The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder

Cornelius Maurer, Valérian Chambon, Sacha Bourgeois-Gironde, Marion Leboyer, Tiziana Zalla

The Influence of Prior Reputation and Reciprocity on Dynamic Trust-building in Adults with and without Autism Spectrum Disorder

Cornelius Maurer[1], Valerian Chambon[1], Sacha Bourgeois-Gironde[1,2], Marion Leboyer [3, 4], Tiziana Zalla[1,4]

1. Institut Jean Nicod, Centre National de la Recherche Scientifique, Département d'Etudes Cognitives, Ecole Normale Supérieure & PSL Research University, Paris, France.
2. Laboratoire d'Economie Mathématique et de Microéconomie Appliquée (LEMMA), Université Paris 2 Panthéon Assas, Paris, France.
3. Institut National de la Santé et de la Recherche Médicale (INSERM), U955, Institut Mondor de Recherche Biomédicale, Psychiatrie Translationnelle (Créteil, France).
4. Fondation FondaMental, Créteil, France.

Correspondence to:

Tiziana Zalla
Institut Jean Nicod- CNRS
Ecole Normale Supérieure
29 rue d'Ulm 75005 Paris, France
Telephone: [+33] (0)1 44 32 26 70
Fax: [+33] (0)1 44 32 26 99
E-Mail: tiziana.zalla@ens.fr

Abstract

The present study was designed to investigate the effects of reputational priors and direct reciprocity on the dynamics of trust building in adults with (N= 17) and without (N= 25) autism spectrum disorder (ASD) using a multi-round Trust Game (MTG). On each round, participants, who played as investors, were required to maximize their benefits by updating their prior expectations (the partner's positive or negative reputation), based on the partner's directed reciprocity, and adjusting their own investment decisions accordingly. Results showed that reputational priors strongly oriented the initial decision to trust, operationalized as the amount of investment the investor shares with the counterpart. However, while typically developed participants were mainly affected by the direct reciprocity, and rapidly adopted the optimal Tit-for-Tat strategy, participants with ASD continued to rely on reputational priors throughout the game, even when experience of the counterpart's actual behavior contradicted their prior-based expectations. In participants with ASD, the effect of the reputational prior never disappeared, and affected judgments of trustworthiness and reciprocity of the partner even after completion of the game. Moreover, the weight of prior reputation positively correlated with the severity of the ASD participant's social impairments while the reciprocity score negatively correlated with the severity of repetitive and stereotyped behaviors, as measured by the Autism Diagnostic Interview–Revised (ADI-R). In line with Bayesian theoretical accounts, the present findings indicate that individuals with ASD have difficulties encoding incoming social information and using it to revise and flexibly update prior social expectations, and that this deficit might severely hinder social learning and everyday life interactions.

Key Words: Trust Game, reciprocation, social interaction, moral judgment.

1. Introduction

Trust is critical for initiating and maintaining cooperative behavior, especially in social interactions characterized by risk and uncertainty (Deutsch, 1958, 1960; Riegelsberger, Sasse & McCarthy, 2005). Understanding the cognitive mechanisms underlying trust behavior and its disturbances is a relevant topic for research in social sciences, neuro-economics, cognitive psychology and psychiatry. Autism spectrum disorder (ASD), a condition characterized by impairments in social interaction and communication, has often been associated with difficulties reading social information from faces and actions, including emotions, intentions, and trustworthiness traits.

1.1. Trust, reciprocity and reputation

Current research on trust has focused on the understanding of factors that signal trustworthiness and influence our decisions to cooperate, such as physical appearance, reciprocity, reputation or group membership (Oosterhof & Todorov, 2008; Delgado et al., 2005; Boero et al., 2009), while fewer studies have investigated how people use and integrate different cues of trustworthiness for decision making in a social interactive situation.

Although cooperation can produce high mutual payoffs, it involves putting at risk one's own resources, such as time, money, or health. Thus, interpersonal trust helps us to deal with the risk of defection, since we usually lack full information about the abilities and intentions of other agents. In probabilistic terms, considering another person as trustworthy means believing that the chance of him/her intending to act in a way that is beneficial to us is high enough to consider cooperation (Gambetta, 2000).

Direct reciprocity has been shown to be a key mechanism for creating trust and fostering human cooperation (King-Casas et al. 2005; Hoffman, Yoeli & Nowak, 2015). In many instances of everyday life, e.g. when dealing with family members or colleagues, repeated exchanges with a partner serve as a reliable predictor of her/his trustworthiness. However, when trying out a new

restaurant, going on a blind date, or dealing with an unknown merchant via e-commerce, we lack a shared history of past interactions (e.g., customer evaluations, reports on previous transactions). Reciprocity also works through reputation: it is an evolved social mechanism designed to foster cooperation in larger human groups and to regulate interactions in complex systems. Reputation is intrinsically associated with moral norms and values, and is a valid surrogate for the interaction-based personal experience. Research using trust games has found preference for partners with good reputations, that is, people are more prone to cooperate with those partners they have observed treating others generously than with those whom they have observed behaving selfishly (Wedekind & Milinski, 2000). Neuroimaging studies have revealed that reputation has a long lasting effect on the evaluation of a person's trustworthiness, as it diminishes the reliance on neurological feedback mechanisms of reward learning (Delgado et al. 2005; Fouragnan et al., 2013). Therefore, it is not surprising that many e-commerce ventures have established global market reputation systems to help their users interact when confronted with the uncertainty of anonymous counterparts over long distances (Nowak & Sigmund, 2005). Nonetheless, there is always the risk that the reputation does not reflect the partner's actual intentions, either because it is misleadingly used for spreading unfounded rumors and manipulating co-players, or because people do not always live up to their reputations as they interact with others. Hence, overall, "Tit-for-Tat" (TfT), based on reciprocation, remains the optimal behavioral strategy in repeated exchanges, as we should trust the partner only as long as she/he reciprocates, and stop reciprocating once our trust is betrayed (Axelrod, 1984). As these studies indicate, decision-making in ecological and complex social situations requires a set of cognitive functions that goes beyond the Theory of Mind (ToM), i.e. the ability to attribute beliefs and other mental states to others. Crucially, cooperation and the decision to trust others rely on the ability to integrate different types of social information and use them in a flexible and adaptive manner during ongoing exchanges.

1.2. Trust and moral evaluation in autism spectrum disorder

Impairments in social interaction and communication are core features of autism spectrum disorder (ASD) (American Psychiatry Association, 2013). Reduced trust and social reciprocity are also commonly reported in individuals with ASD (Volkmar & Klin, 2000), and associated with low levels of blood plasma oxytocin (Modahl et al., 1998; Andari et al., 2010). In a previous study, Adolphs, Sears, & Piven (2001) reported decreased responsiveness to facial cues of trustworthiness in adults with ASD during a cooperation task, while they showed preserved trustworthiness judgments on the basis of biographical stories depicting the person's lifestyle and activities. More recently, using a trust game, Ewing et al. (2015) reported that, if explicitly prompted, children with autism, aged 6 -12 years, were able to behave rationally, that is, in line with partner trustworthiness, when making investment decisions. However, when asked to evaluate trustworthiness from facial appearances, they failed to spontaneously use this information to modulate their decision in ecological contexts. Overall, these findings weaken the hypothesis of a general impairment in trust processing in individuals with ASD and support the notion that they might be unable to use trustworthiness cues from different sources of information in a consistent manner.

Diminished social cognition and behavior in people with ASD are generally described as a deficit in ToM (Baron-Cohen, Leslie & Frith, 1985; Baron-Cohen, 1995). Adults with high-functioning ASD, who exhibit relatively preserved explicit ToM, show difficulties in real-life situations that might reflect an inability to use information about others' intentions. As indicated by previous studies (Zalla & Leboyer, 2011), judgments of intentionality (i.e., whether an agent has acted intentionally or unintentionally) in adults with ASD may be preponderantly informed by moral evaluations of the situation rather than by intentional cues. Specifically, the intentionality judgment in adults with ASD is characterized by an overreliance on moral evaluation of the agent's blameworthy action merely based on the action outcomes (Zalla et al., 2009; Zalla & Leboyer, 2011; Buon et al., 2013). Zalla et al. (2011) have suggested that social normative reasoning is preserved in individuals with ASD, and that their propensity to judge normative transgressions

more seriously and inflexibly reflects a diminished sensitivity to the intentional properties of action, especially when rule violations bring about negative outcomes. As a result, based on these previous findings (Zalla et al., 2009; Zalla and Leboyer, 2011), one might expect that, while in typically developed individuals moral judgments and prior expectations are continuously updated by new incoming information about the agent's action, the enhanced sensitivity to negative and blameworthy outcomes biases decision-making in social interaction in participants with ASD.

1.3. A unified framework for characterizing dynamic trust-building

Recently, Bayesian models have offered a promising framework for the understanding of cognitive functioning in ASD, including abnormal social cognition, enhanced sensations, and sensory precision (Pellicano & Burr, 2012; Lawson, Rees & Friston, 2014; Van de Cruys et al. 2014; Chambon et al., 2017a). The 'Hypo-Priors' hypothesis (Pellicano & Burr, 2012) suggests that sensory atypicalities and difficulties with social interaction in ASD can be explained by a diminished influence of top-down prior expectations, along with enhanced "bottom-up" functioning and increased reliance on sensory evidence. The Predictive Coding theory states that the prominent features of autism stem from the exuberant production of prediction errors (Lawson, Rees & Friston, 2014). According to this theory, cognition is modeled as a hierarchical organization in which expectations (*priors*), formulated at higher hierarchical levels, convey prediction to the lower levels of sensory signals where precision needs to be adequately attenuated. The discrepancy between these sources of information is known as 'prediction error'. Reduced adaptation to numerosity stimuli (Turi et al. 2016), biological motion (van Boxtel, Dapretto & Lu, 2016), objects (Skewes, Jegindø, & Gebauer, 2015), visual illusions (Palmer et al., 2015) and faces (Ewing, et al. 2015) have been presented as evidence for attenuated influence of priors in ASD.

