Impaired theory of mind for moral judgment in high-functioning autism

Joseph M. Moran[1,2], Liane L. Young[1], Rebecca Saxe, Su Mei Lee, Daniel O'Young, Penelope L. Mavros, and John D. Gabrieli

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, and Division of Health Sciences and Technology, Harvard University–Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Nancy G. Kanwisher, Massachusetts Institute of Technology, Cambridge, MA, and approved January 10, 2011 (received for review August 9, 2010)

High-functioning autism (ASD) is characterized by real-life difficulties in social interaction; however, these individuals often succeed on laboratory tests that require an understanding of another person's beliefs and intentions. This paradox suggests a theory of mind (ToM) deficit in adults with ASD that has yet to be demonstrated in an experimental task eliciting ToM judgments. We tested whether ASD adults would show atypical moral judgments when they need to consider both the intentions (based on ToM) and outcomes of a person's actions. In experiment 1, ASD and neurotypical (NT) participants performed a ToM task designed to test false belief understanding. In experiment 2, the same ASD participants and a new group of NT participants judged the moral permissibility of actions, in a 2 (intention: neutral/negative) × 2 (outcome: neutral/negative) design. Though there was no difference between groups on the false belief task, there was a selective difference in the moral judgment task for judgments of accidental harms, but not neutral acts, attempted harms, or intentional harms. Unlike the NT group, which judged accidental harms less morally wrong than attempted harms, the ASD group did not reliably judge accidental and attempted harms as morally different. In judging accidental harms, ASD participants appeared to show an underreliance on information about a person's innocent intention and, as a direct result, an overreliance on the action's negative outcome. These findings reveal impairments in integrating mental state information (e.g., beliefs, intentions) for moral judgment.

Asperger disorder | social cognition | mentalizing

Autism is a neurodevelopmental disorder characterized by persistent difficulties, among others, in the domain of social interaction. Children with autism have substantially delayed maturation of theory of mind (ToM), the ability to infer the contents of other people's minds, including beliefs and intentions (1).* Although adults with high-functioning autistic spectrum disorders (ASD) continue to experience clinical and practical difficulties with understanding other people's beliefs and intentions, these adults typically succeed on standard tests for ToM (2, 3). These tests include first-order false belief tasks, which require the participant to understand that another person has a belief about the world that is both different from the participant's own belief and factually incorrect. The apparent paradox between everyday difficulty in understanding what other people are thinking and success in laboratory tests of ToM suggests that through development, these individuals acquire further compensatory reasoning skills that enable them to succeed on explicit measures of ToM (2, 4).

The development of compensatory skills in ASD is evident on other studies of social reasoning. In these studies, ASD adults make accurate explicit judgments about other people's behaviors, but further probing suggests an enduring atypical ability to analyze other people's minds. For example, ASD adults are accurate in judging whether examples of behavior represent *faux pas* or not (5), as well as whether someone was telling the truth (Strange Stories) (6). In both cases, however, further investigation revealed that ASD patients often generated atypical reasons for their judgment that were inaccurate or inappropriate. The Strange Stories and *faux pas* tasks rely on broad social knowledge, so poor justifications may arise from a lack of general social knowledge rather than a specific deficit in ToM. In another task, Reading the Mind in the Eyes, ASD adults were able to pass first- and second-order false belief tasks ("He thinks that she thinks that X is true"), but made more errors than controls when determining a person's state of mind from the expression in their eyes (7). These studies therefore support the idea that ASD patients think differently about social behavior, even when they pass simple false belief tasks. What remains unclear is the nature of the underlying differences in thought (i.e., what specific kind of information is processed differently in ASD).

Experiments that test spontaneous expectations about human actions more clearly reveal different thought patterns in ASD (7, 8). ASD adults made accurate ToM judgments about the actions of another person on a simple false belief task, but unlike even typically developing infants, they failed to spontaneously anticipate another's person's actions based on false beliefs (as measured by spontaneous anticipatory eye movements) (8). These findings suggest an enduring yet subtle social deficit in adults with ASD; however, the link between these adult social deficits and the childhood delay in explicit reasoning about beliefs and intentions remains unclear.

