Title: Adults with autism over-estimate the volatility of the sensory environment.

Authors:

Rebecca P. Lawson[1, 2, 3 *], Christoph Mathys[1, 4, 5, 6] & Geraint Rees[1, 2].

[1] Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London, WC1N 4BG.

[2] Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, WC1N 3AR.

[3] Department of Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB.

[4] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy

[5] Max Planck UCL Centre for Computational Psychiatry and Ageing, University College London, Russell Square House, 10-12 Russell Square, London, WC1B 5EH.

[6] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and Eidgenössische Technische Hochschule (ETH) Zurich, 8032 Zurich, Switzerland

---

Abstract (148/150)

Insistence on sameness and intolerance of change are part of the diagnostic criteria for Autism Spectrum Disorder (ASD) but there is little research addressing how people with ASD represent and respond to environmental change. Here, we find that behavioural and pupillometric measurements show adults with ASD are less surprised than neurotypical adults when expectations are violated, with reduced surprise predicting greater symptom severity. A hierarchical Bayesian model of learning suggests that in ASD a tendency to over-learn about volatility in the face of environmental change drives a corresponding reduction in learning about probabilistically aberrant events – putatively rendering them less surprising. Participant-specific modelled estimates of surprise about environmental conditions are linked to pupil size in the ASD group, suggesting heightened phasic noradrenergic responsivity in line with neural gain impairments. This study offers novel insight into the behavioural, algorithmic and physiological mechanisms that underlie responses to environmental volatility in ASD.

Keywords: ASD, autism, Bayesian, computational modelling, pupillometry; volatility, uncertainty, noradrenaline, predictive coding, perception

Introduction

When negotiating changeable real-world environments, humans face a set of learning problems involving different forms of uncertainty in which the weighting of new evidence and prior expectations need to be dynamically adjusted. Imagine opening your sock drawer and finding a pineapple inside. How surprised should you be? Under normal circumstances, you would expect to see socks - but if your four-year-old niece is visiting, you might adjust your expectations to suit a more volatile environment, lessening any surprise. However, over-estimating how volatile your bedroom is may result in compromised learning of the association between the cue (sock drawer) and outcome (socks) in the first place. In other words, aberrant representation of volatility may impair the dynamic formation of appropriate prior expectations, rendering *both* the pineapple and the socks mildly surprising. Bayesian theories of perception in ASD[1] propose that reduced weighting of prior expectations, relative to sensory inputs, leads to the perceptual atypicalities associated with the condition[2–7], but no studies to date have actually quantified the learning dynamics by which sensory expectations are formed in ASD. Here we sought to empirically address whether volatility learning is compromised in ASD[6,7].

Computationally, the amount of weight given to a surprising event is determined by its precision (inverse variance, proportional to learning rate:α) with α determining the rate of integration over past events to predict future outcomes. While computational studies of decision-making about rewards and punishments show that participants adapt their rate of learning about action-outcome contingencies in response to changes in environmental volatility [8–10], these models did not fit individual differences in volatility learning. However, knowing whether to disregard an unexpected outcome or take it seriously (i.e. whether to adopt a high or low learning rate about cue-outcome probabilities) depends on the precision of your beliefs about environmental change (i.e. whether you adopt a high or low learning rate about volatility). The recent application of hierarchical learning models has allowed the quantification of individual learning about *both* probabilistic relationships and how these relationships change over time (volatility)[11–14], but no studies have applied these models to understand learning about uncertainty in ASD.

In a state where uncertainty about one's beliefs is high (e.g. in volatile conditions), top-down prior expectations should be suppressed, relative to new bottom-up sensory evidence, in order to promote new learning about the current environmental context[15]. With their broad distribution and extensive connectivity, neuromodulatory systems are ideally placed to facilitate the widespread changes in neural gain necessary to support such a function[16]. Noradrenaline (NA), in particular, is thought to signal contextual change, leading to enhanced bottom-up, thalamocortical transmission of sensory information[17–19]. Recent neurocomputational accounts of autism have proposed that aberrant signalling of volatility could result in pathological neural gain, consistent with the cognitive and perceptual profile of autism such as enhanced perceptual functioning, sensory overload and context insensitivity[4–6,20].

Here, we tested these computational and neurobiological hypotheses by examining how adults with ASD respond to experimentally manipulated changes in their sensory expectations that independently assessed changes in the category of a stimulus, the informativeness of a cue predicting its appearance and changes in these associations over time. To do so we employed a hierarchical Bayesian model that allows us to characterise each individual participants learning "fingerprint"; specifically *simultaneous* learning about multiple different sources of environmental uncertainty [11]. We hypothesised that adults with ASD will show reduced behavioural and neurophysiological responses in contrasts of 'unexpected'

[1] Although we abide by the terminology of the diagnostic and statistical manual (DSM-5) we wish to acknowledge that the term autistic person is preferred by many people on the spectrum [1].

(UE) and 'expected' (E) trials based on the experimental 'ground truth' (e.g. reduced surprise when they ought to have been surprised). This is in line with previous studies showing reduced distinction between repeated and novel stimuli in ASD [21–23]. However, we hypothesise that computational modelling of the actual learning process for each individual will demonstrate an increased tendency to represent and respond to environmental volatility in ASD, compromising learning about probabilistic relationships in the environment. Accordingly, we hypothesise that computational metrics of prediction error, which estimate when each individual was *actually* surprised, will be reflected in pupil responses, indicating aberrant neuromodulatory function in ASD.

Results

We used a modified version of a common probabilistic associative learning task [24] to test the impact of learned expectations and sensory noise on behaviour (reaction times (RT), error rates) and indices of phasic NA function (pupillometry) [25] in adults with ASD (n=24) and age and IQ matched neurotypical adults (NT's; n=25) (Online Methods).

