the posterior probability that all model frequencies are equal; bor] = 5.9e−6; CTL2: M1 pxp = 0.95, bor = 1.3e−5); however, this was not the case for the ASD group, whose performance was better described by two alternative models (bor = 0.016): M2 (pxp = 0.77), in which intent estimates did not stabilize over time, and M5 (pxp = 0.21), in which intent updates are based directly upon Agents' actions without accounting for agent beliefs. As Bayesian model selection is a relative measure, we reran the Bayesian model selection analysis on the subset of models (M1, M2, and M5) identified as most likely in the initial analysis (Figure 4C). This second analysis again confirmed that M1 was the best model for both CTL groups (CTL1: M1 pxp = 0.99, bor = 0.004; CTL2: M1 pxp = 0.91, bor = 0.069), while ASD performance was found to be more heterogeneous, with no model emerging as the most likely (M1 pxp = 0.26, M2 = 0.44, M5 = 0.30, bor = 0.775). Exploratory analyses (data not shown) of ASD participants' individual model fits (based on Bayesian Information Criterion, BIC), did not reveal a clear explanatory pattern for which model best explained the data likely because subgroup sample sizes were too small.

Analysis of Individual Differences

Finally, we examined the relationships between participants' ToM ability (behavioral accuracy and model parameters) and their social functioning as measured by the Autism Diagnostic Observation Schedule (ADOS) [17, 18]. In accordance with our *a priori* hypothesis (that social dysfunction in ASD arises in part from ToM learning impairment), we observed negative correlations (Pearson's r) between behavioral measures of belief accuracy (r = −0.39 $CI_{95\%}$ = [−0.68, −0.03]) and intent learning (r = −0.46 $CI_{95\%}$ = [−0.78, −0.07]) and the Social Affect (SA; but not the Restricted and Repetitive Behaviors [RRB]; see Table S4) component of ASD participants' ADOS Calibrated Severity Scores (CSS) [13]. The relationship with belief accuracy is particularly interesting as it suggests that multiple learning processes (one independent of ASD-related social impairment, and the other not) contribute to performance. The lack of a relationship with RRB CSS scores is consistent with previous findings of intact non-social learning in ASD [19]. A similar pattern of results was seen for model belief and intent learning rates (Table S6), though confidence intervals for these correlations did not exclude zero. In addition to these *a priori* analyses, we conducted post-hoc exploratory analyses of the relationship between ToM ability and a number of laboratory measures that may index real-world social functioning (Tables S5 and S6). Only four correlations had 95% confidence intervals that excluded zero after correcting for multiple comparisons, and all were in the predicted direction: better performance on metrics of our experimental task was positively correlated with Social Network Index (Diversity) [20] and the "Reading the Mind in the Eyes Task" [21] and negatively correlated with Autism Quotient [22]. All other substantial correlations (regardless of whether their 95% CIs excluded zero) were directionally supportive of the interpretation that impairments on our experimental task are associated with impairments in social behavior. These patterns of correlations support the external validity of our task and point to future studies that could identify its real-world correlates with further precision.

DISCUSSION

Using a novel ToM learning task together with model-based analyses, we uncovered a specific impairment in ASD: an impairment in the ability to use an understanding of another person's beliefs in order to learn about their intentions from observing their choices. Individuals with ASD were able to track an Agent's

beliefs about choice context, updating this as the context (normal or reversal mode) changed, and they rationally integrated their estimates of belief and intent to predict the Agent's choice (i.e., they were consistent, even if not accurate). These findings argue against a nonspecific learning or reasoning deficit and instead suggest that ASD may feature impairment in a rather specific component of ToM, which in our task corresponds to inferring other peoples' intentions from their actions while accounting for their beliefs.

There is vigorous debate about the psychological processes that constitute ToM [23, 24] as well as about the neural systems that subserve it [25]. These debates highlight the need to decompose ToM into component processes, which could in turn be related to individual differences and psychopathology. This broad aim has recently received considerable attention (e.g., in the RDoC initiative from the National Institute of Mental Health [26]) and has yielded initial efforts at computational modeling of ToM [8, 10–12, 27–29]. The benefits of model-based approaches for understanding the social impairments in disorders such as ASD are 2-fold. First, by identifying latent factors that are not directly observable in behavior, a model-based analysis can identify new targets for study and intervention. It is also likely that such latent factors provide a closer correspondence to neural processes that can subsequently be investigated with neuroimaging. Second, by decomposing the component processes of ToM and identifying individual variation in specific processes, model-based investigations provide a much more fine-grained characterization of ASD and its possible subtypes, aiding diagnosis and ultimately moving toward personalized medicine. While our model-based approach requires the complexity of a task that can be decomposed and a large number of trials, making it challenging as a clinical instrument (e.g., the approximate time taken to complete the experiment was 2 h), an important future direction would be to design simpler and more compact tasks that might focus just on one component.