Here we sought to develop a unique explicit test of ToM reasoning in adults, which could not be easily solved by compensatory heuristics. Like the traditional false belief task, we sought a task that directly measured reasoning about beliefs and intentions, with simple quantitative response scales, rather than verbal justifications; however, like the *faux pas* task, we sought to tap more-sophisticated aspects of ToM reasoning. To this end, we measured ToM reasoning for moral judgment.

Moral judgment is a complex social cognitive task that relies on ToM (9). Neurotypical (NT) adults weigh a person's intention more heavily than the outcome of their action when evaluating the moral permissibility of an action. For example, NT adults judge attempted but failed harms (e.g., attempted but unsuccessful murder) as more morally blameworthy than accidental harms (e.g., unintentionally killing someone) (9). Such judgments require that participants balance considerations of the agent's beliefs and intention, which depend upon ToM, against considerations of the

Author contributions: J.M.M., L.L.Y., R.S., and J.D.G. designed research; J.M.M., S.M.L., D.O., and P.L.M. performed research; J.M.M., S.M.L., and D.O. analyzed data; and J.M.M., L.L.Y., R.S., and J.D.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]J.M.M. and L.L.Y. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: jmmoran@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1011734108/-/DCSupplemental.

*Belief refers to the understanding that another person holds a particular premise to be true (e.g., the belief that it will rain tomorrow), whereas intention refers to another person's determination to act in a certain way (e.g., the intention to bring an umbrella), based on one or more beliefs.

actual outcomes, which do not depend upon ToM. The association between moral judgment and ToM is supported by neuroimaging evidence. For example, the level of blame participants assign for accidental harms correlates negatively with activation in their right temporoparietal junction (rTPJ) (10), a critical node in the ToM network (11). More rTPJ activation predicts greater consideration of the actor's intentions, and therefore less blame for accidents. Disrupting rTPJ activation using transcranial magnetic stimulation (TMS) also disrupts the use of mental state information for moral judgment (12). Reduced ToM ability in autism is associated with reduced activation of the rTPJ, a key region in these judgments (13). Moral judgments then may provide a sensitive test of enduring deficits of ToM in high-functioning ASD. Narrow compensatory strategies, which generate the correct answer on the *faux pas* and Strange Stories tasks (5, 6), are likely to fail for moral judgments that lack simple correct answers—e.g., when the person's innocent intention conflicts with the action's harmful (accidental) outcome.

Here, across two experiments, we tested ASD adults' performance on a standard false belief task (11) and on a well-characterized moral judgment task (9). We predicted (*i*) that ASD individuals would succeed on a standard test of false beliefs, but (*ii*) that ASD individuals should make atypical moral judgments, especially for accidental harms. Prior research on moral judgment in autism has focused on either the ability to distinguish between intentional moral and conventional harms (14) or moral judgment of intentional negative acts as good or bad (15). Both studies found that ASD individuals distinguished morally acceptable from morally unacceptable acts as reliably as did NT individuals. These studies make clear that ASD and NT individuals possess the same basic understanding of moral right and wrong. Neither study, however, varied the intentionality of immoral acts, or required participants to deploy ToM to make moral distinctions. We therefore hypothesized that ASD individuals, due to ToM deficits, would fail to exculpate accidental harms to the same degree as NT individuals. In other words, ASD individuals should neglect a person's innocent intentions and therefore assign more moral blame for accidental harm.

In experiment 1, participants answered questions about single-paragraph stories that probed either their understanding of a person's false belief, which requires ToM, or a false physical depiction of the world (e.g., a photograph or drawing), which does not require ToM.

In experiment 2, participants read vignettes in a 2 × 2 design: protagonists produced either a negative outcome (someone's death) or a neutral outcome (no harm) based on the belief that they were causing the negative outcome (negative belief) or the neutral outcome (neutral belief). The moral judgments in experiment 2 required both processing beliefs and intentions (whether a person had a reasonable belief or a negative intention), which requires ToM, and processing outcomes (whether there was or was not a negative outcome), which does not require ToM.