In a recent study, Chambon and collaborators (2017a) have shown that diminished influence of prior knowledge about social intentions in adults with ASD might hinder the ability to predict individual intentions in the context of an iteratively interacting game, when direct sensory information is not available. While typically developed (TD) adults exhibited a strong initial

preference for TfT cooperative intentions, over alternative (non-TfT) defecting intentions (Chambon et al., 2011, 2017a, 2017b), adults with ASD showed no initial preference for the TfT mode of reciprocation. Importantly, they progressively acquired a social bias through the extraction of observed regularities by means of a general probabilistic learning mechanism. Interestingly, attenuated social priors predicted the severity of clinical symptoms in the area of social interaction, while the magnitude of social learning inversely correlated with the severity of repetitive and stereotyped behaviors. These results have provided the first empirical evidence that a disturbance in the Bayesian inferential mechanism which integrates prior social knowledge and sensory information might disrupt action prediction and learning in a social context in ASD. Within a Bayesian framework, two sources of information – direct reciprocity (i.e., ongoing exchanges) and indirect reciprocity (i.e., reputational priors) – may combine to produce accurate predictions about others' behavior and regulate collaboration and dynamic trust-building.

1.4. The current study

The aim of the current study was to investigate how these two types of information affect decisions to trust and social learning mechanisms in adults with and without ASD. For this purpose, we implemented a new version of the multi-round Trust Game (MTG) where each player interacted in turn with four partners (counterparts). The participants played the role of investors and each counterpart was introduced to a player with a short biographical story describing the person's lifestyle and professional achievements. Two 'good' counterparts were depicted as having committed praiseworthy actions (positive reputational prior) while the two 'bad' counterparts were depicted as having committed blameworthy actions (negative reputational prior). Crucially, during the MTG, one of the 'good' partners behaved in an individualist manner and one of the 'bad' partners behaved in a collaborative manner, so as to create two congruent and two incongruent profiles. The current MTG models a repeated sequential economic interaction between two agents, where trust is operationalized as the amount of investment the investor shares with the trustee.

So far, few studies have addressed the question of how individuals integrate these two sources of direct and indirect information about reciprocity, i.e., the reputational priors and interpersonal exchanges, and why individuals sometimes diverge from the optimal TfT strategy in repeated trust games. While previous studies have either manipulated trustees' reciprocity by varying the return ratio (Berg, Dickhaut, & McCabe, 1995; Camerer, 2003; King-Casas et al., 2005; Bourgeois-Gironde & Corcos, 2011) or the partner's reputation (Delgado et al., 2005; Behrens et al., 2007; Fouragnan et al., 2013), no study has systematically analyzed how prior reputational information and reciprocal behavior interact and conjointly impact trust over time.

In the present study, our systematic manipulation of congruent and incongruent setups of reputational priors and reciprocity allowed us to investigate, for the first time, the effect of these two factors on the dynamics of trust-building and decision-making in adults with ASD, as compared to TD adults. In the context of the current findings, we hypothesized that all participants would exhibit a behavioral susceptibility to each counterpart's reputational prior in the initial stages of interactions. We predicted that participants with ASD would encounter more difficulties in learning from on-line interactions and adjusting trust and reciprocity decisions on the basis of the two sources of information about a counterpart's reliability. Specifically, based on previous evidence (Zalla et al., 2009; Zalla et al., 2011; Zalla & Leboyer, 2011), we expected that individuals with ASD would be strongly biased by reputational priors, and that difficulties in adjusting their behavior efficiently would arise when confronted with incongruous counterpart profiles. In addition, we expected that the unbalanced interplay between priors (expectations based on initial reputation) and direct reciprocity would correlate with clinical symptoms of autism.

2. Materials and Methods

2.1. Participants.

Participants included 17 adults with ASD, recruited at the Albert Chenevier Hospital in Créteil (France) and 25 TD adults matched on age, gender, education, and Intelligence Quotient (IQ),

as measured by the Wechsler Adult Intelligence Scale (Wechsler, 1999). Participants with ASD scored significantly higher on the Autism-Spectrum Quotient (AQ; Baron-Cohen et al., 2001) relative to the TD group (Table 1).

--Table 1 about here --

Table 1. Means (and standard deviations) of demographic and clinical data for the participants with ASD (ASD) and typically developed (TD).

	ASD	**TD**	***t and p* values**
N (male:female ratio)	16 :1	21 :4	
Age in years	34.2 (8.9)	30.3 (9.4)	t= -1.35, p < .18
Education in years	14.6 ± 3.2	15.4 ± 2.3	t= 1.42, p < .16
Full-scale IQ	107.2 (17.1)	112.9 (9.9)	t= 1.37, p < .18
Verbal IQ	109.2 (16.1)	114.8 (9.3)	t= 1.43, p < .16
Performance IQ	102.8 (19)	107.5 (12.8)	t= 0.95, p < .34
Autism Spectrum Quotient	33.1 (6.8)	14.8 (4.7)	t= -10.36, p <.0001
ADI [A,B,C]*	14.5 (6.3); 8.2 (4.9); 5.1 (2.7)	---	
ADOS [*cut-off* >7]	12.6 (3.9)	---	

* [A] = reciprocal social interaction, [B] = communication, [C] = stereotyped behaviors

Participants with ASD received a clinical diagnosis of autism spectrum disorder based on the criteria of the DSM-IV-TR (American Psychiatric Association, 2000), the Asperger Syndrome Diagnostic Interview (ASDI; Gillberg et al., 2001) and the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000). Semi-structured interviews with parents or caregivers using the Autism Diagnostic Interview (ADI-R, Lord et al., 1994) yielded scores in three content areas: [A] social interaction, [B] communication, and [C] repetitive and stereotyped behaviors, allowing a separate quantification of the severity of the symptomatology. The cut-off points for these domains are 10, 8, and 3, respectively. Additional exclusion criteria for participants with ASD included the absence of co-morbid diagnoses. Exclusion criteria for all participants included IQ level (<70) and a history of major psychiatric or neurological disorder, or any medical condition or treatment affecting brain functions. All participants were native French speakers and had normal or corrected to normal vision. Participants gave informed consent to participation in the study under a protocol approved by the Institutional Ethical committee (INSERM, Institut Thématique Santé Publique; C07-33) and performed in accordance with the ethical standards advised in the Declaration of Helsinki.

2.2. The Multi-round Trust Game task

All participants played a *Multi-round Trust Game task* (MTG), an adapted iterated version of the two players Trust Game (Berg, Dickhaut, & McCabe, 1995) in the role of investors sequentially with four virtual partners, the counterparts (Figure 1a). Every investor played ten consecutive rounds of the Trust Game with the same counterpart before changing partners. To ensure no order effect on reciprocation, presentation orders of the four virtual counterparts were counterbalanced between subjects.

Each round of the Trust Game consisted of an investment, a return, and a feedback phase. In the investment phase the investor received 10 Experimental Monetary Units (EMUs), independently of previous actions. Next, the investor could invest any discreet amount of EMUs *i* in the counterpart.

The investment was quadrupled to $i \times 4$ and sent to the counterpart. After confirming their investment by clicking on the corresponding number on the scale, the screen went black for a delay of 2 s representing the return phase, in which the counterpart would choose the amount of EMUs to be returned to the participant. During the return phase the counterpart could return a ratio of the pie amounting to $r \times i \times 4$ back to the investor (Figure 1b). During the investment phase, the left side of the screen displayed the face of the counterpart and the right side an 11-point scale ranging from zero to ten EMUs, the name of the counterpart, and game instructions. Finally, in the feedback phase, payoff structures were revealed. Participants were informed of: (a) the amount of EMUs the counterpart had reciprocated, (b) their own round score, (c) the counterpart's round score, and (d) their subtotal score, comprising all round scores with this specific counterpart. Participants' round scores were calculated by adding the amount of EMUs they did not invest and the amount of EMU's returned by the counterpart. After confirming their investment by clicking on the corresponding number on the scale, the screen went black for a delay of 2 s representing the return phase, in which the counterpart would choose the amount of EMUs he returned to the participant. The counterpart's round score derived from the difference between the quadrupled investment he/she received and the amount of EMUs returned to the participants: $(1 - r) \times (i \times 4)$ (Figure 1c).

In the feedback phase (6s), participants were informed about the payoff structure, i.e., the amount of EMUs the counterpart had reciprocated, their own round score, counterpart's round score, and their subtotal score, comprising all round scores with this specific counterpart. Participants were informed of their overall score across conditions, and the equivalent amount they would receive in Euros at the end of the experimental session.

The Trust Game task was carried out using PsychoPy software Version 1.64 (Peirce, 2007).

--Figure 1 about here --

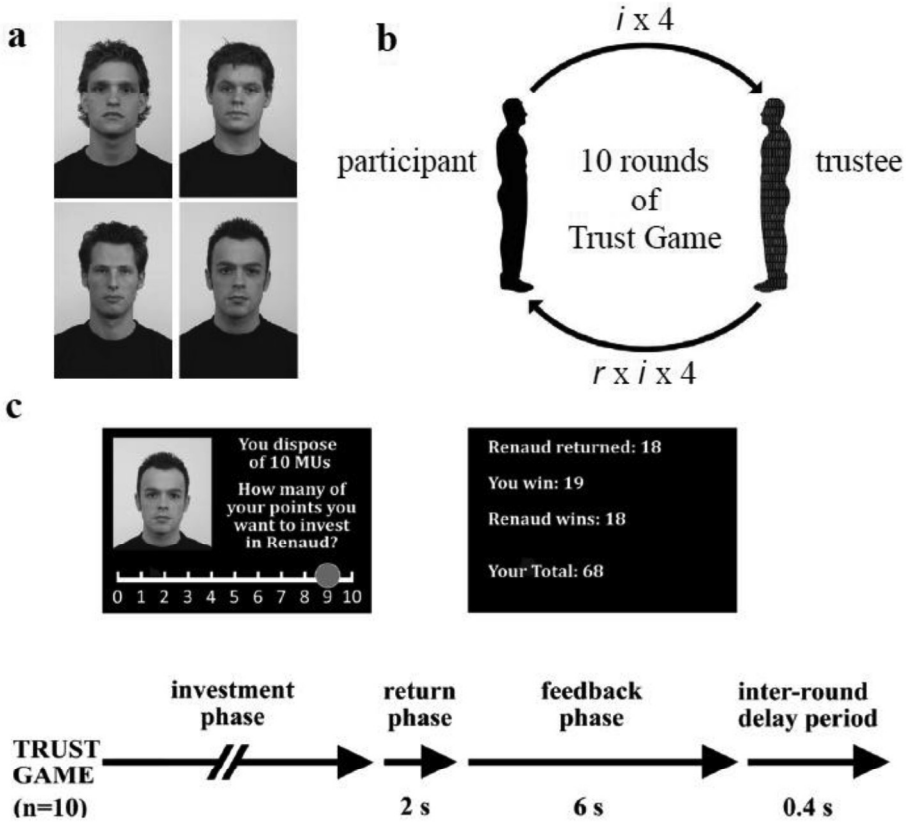


Figure 1. Experimental design. (A) Pictures of the four virtual players. (B) Schematic representation of the MTG. (C) Timeline of one round: Each round of the Trust Game consists of an investment (self-paced), a return (2 s), and a feedback phase (6s). Rounds are separated by an inter-round delay period of 0.4 s.