The current study thus extends the findings of previous model-based studies of social cognition in a number of ways. First, we introduce a model-based task derived from classic ToM tasks (e.g., requiring the representation of true and false beliefs to infer intention) that permits the deconstruction of specific ToM component processes and facilitates the discovery of how they combine to produce ToM. Second, unlike other model-based studies of social cognition, our study examines learning about others in the absence of reward, thereby avoiding possible confounds. Finally, it is the first study to decompose these processes in a population that meets diagnostic criterion on the ADOS, a gold standard in autism research.

The model selection analyses revealed that the computational mechanisms through which individuals with ASD implement ToM learning are more heterogeneous than, and possibly distinct from, those used by healthy controls (whose data were remarkably consistent). In particular, the findings suggest that ASD participants were more likely than CTL participants to follow the Agent's actions (CO) without considering their beliefs (Model 5), and/or their estimates of Agent intent were less likely than those of CTL participants to stabilize over time. While beyond the scope of the current study, a clear next step is to collect data from a much larger sample in a targeted study with a revised task that has been optimized to distinguish between the identified model alternatives. While we would expect heterogeneity to emerge, its source remains an open question that will additionally require longer tasks and test-retest validation, making its full elucidation challenging. It is possible that high-functioning individuals with ASD are homogeneous but noisy in their performance from trial to trial, it is possible that there are distinct subtypes with different individuals behaving according to different models, and it is possible that within-subject variability arises not trial-wise but over longer temporal epochs. Some of these possibilities could be tested under hypotheses also motivated by other studies, for instance, the finding that neural computations in ASD are less reliable [30].

There are important constraints on the generality of our findings [31]: our task was artificial and did not involve actual social interactions with people (to increase experimental control), and the task decomposed an aspect of social cognition that is normally encountered as a complex combination of processes in the real world (to achieve the aim of identifying which component might be disproportionately impaired). These constraints, respectively, make it important for future work to attempt to design ecologically valid tasks that better mimic the real world (perhaps using videos or virtual-reality scenarios) and to verify that our task did not introduce confounding processes that could explain the impairment (a difficult challenge that can only be approached by using a large diversity of different task designs). Indeed, the power to detect impairments depends crucially on the exact nature of the task [32]. In order to further decompose the present findings and ultimately link them to neurobiological computations, it will also be important to disambiguate several possible mechanisms that could account for the results thus far. For instance, does impairment result from an inability to correctly interpret actions in context (disrupting the computation of IO_t), is intention-updating specifically impaired (whereby λ_{Int} is influenced), or is learning intact but the readout of intention information impaired (influencing the use of I_t)? Some of these questions could be illuminated by conducting the task we described in conjunction with neuroimaging, to identify the neural systems engaged by the component processes. Finally, although the present study is limited to high-functioning adults with ASD (necessitated by the demands of the task), the deficit uncovered here could be further tested with simplified tasks in children and lower-functioning individuals. This could provide a finer-grained understanding of how ToM processes are implemented and change throughout development as well as help identify new targets for exploration and intervention. An ultimate clinical goal would be to design a much shorter version of the task that distills its test of specific ToM components (such as the intent-learning process revealed here) and validate its diagnostic and prognostic value.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, six tables, and one methods file and can be found with this article online at https://doi.org/10.1016/j.cub.2018.12.039.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Antonio Rangel for help with initial experimental design; Drs. Daniel McNamee and Jeff Cockburn for help with model-based analyses; and Ghoncheh Ayazi, Marisol Espino, Lynn Paul, Tim Armstrong, Remya Nair, and other members of the Adolphs lab for help with data collection, database management, data analysis, and autism assessment. This work was supported by the National Institute of Mental Health at the National Institutes of Health (K01MH099343 to D.A.S. as well as P50MH094258 and R01MH080721 to R.A.).

AUTHOR CONTRIBUTIONS

D.A.S. and C.A.H. developed the initial experimental concept. I.A.R. and D.A.S. contributed to data collection and analysis. All authors contributed to experimental design and discussion of results. I.A.R., R.A., and D.A.S. contributed extensively to write up of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 24, 2018
Revised: November 22, 2018
Accepted: December 20, 2018
Published: January 24, 2019