Results

Experiment 1. The ASD and NT groups did not differ significantly in either response latencies [main effect of Group: $F(1, 24) = 2.34$, $P > 0.13$; Group × Condition interaction: $F(1, 24) = 0.09$, $P > 0.75$] or accuracy [main effect of Group on percent correct: $F(1, 24) = 1.76$, $P > 0.19$; Group × Condition interaction: $F(1, 24) = 0.72$, $P > 0.41$; Fig. 1] when judging false belief and false photograph conditions. Across participants, responses were quicker for false belief vs. false photograph conditions [main effect of Condition on response latency: $F(1, 24) = 5.31$, $P < 0.03$; mean RT (msec) ± SEM: false belief = 3,126.1 ± 121; false photograph = 3361.7 ± 108], which were also associated with more correct responses [main effect of Condition on percent correct: $F(1, 24) = 8.20$, $P < 0.01$; mean percent correct responses ± SEM: false belief = 93.28 ± 1.69; false photograph = 85.55 ± 2.21]. Performance was below ceiling in both groups, possibly due to the rapid presentation of the stories (10 s).

Experiment 2. Actions with neutral intentions and neutral outcomes were judged more permissible than those with negative intentions and negative outcomes [main effects of Intention ($F[1, 26] = 230.19$, $P < 0.0001$) and outcome ($F[1, 26] = 89.69$, $P < 0.0001$)] (Fig. 2*B*). Accidental harms were judged as more permissible than intentional harms (Intention × Outcome interac-

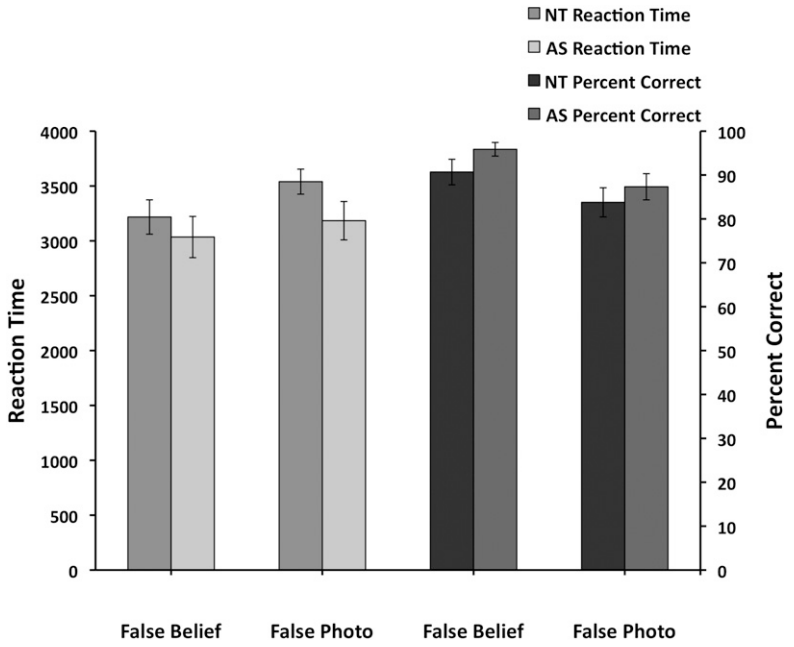

Fig. 1. Experiment 1: Similar ToM performance on a false belief task in ASD and NT adults. NT and ASD groups did not differ on response accuracy in either false belief or false photograph conditions.

