A reverse Turing-test for predicting social deficits in people with Autism

Baudouin Forgeot d'Arc[1,2], Marie Devaine[3,4,5], Jean Daunizeau[3,4,5] *

[1] Département de Psychiatrie, Université de Montréal, Canada

[2] Centre Intégré Universitaire de Santé et Services Sociaux de Nord-de-l'Île-de-Montréal

[3] Université Pierre et Marie Curie, Paris, France

[4] Institut du Cerveau et de la Moelle épinière, Paris, France

[5] INSERM UMR S975

Address for correspondence:

Jean Daunizeau

Motivation, Brain and Behaviour Group

Brain and Spine Institute

47, bvd de l'Hopital, 75013, Paris, France.

Tel: +33 1 57 27 47 19

Mail: jean.daunizeau@gmail.com

Keywords: social cognition, mentalizing, theory of mind, flexibility, computational psychiatry

Word counts: abstract = 221, main text = 5124

Number of figures = 4

Number of tables = 1

Number of supplementary texts = 1

Abstract:

Social symptoms of autism spectrum disorder (ASD) are typically viewed as consequences of an impaired Theory of Mind, i.e. the ability to understand others' covert mental states. Here, we test the assumption that such "mind blindness" may be due to the inability to exploit contextual knowledge about, e.g., the stakes of social interactions, to make sense of otherwise ambiguous cues (e.g., idiosyncratic responses to social competition). In this view, social cognition in ASD may simply reduce to non-social cognition, i.e. cognition that is not informed by the social context. We compared 24 adult participants with ASD to 24 neurotypic participants in a repeated dyadic competitive game against artificial agents with calibrated mentalizing sophistication. Critically, participants were framed to believe that they were competing against humans (social framing) or not (non-social framing), hence the "reverse Turing test". In contrast to control participants, the strategy of people with ASD is insensitive to the game's framing, i.e. they do not constrain their understanding of others' behaviour with the contextual knowledge about the game (cf. competitive social framing). They also outperform controls when playing against simple agents, but are outperformed by them against recursive algorithms framed as human opponents. Moreover, computational analyses of trial-by-trial choice sequences in the game show that individuals with ASD rely on a distinctive cognitive strategy with subnormal flexibility and mentalizing sophistication. These computational phenotypes yield 79% diagnosis classification accuracy and explain 62% of the severity of social symptoms in people with ASD.

Introduction

The diagnosis of autism spectrum disorders or ASD is based on alterations in two cognitive domains [1]: reciprocal social interaction (social deficits) and flexibility of behaviour (non-social deficits). Although one of the most heritable neurodevelopmental conditions [2,3], there is a remarkably small overlap between the genes that are associated with the distinct ASD-like behavioural traits in the general population [4]. This may explain ASD's high clinical heterogeneity [5], and eventually challenge the relevance of research agendas aimed at identifying a unique aetiology for both social and non-social deficits in ASD [6]. Somewhat paradoxically, this also highlights the importance of forging social and non-social neurocognitive theories, which can bridge the gap between biological and clinical observations in ASD [7,8]. This work is a step forward in this direction.

One of the most influential theories about ASD social deficits asserts that these are eventually due to an underlying impairment in Theory of Mind or ToM [9,10], i.e. the ability to understand others' covert mental states. This has been repeatedly evidenced in children using tests of ToM, e.g., false belief understanding [11–13], sarcasm/irony detection [14,15] or moral evaluation [16,17]. However, these tests yield quite unreliable results in older individuals. For example, high-functioning ASD adolescents and adults successfully pass false-belief [18] or facial emotion recognition [19] tests. Further refinements of the "mind blindness" theory of ASD thus suggest that although adults with ASD may succeed in simple mindreading tasks when explicitly instructed to mentalize, they lack some form of implicit and spontaneous ToM [20–22]. Alternatively, one may argue that social deficits in adults with ASD may only become apparent when the task mirrors the demands of ecological social exchanges, which critically rely on highly contextual and interactive signalling [8,23]. This is because social deficits in ASD may be less about inaccurate processing of "socially-salient" stimuli (e.g., facial expressions, speech prosody, etc...), than about the inability to exploit contextual knowledge about, e.g., the stakes of social interactions, to make sense of otherwise ambiguous cues (e.g., idiosyncratic responses to social competition). In this view, social cognition in ASD may simply reduce to non-social cognition, i.e. cognition that is not informed or constrained by the social context.

Recent advances in artificial social cognition [24–26] now enable us to mimic the way human players adapt to others in the context of simple repeated dyadic games. Rather than asking whether this type of algorithm passes the Turing test [27], we ask whether believing it is human or not changes the way people interact with it. The nature of such "reverse Turing test" will be clearer below (see Figure A0 in the Supplementary Text S1). We start with the premise that social interactions induce a specific evolutionary challenge, namely: forecasting others' overt behaviour from learned associations with predictive cues, including past behaviour [28,29]. Critical here is the notion that people may engage with others equipped with a cognitive repertoire composed of many *learning strategies*, each of which may be tied to distinct representations and/or policies. Arguably, somewhere at the end of the spectrum lie ToM-related learning strategies that derive from adopting the intentional stance [26,30–32], whose sophistication increases with the depth of recursive beliefs (as in "I believe that you believe that I believe..."). Nevertheless, learning in social contexts can take less sophisticated forms, ranging from simple heuristics, to trial-and-error learning, to cognitive shortcuts of ToM that simply care about others' overt reaction to one's own actions [33]. The ability to compliantly draw learning strategies from one's cognitive repertoire is what we term flexibility. Importantly, mathematical modelling can be used to turn a given learning strategy into a learning rule (i.e. the precise way in which agents adapt to the history of past actions and outcomes). In appropriate experimental contexts (e.g., dyadic games), this endows learning strategies with a specific behavioural signature that can be disclosed from quantitative analyses of trial-by-trial choice sequences [26]. One can then measure and compare the computational properties of people's learning repertoire, in particular: its ToM-sophistication and its flexibility. In what follows, were refer to these as "computational phenotypes" of social cognition. We then ask whether people with and without ASD differ with respect to these computational phenotypes, which we infer from observed trial-by-trial choices in dyadic interactive games against artificial players. Critically, participants are not told about the algorithmic nature of their opponents. Rather, we have them believe either that they are competing against each other (social framing) or that they are gambling like in a casino (non-social framing). We focus on peoples' ability to alter their behavioural strategy as a function of whether or not they think they are competing against someone, hence the "reverse Turing test". We predict that, in contrast to control participants, adults with ASD would not be able to constrain their understanding of others' behaviour with the contextual knowledge about the game (cf. competitive social framing), hence failing our "reverse Turing test".

Results

We asked 24 adult participants with ASD and 24 control participants to play repeated games against artificial "mentalizing" opponents, which differ in their ToM sophistication (hereafter: *k-ToM* agents, see below). In total, each participant played 4x2x2=16 games (4 opponent types, 2 framing conditions, 2 repetitions), where each game consisted in 60 successive trials. To succeed, subjects had to anticipate and predict the behaviour of their opponent, who hid himself in one out of two possible locations at each trial (see Figure 1 below).