REFERENCES

1. Leslie, A.M. (1987). Pretense and representation: The origins of "theory of mind." Psychol. Rev. *94*, 412–426.
2. Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behav. Brain Sci. *1*, 515–526.
3. Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition *13*, 103–128.
4. Perner, J., and Lang, B. (1999). Development of theory of mind and executive control. Trends Cogn. Sci. *3*, 337–344.
5. Baron-Cohen, S. (1995). Mindblindness: An Essay on Autism and Theory of Mind, Revised Edition (The MIT Press).
6. Frith, U. (2003). Autism: Explaining the Enigma, Second Edition (Malden, MA: Wiley-Blackwell).
7. Stanley, D.A., and Adolphs, R. (2013). Toward a neural basis for social behavior. Neuron *80*, 816–826.
8. Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. Nature *456*, 245–249.
9. Joiner, J., Piva, M., Turrin, C., and Chang, S.W.C. (2017). Social learning through prediction error in the brain. Npj Sci. Learn. *2*, 8.
10. Sevgi, M., Diaconescu, A.O., Tittgemeyer, M., and Schilbach, L. (2016). Social Bayes: Using Bayesian Modeling to Study Autistic Trait-Related Differences in Social Cognition. Biol. Psychiatry *80*, 112–119.
11. Diaconescu, A.O., Mathys, C., Weber, L.A.E., Daunizeau, J., Kasper, L., Lomakina, E.I., Fehr, E., and Stephan, K.E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. PLoS Comput. Biol. *10*, e1003810.
12. Yoshida, W., Dziobek, I., Kliemann, D., Heekeren, H.R., Friston, K.J., and Dolan, R.J. (2010). Cooperation and heterogeneity of the autistic mind. J. Neurosci. *30*, 8815–8818.
13. Hus, V., and Lord, C. (2014). The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. J. Autism Dev. Disord. *44*, 1996–2012.
14. Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. USA *104*, 8235–8240.
15. Cumming, G. (2014). The new statistics: why and how. Psychol. Sci. *25*, 7–29.
16. Rigoux, L., Stephan, K.E., Friston, K.J., and Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. Neuroimage *84*, 971–985.
17. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J. Autism Dev. Disord. *30*, 205–223.
18. Lord, C., Rutter, M., DiLavore, P.C., Risi, S., Gotham, K., and Bishop, S.L. (2012). ADOS-2: Autism Diagnostic Observation Schedule, Second Edition. Part 1: Modules 1–4.
19. Lin, A., Rangel, A., and Adolphs, R. (2012). Impaired learning of social compared to monetary rewards in autism. Front. Neurosci. *6*, 143.
20. Cohen, S., Doyle, W.J., Skoner, D.P., Rabin, B.S., and Gwaltney, J.M., Jr. (1997). Social ties and susceptibility to the common cold. JAMA *277*, 1940–1944.
21. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J. Child Psychol. Psychiatry *42*, 241–251.
22. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J. Autism Dev. Disord. *31*, 5–17.
23. Apperly, I.A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. Q J Exp Psychol (Hove) *65*, 825–839.
24. Mitchell, J.P. (2005). The false dichotomy between simulation and theory-theory: the argument's error. Trends Cogn. Sci. *9*, 363–364, author reply 364.
25. Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., and Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. Trends Cogn. Sci. *19*, 65–72.
26. Insel, T.R. (2014). The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. Am. J. Psychiatry *171*, 395–397.
27. De Martino, B., O'Doherty, J.P., Ray, D., Bossaerts, P., and Camerer, C. (2013). In the mind of the market: theory of mind biases value computation during financial bubbles. Neuron *79*, 1222–1231.
28. Boorman, E.D., O'Doherty, J.P., Adolphs, R., and Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. Neuron *80*, 1558–1571.
29. Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. USA *105*, 6741–6746.

30. Dinstein, I., Heeger, D.J., Lorenzi, L., Minshew, N.J., Malach, R., and Behrmann, M. (2012). Unreliable evoked responses in autism. Neuron *75*, 981–991.
31. Simons, D.J., Shoda, Y., and Lindsay, D.S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. Perspect. Psychol. Sci. *12*, 1123–1128.
32. Senju, A., Southgate, V., White, S., and Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. Science *325*, 883–885.
33. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.
34. Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. *10*, 437–442.
35. Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., and McLennan, J. (1989). Autism diagnostic interview: a standardized investigator-based instrument. J. Autism Dev. Disord. *19*, 363–387.
36. Rutter, M., Bailey, A., and Lord, C. (2003). The Social Communication Questionnaire. Los Angeles: Western Psychological Services.
37. Weschler, D. (1999). WASI: Wechsler abbreviated scale of intelligence (Pearson).
38. Weschler, D. (2011). WASI-II Wechsler abbreviated scale of intelligence, Second Edition (Pearson).
39. Weschler, D. (1997). WAIS-III, Wechsler adult intelligence scale, Third Edition (Pearson).
40. Weschler, D. (1981). WAIS-R: Wechsler adult intelligence scale, Revised Edition (The Psychological Corporation).
41. Weschler, D. (1991). WISC-III Wechsler intelligence scale for children, Third Edition (The Psychological Corporation).
42. Baron-Cohen, S., and Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. J. Autism Dev. Disord. *34*, 163–175.
43. Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., and Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 361–374.
44. Rescorla, R.A., and Wagner, A.W. (1972). Classical Conditioning II: Current Theory and Research, First Edition, A. Black, and W.F. Prokasy, eds. (Appleton Century Crofts).
45. Cumming, G. (2012). Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis (New York: Routledge).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw and analyzed data	This Paper	https://osf.io/ahp5q/
Software and Algorithms		
Charity and Mentalizer task programs and materials.	This Paper	https://osf.io/ahp5q/
MATLAB	Mathworks, Natick, MA, USA	https://www.mathworks.com/products/matlab.html; RRID: SCR_001622
Psychtoolbox 3	[33, 34]	http://psychtoolbox.org/; RRID: SCR_002881
mfit code for Bayesian hierarchical modeling and model comparison.	Sam Gershman	https://github.com/sjgershm/mfit
Code for data analysis and ToM modeling.	This Paper	https://osf.io/ahp5q/