tion [$F(1, 26) = 18.14$, $P < 0.0001$]). Critically, group differences were observed in a Group × Intention interaction [$F(1, 26) = 5.40$, $P < 0.03$]: NT participants judged actions in neutral Intention vignettes as more permissible than did ASD participants (NT 5.54 ± 0.25; ASD 4.59 ± 0.27). This Group × Intention interaction was driven by the difference between NT and ASD participants' judgments of accidental harms: the ASD group judged accidental harm as less morally permissible than the NT group [Bonferroni-corrected t test, two-tailed, t (corrected df 20.34) = 2.87, $P < 0.009$]. The ASD and NT groups did not differ reliably on any other kind of moral judgments, including judgments of neutral scenarios (neutral outcome, neutral intent), attempted harm (neutral outcome, harmful intent), or intentional harm (harmful outcome, harmful intent) vignettes (all other t's < 1.3, P's > 0.2). Further, whereas NT participants rated attempted harm (neutral outcome, harmful intent) as less morally permissible than accidental harm [harmful outcome, neutral intent; within-group paired t test, $t(14) = 6.24$, $P < 0.001$], ASD participants did not reliably differentiate between attempted and accidental harm [within-group paired t test, $t(12) = 1.76$, $P > 0.10$].

Discussion

Here we show compromised ToM for moral judgment in adults with ASD, who successfully answered questions about mental states in a standard false belief task. The ASD and NT participants demonstrated nearly identical ability to understand simple false beliefs in other people: they did not differ in accuracy or latency to make judgments about false beliefs. In experiment 2, however, NT participants exculpated protagonists for accidental harms caused on the basis of innocent intentions, whereas ASD individuals were less willing to make such exculpatory moral judgments. In judging accidental harms, ASD participants, relative to NT participants, appeared to show an underreliance on the information about innocent intentions and, as a result, an overreliance on negative outcomes. Making moral judgments about an action based on the analysis of a person's intentions requires ToM. Thus, these findings reveal a ToM deficit in ASD adults that influenced explicit moral judgments.

This selective difference in moral judgments involving ToM occurred despite many other similarities between the ASD and NT groups. ASD and NT participants showed similar behavioral patterns in their strong condemnation of intentional harm, intermediate condemnation for attempted harm, and lack of condemnation for scenarios in which both intentions and outcomes were neutral. Further, the two groups did not differ in IQ. Indeed, the mean IQ of the ASD group was well above average (mean of 120), which demonstrates a strong dissociation between overall intelligence and moral judgments that require an analysis of another person's intentions.

In several respects the pattern of results displayed by the ASD adults mirrors that displayed by typically developing children (16, 17). Three-year-old NT children systematically make the wrong prediction on standard false belief tasks, consistent with an immature ToM (18), but by age 4 or 5, NT children are at ceiling in predicting and explaining actions in terms of false beliefs (19). Even at age 4, however, NT children are likely to assign more moral weight to outcomes vs. intentions when evaluating actions (16). For instance, 4-y-old children will judge a person who helpfully attempts to direct a lost traveler to his destination but accidentally misdirects him as more naughty than a person who attempts to misdirect a lost traveler but fails (20). The proclivity to use belief information to exculpate people for accidental harms increases in NT children from ages 5 to 11 (21). Other sophisticated moral judgments of intentions (e.g., judgments about actions that knowingly or unknowingly interfere with