Opponents either followed a predetermined pseudo-random sequence with a 65% bias for one hand (*RB*), or were designed to deceive the participants from learned anticipations of their behaviour (*0-ToM, 1-ToM* and *2-ToM*). The difference between *k-ToM* opponents lies in how they learn from the past history of participants' actions, where *k* refers to their calibrated ToM sophistication. In brief, *0-ToM* does not try to interpret the participants' action sequence in terms of a strategic attempt to win. Rather, it simply assumes that abrupt changes in the participants' behaviour are a priori unlikely. It thus tracks the evolving frequency of participants' actions, and chooses to hide the reward where it predicts the opponent will not seek. It is an extension of “fictitious play” learning [34], which can exploit participants' tendency to repeat their recent actions. In contrast, *1-ToM* is equipped with (limited) artificial mentalizing, i.e. it attributes simple beliefs and desires to participants. More precisely, it assumes that participants' actions originate from the strategic response of a *0-ToM* agent that attempts to predict its own actions. Note that the computational sophistication of artificial mentalizing is not trivial, since *1-ToM* has to explicitly represent and update its (recursive) belief about its opponents' beliefs. Practically speaking, *1-ToM* learning essentially consists in an on-line estimation of *0-ToM*'s parameters (e.g., learning rate and behavioural temperature) given the past history of both players' actions. This makes *1-ToM* a so-called “meta-Bayesian” agent [26,35] that can outwit strategic opponents that do not mentalize when competing in the game (such as *0-ToM*). Although *1-ToM* is mentalizing, it is not capable of dealing with other mentalizing agents. This is the critical difference between *1-ToM* and *2-ToM*. At this point, suffices to say that *2-ToM* is an artificial mentalizing agent that can learn to predict how other mentalizing agents (such as *1-ToM*) will behave.

Critically, participants were not cued about opponent conditions. This implies that they had to adapt their behaviour according to their understanding of the history of past actions and outcomes. In addition, except in the control (*RB*) condition, there is no possibility to learn the correct answer from simple reinforcement. This is because *k-ToM* artificial learners exhibit no systematic bias in their response. Further details regarding the experimental protocol as well as *k-ToM* artificial agents can be found in the Methods section below.

Figure 2 below summarizes the performance results, in terms of the net rate of correct answers in each of 4x2 conditions, for both (control and ASD) groups.

One can see that the performance patterns are markedly different between NT and ASD participants. To begin with, the performance of NT participants qualitatively reproduces previous experiments with healthy human adults [26]. In brief, in the non-social framing condition, NT participants eventually lose against artificial mentalizing agents (*1-ToM* and *2-ToM*) whereas they maintain their earnings in the social framing condition. The ASD group however, seems to show no effect of the framing manipulation, i.e. their performance pattern across opponents is the same, irrespective of whether they believe that they are competing against other people or not. Interestingly, they seem to lose against artificial mentalizing agents (as NT controls in the non-social framing condition), but they outperform NT controls against non-mentalizing learning agents (*0-ToM*). We performed a pooled variance ANOVA to assess the statistical significance of these observations. We found a significant three-way interaction between group (ASD vs NT), opponent and framing (F[3,690]=3.6, p=0.014, R^2=1.5%), a significant interaction between group and opponent

(F[3,690]=9.5, $p<10^{-4}$, R^2=4.0%) and a main effect of opponent (F[3,690]=33.7, $p<10^{-4}$, R^2=12.8%). We then looked more closely at the three-way interaction using post-hoc tests. In the NT group, there was a main effect of opponent (F=4.5, p=0.004), no main effect of framing (F=2.6, p=0.11) but a significant interaction opponent x framing (F=3.7, p=0.011). In the ASD group, there was a main effect of opponent (F=38.7, $p<10^{-4}$) but no main effect of framing (F=0.5, p=0.46) nor interaction (F=1.3, p=0.27). In other terms, only NT participants show the opponent x framing interaction. This is due to the fact that NT participants perform better in the social than in the non-social framing only against artificial mentalizing agents ($p<10^{-4}$). Now focusing on performances against artificial mentalizing agents, there was a significant interaction between group and framing (p=0.001). This is because against *1-ToM* and *2-ToM*, NT participants perform significantly better than ASD people against artificial mentalizing agents in the social framing ($p<10^{-4}$) but not in the non-social framing (p=0.65). Besides, ASD participants perform significantly better than NT participants against *0-ToM* ($p<10^{-4}$), and this effect does not depend upon the game's framing (p=0.46).

At this point, we asked whether we could classify ASD and NT participants based upon their performance patterns in the task. Averaging performances over repetitions yielded a feature space of 8 dimensions (4 opponent types, 2 framings), which was then fed to a classifier based upon logistic regression [36]. Test classification accuracy was evaluated using a simple leave-one-out cross-validation scheme. The classifier achieved 73% of correct out-of-sample classifications, which is statistically better than chance (p=0.001). This will serve as a reference point for evaluating the added-value of computational phenotypes.

One of the main differences between NT and ASD participants is thus that the latter seem to be insensitive to the framing manipulation. This interpretation, however, neglects the possibility that distinct leaning strategies may eventually yield similar performances in the game. In other terms, performance measures are potentially blind to learning strategies, which can only be inferred from analyses of trial-by-trial action sequences in the game. We thus considered a set of eight distinct learning models that constitute peoples' potential learning repertoire. Each of these learning models provides a probabilistic prediction of observed peoples' trial-by-trial choice sequences. We then performed a subject-specific bayesian model comparison of these models, and evaluated both the flexibility - $\hat{f}$ - and the ToM-sophistication - $\hat{k}$ - of peoples' learning repertoires. In what follows, we refer to these as peoples' "computational phenotypes". We refer the interested reader to the Methods section.

We first asked whether control and ASD participants would show differences in their repertoire's ToM-sophistication. Figure 3 below shows the repertoire's ToM-sophistication $\hat{k}$ averaged across repetitions, across opponent conditions and across participants, for each group and for both framing conditions.

A simple ANOVA shows no evidence for an interaction between group and framing (F[1,46]=0.6, p=0.42, R^2=1.4%), no main effect of framing (F[1,46]=1.8, p=0.18, R^2=3.8%), but a significant group effect (t[46]=1.9, p=0.03, R^2=7.3%). Post-hoc tests show that this group difference is mostly driven by the social framing condition: whereas there is no significant difference between the groups in the non-social condition (t[46]=1.1, p=0.13, R^2=2.7%), there is a strong group difference in the social framing condition (t[46]=1.9, p=0.03,

R^2=7.5%). In other words, only in the social framing do control participants exhibit higher ToM-sophistication than ASD participants.

We then investigated whether control and ASD participants show differences in their repertoire's flexibility. Figure 4 below shows the repertoire's flexibility $\hat{f}$, both across framings and across repetitions.

Here again, there is no significant interaction between group and condition type (F[1,46]=0.55, p=0.46, R^2=1.2%), but there is a significant main effect of condition type (F[1,46]=5.54, p=0.02, R^2=10.7%) and a main effect of group (t[46]=3.4, p=0.001, R^2=20.4). Post-hoc tests show that this group difference in repertoire's flexibility is strong both across framings (t[46]=3.4, p=0.001, R^2=20.7%) and across repetitions (t[46]=2.8, p=0.004, R^2=14.4%). Also, ASD participants show no difference in repertoire's flexibility when considered across framings or across conditions (p=0.26). This contrasts with control participants, who exhibit a significantly greater repertoire's flexibility across framings than across repetitions (p=0.03).