CONTACT FOR RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dr. Damian A. Stanley (dstanley@adelphi.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants (53 psychiatrically and neurologically healthy adults [CTL], 40 male and 13 female; 26 high-functioning adults with autism spectrum disorder [ASD], 21 male and 5 female) (CTL mean age = 31.6, range = [21-59]; ASD mean age = 29.5, range = [20-59]) were recruited from an existing pool in our laboratory. ASD participants were required to be verbal English speakers, and to have met DSM-V/ICD-10 diagnostic criteria for autism spectrum disorder. All met the cutoff scores for ASD on the Autism Diagnostic Observation Schedule-2 (ADOS-2) revised scoring system for Module 4 [13], as well as the Autism Diagnostic Interview-Revised [35] (ADI-R) or Social Communication Questionnaire [36] (SCQ) when an informant was available. All participants also possessed a full-scale IQ score (FSIQ) above 85, determined with one of the versions of the Wechsler Adult Intelligence Scale: the Wechsler Abbreviated Scale of Intelligence, 1st or 2nd edition [37, 38]; the Wechsler Adult Intelligence Scale, 3rd or revised edition [39, 40]; or the Wechsler Intelligence Scale for Children, 3rd edition [41] (1 participant). CTL participants were matched to the ASD group on age, gender, years of education, and IQ (Table S1), and had no family history of ASD. Two CTL and three ASD participants were excluded for not performing the task(s) correctly, based on self-report and experimenter observation. In addition, we tested an independent group of 14 healthy adult subjects, a subset of whom provided the real portion of the Agent data used as feedback in the Mentalizer task (see below); data from these 14 subjects was not used in any of the analyses reported here. All participants had normal or corrected-to-normal vision, gave written informed consent under a protocol approved by the Institutional Review Board of the California Institute of Technology, and were paid for participating in the study.

METHOD DETAILS

Procedure

On the day of the experiment, participants first completed a demographics questionnaire and a "Day of Visit" screening questionnaire containing questions on sleep, drugs or medications taken, and measures of mood. Participants then received an extensive powerpoint briefing (see Methods S1) as well as practice in the presence of the experimenter to familiarize them with the structure and rules of the Charity task (see below, Figure S1). Their comprehension was then assessed with a 4 question quiz (see Methods S1) and for questions that were answered incorrectly, the relevant logic of the task was again discussed with the participant. Participants then completed the Charity task (the final check on their understanding of the task was their behavioral performance, see below). Following the Charity task, participants received a second powerpoint briefing (see Methods S1) describing the Mentalizer task (see below), and then completed that task.

Upon completion of the Mentalizer task, participants filled out a questionnaire in which they indicated the charity-specific preferences (i.e., to donate or take) of the individual they were learning about, and then provided information about any strategies they used during the experiment. The experimenter then randomly selected one trial from each of the Charity and Mentalizer tasks

and, depending on the task, paid the participant the true outcome (Charity task) and/or a reward of 5$ per correct answer (3 possible) and $0.50 for each correct probe trial (Mentalizer task). Finally, participants completed any surveys of interest that had not been completed on a previous visit to the laboratory (see below).

Tasks

The main experimental protocol (the Mentalizer task; Figure 1) consisted of participants learning about another individual's (the Agent) preferences and beliefs from observing that individual's choices in a charitable giving task in varying contexts (the Charity task; Figure S1). To ensure that participants understood the Charity task and the choices that the Agent was making (i.e., to better enable them to take the perspective of the Agent), they completed the Charity task themselves before doing the Mentalizer task.

Charity task (Figure S1)

Participants and Agents performed a charitable giving task in which they decided whether to give money to one of three charities (donate), or take it for themselves (take). On each trial, the participant (or Agent) was shown a picture of one of three charities on one side of the screen, with "$10" displayed underneath it, and the word "you" displayed on the other side of the screen, with a dollar amount (ranging from $7-$13) displayed underneath it. The participant then made a choice (donate or take). The computer program had two modes (or contexts), "normal" and "reversal." In "normal" mode, the program intervened and reversed the participants' choices on 36% of trials (leaving them unaltered on 64% of trials). In "reversal" mode the opposite was true, the program reversed the participants' choices on 64% of trials (leaving choices unaltered on 36% of trials). Therefore, to obtain their desired outcomes most of the time, participants would have to reverse their decisions (i.e., choose what they didn't want) when the program was in "reversal" mode. Importantly, participants were not explicitly aware of the current program mode (except during practice; see below), and instead needed to track it by observing how often the computer was reversing their decisions. Participants were instructed that the mode was "stable across multiple trials" (in actuality 3-12 trials; see Agent data creation below for details), giving them time to learn, and that every so often it would change. Following the participant's choice, the computer's action was displayed (a blue straight arrow indicated that the choice happened as intended, a yellow curved arrow indicated that the choice had been reversed), and non-chosen dollar amount was removed from the screen. Finally, the participant answered 2 follow-up questions: "Was this the outcome you wanted"? (select thumbs up icon for "yes," thumbs down icon for "no") and "what do you think is the mode now"? Answers were given in the form of 4-alternative-forced-choice across icons representing "Definitely Reversal," "Maybe Reversal," "Maybe Normal," "Definitely Normal." All analyses binarized these and all other responses (e.g., reversal, normal). Presentation of all screens was self-paced. The three charities (The Southeast Alaska Conservation Council, Canine Assistants, Pasadena Humane Society and SPCA) were selected based on a previous charitable giving experiment [19] which found that they were equally preferred by individuals with ASD and matched controls. To facilitate understanding of the mode structure, participants completed a series of practice trials (21-25) in which they were explicitly told the program mode ("normal" or "reversal") via the display of a mode-specific icon.