2.3. Experimental design

In the present version of the MTG, we manipulated orthogonally the counterparts' prior reputational information (positive or negative) and reciprocity (cooperative or individualistic) giving rise to four different experimental conditions: positive/cooperative, positive/individualistic, negative/cooperative, and negative/individualistic.

Prior to the beginning of the MTG, the four male counterparts were introduced by their photos, first names, and short biographical stories. The counterparts' photos were emotionally neutral and judged equal for attractiveness and trustworthiness (*Supporting Information, Methods and*

Materials, Table S1). The biographical stories depicted the person's lifestyle and professional achievements that could be either morally praiseworthy or blameworthy so as to create two positive and two negative reputational profiles. Two biographies described counterparts with positive human and social attitudes (e.g., a physician working for the non-governmental organization *Doctors Without Borders,* curing the sick in developing countries, who spends his time back in France mainly with his family and friends), and two biographies depicted counterparts with egotistic and negative social attitudes (e.g., the owner of a call center specializing in the sale of credit to pensioners from modest backgrounds, who uses all the revenues from his business to buy new racing cars, while his numerous interns work without a salary) (*SI, Methods and Materials, Biographical stories*).

Unbeknownst to the participants, virtual counterparts differed in their level of reciprocity: two counterparts were programmed to play a cooperative strategy, returning higher sums than the participants initially invested ($r \times i \times 4 > i$), and the remaining two were destined to play, in every round, an individualistic strategy, never returning more EMUs than the participants invested ($r \times i \times 4 \leq i$). To improve ecological validity (Lis et al. 2016; Shore & Heerey, 2013), counterparts varied their return ratio r randomly across rounds within predefined margins. Cooperative counterparts returned between 37.5% and 50% of the quadrupled investment $i \times 4$ (46.4%±0.03), and individualistic counterparts returned between 12.5% and 25% (17.92%±0.04) of $i \times 4$.

2.4. Procedure

Prior to each round, participants were endowed with 10 Experimental Monetary Units (EMUs), and instructed that they would play as investors with each counterpart during 10 consecutive rounds. They were informed that the objective of the game was to maximize one's own winnings while playing with the partners. They were told that, on each round, they could either keep the 10 EMUs or share any discreet amount of EMUs with the partner who received a quadrupled amount of EMUs. Then, the partner could reciprocate with the participant by sharing either more or fewer

EMUs than the latter initially invested. At the end of the experiment, they would be paid 1 € for every 75 EMUs earned during the game. Participants were introduced to the logic of the MTG without having it named it as such. The words "trust" or "trustworthiness" were never mentioned during the session and the experiment. Prior to the onset of the MTG, to familiarize participants with the task, they played two training rounds of the game (*SI, Methods and Materials, Procedure*).

After completing the MTG, participants were asked to judge the counterpart's trustworthiness on an 11-point Likert-scale ranging from 0 (not at all trustworthy) to 10 (very trustworthy) and on the counterpart's average return ratio (i.e., the ratio between the actual amount of EMUs returned to the participants and the amount of EMUs the counterpart could win by not returning anything) on a 11-point-scale, ranging from 0% to 100%. After completing the experiment, participants were paid 1 € for every 75 EMUs they had earned during the MTG.

3. Results

For all analyses, a $p < 0.05$ was taken as the criterion for significance, and an eta squared (η^2_p) was used as a measure of effect size. For all post hoc tests, Bonferroni correction was applied.

3.1. Total sum of investments

A two-sample t-test ($t(40) = -3.21$, $p < .01$, $r = .45$) revealed that the ASD group (194.29±38.10) compared to the TD group (237.96±46.36) globally invested significantly fewer EMUs. Since we reported significant intergroup difference in overall investments, we normalized the investments across groups by calculating an investment *ratio* for every round and participant by dividing the amount of EMUs invested in round x by the average amount of EMUs invested across all rounds of Trust Game and counterparts. This ratio measured whether the investment of a participant was above or below the individual mean investment over the course of the MTG:

$$\{\text{investment ratio in round x}\} = \frac{\{amount\ of\ EMUs\ invested\ in\ round\ x\}}{\{average\ amount\ of\ EMUs\ invested\ across\ all\ rounds\}}$$

3.2. Mean investments

To determine whether the counterparts' reputational status and reciprocity differentially influence decisions to trust or distrust in the two groups of participants, we ran a 2 × 2 × 2 repeated-measures ANOVA on mean investment ratios, with Group (ASD and TD) as a between-subject factor, and Reputational prior (positive and negative) and Reciprocity (cooperative and individualistic) as within-subject factors *(SI, Analyses and Results, Analysis on non-normalized investment data).*

For the Reciprocity factor, the first round was excluded because no information about the counterpart's reciprocity was available: the mean investment ratio was therefore computed as the mean over *nine* (rather than ten) rounds.

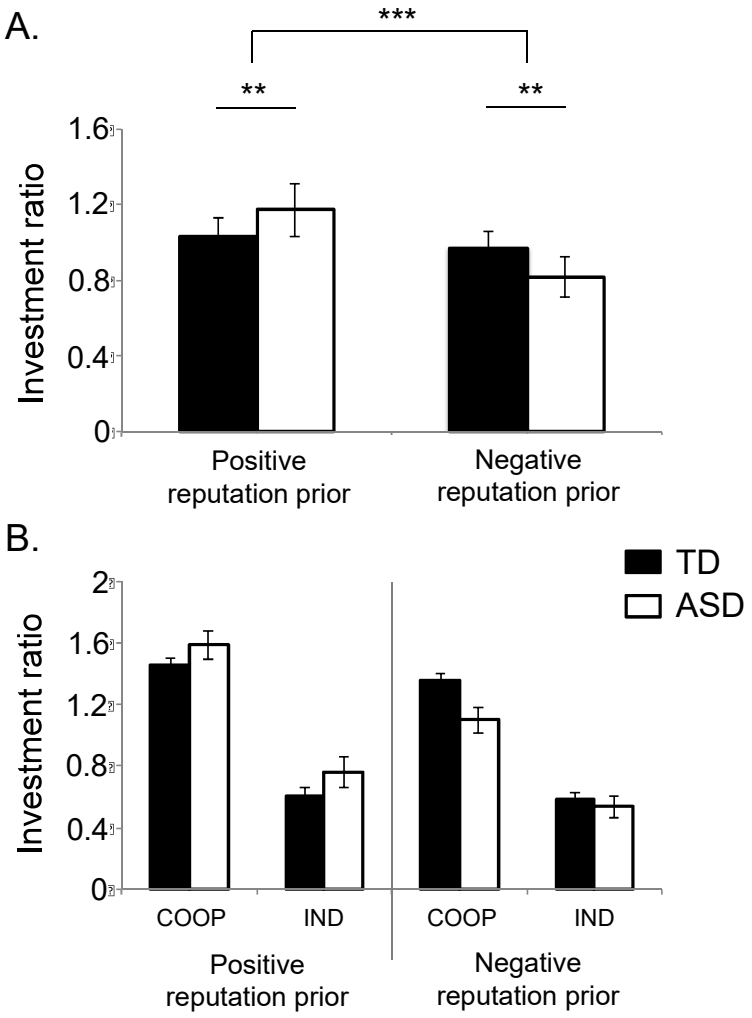
We found no main effect of group ($F(1,40) = 1.5$, $p = 0.23$), but significant main effects of Reputational prior ($F(1,40) = 28.40$, $p < .001$, $\eta^2_p = .42$), and Reciprocity ($F(1,40) = 90.02$, $p < .001$, $\eta^2_p = .69$) factors, as well as significant Reputational prior × Reciprocity ($F(1,40) = 7.29$, $p < .05$, $\eta^2_p = .15$) and Group × Reputational prior interactions ($F(1,40) = 13.72$, $p < .001$, $\eta^2_p = .25$, see Figure 2, A). Neither the Group × Reciprocity ($F(1,40) = 0.48$, $p = .49$) nor the 3-way Group × Reputational prior × Reciprocity ($F(1,40) = 2.43$, $p = 0.12$) interactions were significant.

Decomposing the Reputational prior × Reciprocity interaction, post-hoc comparisons revealed that when playing with *individualistic* counterparts, the Reputational prior had no significant effect on investment ratios as participants equally endowed counterparts with positive and negative reputational priors (0.67 vs. 0.56, $p = .13$). Conversely, Reputational prior had a strong significant effect on investments when playing with *cooperative* counterparts, as participants invested more with counterparts with a positive reputation than with counterparts with a negative reputation (1.51 vs. 1.25, $p < .0001$) (Figure 2, B). Importantly, the Group × Reputational prior interaction revealed that, relative to TD participants, participants with ASD endowed greater investment ratios to counterparts with a *positive* reputation (1.17 vs. 1.03, $p = .005$) and lower investment ratios to counterparts with a *negative* prior reputation (0.82 vs. 0.97, $p = .002$) irrespective of reciprocity

(**Figure 2, A**). Overall, the ASD group endowed significantly greater investment ratios to counterparts with a positive prior reputation than to counterparts with a negative prior reputation ($p < .001$), while this difference was not significant in TD ($p > .05$).

-- Figure 2 about here --

Figure 2. Mean investment ratios across conditions in ASD and TD participants. (A) Mean ± SEM of participants' investment ratios as a function of counterpart's Reputational prior (Positive and Negative reputation prior). **= $p < 0.01$; ***= $p < 0.001$. (B) Mean ± SEM of participants' investment ratios as a function of counterpart's Reputational prior and Reciprocity (Coop = Cooperative and Ind = Individualistic).



3.3. Weights of Reputational prior and Reciprocity on investments

To better characterize participants' sensitivity to prior expectations derived from reputation or reciprocity factors, we calculated the weights of "reputational prior" and "reciprocity". The weight of "reputational prior" was derived by subtracting the mean investment ratio endowed to counterparts with a *negative* reputation from the mean investment ratio endowed to counterparts with a *positive* reputation in the first and last rounds. Calculating the weight of counterparts' reputational prior on trust-learning allowed us to estimate the effects of this factor and to compare it across the two groups throughout the game. We reasoned that the more participants were sensitive to the reputational priors, the greater the impact of reputation valence on investments, i.e., the greater the difference between investment ratios endowed to counterparts with a *positive* vs. a *negative* reputation. The weight of the "reciprocity" was calculated by subtracting the mean investment ratio endowed to individualistic vs. cooperative counterparts in the second and last rounds. Again, we reasoned that the more participants were sensitive to reciprocity information, the greater the expected difference between the investment ratio endowed to *cooperative* vs. *individualistic* counterparts.