If only, this computational analysis confirms that ASD participants are relatively insensitive to the game's framing. But do these computational phenotypes provide clinically useful information, above and beyond performance scores? We address this question by assessing the accuracy of a diagnosis classifier relying upon peoples' flexibility $\hat{f}$ and ToM-sophistication $\hat{k}$. To begin with, we classified participants based upon computational phenotypes alone. The classifier only reached 67% of correct out-of-sample classifications. This is statistically significant (p=0.014), but worse than classification accuracy based upon performance patterns alone. However, when pooling performance patterns and computational phenotypes together, the classifier now yielded 79% of correct out-of-sample classifications ($p<10^{-4}$). This is important, since it means that computational phenotypes bring additional, diagnosis-relevant, information.

Finally, we asked whether we could predict, from estimated computational phenotypes, inter-individual variations in symptom severity among ASD participants. More precisely, we focused on the 'social' and ‘stereotyped behavior’ subscores of the ADOS scale, which quantify social and non-social deficits, respectively. We found that inter-individual differences in computational phenotypes predict social deficits with high accuracy (F[4,15]=6.1, p=0.004, R^2=62.1%) but not non-social deficits (F[4,15]=1.5, p=0.25, R^2=28.8%).

Discussion

In this work, we have performed a matched comparison of social and non-social behavioural adaptation in individuals with and without autism. In contrast to typically developed individuals, individuals with ASD do not change the way they play according to whether or not they believe they are competing with other humans. Typical individuals outperform people with ASD only when they think they are competing with another human being (and while playing against learning algorithms equipped with artificial mentalizing). However, people with ASD outperform typical individuals against non-mentalizing learning algorithms, irrespective of the task framing (social and non-social). The learning repertoire of individuals with ASD exhibits less flexibility and less ToM-sophistication, especially when the task is framed as a social game. Taken together, performance patterns and computational phenotypes correctly classify up to 79% of the participants according to their diagnosis. In addition, computational phenotypes predict 62% of the variance of the severity of social symptoms in ASD people.

Maybe the most striking result of our work is that people with ASD fail our "reverse Turing test", in the sense that their cognitive strategy is insensitive to the game's framing. Recall that we demonstrated this in four different ways: (i) ASD participants show no difference between performance or ToM-sophistication scores between framing conditions (cf. Fig. 2), (ii) performance variations induced by opponent types in different framing conditions are significantly correlated (see section 4 in Supplementary Text S1), (iii) model-free decompositions of their trial-by-trial choice sequences show no effect of framing (see section 5 in Supplementary Text S1), and (iv) their learning repertoire exhibits very low flexibility across framing conditions (cf. Fig. 4). Importantly, participants' debriefing showed that the framing manipulation was similarly credible in both groups of subjects (see section 2 in Supplementary Text S1). In line with social motivational theories of autism [37], one may argue that, in contrast to control participants, ASD participants may not have been interested enough to invest the cognitive effort required for improving their performance in the social framing condition. Such global motivational and/or attentional interpretations are unlikely however, because ASD participants actually outperform controls against *0-ToM* in the social framing condition. In addition, financial incentive manipulations have no effect on performance in the game (see section 3 in Supplementary Text S1). In any case, our computational results rather suggest that people with ASD rely on a very limited learning repertoire, which they deem reliable in both social and non-social contexts. It is interesting to note that the model that best captures trial-by-trial choice sequences of ASD players, in both framing conditions, is the so-called "influence learning" strategy [33]. From a computational standpoint, this model possesses broad adaptive fitness because it essentially is a generic way of learning in reactive environments (i.e. environments that react to one's actions). In other words, influence learning can be seen as an all-purpose cognitive toolkit that would be expected to perform well in a wide range of contexts, excluding challenging social interactions (cf. pattern of performances against *RB*, *0-ToM*, *1-ToM* and *2-ToM* in section 9 in the Supplementary Text S1). Obviously, our experimental claim does not go as far as to assert that the cognitive repertoire of ASD people is generally limited to influence learning. Nevertheless, it provides a remarkable example of how people in the autism spectrum may solve the unavoidable trade-off between behavioural adaptability and cognitive complexity.

This type of trade-off is arguably steepest in ecological social contexts. Not only may subtle signals (e.g., facial expressions, speech prosody, etc...) reflect profoundly different mental states, but the stakes of typical social exchanges may be dynamic, partially implicit, multiple and even conflicting (e.g., impose a deal and induce sympathy). This implies that flexibility may be a critical feature of typical social cognition [38]. In this work, we provide two independent pieces of evidence in favour of this notion. First, in the ASD group, the severity of social symptoms is partially predicted by our measure of repertoire flexibility. Second, in the NT group, flexibility (between repetitions) is significantly higher in the social than in the non-social framing condition (see section 7 in Supplementary Text S1). On the one hand, these results contribute to the ongoing debate regarding the specificity of social cognition [23,39]. In brief, social cognition is special, if only because its flexibility is enhanced (notwithstanding its sophistication). On the other hand, they also bridge a gap between social and non-social theories of ASD. Recall that the latter typically take weak context-sensitivity as a feature of ASD cognition [40,41]. Importantly, our results tend to support the view that social deficits in ASD may be but a limiting case of the failure to account for the social context when drawing inferences about others.

Methods

Ethics statement

Behavioural assessments were performed in accordance with institutional ethical guidelines, which comply with the guidelines of the declaration of Helsinki. The research protocol was approved by the Ethical Committee of the Hôpital Rivière-des-Prairies, Montréal, where the tests were performed.

Experimental methods

Participants: n=24 adults with ASD without mental nor language deficiency and n=24 NT control subjects participated in the study. All subjects were French speakers (Québec), and both groups were matched in terms of gender balance (ASD: 21 males, NT: 21 males), age (ASD: 25.5 y.o. ± 5.7; NT: 27.9 y.o. ± 8.6) and IQ (ASD: 104 ± 17; NT: 106 ± 14). ASD participants were assessed with ADOS-G and met DSM-5 criteria for ASD. NT participants went through a semi-structured interview to screen for any psychiatric treatment history, learning disorders, personal or family history (2 degrees) for mood disorder, ASD or schizophrenia. No included participant reported strong depressive symptoms (Beck depression Inventory score<20). All participants gave their informed consent, were fully debriefed at the end of the experiment, and received a financial compensation for their participation.

The behavioural task consists of a computerized game (60 trials each) with two framing conditions. In the *social* condition, the task was framed as an online competitive game with another participant. In the *non-social* condition, it was framed as a betting -casino-like-game. In fact, both games were played against four different learning algorithms with different artificial mentalizing sophistication (ranging from a random sequence with a bias to so-called *2-ToM* agents). At each trial, subjects had 1300 ms to make a binary choice (the place to hide or the slot machine to try), which was fed to the learning algorithms to compute online predictions of the participant's action at the next trial. In total, each participant performed 2×4×2=16 games (2 framings, 4 opponent types, 2 repetitions) in a pseudo-randomized order. We refer the interested reader to the Supplementary Text S1 for more details regarding the experimental protocol.

Computational modelling of learning strategies

In this section, we give a brief overview of the set of candidate learning models, with a particular emphasis on *k-ToM* models (because these are also used as on-line algorithms during the experimental phase). We will consider repeated dyadic (two-players) games, in which only two actions are available for each player (the participant and his opponent). Hereafter, the action of a given agent (resp., his opponent) is denoted by a^{self} (resp., a^{op}). A game is defined in terms of its payoff table, whose entries are the player-specific utility $U\left(a^{self},a^{op}\right)$ of any combination of players' actions at each trial. In particular, competitive social interactions simply reduce to anti-symmetric players' payoff tables (see Table 1 below).