Mentalizer task (Figure 1)

In the Mentalizer task, participants (Mentalizers) watched an Agent perform the Charity task while learning about and tracking the Agent's beliefs (about program mode) and intentions (to donate or take) from trial to trial. Whereas the Agent had to infer the current program mode ("normal"or "reversal") based on the reversal history, the Mentalizer was always explicitly informed about the true mode. This task structure created opportunities for the Mentalizer to represent the Agent's false beliefs about the program mode, a gold standard of traditional ToM tasks [14]. Thus, participants had to represent the Agent's beliefs (true or false), and integrate that representation with their expectation of Agent intent (to donate or take) in order to correctly predict the Agent's actions. On each trial the Mentalizer was asked to answer 3 questions: 1) does the Agent *intend* to donate to the charity or not? ("donate" or "take"), 2) what mode does the Agent *believe* the program to be in? ("normal" or "reversal"), and 3) what *choice* will the Agent make? ("donate" or "take"; the order of questions 1 and 2 is random). Following these predictions, the Mentalizer was first shown what the Agent chose, and then whether the computer reversed the choice or not, and where the money ended up. At random points during the experiment, probe trials required the Mentalizer to indicate the current mode of the program, ensuring that they kept it in mind while tracking the Agent's true and false beliefs. Agent choices were simulated by combining beliefs about the mode (normal or reversal) from real participants on the Charity task with simulated idiosyncratic preferences for the different charities (choices were always 100% consistent with simulated preferences; see simulation of Agent choices below).

Background Assesment

Because participants were part of a database maintained in our laboratory at the California Institute of Technology, they had all completed a battery of different measures related to Social Cognition (see Table S1 for average scores). When available in the database, we obtained scores for the "reading the mind in the eyes" task [21] (RMET), Autism Spectrum Quotient [22] (AQ), the Empathy Quotient [42] (EQ), the Systematizing Quotient [43] (SQ), and the Social Network Index [20] (SNI). When not available in the database, the missing measures were administered on the day of testing.

Autism Diagnostic Observation Schedule

All ASD participants completed the Autism Diagnostic Observation Schedule (ADOS, Module 4; [17]), a method of quantifying the severity of behavioral symptoms of autism in an individual. The main component of an ADOS evaluation is a structured 1-hour interview of a participant by a trained experimenter. A video of this interaction was scored to consensus by trained raters [18] on several metrics. These metrics relied upon the standard algorithm, and on an algorithm that has been developed more recently [13] to include a Social Affect (SA) domain and Calibrated Severity Scores (CSS). Data analysis focused on CSS as recommended; however, for

completeness we provide a summary of both raw and CSS data using the updated algorithm (ADOS2) for our ASD participants (Table S2) as well as their raw scores for sections A and B of the standard algorithm (ADOS).

Apparatus

All analyses were performed in MATLAB (Mathworks Inc., Natick, MA, USA), and behavioral data was collected on an iMac computer using MATLAB and Psychtoolbox 3 [33, 34].

Charity task trial order and Agent generation

Trial orders (i.e., assignment of trial mode, reversal outcome, charity, and monetary outcomes) for the Charity task were constructed to facilitate stability of participant beliefs and increase the probability of false beliefs. Agent choices presented during the Mentalizer task were generated by combining the Charity task beliefs of real participants with simulated preferences for charities. Choices were simulated to ensure that they were 100% consistent with agent preferences as it was determined in early testing that lower agent choice consistency rates significantly impaired performance. Agent behavior was generated according to two algorithms designed to meet the above constraints. No differences were identified in participant performance across the algorithms.

Version 1

The algorithm for the generation of version 1 trial orders was developed over a series of pilot studies that investigated how manipulation of mode contingencies and durations influenced participant beliefs with a focus on prolonged periods of belief stability and maximizing false belief rate. The algorithm that came out of this initial investigation is detailed below.