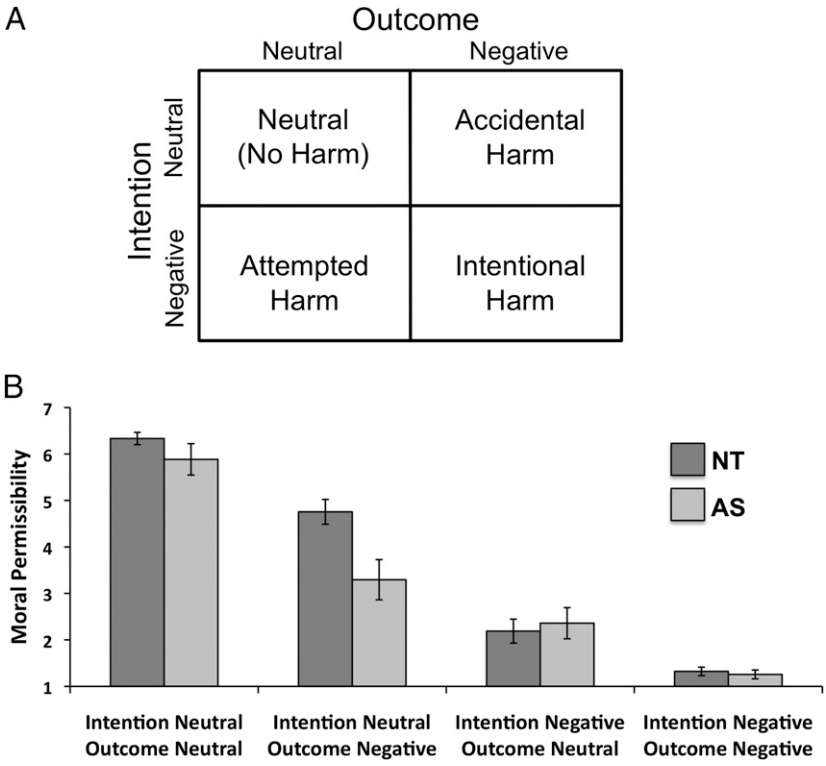

Fig. 2. Experiment 2: Different moral judgments about accidental harms between ASD and NT adults. (*A*) Experiment 2 followed a 2 (Group: ASD/NT) × 2 (Intention: neutral/negative) × 2 (Outcome: neutral/negative) design. (*B*) ASD participants rated accidental harms as less morally permissible than NT participants. All other ratings did not differ between groups.

someone else's plans) also show developmental change through late childhood (17).

A possible explanation for why children as young as age 4 are able to pass standard false belief tasks, but fail to make mature moral judgments, is that children may be able to encode and represent beliefs before they are able to use belief information fully and flexibly, in concert with outcome information, for moral judgment (22). Thus, children who are able to understand that people have mental states independent of physical reality (ToM) still persist in condemning accidental harms. Exculpation for accidents requires an especially robust mental state representation to override a (possibly prepotent) response to the salient information about actual harm (9, 23).

The typical developmental pattern may provide a model for the differences observed in our participants with ASD. ASD individuals, like typical 5-y-olds, gave the correct answers to simple false belief questions, but persisted in overweighting outcomes when making moral judgments. One possibility is that early maturing aspects of ToM, such as understanding false beliefs, may be delayed in ASD (2, 4, 24), whereas later-maturing aspects of ToM, such as exculpation of accidental harm, may never fully develop in even high-functioning ASD. On this view, ASD causes a delay, rather than a disruption, in ToM development. By contrast, an alternative hypothesis is that ASD individuals develop atypical compensatory mechanisms for solving simple false belief tasks, which do not easily encompass the more subtle demands of moral judgment. This view is supported by evidence that when children with ASD succeed on simple ToM tasks, they do so not just late, but in an atypical order, suggesting that their success reflects the operation of a different thought process (25).

Our findings serve as a clear demonstration that making moral judgments that rely on ToM causes measureable difficulties even in high-functioning ASD adults. There are no truly correct or incorrect judgments in this task. Indeed, NT participants differ in the amount of blame they assign to accidental harm (10). Aphorisms capture both the importance of innocent intentions ("It's the thought that counts") and the notion that intentions are often not enough ("The road to hell is paved with good intentions"). Scenarios in the present task also included other factors that subtly but systematically affect the assessment of beliefs and intentions (e.g., Is it reasonable to believe white powder to be sugar in an unfamiliar chemical factory?). Nevertheless, on average, the ASD group weighed beliefs and intentions less than the typical control group—a difference that could lead to a difficulty for ASD individuals in their everyday interactions with other people. Moral judgments that pit mental states against outcomes may therefore constitute a sort of stress test of ToM, and reveal an enduring deficit in ToM-dependent judgments even among very high-functioning individuals with ASD.