To assess the respective weight of Reputational prior and Reciprocity on participants' investments, investment ratios were subjected to a 2 × 2 × 2 repeated-measures ANOVA with Group (TD vs. ASD) as a between-subject factor, and Round (initial vs. last round) and Type of Weight (Reputational prior vs. Reciprocity) as a within-subject factors *(SI Analyses and Results, Trial-by-trial analyses).*

We found no significant effect of Group ($F(1,40) = 0.71$, $p = 0.4$), but a significant effect of the Type of Weight ($F(1,40) = 12.77$, $p < .001$, $\eta^2_p = .24$) and a significant Group × Type of Weight interaction ($F(1,40) = 6.48$, $p < .05$, $\eta^2_p = .14$). The weight of the Reputational prior was *greater* in ASD than in TD participants (0.49 vs. 0.19, $p < .001$), while the weight of Reciprocity on investment was *equal* in the two groups (0.58 vs. 0.73, $p = 0.32$). Moreover, for the ASD group, the

weights of Reputational prior and Reciprocity on the investment ratios were equal (reputational prior = 0.50 vs. reciprocity = 0.59, p = .56) whereas, for the TD group, the weight of Reciprocity (cooperative > individualistic) on the investments was significantly greater than the weight of Reputational prior (positive > negative) (Reciprocity = 0.73 vs. Reputation = 0.20, p < .0001) (Figure 3, A).

-- Figure 3 about here --

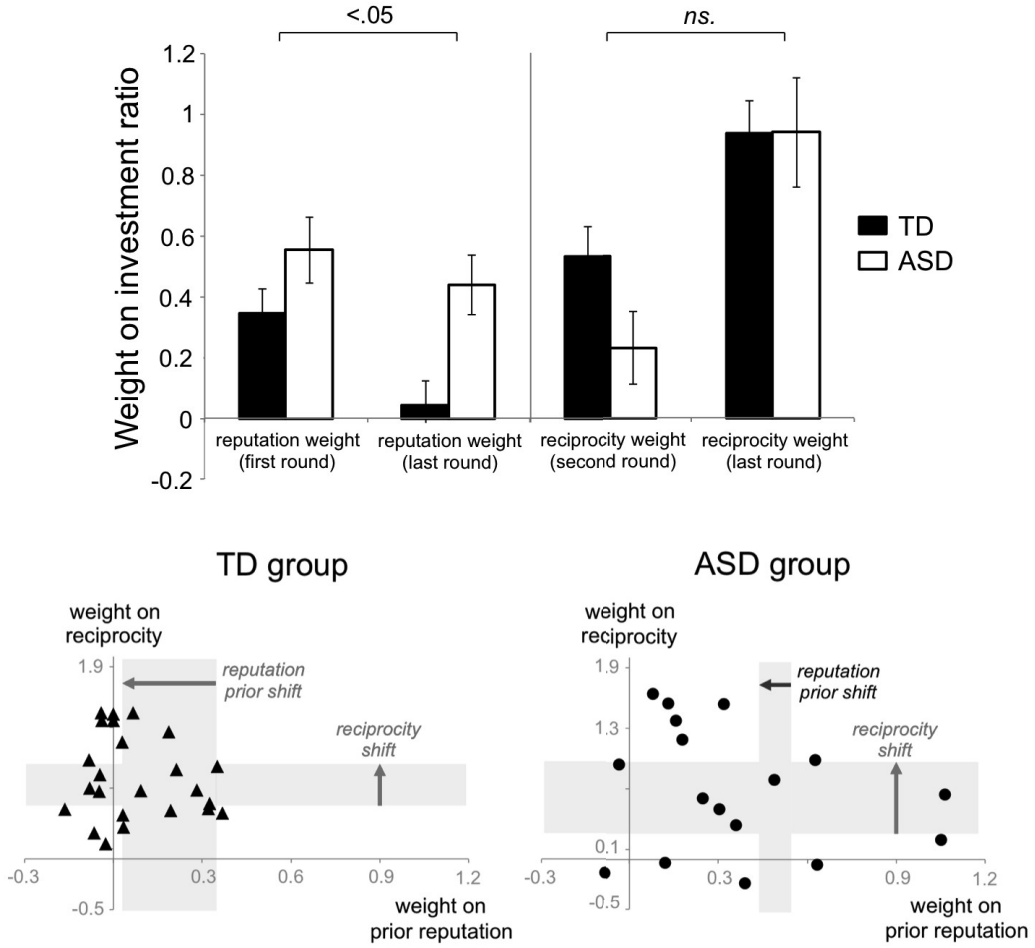


Figure 3. (A)The Weight of Reputational Prior and Reciprocity on mean investment ratios in the initial and final rounds in TD (black bars) and ASD (white bars) groups. Left panel: for the reputation factor, the Round × Group interaction is significant ($p < 0.05$). Right panel: for the reciprocity factor, the Round × Group interaction is not significant (*ns.*). All error bars indicate

SEM. (B) Scatterplot representing the weights of reputation prior (*x*-axis) and reciprocity (*y*-axis) on investment ratios throughout the MTG in TD (left panel, triangles: mean investment ratios) and ASD (right panel, circles: mean investment ratios) participants. The grey surface and the arrow represent the magnitude and the direction of a shift in the weight of the reputation factor (vertical surface) on investments from first to last round, and in the weight of the reciprocity factor (horizontal surface) from second to last round.

We also found a significant main effect of Round ($F(1,40) = 6.06$, $p = .02$, $\eta^2_p = .13$) but no significant Group × Round interaction effect ($F(1,40) = 2.97$, p= .09). The type of Weight × Round interaction was significant ($F(1,40) = 25.1$, $p < .0001$, $\eta^2_p = .39$) revealing that Reputational prior and Reciprocity differentially influenced investments in the initial and final rounds. While the weight of Reputational prior on investment ratios progressively *decreased* throughout the game ("reputation shift") so that the mean of investment ratio was greater in the initial round than in the final round (0.43 vs. 0.20, $p = .02$), the weight of Reciprocity *increased* throughout the game ("reciprocity shift") and was significantly greater in the last round, relative to the initial round (0.41 vs. 0.93, $p = .0002$). Interestingly, a significant reciprocity shift (vertical red arrows) was found in both groups: experiencing *reciprocity* (i.e., repeated interactions with either cooperative or individualistic counterparts) made participants invest more in cooperative than individualistic partners. In contrast, a significant shift in the "reputational" weight was only found in TD participants (horizontal red arrow), not in ASD (horizontal black arrow). Thus, while prior reputation was less and less taken into account by TD participants throughout the task, this effect of reputation remained stable at a high level in ASD (Figure 3, B).

Moreover, the weights of Reputational prior and Reciprocity on investment ratios significantly differed in the final round, due to the weight of Reciprocity having a greater influence than the weight of Reputational prior (0.93 vs. 0.20, $p < .0001$). Note that this effect was not modulated by the Group factor, as shown by a non-significant three-ways interaction effect ($F(1,40) = 0.15$, $p =$

0.7), revealing that the degree of a counterpart's reciprocity (i.e., cooperative or individualistic) influenced investments similarly in both groups ($p = 0.09$) (*(SI Analyses and Results, Figure S1*).

To recap, we found a similar effect of experiencing reciprocity (i.e., repeated interactions with either cooperative or individualistic counterparts) across the two groups. Both TD and ASD participants modified their investments according to whether the partner was cooperative (returned higher sums than the participants initially invested) or individualistic (returned lower sums than the participants initially invested). However, we found a stronger effect of reputational priors in ASD relative to TD participants. Thus, while the effect of Reputation progressively diminished in TD participants as a result of experiencing (repeated) interactions with counterparts, the weight of Reputation was never overcome by the weight of Reciprocity in ASD and overall was greater in ASD relative to TD, despite the experiencing of similar interactions by the two sets of participants.

3.4. Judgments of trustworthiness and reciprocity

To evaluate the effects of Reputational prior and Reciprocity on participants' judgments of counterparts' trustworthiness and reciprocity after completing the MTG, we ran two separate 2 × 2 × 2 repeated-measures ANOVA, with Group (ASD and TD) as between-subject factor, and Reputational Prior (positive and negative) and Reciprocity (cooperative and individualistic) as within-subject factors.

We found significant main effects of Reputational prior ($F(1,40) = 37.1$, $p < .0001$, $\eta^2_p = .42$) and Reciprocity ($F(1,40) = 44.2$, $p < .0001$, $\eta^2_p = .48$) factors, and a significant Group × Reputational prior interaction effect ($F(1,40) = 9.8$, $p < .003$, $\eta^2_p = .20$). All participants rated cooperative counterparts as more trustworthy than individualistic counterparts (6.2±2.8 and 3.1±2.6, respectively; $p < .0001$). The effect of the Reputational prior is further qualified by the Group × Reputational prior interaction revealing that ASD participants judged counterparts with a negative reputation as less trustworthy than did TD participants (2.7±2.4 and 4.3 ±3.1, respectively; $p < .01$). They also rated counterparts with a positive reputation as more trustworthy than those with a

negative reputation (6.2 ±2.9 and 2.7±2.4, respectively; p <.0001), while this difference was not significant in the TD group (5.4 ±3 and 4.3 ±3.1, respectively; p = .08) (*SI, Analyses and Results, Figure S2).*

Concerning the ex-post judgment of counterparts' reciprocity, as measured by the amount of EMUs returned by the counterpart, the 2 × 2 × 2 repeated-measures ANOVA yielded significant main effects of Reputational prior (F(1,40) = 9.41, p =.004, η^2_p = .19) and Reciprocity (F(1,40) = 17.5, p = .0002, η^2_p = .30), as well as a significant Group × Reputational prior interaction (F(1,40) = 4.49, p < .05, η^2_p = .10). Participants judged cooperative counterparts as having reciprocated more than individualistic ones (46.90±16.89 and 29.52±14.97, respectively; p <.0001). Bonferroni post-hoc comparisons revealed that the Group × Reputational prior interaction effect was due to the ASD group judging the returned investments of counterparts with a negative reputation as being significantly lower than those of counterparts with a positive reputation (30.9 ±12.8 and 46.2±15.3, respectively; p = .01). This difference was not significant in the TD group (36.6±12.8 and 39.4 ±14, respectively; p = 0.1) (*SI, Analyses and Results, Figure S2).*

3.5. Relationship with clinical symptoms

Regression analyses were conducted to assess whether an abnormal dependence on Reputational priors was predictive of the severity of autistic symptoms, in the areas of repetitive behavior, communication, and social interaction, as measured by the Autism Diagnostic Interview–Revised (ADI-R). For each clinical sub-score, we conducted regression analyses using *i)* the 'reputation score' (i.e., the initial weight on prior reputation) or *ii)* the 'reciprocity score' (i.e., end weight – initial weight of Reciprocity on investment ratio) as predictor variables. One participant was excluded from the analysis because his 'reputation score' was greater than two standard deviations below the group mean. We used either raw scores (simple linear regressions) or their transformed values (simple non-linear regressions with logarithmic, polynomial, or exponential transformations). Models with the highest adjusted R^2 and a p<0.05 are reported.