Generation of version1 trial orders: For the Charity task, two sets of mode (normal/reversal) and reversal outcome (participant choice executed or reversed) histories (105 trials total) were created with the constraints that mode durations (in number of trials) were chosen from two uniform distributions (3-5 and 8-12 trials) in a manner that ensured approximately equal representation of each mode across the full experiment. This configuration was developed over a series of pilot studies and chosen because it increased the probability of false belief trials (i.e., trials in which participants' beliefs about the program mode were incorrect). Then reversal outcomes (executed/reversed) for all trials in a given mode were set to match the contingencies of their particular mode (65% executed for normal mode, 35% executed for reversal mode) and randomly distributed across the trials belonging to that mode. Finally, the charity and agent monetary outcome ($7-$13; uniform distribution) order for each history was pseudo-randomly generated with the following constraints: The exact same combination of charity and agent monetary outcome could not occur on consecutive trials and neither charity type, nor agent monetary outcome, could occur on more than 2 consecutive trials. The charity monetary outcome was kept constant across all trials ($10). One of the resulting two trial orders was randomly assigned to each participant completing the Charity task.

Creation of version 1 agents: To create the Agent data for the Mentalizer task, two additional trial orders were created using the above algorithm and 6 participants were recruited to complete the Charity task as potential agents. Participants were recruited from the population surrounding the California Institute of Technology in Pasadena, CA and were paid $20/hr. The belief data for each participant were fit with a learning model that used a basic Rescorla-Wagner [44] update rule with an unbiased initial belief (i.e., $p = 0.5$). Based on this analysis, two participants (one for each trial order) for whom the belief model predicted a similar level ($\sim$70%) of actual participant beliefs were chosen to serve as agents for the Mentalizer task. For each of these agents, an idiosyncratic set of charity preferences was randomly assigned such that one charity was preferred and the other two disliked. Using these charity preferences (as well as the inverse set of preferences for each agent) two sets of simulated charity choices were created for each agent by combining the real beliefs of the participants from the final 80 trials of their charity data (i.e., after practice) with the randomly assigned charity preferences so that agent behavior was 100% consistent with their actual belief and simulated preference on a given trial. Participants completing the Mentalizer task were randomly assigned one of the resulting 4 (2 agents X 2 inversely-related charity preferences) simulated agents to observe.

Version 2

We developed a second algorithm for generating Charity task trial orders and simulating agent data that had the following goals: 1) elicitation of stable periods of participant (and therefore agent) belief, 2) maximization of potential for false beliefs (i.e., trials on which participant holds incorrect beliefs about the mode), and 3) maximization of agent learnability.

Generation of version 2 trial orders: First, 10,000 sets of mode (normal/reversal) and reversal outcome (executed/reversed) histories (84 trials total) were created. Mode durations (in number of trials) were chosen from two uniform distributions (3-5 trials and 8-12 trials) in a manner that ensured approximately equal representation across all trials and modes. Each mode was in effect for 42 trials overall. The outcomes of each trial (whether or not the Agent's decision was executed or reversed) were generated by setting the total number of executed trials to 27/42 (64.3%), for normal mode and 15/42 (35.7%) for reversal mode, then randomly distributing the executed outcomes across all trials in the respective mode. Probabilistic belief estimates were then simulated using a basic Rescorla-Wagner [44] update rule with an unbiased initial belief (i.e., $p = 0.5$) and a learning rate of 0.3. This yielded 10,000 sets of modes, reversal outcomes, and simulated agent beliefs.

For each simulated belief set, a metric of belief stability was calculated by first binarizing probabilistic beliefs and then calculating the number of belief switches (i.e., changing of belief from one mode to another) across all 84 trials. The 95% of sets with the least stable agent beliefs (i.e., with the most switches) were then discarded. Of the remaining 500 sets, those with binarized belief histories that disagreed with the true mode (i.e., false beliefs) on 30% to 40% of trials were identified ($n = 248$) and the rest discarded.

The remaining sets were then assigned charity information (28 trials per charity). For each set, 100 permutations of the charity and agent monetary outcome order ($7-$13; uniform distribution) were pseudo-randomly generated with the following constraints: The exact same combination of charity and agent monetary outcome could not occur on consecutive trials and neither charity type, nor agent monetary outcome, could occur on more than 2 consecutive trials. The charity monetary outcome was kept constant across all trials ($10). Idiosyncratic agent preferences for the charities were then pseudo-randomly assigned such that simulated agents either preferred one charity and disliked the other two, or they preferred two charities and disliked one. Finally, agent choices were simulated by combining the binarized simulated beliefs with the agent preferences so that agent behavior was 100% consistent with their belief and preference on a given trial.

To assess simulated agent learnability, for each of the 100 charity order permutations within the 248 simulated agent belief and choice histories we simulated Mentalizer belief and intent learning. Belief learning was modeled with a simple Rescorla-Wagner update rule with the learning rate set to 0.681. Intent learning was modeled using the simulated Mentalizer beliefs to interpret the Agent choices and a Bayesian update rule (a simple beta distribution weighted by a learning rate) with a learning rate of 0.5575. Both model learning rates were estimated from fits to pilot data (participants from version 1). The Bayesian model was an early version (and performed similarly) to the RW_overT model for Mentalizer intent learning which stabilizes over time (M1 Figures 3 and S2). Intent model accuracy (binarized intent predictions compared to simulated Agent charity preferences) was calculated for the 4th quarter of trials and sets for which learning was less than 100% accurate were discarded.