Hider Seeker	Left	Right
Left	1,0	0,1
Right	0,1	1,0

Table 1: Competitive payoff table.

Participants play the role of the seeker, the opponent is the hider.

By convention, actions a^{op} and a^{self} take binary values encoding the first ($a=1$) and the second ($a=0$) available options. According to Bayesian decision theory, agents aim at maximising expected payoff $V=E\left[U\left(a^{self},a^{op}\right)\right]$, where the expectation is defined in relation to the agent's uncertain predictions about his opponent's next move. This implies that the form of the decision policy is the same for all agents, irrespective of their ToM sophistication. Here, we consider that choices may exhibit small deviations from the rational decision rule, i.e. we assume agents employ the so-called "softmax" probabilistic policy:

$$P\left(a^{self}=1\right)=\frac{1}{1+\exp\left(-\frac{\Delta V}{\beta}\right)} \tag{1}$$

where $P\left(a^{self}=1\right)$ is the probability that the agent chooses the action $a^{self}=1$, ΔV is the expected payoff difference (between actions $a^{self}=1$ and $a^{self}=0$), and β is the so-called behavioural "temperature" (which controls the magnitude of deviations from rationality). The sigmoidal form of Equation 1 simply says that the probability of choosing the action $a^{self}=1$ increases with the expected payoff difference ΔV, which is given by:

$$\begin{aligned}\Delta V &= p^{op}\left(U(1,1)-U(0,1)\right)+\left(1-p^{op}\right)\left(U(1,0)-U(0,0)\right)\\ &=2p^{op}-1\end{aligned} \tag{2}$$

where p^{op} is the probability that the opponent will choose the action $a^{op}=1$, and the second line derives from inserting the above payoff matrix (Table1). In brief, Equation 2 simply says that participants are rewarded for correctly guessing where their opponent is hiding.

Let us now summarize the mathematical derivation of *k-ToM* models, which essentially differ in how they estimate p^{op} from the repeated observation of their opponent's behaviour. We will see that *k* indexes a specific form of ToM sophistication, namely: the recursive depth of learners' beliefs (as in "I believe that you believe that I believe..."). Note that *k-ToM*'s learning rule can be obtained recursively, starting with *0-ToM* [29].

By convention, a *0-ToM* agent does not attribute mental states to his opponent, but rather tracks his overt behavioural tendency without mentalizing. More precisely, *0-ToM* agents simply assume that their opponents choose the action $a^{op}=1$ with probability $p^{op}=s\left(x_t\right)$, where the unknown log-odds x_t varies across trials t with a certain volatility σ^0 (and s is the sigmoid function). Observing his opponent's choices gives *0-ToM* information about the hidden state x, which can be updated trial after trial using Bayes rule, as follows:

$$\begin{aligned}\mu_t^0 &\approx \mu_{t-1}^0+\Sigma_t^0\left(a_t^{op}-s\left(\mu_{t-1}^0\right)\right)\\ \Sigma_t^0 &\approx \frac{1}{\dfrac{1}{\Sigma_{t-1}^0+\sigma^0}+s\left(\mu_{t-1}^0\right)\left(1-s\left(\mu_{t-1}^0\right)\right)}\end{aligned} \tag{3}$$

where μ_t^0 (resp. Σ_t^0) is the approximate mean (resp. variance) of *0-ToM*'s posterior distribution $p\left(x_t^0 \middle| a_{1:t}^{op}\right)$. Inserting $\hat{p}_{t+1}^{op} = E\left[s\left(x_{t+1}\right) \middle| a_{1:t}^{op}\right]$ into Equation 1 now yields *0-ToM*'s decision rule. Here, the effective learning rate is the subjective uncertainty Σ^0, which is controlled by the volatility σ^0. At the limit $\sigma^0 \rightarrow 0$, Equation 3 converges towards the (stationary) opponent's choice frequency and *0-ToM* essentially reproduce "fictitious play" strategies [34].

0-ToM's learning rule is the starting point for a *1-ToM* agent, who considers that she is facing a *0-ToM* agent. This means that *1-ToM* has to predict *0-ToM*'s next move, given his beliefs and the choices' payoffs. The issue here is that *0-ToM*'s parameters (volatility σ^0 and exploration temperature β) are unknown to *1-ToM* and have to be learned, through their non-trivial effect on *0-ToM*'s choices. At trial $t+1$, a *1-ToM* agent predicts that *0-ToM* will chose the action $a^{op}=1$ with probability $p_{t+1}^{op,0} = s \circ v^0\left(x_t^0, a_{1:t}\right)$, where the hidden states x_t^0 lumps σ^0 and β together and the mapping v^0 is derived from inserting *0*-ToM's learning rule (Equation 3) into Equations 1-2. Similarly to *0-ToM* agents, *1-ToM* assumes that the hidden states x_t^0 vary across trials with a certain volatility σ^1, which yields a meta-Bayesian learning rule similar in form to *0-ToM*'s, but relying on first-order meta-beliefs (i.e. beliefs about beliefs). In brief, *1-ToM* eventually learns how her (*0-ToM*) opponent learns about herself, and acts accordingly (cf. Equations 1-2).

1-ToM agents are well equipped to deal with situations of observational learning. However, when it comes to reciprocal social interactions, one may benefit from considering that others are also using ToM. This calls for learning strategies that rely upon higher-order meta-beliefs. By construction, *k-ToM* agents ($k \geq 2$) consider that their opponent is a κ-*ToM* agent with a lower ToM sophistication level (i.e.: $\kappa < k$). Importantly, the sophistication level κ of *k-ToM*'s opponent has to be learned, in addition to the hidden states x^{κ} that control the opponent's learning and decision making. The difficulty for a *k-ToM* agent is that she needs to consider different scenarios: each of her opponent's possible sophistication level κ yields a specific probability $p_{t+1}^{op,\kappa} = s \circ v^{\kappa}\left(x_t^{\kappa}, a_{1:t}\right)$ that she will choose action $a^{op} = 1$. The ensuing meta-Bayesian learning rule entails updating *k-ToM*'s uncertain belief about her opponent's sophistication level κ and hidden states x^{κ}:

$$
\begin{aligned}
&\lambda_t^{k,\kappa} \approx \left[\frac{\lambda_{t-1}^{k,\kappa}\ p_t^{op,\kappa}}{\sum_{\kappa'<k} \lambda_{t-1}^{k,\kappa'}\ p_t^{op,\kappa'}}\right]^{a_t^{op}} \left[\frac{\lambda_{t-1}^{k,\kappa}\left(1-p_t^{op,\kappa}\right)}{\sum_{\kappa'<k} \lambda_{t-1}^{k,\kappa'}\left(1-p_t^{op,\kappa'}\right)}\right]^{1-a_t^{op}} \\
&\mu_t^{k,\kappa} \approx \mu_{t-1}^{k,\kappa} + \lambda_t^{\kappa}\ \Sigma_t^{k,\kappa}\ W_{t-1}^{\kappa}\left(a_t^{op} - s \circ v^{\kappa}\left(\mu_{t-1}^{k,\kappa}\right)\right) \\
&\Sigma_t^{k,\kappa} \approx \left[\left(\Sigma_{t-1}^{k,\kappa} + \sigma^k\right)^{-1} + s' \circ v^{\kappa}\left(\mu_{t-1}^{k,\kappa}\right)\ \lambda_t^{\kappa}\ W_{t-1}^{\kappa\ T} W_{t-1}^{\kappa}\right]^{-1}
\end{aligned}
\tag{4}
$$

where $\lambda_t^{k,\kappa}$ is *k-ToM*'s posterior probability that her opponent is *κ-ToM*, and W^{κ} is the gradient of v^{κ} with respect to the hidden states x^{κ}. Equation 4 also captures *1*-ToM's learning rule, when setting $\lambda_t^{1,0}$ □ 1. Note that although the dimensionality of *k-ToM*'s beliefs increases with *k*, *k-ToM* models do not differ in terms of the number of their free parameters. More precisely, *k-ToM*'s learning and decision rules are entirely specified by their prior volatility σ^k and behavioural temperature β.