We found a significant positive correlation between the Reputation score and ADI sub-score measuring social interaction disorders (R^2 = 0.27, p = 0.032) indicating that the greater the effect of the Reputational prior, the more severe the social disturbances in ASD participants. The Reciprocity score correlated negatively with the ADI score for repetitive and stereotyped behaviors (R^2 = 0.32, p = 0.016) so that the greater the effect of Reciprocity on mean investment ratio, the less repetitive and restricted the behaviors **(Figure 4)**.

-- Figure 4 about here --

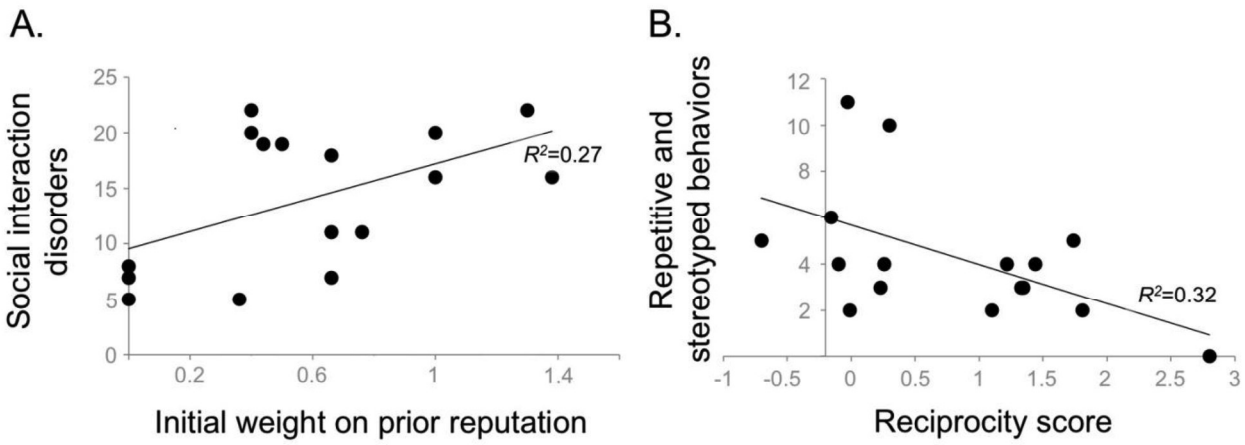


Figure 4. Regression analyses between cognitive variables and clinical symptoms in ASD participants. (A) The reputational score (i.e., the initial weight on prior reputation) positively correlated with the severity of social interaction disorders: the greater the influence of prior reputation, the more severe the symptomatology in the domain of social interaction. (B) The magnitude of the reciprocity score negatively correlated with the severity of repetitive and stereotyped behaviors: the greater the effect of reciprocity on investments over time, the less repetitive and restricted the behaviors exhibited by ASD participants.

We also observed a correlation trend between the Reciprocity score and the ADI sub-score for communication (R^2 = 0.19, p = 0.08), so that the greater the effect of Reciprocity, the more severe communication impairments were in participants with ASD**.** No significant correlation was found between the Reciprocity score and the ADI sub-scores in social interaction (R^2 = 0.008, p = 0.73), and between the Reputation score and the ADI sub-scores measuring disturbances in the domains of repetitive and stereotyped behaviors (R^2 = 0.003, p = 0.81) and communication (R^2 = 0.16, p = 0.1).

In addition, we performed regression analyses using the AQ scores against Reputation and Reciprocity scores in all participants. We found that the AQ score tended to correlate positively with the Reputation score (R^2 = 0.31, p = 0.1) while the correlation between the AQ score and the Reciprocity score was not significant (R^2 = 0.14, p = 0.61) (*SI Analyses and Results, Figure S3).*

4. Discussion

Accurate estimation of another person's trustworthiness is essential in many instances of social interaction. Reputation and previous reciprocal interactions have been shown to deeply influence the way people perceive and interact with each other (Delgado et al., 2005; King-Casas et al., 2005; Boero et al., 2009; Fouragnan et al., 2013; Ewing et al., 2015). In the present study, we orthogonally manipulated the Reputation Prior (positive or negative) and the Reciprocity (cooperative or individualistic) factors during a modified version of MTG to investigate the effects of knowledge of the partner's prior reputation (positive or negative) and reciprocity (cooperative or individualistic behavior) on trust-learning, in adults with and without ASD.

As expected, both individuals with and without ASD were sensitive to the counterpart's reputational priors and reciprocity, and all participants were influenced by the reputational priors in their initial decision to trust or distrust the counterpart, in that they were prone to invest more with counterparts with a positive reputation than with those with a negative reputation. Crucially, following the very first exchanges, TD participants adapted their investment strategies mainly on the basis of the counterpart's cooperative or individualist attitudes, by adopting a TfT strategy. At the end of MTG, for TD individuals, the reputational prior was no longer effective and their decision to trust or not trust was mainly affected by the counterpart's direct reciprocity. Consistently, their final judgments of trustworthiness and reciprocity essentially resulted from the counterpart's actual behavior, that is from his or her propensity to reciprocate generosity. These findings are consistent with a previous study showing that while reputation would affect the very first rounds of a Prisoner's Dilemma game, after a couple of rounds the personal experience obtained with the co-player about his/her propensity to reciprocate becomes more decisive (Wedekind & Braithwaite, 2002).

The present results also show that, overall, participants with ASD allocated smaller investments to counterparts, suggesting difficulty or reluctance to develop trust in others, in accordance with previous studies suggesting that ASD is associated with reduced experience of trust or difficulties in

estimating trustworthiness in others (Lis et al. 2016; Li, Zhu & Gummerum, 2014). More importantly, although ASD participants were sensitive to reciprocity, the weight of reciprocity on the investment strategy never overcame the weight of the reputational prior, which continued to exert a strong impact on their trust behavior. More precisely, unlike TD participants, participants with ASD were highly reluctant to reciprocate with participants with incongruous profiles: they never completely trusted the counterparts with a negative reputation, even if they repeatedly showed collaborative behavior, and they continued to trust more the counterparts with a positive reputation, who received greater amounts of investment, even if they showed uncooperative behavior.

Calculating the weights of counterparts' reputational prior and reciprocity on trust-learning allowed for estimating the respective effects of these factors and for comparing them in the two groups of participants throughout the game. The results confirmed that while both groups were equally responsive to direct reciprocity, and modified their investments according to whether the partner was cooperative (returned higher sums than the participants initially invested) or individualistic (returned lower sums than the participants initially invested), the weight of reputational prior was significantly stronger in ASD than in TD participants. Consistently, while for the ASD group the weights of Reputational prior and Reciprocity on the investments were equal, for the TD group the weight of Reciprocity was stronger than the weight of Reputational prior so that, by the end of the game, the Reciprocity weight predominated and progressively extinguished the effect of Reputation.

In the present study, reputational information about the counterparts was explicitly provided to create prior expectations (or beliefs) about a partner's collaborative or defective attitudes. However, although reputational information can sometimes be a surrogate when direct or prior personal experience is lacking or unreliable, there is always the risk that reputation does not accurately reflect the agent's actual intentions. Participants were thus expected to use this information and to combine it with direct personal experience about a counterpart's behavior to form new expectations

about his/her actual propensity to reciprocate. In our study, participants could revise their prior beliefs about a partner's trustworthiness and maximize their benefits during the MTG, while over-adjusting or under-adjusting their own investment, based on reciprocation. Hence, TfT turns out to be the optimal behavioral strategy for repeated interactions, as we would trust another agent only as long as he/she reciprocates and stop trusting once our trust is betrayed (Axelrod, 1984).

Remarkably, the increased and long-lasting influence of reputational priors also affected the ex-post judgments of a counterpart's trustworthiness and reciprocity. At the end of the game, all participants estimated that cooperative counterparts were more trustworthy than individualistic ones. However, individuals with ASD consistently judged counterparts with negative reputations to be less trustworthy than did TD participants. They also rated counterparts with a negative reputation as being less trustworthy than the ones with a positive reputation, regardless of their reciprocity. Similarly, while all participants estimated that cooperative counterparts reciprocated more than individualistic ones, only the ASD group reported that the returned investment by counterparts with a positive reputation was significantly greater than that by counterparts with a negative reputation.

Interestingly, the Reputation score positively correlated with social disturbances, as measured by the ADI-R sub-score in participants with ASD, and with the severity of autistic traits in all participants, as measured by the AQ score, suggesting a relationship between an abnormal weight of reputational priors and social impairments, and more generally, with the severity of the autistic symptoms. In addition, we found that, while reduced reciprocity might be a predictor of communication impairment, a significant correlation indicates a stronger relationship between reduced reciprocity and the severity of repetitive/ stereotyped behaviors in ASD.

These findings raise an important issue regarding the causal links between the increased effect of reputational prior knowledge, observed in the present study, and social and behavioral impairments in ASD, which might impede these persons from learning through interpersonal interactions. Although the present study suggests that increased effect of reputational priors, or "Hyper-Priors", might predict deficits in social interaction and behavior in ASD, when taken together, current

findings converge on the notion that the unbalanced interplay of prior knowledge and sensory information would be responsible for the abnormal or atypical cognitive processing in ASD. Interestingly, the present results also suggest a relationship between an impairment in this Bayesian computational mechanism which relies on two sources of information - prior knowledge and sensory information - and behavioral inflexibility in ASD, as already reported by a previous study (Chambon et al., 2017a). Faced with this complex scenario, Bayesian approaches to human brain functions may provide a unified framework that links distinct, and apparently unrelated, autistic characteristics and impairments, encompassing the cognitive, sensory, behavioral, and clinical domains.