Following this, for each of the 248 Agent belief and choice histories, we calculated the average strength (distance from p = 0.5) of the 4th quarter of Mentalizer intent probabilistic model predictions for each of the remaining charity order permutations and selected the set with the highest mean value, yielding 248 Agent belief histories with their associated mode and reversal outcome histories. Of these 248, the top 20 (based on the same strength metric as above) were selected as candidate mode and outcome histories. Finally, for each of the 20 orders, practice trials (n = 21) were selected randomly from another order were added to the front of each order, making each 105 trials long (21 practice, 84 test).

Creation of version 2 agents: In order to generate real (i.e., human) Agent beliefs for use in the Mentalizer experiment, 8 participants were recruited to complete multiple runs (4 histories per session; 1-4 sessions per participant) of the Charity task using the 20 generated mode and reversal histories. Data collection continued until behavioral data for each of the 20 histories had been collected from 4 different participants. Participants were recruited from the California Institute of Technology Summer Undergraduate Research Fellowship program and were paid $20/hr. For each mode and reversal history, we calculated the difference between the average number of times participants believed the mode switched over the course of the experiment and the true number of mode switches (average minus true). Histories were rank ordered, with those that had less behavioral than actual switches (an indicator of higher false belief rates) ranked highest, and the top 10 histories were selected.

The specific belief behavioral data for each history was chosen by selecting the participant whose belief accuracy most closely resembled the average belief accuracy of all 4 participants for that history. Charity preferences for each agent history were simulated in the same manner as in version 1 above. Note: a programing error meant that, for version 2 only, charity preferences were fixed such that agents preferred the Pasadena Humane Society and did not prefer the other 2 charities, given that this was consistent across populations, unknown to the participants, and the charities were matched for interest-level, we are confident this did not bias our findings. Finally, 2 Agents were selected at random for use with data collection in order to match the number of Agents used in Version 1 of the task.

ToM Learning Model

Mentalizer Belief (i.e., whether the current program mode was ''normal'' or ''reversal'') and Intent (i.e., preference for the different charities) estimates were modeled using a combination of modified reinforcement learning models [44] (Figure 3A). The belief model used a simple Rescorla-Wagner update rule and was flexible in order to accommodate switches of program mode (and therefore, belief) over the course of the experiment. The equation for the belief model was:

$$B_{t+1} = B_t + (\lambda_{Bel}) * MPE_t;$$

$$MPE_t = (MO_t - B_t)$$

in which the Mentalizer's probabilistic estimate that the Agent believes the program is in normal mode on the next trial (B_{t+1}) is equal to their probabilistic estimate (B_t; range [0 to 1]) on trial t, plus their mode prediction error on trial t (MPE_t; range [-1 to 1]), scaled by their learning rate (λ_{Bel}; range [0 to 1]). Mode prediction error (MPE_t) was equal to the reversal outcome they observed on trial t (i.e., did the computer reverse their choice or not; MO_t; 0 = reversed, 1 = not reversed) minus their estimate of Agent Belief on trial t (B_t).

The basic equation for the Intent model (see also Figure 3A) was similar to that of the Belief model with a slight adjustment. Given that one's preferences toward different charities are likely to be relatively stable, we modified the equation so that the influence of prediction errors on probabilistic estimates of Agent Intent attenuated over time. In addition, to allow for correct interpretation of choice behavior, the Choice outcomes were modified to take Agent Belief into account, creating Intent outcomes. For each charity there was a separate model tracking intent, and the trial number t refers to the number of trials specific to that charity. The equation for the Intent model was as follows:

$$I_{t+1} = I_t + (\lambda_{Int}/t) * IPE_t;$$

$$IPE_t = (IO_t - I_t);$$

$$IO_t = |CO_t + \text{round}(B_t) - 1|$$

In which I_t is the Mentalizer's probabilistic estimate (range: [0 to 1]) that the Agent intends to donate to a given charity on trial t (t being specific to the charity). IPE_t is the Intent Prediction Error on trial t, which is equal to the difference between the Intent outcome on trial t (IO_t) and I_t. Because the Agent's belief about program mode varies across the experiment, IO_t is a combination of the choice outcome on trial t (CO_t) and the Agent's current belief (B_t; this ensures that the intent outcome is the reverse of the choice outcome when the Agent believes the program is in reversal mode).

Finally, Mentalizer predictions of Agent choices to donate on trial t (C_t) were modeled by combining the Mentalizer Belief (B_t) and Intent (I_t) estimates according to the following equation:

$$C_t = |1 - \text{round}(B_t) - I_t|$$

Which produces a choice prediction that is probabilistic and equal to I_t when the Agent believes the program is in Normal mode and $1-I_t$ when the Agent believes the program is in reversal mode.