Formally speaking, only *k-ToM* agents with *k*≥1 are mentalizing about others' covert mental states, i.e. represent and update others' beliefs. They can do this because they adopt the intentional stance [32], i.e. they assume that p^{op} is driven by their opponent's hidden beliefs and desires. More precisely, they consider that the opponent is himself a Bayesian agent, whose decision policy $p^{op} = P\left(a^{op} = 1\right)$ is formally similar to Equation 1. This makes *k-ToM* meta-Bayesian learners [35] that relies upon recursive belief updating ("I believe that you believe that I believe..."). Critically, the recursion depth *k* induces distinct ToM sophistication levels, whose differ in terms of how they react to the history of players' actions in the game.

With the exception of *0-ToM*, we so far only described sophisticated learning models that are capable of (artificial) ToM. But clearly *0-ToM* is not the only way people may learn in social contexts without mentalizing. We thus consider below other learning strategies that may populate peoples' learning repertoire.

First, let us consider a heuristic learning model, whose sophistication somehow lies in between *0-ToM* and *1-ToM*. In brief, "influence learning" adjusts a *0-ToM*-like learning rule to account for how her own actions may influence her opponent's behaviour [33]:

$$p_{t+1}^{op} = p_t^{op} + \eta \underbrace{\left(a_t^{op} - p_t^{op}\right)}_{\text{prediction error}} + \lambda \underbrace{p_t^{op}\left(1 - p_t^{op}\right)\left(1 - 2a_t^{self} - \beta\, s^{-1}\left(p_t^{op}\right)\right)}_{\text{"influence" adjustment term}} \quad (5)$$

where η (resp. λ) controls the relative weight of its prediction error (resp. the "influence" adjustment term). Numerical simulations show that, in a competitive game setting, *Inf* wins over *0-ToM* but loses against *k-ToM* players with *k*≥1. In other terms, although it is in principle able to adapt to reactive environments, *Inf* cannot successfully compete with learners endowed with mentalizing [28].

Second, participants may learn by trial and error, eventually reinforcing the actions that led to a reward. Such learning strategy is the essence of classical conditioning, which is typically modelled using reinforcement learning or *RL* [42]. In this perspective, participants would directly learn the value of alternative actions, which bypasses Equation 2. More precisely, an *RL* agent would update the value of the chosen option in proportion to the reward prediction error, as follows:

$$\begin{cases} V_{t+1}^{i} = V_{t}^{i} + \alpha\left(R_{t} - V_{t}^{i}\right) & \text{if action } a_{t}^{self} = i \text{ was chosen} \\ V_{t+1}^{i} = V_{t}^{i} & \text{otherwise} \end{cases} \qquad (6)$$

where $R_{t} = U\left(a_{t}^{self}, a_{t}^{op}\right)$ is the last reward outcome and α is the (unknown) learning rate. At the time of choice, RL agents simply tend to pick the most valuable option (cf. Equation 1).

Third, an even simpler way of adapting one's behaviour in operant contexts such as this one is to repeat one's last choice if it was successful and alternate otherwise. This can be modeled by the following update in action values:

$$\begin{cases} V_{t+1}^{i} = R_{t} & \text{if action } a_{t}^{self} = i \text{ was chosen} \\ V_{t+1}^{i} = -R_{t} & \text{otherwise} \end{cases} \qquad (7)$$

This strategy is called win-stay/lose-switch (*WS*), and is almost identical to the above *RL* model when the learning rate is $\alpha = 1$. Despite its simplicity, *WS* can be shown to have remarkable adaptive properties [43].

Last, the agent may simply act randomly, which can be modeled by fixing the value difference to zero ($\Delta V = 0$). Although embarrassingly simple, this probabilistic policy eventually prevents one's opponent from controlling one's expected earnings. It thus minimizes the risk of being exploited at the cost of providing chance-level expected earnings. It is the so-called "Nash equilibrium" of our "hide and seek" game. Since we augment this model with a potential bias for one of the two alternative options (as all the above learning models), we refer to it as *biased Nash* or *BN*.

Empirical estimates of computational phenotypes

Our working hypothesis is that people may not always rely on the same learning model across different game sessions or conditions. Rather, they select a learning strategy from among a repertoire, whose flexibility and ToM sophistication define our computational phenotypes. The empirical estimation of these thus consists of three steps. First, we perform a statistical (Bayesian) comparison of learning models [44]. For each subject, we fit trial-by-trial actions sequences $a_{1:60}$ with each learning model ($m \in$ {*BN*, *WSLS*, *RL*, *0-ToM*, *Inf*, *1-ToM*, *2-ToM*, *3-ToM*}) using a variational-Laplace approach [45,46]. This eventually yields 8x48x4x2x2=6144 model evidences $p\left(a_{1:60} \middle| m\right)$ (8 models, 48 participants, 4 opponent conditions, 2 framing conditions, 2 repetitions).

Second, we define the *repertoire's flexibility* $\hat{f}^{(1,2)}$ (between conditions 1 and 2) in terms of the posterior probability that a given participant employs different learning strategies across two conditions:

$$\hat{f}^{(1,2)} = p\left(m^{(1)} \neq m^{(2)} \middle| a_{1:60}^{(1)}, a_{1:60}^{(2)}\right) = 1 - \sum_{m} p\left(m^{(1)} = m \middle| a_{1:60}^{(1)}\right) p\left(m^{(2)} = m \middle| a_{1:60}^{(2)}\right) \qquad (8)$$

where $m^{(1)}$ (resp. $m^{(1)}$) is the participants' learning strategy in the first (resp. second) condition, $p\left(m^{(1)} = m \middle| a_{1:60}^{(1)}\right)$ (resp. $p\left(m^{(2)} = m \middle| a_{1:60}^{(2)}\right)$) is the probability that the participant had a learning strategy m given his trial-by-trial choice sequence $a_{1:60}^{(1)}$ (resp. $a_{1:60}^{(2)}$) in condition 1 (resp. condition 2). Note that we measure the *repertoire's flexibility* $\hat{f}$ both across framings and across repetitions.

Third, we define the *repertoire's ToM-sophistication* $\hat{k}$ in terms of the expected depth of recursive belief update:

$$\hat{k} = E\left[k \middle| a_{1:60}\right] = \sum_{k} k \; p\left(k \middle| a_{1:60}\right) \tag{9}$$

where $p\left(k \middle| a_{1:60}\right) = p\left(m = "k - ToM" \middle| a_{1:60}\right)$ is the posterior probability of model *k-ToM* given the participant's trial-by-trial choice sequence $a_{1:60}$. Note that we restrict the summation in Equation 9 to *k-ToM* models, because the depth *k* of recursive beliefs is not defined for the other learning models. Note that we measure the *repertoire's ToM-sophistication* $\hat{k}$ in both framing conditions (social and non-social).