The current findings relate to evidence about the neural basis of ToM for moral judgment. Given prior evidence that rTPJ activation is uniquely correlated with individual differences in moral judgment of accidental harms (10), dysfunction in this region may be the mechanism by which ASD individuals fail to exculpate accidental harms. The rTPJ might therefore be recruited when participants must use intention information to overcome a salient negative outcome. As noted above, ASD individuals also show a correlation between rTPJ activation and independently measured ToM ability (13); this observation further strengthens the hypothesis that intact rTPJ processing may be necessary for moral judgments that depend on ToM (i.e., exculpation of accidental harm). Future neuroimaging research using this task in an ASD population should be able to determine whether impaired exculpation of accidental harms is associated with reduced or nonspecific rTPJ functioning.

These findings also provide unique information about the brain organization of component processes of moral judgment. Moral judgment impairments have been documented in patients with focal damage to bilateral ventromedial prefrontal cortex (VMPC) (26, 27). Such patients have socioemotional deficits that may be due to inability to generate typical emotional responses to abstract mental state information (e.g., harmful intentions) (28). Patients with VMPC damage were tested on the same scenarios as in the current study and showed a selective deficit on failed attempts to harm, endorsing failed attempts as morally permissible. Convergent fMRI evidence indicates a correlation between blame for failed attempts and VMPC activity (10). In contrast to the current ASD participants, who had difficulty exculpating accidents based on neutral intent, VMPC participants appeared able to encode intent information but not to respond emotionally to this information; they therefore did not condemn harmful intentions in the absence of actual harm. These two findings constitute a double dissociation, and as such increase confidence in the notion that the differences in accidental harm showed by the present ASD participants are selective and not due to extraneous variables such as task difficulty. It is difficult to identify the specific brain basis of the altered moral judgments in the atypical neurodevelopment of ASD, but the combination of findings indicates that distinct components of moral judgment are associated with distinct neural systems.

This study focused on moral judgments, but it is likely that the ASD weakness in using mental state information in the face of conflicting information would apply broadly to judgments about other people. Future research ought to examine this possibility directly, as well as examine individual differences in larger ASD groups. The present findings are consistent with the observation that ASD individuals are impaired at implicit but not explicit ToM (8). Critically, these findings also extend our understanding of that impairment into actual judgment, where impairments in high-functioning ASD have been previously difficult to detect in the laboratory.

Materials and Methods

Participants. This study was conducted according to the principles expressed in the Declaration of Helsinki. The study was approved by the Committee on the Use of Humans as Experimental Subjects of the Massachusetts Institute of Technology (MIT protocol no. 0608001876). All participants provided written informed consent for the collection of samples and subsequent analysis.

Participants were recruited from the local MIT community (NT participants) or via advertisements placed with the Asperger's Association of New England (ASD participants).

Experiment 1. The NT and ASD groups did not differ significantly on age [NT (mean ± SEM) = 28.00 ± 1.47; ASD = 33.15 ± 2.76; $t(24) = 1.47$, $P > 0.11$], sex (NT: six women, seven men; ASD: four women, nine men), or IQ [NT: 118.8 ± 2.28; ASD: 120.46 ± 3.15; $t(24) = 0.41$, $P > 0.68$].

Experiment 2. The NT and ASD groups did not differ significantly on age [NT (mean ± SEM) = 31.67 ± 1.52; ASD = 33.15 ± 2.76; $t(26) = 0.47$, $P > 0.64$], sex (NT: six women, nine men; ASD: four women, nine men), or IQ [NT: 114.8 ± 3.48; ASD: 120.46 ± 3.15; $t(26) = 1.21$, $P > 0.24$].