Within the Bayesian framework, Pellicano and Burr (2012) propose that perceptual experience in ASD can be explained by reduced priors, or *Hypo-Priors*, while the predictive coding account (Lawson et al., 2014) claims that cognitive impairments and perceptual atypicalities in ASD are rather explained by an aberrant increased precision of sensory information, relative to prior expectations. Recently, Chambon et al. (2017a) have found that TD adults relied more on prior knowledge and were spontaneously biased towards the TfT interaction mode when asked to make predictions about other people's behavior. Conversely, adults with ASD exhibited weaker prior social expectations and no spontaneous predisposition for the TfT strategy of interaction. Unlike the Chambon et al. (2017a) study, in the present work, all participants were sensitive to the prior reputational information, which was explicitly provided before the MTG, as this strongly affected the decision to trust in their initial exchange, when direct information about the counterpart's attitude was not available. However, while in TD participants prior expectations were updated and behavior strategy was flexibly modulated by direct information about the partner's fairness, participants with ASD remained over-sensitive to the reputational priors throughout the game, and never adopted the TfT interaction strategy in those conditions in which reputation and direct reciprocity generated incongruous profiles for a partner. In a slight variation of the predictive coding model, Van de Cruys and collaborators (2014) showed that perceptual operations in people

with ASD are not characterized by *hypo-priors*, but would rather reflect difficulties in flexibly adjusting precision. During on-line social interaction, as is the case in our version of the MTG, greater precision at higher hierarchical levels can hinder social learning as each new experience would generate large prediction errors resulting in a diminished ability to update trustworthiness based on direct social exchanges (Van de Cruys et al., 2014).

Overall, these findings suggest that impairments in social interaction and cognition in ASD result from the unbalanced interplay between top-down processing based on prior knowledge and bottom-up sensory driven processing, which specifically affects the domain of social interaction. Thus, rather than a general abnormal effect of prior knowledge (*Hyper*- or *Hypo-Priors*), a specific deficit in the Bayesian inferential mechanisms that adaptively integrate social prior expectations with sensory information would account for difficulties with social learning and decision making in ASD.

It is noteworthy that, unlike Chambon et al., (2017a)'s study, in which social priors were induced implicitly, in the present study the social reputational priors were provided explicitly. This might explain why individuals with ASD, who are often impaired in processing implicit social knowledge (see Senju et al., 2009; Zalla et al., 2014), could be more responsive to prior social knowledge when it is made salient. In a previous study, Zalla et al. (2014) found that while social stereotyped knowledge automatically enhances detection of the speaker's communicative intents and attitudinal features in TD adults, it is not integrated and used in pragmatic communicative processes in adults with ASD when this knowledge is processed implicitly. These findings are consistent with the notion that difficulties arise for individuals with ASD when they have to implicitly draw inferences about mental states and social attitudes from general knowledge. Thus, reduced automaticity in processing social information might explain some of the impairments in rapid communication and on-line social interaction. It is likely that only when social information is encoded overtly or activated through explicit and controlled processes, can it be used for social reasoning and behavior by individuals with ASD, as in the present study.

Human decision is driven by multiple goals and motivations. The ecological validity of trust games has been extensively assessed in previous studies showing that trust behavior is predicted by the partner's reputation (Camerer, 2003; Glaeser et al., 2000) as well as by his/her reliability in real-life situations and social acceptance (Rotenberg et al., 2005). Reputation, which is built on adherence to shared moral values, can be an effective tool for predicting another's reciprocation and fairness. For example, research on trolley dilemmas has suggested that decision makers do not care exclusively about expected costs and benefits, but also give weight to other considerations, including evaluative judgments about the moral rightness or wrongness of potentially offensive actions, distinct from their expected consequences (Bennis, Medin & Bartels, 2010).

Previous studies on moral cognition have found that ASD individuals were highly responsive to moral normative violations and that their moral judgments were more severe than those of typical individuals (Zalla et al., 2011; Buon et al., 2013). When asked to judge involuntary offensive actions that are socially inappropriate, such as the Faux Pas, individuals with ASD attached more importance to normative transgressions than to the agent's intention, or the victim's emotional state generated by the situation (Zalla et al., 2009). Because of attenuated social priors, social interactions in a dynamically changing social world would be a real challenge for individuals with ASD. Hence, relying on stable representations, such as social reputation grounded on moral judgments, could be used as a compensatory mechanism to improve their capacity to predict future encounters and supplement their lack of flexibility in adjusting to rapid changes in the social world.

The inflexible overreliance on a partner's reputation, as revealed by the present study, is consistent with previous evidence showing reduced cognitive flexibility in ASD (Hill & Bird, 2006). In accordance with this interpretation, recent studies on trustworthiness indicate that children with ASD are less flexible in revising their initial trust judgment after deception, as compared to TD children (Yi et al., 2013), and show more trust toward an unknown person, as compared to TD children (Yi et al., 2014).

In conclusion, using a MTG, the present study has investigated how people with and without ASD dynamically build trustworthiness and develop a behavioral strategy based on the interplay between prior knowledge about a partner's social reputation and direct personal interaction. We show that adults with ASD have difficulties with dynamically integrating these two distinct sources of information, relying strongly and inflexibly on prior reputational information about their partners and less on feedback learning. The weight of counterpart reciprocity was lower in participants with ASD, insofar as they rely more on reputational priors when making investment decisions. Reduced reciprocity was related to the clinical severity of ASD symptoms in the domains of repetitive and restricted behaviors, and communication. In line with Bayesian theoretical accounts, the present results indicate that difficulties with social interaction in adults with high-functioning ASD appear to stem from impairments in building and updating prior social knowledge, which severely hinder social learning based on direct evidence and accurate probability judgments about incoming events. Thus, when asked to estimate the amount of investments by the partner, participants with ASD exhibit inaccurate memory about reciprocation. The fact that adults with ASD erroneously remembered the amount of return EMUs supports the hypothesis that enhanced adherence to reputational knowledge reflects difficulties with updating newly acquired relevant social information about someone's trustworthiness during on-line interaction.

Finally, we would like to acknowledge some limitations of the present study. The first concerns the small sample size, including mainly male participants. Further studies are needed to replicate these results on a larger group of individuals with ASD, possibly including more female participants, to investigate sex/gender differences in social abilities using a trust game. Recent research has reported that females with ASD have superior socio-emotional skills and friendship stability compared to males with ASD, partly explaining the male bias in autism prevalence (Head, McGillivray & Stokes, 2014). Further studies using different experimental paradigms are needed to support the hypothesis of a specific unbalanced interaction of top-down and bottom-up processing of social information in autism. In addition, since the current study did not balance the presentation

order of reputation and reciprocity information, a follow-up study should assess whether difficulties in updating acquired information will also occur when reciprocity signals precede reputational information.

Since little is known about trust and sociability in younger individuals with autism, future research should also be conducted on children and adolescents to detect relevant developmental changes in trust-building and reciprocation in both ASD and typical population.

Acknowledgements

We gratefully acknowledge the commitment of the participants and their families to the pursuit of research in autism. This research was supported by Fondation FondaMental to TZ and ML. T.Z., S.B.-G. and V.C. were supported by ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* (program 'Investissements d'Avenir'). V.C. was supported by ANR-16-CE37-0012-01 ("Jeune Chercheur"). Designed research: T.Z, S B-G, and C.M. Performed research: C.M and M.L. Analyzed data: V.C., C.M. and T.Z. Wrote the paper: T.Z., C.M. and V.C. There was no conflict of interest.

References

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of cognitive neuroscience, 13*(2), 232-240.

American Psychiatry Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). DSM-IV-TR (Text Revision). Washington, DC: American Psychiatry Association.

Andari, E., Duhamel, J. R., Zalla, T., Herbrecht, E., Leboyer, M., & Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proceedings of the National Academy of Sciences, 107*(9), 4389-4394.

Axelrod, R. (1984). *The Evolution of Cooperation.* New York, NY: Basic Books.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition, 21*(1), 37-46.

Baron-Cohen, S. (1995). Mindblindness. An essay on autism and theory of mind. Cambridge: MIT Press.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders, 31*(1), 5-17.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience, 10*(9), 1214-1221.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science, 5*(2), 187-202.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior, 10*(1), 122-142.

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2009). Reputational cues in repeated trust games. *The Journal of Socio-Economics, 38*(6), 871-877.

Bourgeois-Gironde, S., & Corcos, A. (2011). Discriminating strategic reciprocity and acquired trust in the repeated trust-game. *Economics Bulletin, 31*(1), 177-188.

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. *Journal of autism and developmental disorders, 43*(2), 458-470.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Chambon, V., Domenech, P., Pacherie, E., Koechlin, E., Baraduc, P., & Farrer, C. (2011). What are they up to? The role of sensory evidence and prior knowledge in action understanding. *PloS One, 6*(2), e17133.

Chambon, V., Farrer, C., Pacherie, E., Jacquet, P. O., Leboyer, M., & Zalla, T. (2017a). Reduced sensitivity to social priors during action prediction in adults with autism spectrum disorders. *Cognition, 160*, 17-26.

Chambon V., Domenech P., Jacquet P.O., Barbalat G., Bouton S., Pacherie E., Koechlin E., Farrer C. (2017b). Neural coding of prior expectations in hierarchical intention inference. Scientific Reports, 7(1):1278.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature neuroscience, 8*(11), 1611-1618.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution, 2*, 265–279.

Deutsch, M. (1960). Trust, trustworthiness, and the F scale. *The Journal of Abnormal and Social Psychology, 61*(1), 138-140.

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Appearance-based trust behaviour is reduced in children with autism spectrum disorder. *Autism, 19*(8), 1002-1009.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience, 33*(8), 3602-3611.

Gambetta, D. (2000). Can We Trust Trust? In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 213-237). Oxford: electronic edition, University of Oxford.

Gillberg, C., Gillberg, C., Råstam, M., & Wentz, E. (2001). The Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI): a preliminary study of a new structured clinical interview. *Autism, 5*(1), 57-66.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics, 115*(3), 811-846.

Head AM, McGillivray JA & Stokes MA. (2014) Gender differences in emotionality and sociability in children with autism spectrum disorders. *Molecular Autism*. Feb 28;5(1):19.