All statistical analyses were performed using the VBA toolbox [36].

Acknowledgements

Authors thank Alexandra Duquette and Patricia Jelenic for contributing with data collection, and Pr. Laurent Mottron for enabling the recruitment of ASD patients.

Financial disclosure

BFA acknowledges support from "Fondation Les Petits Trésors de l'Hôpital Rivière des Prairies" and Fonds de Recherche en Santé du Québec (FRQS). MD and JD have nothing to declare.

References

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, D.C: American Psychiatric Association; 2013.

2. Leekam S. Social cognitive impairment and autism: what are we trying to explain? Phil Trans R Soc B. 2016;371: 20150082. doi:10.1098/rstb.2015.0082

3. Waye MMY, Cheng HY. Genetics and epigenetics of autism: A Review. Psychiatry Clin Neurosci. 2017; doi:10.1111/pcn.12606

4. Ronald A, Happé F, Bolton P, Butcher LM, Price TS, Wheelwright S, et al. Genetic heterogeneity between the three components of the autism spectrum: a twin study. J Am Acad Child Adolesc Psychiatry. 2006;45: 691–699. doi:10.1097/01.chi.0000215325.13058.9d

5. Lord C, Cook EH, Leventhal BL, Amaral DG. Autism Spectrum Disorders. Neuron. 2000;28: 355–363. doi:10.1016/S0896-6273(00)00115-X

6. Happé F, Ronald A, Plomin R. Time to give up on a single explanation for autism. Nat Neurosci. 2006;9: 1218–1220. doi:10.1038/nn1770

7. Frith U. Why we need cognitive explanations of autism. Q J Exp Psychol. 2012;65: 2073–2092. doi:10.1080/17470218.2012.697178

8. Sinha P, Kjelgaard MM, Gandhi TK, Tsourides K, Cardinaux AL, Pantazis D, et al. Autism as a disorder of prediction. Proc Natl Acad Sci. 2014;111: 15220–15225. doi:10.1073/pnas.1416797111

9. Adams MP. Explaining the theory of mind deficit in autism spectrum disorder. Philos Stud. 2013;163: 233–249.

10. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a “theory of mind”? Cognition. 1985;21: 37–46.

11. Frith U, Happé F. Autism: beyond “theory of mind.” Cognition. 1994;50: 115–132.

12. Girli A, Tekin D. Investigating false belief levels of typically developed children and children with autism. Procedia - Soc Behav Sci. 2010;2: 1944–1950. doi:10.1016/j.sbspro.2010.03.261

13. Grant CM, Grayson A, Boucher J. Using Tests of False Belief with Children with Autism: How Valid and Reliable are they? Autism. 2001;5: 135–145. doi:10.1177/1362361301005002004

14. Wang AT, Lee SS, Sigman M, Dapretto M. Neural basis of irony comprehension in children with autism: the role of prosody and context. Brain J Neurol. 2006;129: 932–943. doi:10.1093/brain/awl032

15. Zalla T, Amsellem F, Chaste P, Ervas F, Leboyer M, Champagne-Lavau M. Individuals with Autism Spectrum Disorders Do Not Use Social Stereotypes in Irony Comprehension. PLOS ONE. 2014;9: e95568. doi:10.1371/journal.pone.0095568

16. Fadda R, Parisi M, Ferretti L, Saba G, Foscoliano M, Salvago A, et al. Exploring the Role of Theory of Mind in Moral Judgment: The Case of Children with Autism Spectrum Disorder. Front Psychol. 2016;7. doi:10.3389/fpsyg.2016.00523

17. Margoni F, Surian L. Mental State Understanding and Moral Judgment in Children with Autistic Spectrum Disorder. Front Psychol. 2016;7. doi:10.3389/fpsyg.2016.01478

18. Happé FGE. The Role of Age and Verbal Ability in the Theory of Mind Task Performance of Subjects with Autism. Child Dev. 1995;66: 843–855. doi:10.1111/j.1467-8624.1995.tb00909.x

19. Ponnet K, Buysse A, Roeyers H, De Clercq A. Mind-reading in young adults with ASD: does structure matter? J Autism Dev Disord. 2008;38: 905–918. doi:10.1007/s10803-007-0462-5

20. Frith CD, Frith U. Implicit and Explicit Processes in Social Cognition. Neuron. 2008;60: 503–510. doi:10.1016/j.neuron.2008.10.032

21. Schneider D, Nott ZE, Dux PE. Task instructions and implicit theory of mind. Cognition. 2014;133: 43–47. doi:10.1016/j.cognition.2014.05.016

22. Zwickel J, White SJ, Coniston D, Senju A, Frith U. Exploring the building blocks of social cognition: spontaneous agency perception and visual perspective taking in autism. Soc Cogn Affect Neurosci. 2011;6: 564–571. doi:10.1093/scan/nsq088

23. Schönherr J. What's so Special About Interaction in Social Cognition? Rev Philos Psychol. 2017;8: 181–198. doi:10.1007/s13164-016-0299-y

24. Baker CL, Saxe RR, Tenenbaum JB. Bayesian theory of mind: Modeling joint belief-desire attribution. In Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society. 2011. pp. 2469–2474.

25. Yoshida W, Dolan RJ, Friston KJ. Game theory of mind. PLoS Comput Biol. 2008;4: e1000254. doi:10.1371/journal.pcbi.1000254

26. Devaine M, Hollard G, Daunizeau J. The social Bayesian brain: does mentalizing make a difference when we learn? PLoS Comput Biol. 2014;10: e1003992. doi:10.1371/journal.pcbi.1003992

27. Turing AM. Computing Machinery and Intelligence. Parsing the Turing Test. Springer, Dordrecht; 2009. pp. 23–65. Available: https://link.springer.com/chapter/10.1007/978-1-4020-6710-5_3

28. Devaine M, San-Galli A, Trapanese C, Bardino G, Hano C, Jalme MS, et al. Reading wild minds: A computational assay of Theory of Mind sophistication across seven primate species. PLOS Comput Biol. 2017;13: e1005833. doi:10.1371/journal.pcbi.1005833

29. Devaine M, Hollard G, Daunizeau J. Theory of mind: did evolution fool us? PloS One. 2014;9: e87619. doi:10.1371/journal.pone.0087619

30. Frith CD. The role of metacognition in human social interactions. Philos Trans R Soc B Biol Sci. 2012;367: 2213–2223. doi:10.1098/rstb.2012.0123

31. de Weerd H, Verbrugge R, Verheij B. How much does it help to know what she knows you know? An agent-based simulation study. Artif Intell. 2013;199–200: 67–92. doi:10.1016/j.artint.2013.05.004

32. Dennett DC. The Intentional Stance. Reprint edition. Cambridge (USA): MIT Press; 1989.

33. Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci. 2008;105: 6741–6746.