All participants were prescreened using the social communication questionnaire (SCQ) (29) for a possible ASD. NTs (experiment 1: 4.70 ± 1.15; experiment 2: 5.44 ± 1.44) scored significantly lower than ASDs [14.54 ± 1.88; experiment 1: $t(24) = 3.93$, $P < 0.001$; experiment: $t(26) = 3.52$, $P < 0.001$]. ASD participants underwent both the Autism Diagnostic Observation Schedule (ADOS) (30, 31) and impression by a clinician trained in both ADOS administration and diagnosis of ASDs (Karen Shedlack, Massachusetts General Hospital, Boston). All ASD participants received a diagnosis of either Asperger syndrome or autism based on total ADOS score (communication and social) and on clinical impression based upon the diagnostic criteria of the DSM-IV (32). Participants were paid for their participation and gave their informed consent in accordance with procedures outlined by the MIT Committee on the Use of Humans as Experimental Subjects.

Procedures. ***Experiment 1.*** We investigated ToM in a 2 (Group: ASD/NT) × 2 (Condition: false belief/false photograph) design. Participants viewed single-paragraph stories (see *SI Materials and Methods* for examples), presented for 10 s, and answered a two-alternative forced-choice question (presented for 6 s) poststory presentation, which probed their understanding of the

story. In the false belief condition, participants had to answer a question regarding an incorrect belief held by a person about the physical state of the world. In the false photograph condition, which served as a control, participants answered a question regarding an incorrect physical representation (usually a photograph) of the world. Twelve stories were presented in each condition. In both conditions, participants must engage in counterfactual thinking; in only the false belief condition did participants have to consider a false mental representation of the world. Percent correct and response latencies across groups and conditions served as the dependent measures.

Experiment 2. We investigated moral judgment in a 2 (Group: ASD/NT) × 2 (Intention: neutral/negative) × 2 (Outcome: neutral/negative) design. Participants viewed four kinds of moral scenarios (six per category; see *SI Materials and Methods* for examples): (*i*) Neutral (intention neutral, outcome neutral); (*ii*) Accidental Harm (intention neutral, outcome negative); (*iii*) Attempted Harm (intention negative, outcome neutral); and (*iv*) Intended Harm (intention negative, outcome negative (Fig. 2*A*). Each scenario was presented in a sequence of background, foreshadow, intention, action, and outcome information. The entire scenario text remained on screen while participants made their judgments, to minimize working memory load. Following presentation of each scenario, participants rated the moral permissibility of the action on a seven-point Likert scale (1 = completely morally forbidden, 7 = completely morally permissible).

ACKNOWLEDGMENTS. We thank M. Singh and A. Qureshi for assistance with data collection and analysis; the Asperger's Association of New England and all of our participants; and Karen Shedlack for clinical evaluations.

1. Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a "theory of mind"? *Cognition* 21:37–46.
2. Frith U, Morton J, Leslie AM (1991) The cognitive basis of a biological disorder: Autism. *Trends Neurosci* 14:433–438.
3. Bowler DM (1992) "Theory of mind" in Asperger's syndrome. *J Child Psychol Psychiatry* 33:877–893.
4. Frith U (2004) Emanuel Miller lecture: Confusions and controversies about Asperger syndrome. *J Child Psychol Psychiatry* 45:672–686.
5. Zalla T, Sav AM, Stopin A, Ahade S, Leboyer M (2009) Faux pas detection and intentional action in Asperger Syndrome. A replication on a French sample. *J Autism Dev Disord* 39:373–382.
6. Happé FGE (1994) An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J Autism Dev Disord* 24:129–154.
7. Baron-Cohen S, Jolliffe T, Mortimore C, Robertson M (1997) Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *J Child Psychol Psychiatry* 38:813–822.
8. Senju A, Southgate V, White S, Frith U (2009) Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science* 325:883–885.
9. Young L, Cushman F, Hauser M, Saxe R (2007) The neural basis of the interaction between theory of mind and moral judgment. *Proc Natl Acad Sci USA* 104:8235–8240.
10. Young L, Saxe R (2009) Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47:2065–2072.
11. Saxe R, Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage* 19:1835–1842.
12. Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc Natl Acad Sci USA* 107:6753–6758.
13. Kana RK, Keller TA, Cherkassky VL, Minshew NJ, Just MA (2009) Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Soc Neurosci* 4:135–152.
14. Blair RJ (1996) Brief report: Morality in the autistic child. *J Autism Dev Disord* 26: 571–579.
15. Leslie AM, Mallon R, DiCorcia JA (2006) Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Soc Neurosci* 1:270–283.
16. Piaget J (1965) *The Moral Judgment of the Child* (Free Press, New York).
17. Mant CM, Perner J (1988) The child's understanding of commitment. *Dev Psychol* 24: 343–351.
18. Wimmer H, Perner J (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–128.
19. Leslie AM, Thaiss L (1992) Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition* 43:225–251.
20. Baird JA, Astington JW (2004) The role of mental state understanding in the development of moral cognition and moral action. *New Dir Child Adolesc Dev* 103: 37–49.
21. Shultz TR, Wright K, Schleifer M (1986) Assignment of moral responsibility and punishment. *Child Dev* 57:177–184.
22. Young L, Saxe R (2008) The neural basis of belief encoding and integration in moral judgment. *Neuroimage* 40:1912–1920.
23. Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44:389–400.
24. Baron-Cohen S (1989) The autistic child's theory of mind: A case of specific developmental delay. *J Child Psychol Psychiatry* 30:285–297.
25. Peterson CC, Wellman HM, Liu D (2005) Steps in theory-of-mind development for children with deafness or autism. *Child Dev* 76:502–517.
26. Ciaramelli E, Muccioli M, Làdavas E, di Pellegrino G (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Soc Cogn Affect Neurosci* 2:84–92.
27. Koenigs M, et al. (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446:908–911.
28. Young L, et al. (2010) Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron* 65:845–851.
29. Rutter M, Bailey A, Lord C (2003) *SCQ: Social Communication Questionnaire* (Western Psychological Services, Los Angeles).
30. Lord C, Rutter M, DiLavore PC, Risi S (2002) *Autism Diagnostic Observation Schedule* (Western Psychological Services, Los Angeles).
31. Lord C, et al. (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.
32. American Psychiatric Association (APA) (2000) *Diagnostic and Statistical Manual of Mental Disorders* (Amer Psychiatric Assoc, Washington, DC), revised 4th Ed.

Supporting Information

Moran et al. 10.1073/pnas.1011734108

SI Materials and Methods

Experiment 1: Example Theory of Mind Stories. ***False belief.*** Jenny put her chocolate away in the cupboard. Then she went outside. Alan moved the chocolate from the cupboard into the fridge. Half an hour later, Jenny came back inside. Jenny expects to find her chocolate in the . . .

cupboard fridge

False photograph. The biography describes the room as it was in 1965. Originally the walls were covered in dark wallpaper. By 1965 the paper had been stripped and replaced with cream paint. The biography says that the room was . . .

light dark

Experiment 2: Example Moral Judgment Stories. ***Accidental harm.*** Grace and her friend are taking a tour of a chemical plant. When Grace goes over to the coffee machine to pour some coffee, Grace's friend asks for some sugar in hers. There is white powder in a container by the coffee.

The white powder is a very toxic substance left behind by a scientist, and therefore deadly when ingested in any form.

The container is labeled "sugar," so Grace believes that the white powder by the coffee is sugar left out by the kitchen staff.

Grace puts the substance in her friend's coffee. Her friend drinks the coffee and dies.

Putting the substance in was:

1 2 3 4 5 6 7

Forbidden Permissible

Attempted harm. Dan is giving a visitor a tour of a laboratory. Before visitors enter the testing room, all test tubes containing disease antigens must be contained in a chamber by flipping a switch. A repairman has just come to fix the switch, which had been broken.

The switch has been successfully repaired, so the test tubes are quite safely contained. Thus, anybody who enters the room will be safe and unexposed.

Dan believes that the switch is still broken after a conversation with the repairman, so he believes it is not safe for the visitor to enter.

Dan tells the visitor to enter the testing room. The visitor does not contract any disease and is fine.

Telling the visitor to enter was:

1 2 3 4 5 6 7

Forbidden Permissible