Hill, E. L., & Bird, C. M. (2006). Executive processes in Asperger syndrome: Patterns of performance in a multiple case series. *Neuropsychologia, 44*(14), 2822-2835.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences, 112*(6), 1727-1732.

King-Casas B, et al. (2005) Getting to know you: reputation and trust in a two-person economic exchange. *Science 308*(5718):78-83.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and emotion, 24*(8), 1377-1388.

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in human neuroscience, 8*, 302.

Li, J., Zhu, L., & Gummerum, M. (2014). The relationship between moral judgment and cooperation in children with high-functioning autism. *Scientific reports, 4*, 4314.

Lis, S., Baer, N., Franzen, N., Hagenhoff, M., Gerlach, M., Koppe, G., ... & Kirsch, P. (2016). Social interaction behavior in ADHD in adults in a virtual trust game. *Journal of attention disorders, 20*(4), 335-345.

Lord, C., Rutter, M., & Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders, 24*(5), 659-685.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders, 30*(3), 205-223.

Modahl, C., Green, L. A., Fein, D., Morris, M., Waterhouse, L., Feinstein, C., & Levin, H. (1998). Plasma oxytocin levels in autistic children. *Biological psychiatry, 43*(4), 270-277.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688-2692.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092.

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: a predictive processing perspective. *Proceedings of the Royal Society of London B: Biological Sciences, 282*(1802), 20141557.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences, 16*(10), 504-510.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods, 162*(1), 8-13.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies, 62*(3), 381-422.

Rotenberg, K. J., Fox, C., Green, S., Ruderman, L., Slater, K., Stevens, K., & Carlo, G. (2005). Construction and validation of a children's interpersonal trust belief scale. *British Journal of Developmental Psychology, 23*(2), 271-293.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science, 325*(5942), 883-885.

Skewes, J. C., Jegindø, E. M., & Gebauer, L. (2015). Perceptual inference and autistic traits. *Autism, 19*(3), 301-307.

Shore, D. M., & Heerey, E. A. (2013). Do social utility judgments influence attentional processing?. *Cognition, 129*(1), 114-122.

Turi, M., Karaminis, T., Pellicano, E., & Burr, D. (2016). No rapid audiovisual recalibration in adults on the autism spectrum. *Scientific reports, 6*, 21756.

van Boxtel, J. J., Dapretto, M., & Lu, H. (2016). Intact recognition, but attenuated adaptation, for biological motion in youth with autism spectrum disorder. *Autism Research, 9*(10), 1103-1113.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological review, 121*(4), 649-675.

Volkmar, F. R., & Klin, A. (2000). Diagnostic Issues in Asperger Syndrome. In A. Klin, F. Volkmar, & S. S. Sparrow (Eds.), *Asperger's syndrome* (pp. 25–71). New York: Guilford.

Yi, L., Pan, J., Fan, Y., Zou, X., Wang, X., & Lee, K. (2013). Children with autism spectrum disorder are more trusting than typically developing children. *Journal of experimental child psychology, 116*(3), 755-761.

Yi, L., Fan, Y., Li, J., Huang, D., Wang, X., Tan, W., ... & Lee, K. (2014). Distrust and retaliatory deception in children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 8*(12), 1741-1755.

Wechsler D., (1999) Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio, TX: Harcourt Assessment.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science, 288*(5467), 850-852.

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current biology, 12*(12), 1012-1015.

Zalla, T., Sav, A. M., Stopin, A., Ahade, S., & Leboyer, M. (2009). Faux pas detection and intentional action in Asperger Syndrome. A replication on a French sample. *Journal of Autism and Developmental Disorders, 39*(2), 373-382.

Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology, 2*(4), 681-698.

Zalla, T., Barlassina, L., Buon, M., & Leboyer, M. (2011). Moral judgment in adults with autism spectrum disorders. *Cognition, 121*(1), 115-126.

Zalla, T., Amsellem, F., Chaste, P., Ervas, F., Leboyer, M., & Champagne-Lavau, M. (2014). Individuals with autism spectrum disorders do not use social stereotypes in irony comprehension. *PloS One, 9*(4), e95568.

Supporting Information

Methods and Materials

Table S1. Rating of facial stimuli.

In a preliminary study, an independent sample of 32 TD adults (20 females; mean age 26.5±5.4 years) were asked to rate 17 faces from the Caucasian male neutral frontal set of the Radboud Faces Database (Langner et al. 2010) for perceived trustworthiness and attractiveness using a five-point Likert-scale ranging from 1 (not at all), and 5 (highly) trustworthy/attractive. The four facial stimuli used in our study did not differ in trustworthiness and attractiveness (Rafd090_07, Rafd090_23, Rafd090_30, Rafd090_71).

	Face 1	Face 2	Face 3	Face 4	*T*-test (Bonferroni corrected)
Trustworthiness	3.1±0.9	3.1±0.8	3.0±0.8	2.9±1.0	all $p > .05$
Attractiveness	2.8±0.9	2.7±0.6	3.0±0.9	3.2±0.8	all $p > .05$

Biographical stories

Each face representing a counterpart was randomly associated with a first name and a biography. Prior to the onset of the MTG, participants were required to correctly recognize the four counterparts by their photos, first names and biographical stories. Each counterpart was presented for at least 20-30 seconds to allow participants to encode and memorize the relevant information. To ensure that they were able to correctly identify the four counterparts, prior to the onset of the game, they were required to correctly match each photo with the corresponding biographical description. The matching task was repeated until the participant successfully identified the four counterparts.

The four biographies were previously evaluated for trustworthiness by 20 adults (10 females) using an 11-point Likert-scale ranging from 0 (not at all trustworthy) to 10 (very trustworthy). The two positive profiles significantly differed from the two negative profiles in trustworthiness ($p<.0001$) while the two negative and the two positive profiles did not differ from one another.

1. Positive reputational profile 1

Renaud is a physician working for the non-governmental organization “Doctors without Borders”. He takes part in various missions around the globe improving the health and quality of life of invalids in Third World countries ravaged by conflicts. He has accepted important personal and financial sacrifices. While he is at home in France, he devotes most of his time to his family and friends.

2. Positive reputational profile 2

Clément is a teacher at a local elementary school with a high rate of disadvantaged children. Thanks to his innovative style of teaching, his students benefit from a successful learning process. Lately, his pedagogical methods have earned him the prize for “teacher of the year” by the local education authority.

3. Negative reputational profile 1

Pierre is a real-estate agent. Recently, after a promotion, he divorced his former wife, who is now raising their two daughters. In fact today he dedicates his whole time to advancing his work and carreer. When he became department manager of his agency, he did not hesitate to get rid of three employees and double his salary.

4. Negative reputational profile 2

Thibaud is the owner of a call center specializing in sale on credit. The main part of his revenue comes from pensioners of modest backgrounds. He employs many interns, whom he does not pay for their work. He is a big fan of motor sports and plans to use the benefits from his firm soon to buy a brand-new racing car.

Procedure

At the beginning of the experiment, participants were informed that they were about to play an economic game in the role of the investor with four different partners via a personal computer interface. No further information about the partners and their location was provided.

The two training rounds involved the experimenter showing the participant a schematic representation of the investment phase of the MTG (Figure 1c). The participant was then asked how many EMUs he would like to share with the training partner. Depending on the participant's choice the experimenter would summarize how many EMUs the training partner was going to receive after quadrupling the investment. Next, the experimenter explained that the training partner could either return more EMUs or less EMUs than the participant had invested, exemplifying both options and its consequences for the payoff structure of the participant and the training partner. After selecting the investment of the second training round the participant was asked to calculate how many EMUs the training partner would receive. Next, the participant was asked to calculate her own payoff and that of the training partner for two possible scenarios, i.e. the training partner returns double the investment or half the investment.

Analyses and Results

Analyses on non-normalized investment data

We performed a 2 × 2 × 2 repeated measures ANOVA on non-normalized total investments, with Group (ASD and TD) as a between-subject factor, and Reputational prior (positive and negative) and Reciprocity (cooperative and individualistic) as within-subject factors. The results confirm what we observed in our previous analyses on normalized data (investment ratios), except a now non-significant Reputational prior × Reciprocity interaction ($F(1,40) = 3.83$, $p = .06$).

For the Reciprocity factor, the first round was excluded because no information about the counterpart's reciprocity was available: the total non-normalized investment was therefore computed as the sum over *nine* (rather than ten) rounds. We found a significant main effect of group ($F(1,40) = 9.44$, $p < 0.01$, $\eta^2_p = .19$), Reputational prior ($F(1,40) = 22.84$, $p < .001$, $\eta^2_p = .36$), and Reciprocity ($F(1,40) = 101.21$, $p < .001$, $\eta^2_p = .72$) factors, as well as a significant Group × Reputational prior interaction ($F(1,40) = 10.41$, $p < .01$, $\eta^2_p = .21$). Neither the Group × Reciprocity ($F(1,40) = 3.21$, $p = .07$), Reputational prior × Reciprocity ($F(1,40) = 3.83$, $p = .06$), nor the 3-way Group × Reputational prior × Reciprocity ($F(1,40) = 0.72$, $p = 0.4$) interactions were significant. Decomposing the Group × Reputational prior interaction revealed that, overall, the ASD group endowed, over nine rounds, significantly greater investments to counterparts with a positive prior reputation than to counterparts with a negative prior reputation ($p < .001$), while this difference was not significant in TD ($p > .05$). Furthermore, relative to TD participants, participants with ASD tended to endow, over nine rounds, lower investments to counterparts with a *positive* reputation (52.53 vs. 55.88, $p = .36$) and significantly lower investments to counterparts with a *negative* prior reputation (37.19 vs. 52.94, $p < .001$), irrespective of reciprocity.

Trial-by-trial analyses

We performed a 2 (Reputation: positive vs. negative) × 10(rounds) × 2(Group: TD vs. ASD) ANOVA with including the 10 rounds in the analysis, and a 2 (Reciprocity: cooperative vs.

individualistic) × 9(rounds) × 2(Group: TD vs. ASD) ANOVA. The results are consistent with what we reported in our previous analyses.