34. Berger U. Brown's original fictitious play. J Econ Theory. 2007;135: 572–578. doi:10.1016/j.jet.2005.12.010

35. Daunizeau J, den Ouden HEM, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. Observing the observer (I): meta-bayesian models of learning and decision-making. PloS One. 2010;5: e15554. doi:10.1371/journal.pone.0015554

36. Daunizeau J, Adam V, Rigoux L. VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput Biol. 2014;10: e1003441. doi:10.1371/journal.pcbi.1003441

37. Chevallier C, Kohls G, Troiani V, Brodkin ES, Schultz RT. The Social Motivation Theory of Autism. Trends Cogn Sci. 2012;16: 231–239. doi:10.1016/j.tics.2012.02.007

38. Adolphs R. The neurobiology of social cognition. Curr Opin Neurobiol. 2001;11: 231–239. doi:10.1016/S0959-4388(00)00202-6

39. Frith CD, Frith U. Mechanisms of social cognition. Annu Rev Psychol. 2012;63: 287–313. doi:10.1146/annurev-psych-120710-100449

40. Happé F, Frith U. The weak coherence account: detail-focused cognitive style in autism spectrum disorders. J Autism Dev Disord. 2006;36: 5–25. doi:10.1007/s10803-005-0039-0

41. Haker H, Schneebeli M, Stephan KE. Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? Front Psychiatry. 2016;7: 107. doi:10.3389/fpsyt.2016.00107

42. Rescorla RA, Wagner AR. "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement,." Class Cond II Curr Res Theory. 1972;Vol. 2.

43. Nowak M, Sigmund K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature. 1993;364: 56–58. doi:10.1038/364056a0

44. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - revisited. NeuroImage. 2014;84: 971–985. doi:10.1016/j.neuroimage.2013.08.065

45. Daunizeau J, Friston KJ, Kiebel SJ. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Phys Nonlinear Phenom. 2009;238: 2089–2118. doi:10.1016/j.physd.2009.08.002

46. Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. NeuroImage. 2007;34: 220–234. doi:10.1016/j.neuroimage.2006.08.035

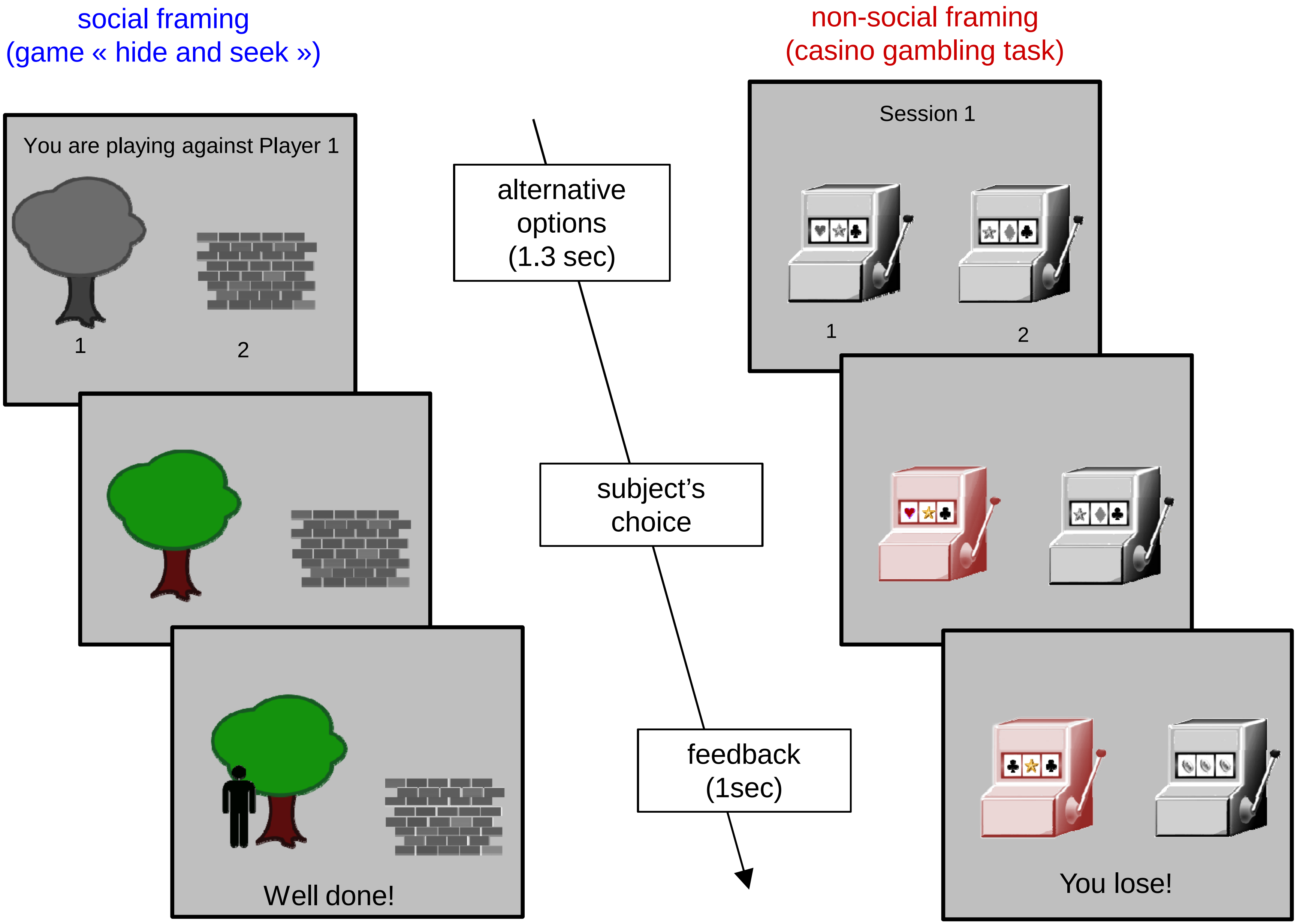
Figure legends

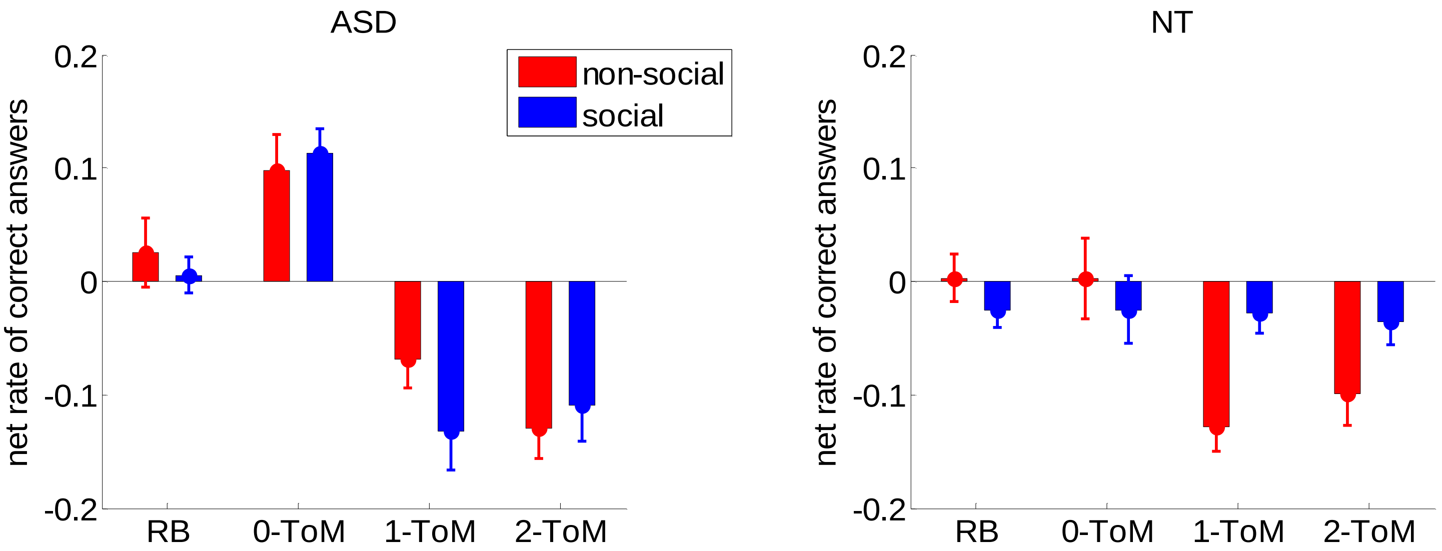
Figure 1: Experimental protocol. Left: social framing ("hide-and-seek" game). Right: non-social framing (Casino game). At each trial, participants have 1300 msec to pick one of the two options (social framing: wall or tree, non-social framing: left or right slot machine). Feedback is displayed for 1 sec; and includes the trial outcome (win or loss) and the actual winning option (social framing: character picture, non-social framing: three identical items).