Following the generation of estimates of B_t, I_t, and C_t, participant response probabilities were calculated according to the following softmax function (in which E_t is used to represent binarized versions of any of the estimates and $p(R_t)$ is the probability of the Mentalizer's given response on trial t).

$$P(R_t) = \frac{e^{(Et)}}{e^{(Et)} + e^{(1-Et)}}$$

Alternative ToM learning models: While we focused our analyses on our *a priori* model (M1; Figure 3), we also ran a number of other models (see Figures 4A and S2) to assess whether differences in performance could be explained by other strategies (for instance, were individuals simply using the actual program mode to interpret Agent choices, rather than tracking Agent beliefs (M3; Figure 4A). In Figure S2 we provide the model equations and a simple explanation for each. Ultimately, we investigated seven different models. We note that no model was able to better explain Mentalizer performance in the CTL groups than our initial *a priori* model (Figure 4B).

QUANTIFICATION AND STATISTICAL DETAILS

Statistical significance

In keeping with recent recommendations on reporting statistical analyses [15, 45], we eschew reporting of p values and instead report effect sizes (often as mean differences) and bootstrapped 95% confidence intervals throughout. Effects with non-overlapping confidence intervals are interpreted as statistically meaningful differences.

Behavioral Analysis

Charity Task

Three metrics were chosen to assess participant behavior in the charity task (Figure S1):

Belief Accuracy: Overall accuracy (compared to actual program mode) of participant responses when estimating program mode (normal or reversed).

Intended Outcome: Proportion of trials that the participant indicated that the outcome of the trial was what they intended; i.e., the participant wanted the money to go to the charity and the money went to the charity (this is not the same as #1, since belief about program mode is inferred from history over trial outcomes, whereas whether the outcome was as intended is susceptible to the trial-wise probabilistic nature of the program mode).

Consistency: Proportion of trials on which the answers the participant provided for belief and for intended outcome were consistent with the choice reversal outcome of the trial; e.g., a trial was consistent if the participant believed the mode to be normal, the computer *did not* reverse their choice, and the participant indicated they got the outcome they intended; a trial was inconsistent if the participant believed the mode to be normal, the computer *did* reverse their choice, but the participant indicated they received the outcome they intended.

Mentalizer Task

Participant accuracy (i.e., binarized responses compared to binarized Agent behavior) on the Mentalizer task was assessed for each of the three responses provided on every trial (Belief, Intent, and Choice; Table S3). In addition, we calculated within-trial consistency (i.e., whether a participant's choice response on a given trial was consistent with their belief and intent responses for that trial; Figure 2D and Table S3). Because Belief expectations reversed repeatedly over the course of the experiment, and Consistency is a measure of within-trial logic (i.e., no learning component), we calculated overall accuracy for these behaviors (chance was 50%). In contrast, if participants learned about the Agents, then their Intent and Choice performance should increase over time.

Therefore, to examine ToM learning we compared average accuracy for Intent and Choice on the last 30 trials to that on the first 30 trials (Table S3).

ToM Model Fitting and Analysis

Learning Rate Estimation Procedure: Behavioral data for each group (CTL1, CTL2, and ASD) were fit with a Bayesian hierarchical model (MATLAB code developed by S. Gershman and available online at https://github.com/sjgershm/mfit; download date 8/1/2018; code available at https://osf.io/ahp5q/) in which group priors were estimated for both Belief (λ_{Bel}) and Intent (λ_{Int}) learning rates and individual learning rates were subsequently estimated under these priors. 95% confidence intervals of learning rate means were calculated by bootstrapping (5000 bootstraps) the individual fit results.

Out-of-sample Model Predictive Performance

To assess model predictive performance (Figure 3C), the group parameters estimated from the fits of the *a priori* model (M1) from each of the control groups (CTL1 and CTL2; each matched to ASD on Age, Gender, Years of Education, and IQ; Table S1) were used to calculate model predictive accuracy (i.e., how accurate model predictions were at predicting participant behavior) for each participant in the other (out-of-sample) group (CTL 2 and CTL1, respectively) as well as for each participant in the ASD group. 95% confidence intervals were calculated by bootstrapping (5000 bootstraps) the participant predictive accuracies.

Correlation with other measures

In addition to assessing model performance, we were interested in exploring the correlation between participants' individual differences in performance (participant accuracy, model learning rates from individual fits) and participants' individual differences on intelligence (FSIQ) and social (AQ, EQ, SQ, RMET, and SNI) measures (Tables S5-6). To ensure no bias in correlational analyses involving model parameters, all data from the mentalizer task (CTL1, CTL2, and ASD) were combined and fit with a single hierarchical model (using the *a priori* model M1).

DATA AND SOFTWARE AVAILABILITY

Complete raw data, experiment code (PsychToolBox 3 [33, 34]), materials, and analysis scripts (MATLAB), are available from the authors and can be found online at https://osf.io/ahp5q/.