For the reputation analysis, the main effect of round was highly significant ($F(9, 360) = 8.7$, $p < 0.001$, $\eta^2_p = .17$) with participants giving more at the end of the game than at the beginning, suggesting a strong effect of experiencing reciprocity on subsequent trials. We also found a significant main effect of the reputation factor ($F(1, 40) = 35.5$, $p < 0.001$, $\eta^2_p = .47$) as overall participants were more likely to invest on partners with a positive vs. a negative reputation (Bonferroni post-hoc tests comparing positive vs. negative reputation: 1.1 vs. 0.89, $p < 0.001$). The reputation × group interaction was significant ($F(1,40)= 14.2$, $p=0.005$, $\eta^2_p = .47$: 0.26). Participants with ASD, relative to TD, made greater donations to partners with a positive initial reputation (ASD vs. TD: 1.18 vs. 1.04, $p = 0.004$) and lower donations to partners with a negative initial reputation (ASD vs. TD: 0.81 vs. 0.96, $p = 0.002$). As for the reputation × round interaction, it was also significant ($F(9, 360) = 2.1$, $p = 0.03$, $\eta^2_p = .04$): participants gave more to partners with a negative, relative to a positive, reputation as the number of rounds increased (positive reputation, first vs. last round : 0.97 vs. 1.18, $p = 0.02$; negative reputation, first vs. last round; 0.65 vs. 0.97, $p= < 0.001$). Note that this effect is mostly due to the fact that donations to partners with a positive reputation reached a ceiling rapidly during the game (at the 5th round), whereas donations to *individualistic* partners with a *positive* reputation rapidly decreased with the number of rounds. Finally, no main effect of group, and no round × group, or reputation × group, or reputation × round × group interaction effect, were found.

The analysis conducted on the Reciprocity score also parallels our previous analyses. We found a significant main effect of the reciprocity factor ($F(1, 40) = 89.3$, $p < 0.001$, $\eta^2_p = .69$). Thus, participants were more likely to invest on cooperative rather than individualistic partners (Bonferroni post-hoc tests: 0.69 vs. 0.30, $p < 0.001$). We also found a significant reciprocity × round interaction ($F(8, 320) = 8.14$, $p < 0.001$, $\eta^2_p = .17$), with participants giving higher donations to cooperative partners in the last rounds of the game than at the beginning (cooperative partners,

second vs. last rounds: 0.60 vs. 0.72, $p < 0.001$; individualistic partners, second vs. last rounds: 0.39 vs. 0.27, $p = 0.05$). We found no main effect of group, and no significant reciprocity × group, or round × group, or reciprocity × round × group, interaction effects.

Figure S1. Round by round investments in TD (Upper panel) and ASD (Lower panel) participants. Mean ± SE of participants' investments of EMUs are shown in the four counterparts, as a function of the Reputational prior (Positive/Negative) and Reciprocity (Cooperative/Individualistic).

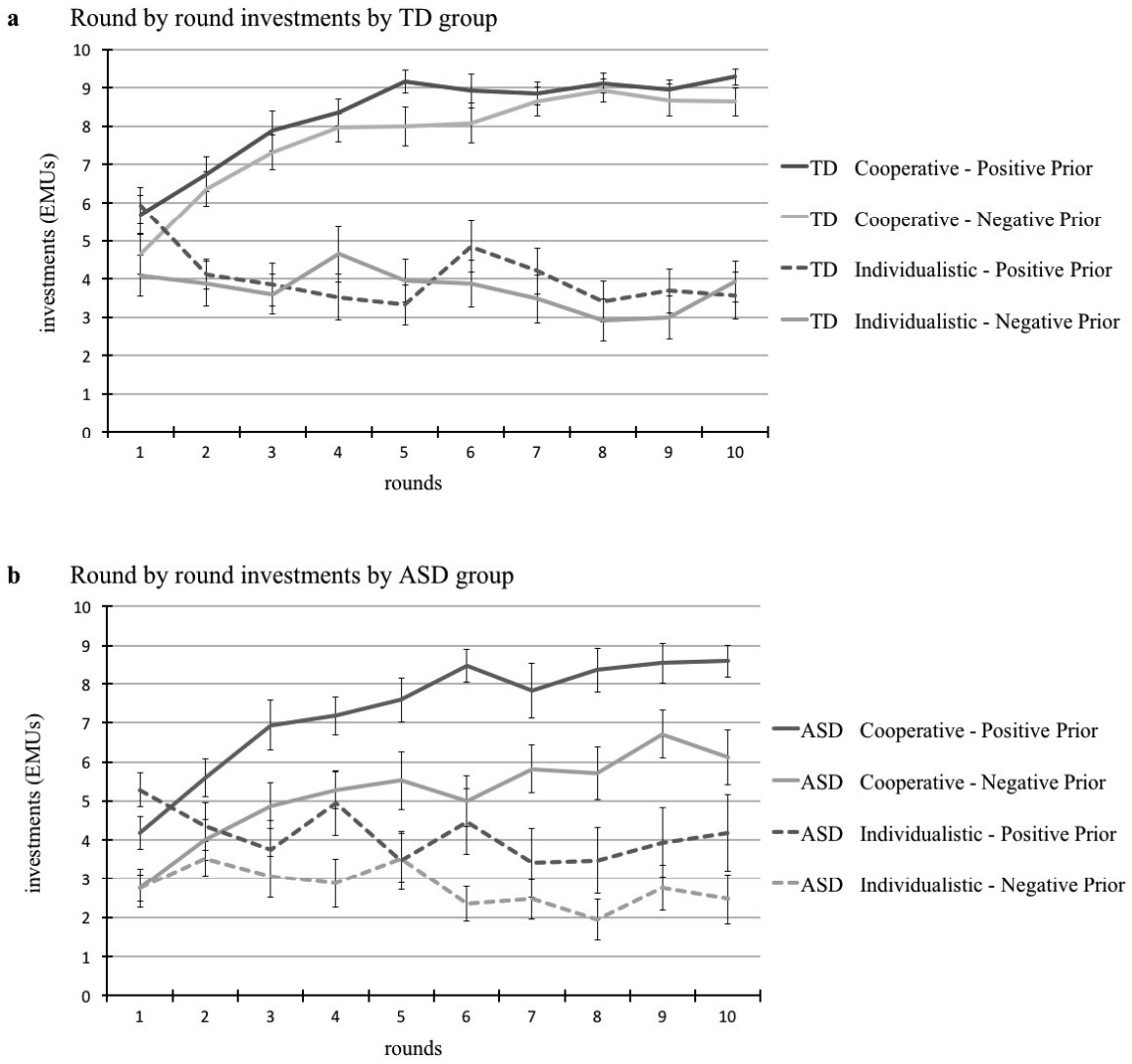


Figure S2. Judgments of trustworthiness (Upper panel) and reciprocity (returned investment) (Lower panel). Mean ± SE judgments are broken down for counterpart's Reputational prior (Positive/Negative) and Reciprocity (Cooperative/Individualistic) for TD and ASD participants.

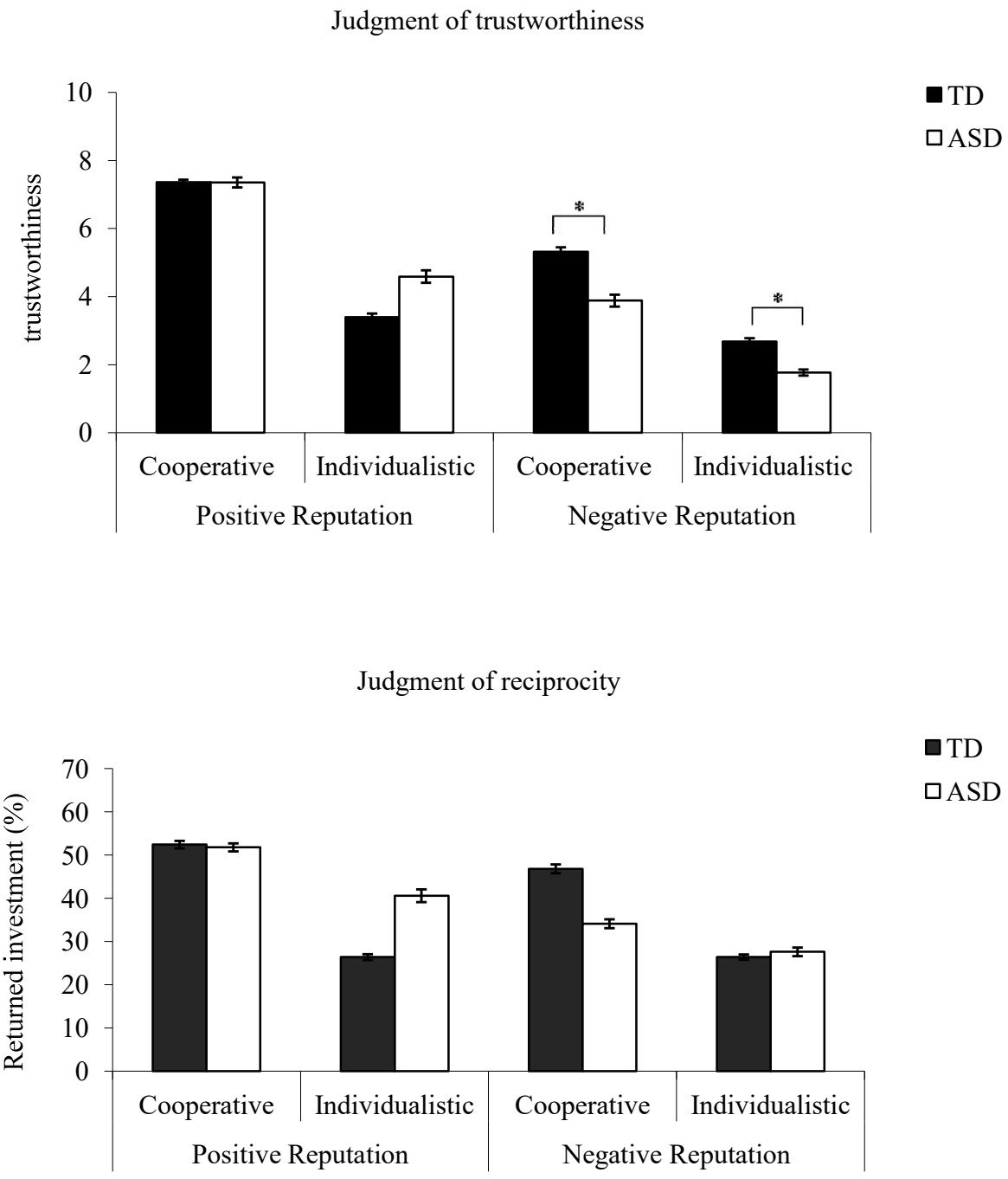


Figure S3. Regression analyses between the Autistic Spectrum Quotient (AQ) and the initial weight on prior reputation (left panel) and the reciprocity score (right panel) in all participants.