Figure 2: Behavioural performance results. Group average net rate of correct answers (y-axis) against the four opponent types (x-axis) for both framing conditions (blue: social, red: non-social) in both ASD (left) and control (right) participants. Note: The net rate of correct answers is defined as (n+-n-)/(n++n-), where n+ and n- be the number of correct and incorrect responses, respectively. In this and all subsequent figures, error bars depict the standard error around the mean.

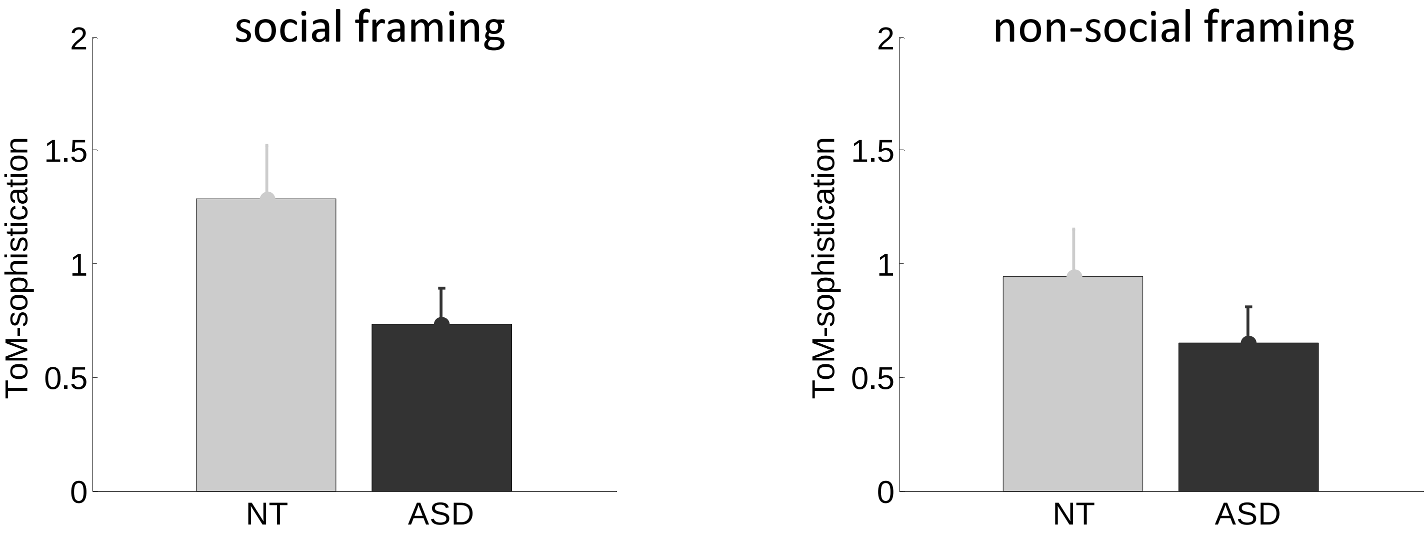
Figure 3: Model-based analysis of trial-by-trial choice sequences: ToM sophistication scores. ToM sophistication scores are shown as a function of framing conditions (left: social, right: non-social) for both control (gray) and ASD participants (back).

Figure 4: Model-based analysis of trial-by-trial choice sequences: repertoire's flexibility. The repertoire's flexibility is shown across framing conditions (left) and across repetitions (right) for both control (gray) and ASD participants (back).


social framing
(game « hide and seek »)
You are playing against Player 1
1
2
Well done!
alternative
options
(1.3 sec)
subject's
choice
feedback
(1sec)
non-social framing
(casino gambling task)
Session 1
1
2
You lose!

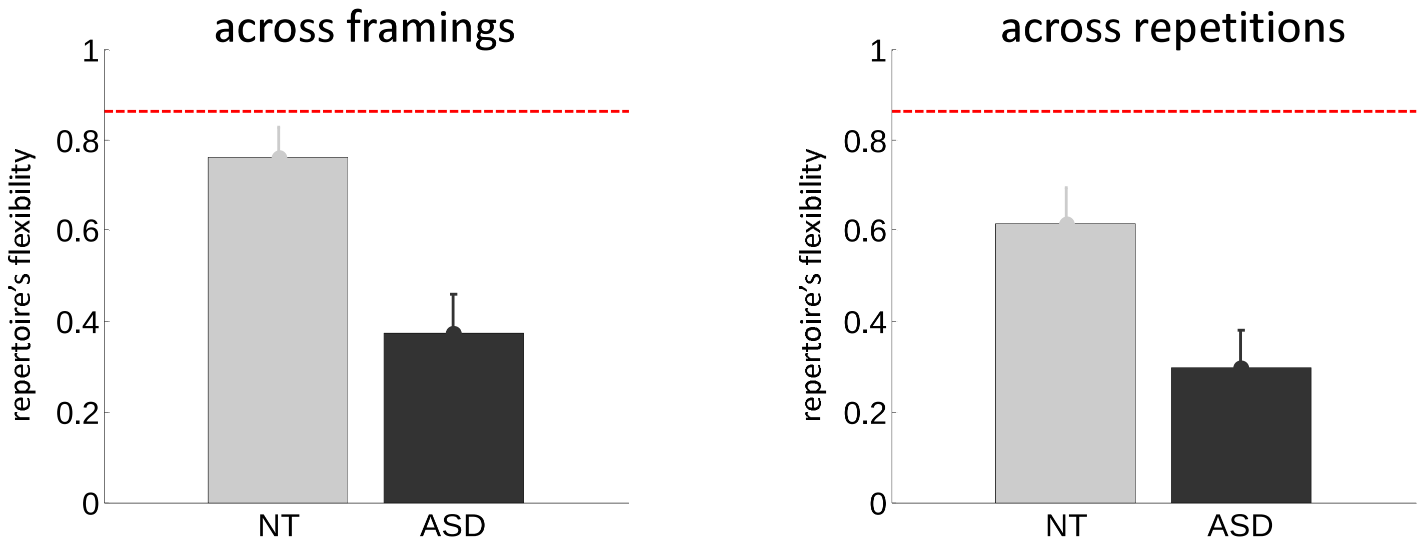


ASD
NT
non-social
social
net rate of correct answers
0.2
0.1
0
-0.1
-0.2
RB
0-ToM
1-ToM
2-ToM



social framing

non-social framing


across framings
across repetitions
repertoire's flexibility
repertoire's flexibility
0
0.2
0.4
0.6
0.8
1
NT
ASD
NT
ASD