Participants performed binary classification of images as either faces or houses, and images had either high (H), medium (M) or no (N) noise added. A tone preceding each image was either highly, weakly or not predictive of a given outcome, and these image-tone associations changed across time (Figure 1) such that trials can be categorised as expected (E), unexpected (UE) or neutral (N). This created a ground truth structure to the environment that participants had to implicitly learn. In contrast to reinforcement learning[26,27], implicit motor learning [28] and serial reaction time [29] tasks that have examined sensitivity to probability manipulations in ASD, this task addresses perceptual associative learning and explicitly manipulates three different forms of uncertainty (categorical sensory uncertainty, probabilistic uncertainty and environmental uncertainty). When a participant receives an unexpected outcome, this may reflect a probabilistically aberrant event or it may signal that the environmental context has changed. To quantify individual learning about these different forms of uncertainty, RTs were modelled using a Bayesian belief update scheme [11] (Online Methods). The model inferred on participant's beliefs about these quantities as reflected in the sequence of cue-outcome associations each participant received and their trial-by-trial responses and response times.

Behaviour

First, we examined behavioural responses where expected (E) and unexpected (UE) trials were categorised according to the ground truth.

Reaction times

Reaction Times (RTs) were submitted to a 3x3 mixed ANOVA with within-subjects factors of expectedness (expected, neutral, unexpected), noise (high, medium, no) and a between participants factor of group (ASD, NT). There was a significant main effect of expectedness ($F(2,94)=25.48$, $P<0.001$) and noise ($F(2,94)=13.60$, $P<0.001$) indicating that RTs were slower for unexpected and high noise stimuli relative to expected and low noise stimuli. A significant main effect of group ($F(1,47)=4.83$, $P=0.03$) indicates that in general the ASD participants were slower to respond than the NT participants. Crucially, only the expectedness*group interaction was significant in this analysis ($F(2,94)=4.47$, $P=0.014$; Figure 2a). The noise*group ($F(2,94)=0.06$, $P=0.94$), noise*expectedness ($F(4,188)=0.47$, $P=0.76$) and expectedness*noise*group interactions were not significant ($F(4,188)=1.31$, $P=0.28$). This suggests that for both groups increasing sensory noise results in slower RT (Figure S1) but adults with ASD only show reduced modulation of RT as a function of learned expectations. This is consistent with reduced influence of prior information on perception and action in ASD[2], although future studies should

explore how learned expectations affect perceiving structure in true noise or 50/50 composite images where reliance on prior beliefs should be greater.

Results were unchanged when the identical analysis was carried out on log reaction times (Table S1).

Subtracting RTs to UE from those to E outcomes provides a low-level index of "surprise" which is significantly greater than zero in both groups (ASD, $t(23)=4.66$, $P<0.001$; NT, $t(24)=7.25$, $P<0.001$) but attenuated in the ASD group relative to the NT group ($t(47)=3.51$, $P=0.001$; Figure 2b). This suggests less distinction between UE and E outcomes in ASD, though this is conditioned upon adequate learning of the ground truth.

To ensure that the group difference in UE-E RT persists over and above participants mean 'baseline' RT and error rates we conducted a linear regression to predict UE-E RT with group (ASD, NT), mean RT and mean errors as predictors. This model was significant overall ($F(3,48)=5.58$, $P=0.002$) and the only significant predictor of UE-E RT difference was group ($t=-2.87$, $P=0.008$). Mean RT ($t=-1.08$. $P=0.28$) and mean errors ($t=1.06$, $P=0.29$) were not significant predictors. Importantly, this analysis demonstrates that the diminished effects of behavioural 'surprise' in ASD participants, persist even when the variance associated with general response speeds and accuracy are included in the model.

Additional analyses confirmed that this key finding of group differences in UE-E RTs remains present when control analyses account for the effects of speed-accuracy trade-off (Figure S2) and general group differences in caution of responding (Figure S3). We do, however, recognise that slower overall responses (and higher accuracy) in the ASD group may indicate a tendency to manage uncertainty with increased response thresholds, which could be tested using drift diffusion models[30,31] in future studies where error rates are higher by design.

Error rates

The same analysis as above was conducted for error rates. There was a significant main effect of expectedness ($F(1.5,70.5)=11.71$, $P<0.001$) and a significant group*expectedness interaction ($F(2,94)=6.34$, $P=0.003$) indicating that NT group made more errors on unexpected, relative to expected trials, whereas the ASD group did not (Figure 2c). The main effect of noise was not significant in this analysis ($F(1.7,78.8)=0.08$, $P=0.92$) and neither was the noise*group ($F(2,94)=0.29$, $P=0.75$), noise*expectedness ($F(4,188)=0.76$, $P=0.55$) and expectedness*noise*group interactions ($F(4,188)=1.28$, $P=0.28$).

Results were very similar when the identical analysis was carried out on log error rates (Table S1).

Subtracting % errors to UE from those to E outcomes provides a low-level index of surprise which is only significantly greater than zero in the NT group (ASD, $t(23)=1.11$, $P=0.28$; NT, $t(24)=3.65$, $P=0.001$) and attenuated in the ASD group relative to the NT group ($t(33.4)=2.83$, $P=0.007$; Figure 2d).

Relation to symptoms

To explore the relationship between behavioural surprise and ASD symptom severity we conducted a multiple linear regression predicting the UE-E RT measure with Autism Diagnostic Observation Scale (ADOS-2) communication, social reciprocal interaction scores, and also IQ as predictors. This model was significant ($F(3,23)=3.28$, $P=0.04$) and communication score was the only significant predictor ($t=-2.57$, $P=0.018$; Figure 3). IQ ($t=1.45$, $P=0.16$) and social reciprocal interaction scores ($t=0.95$, $P=0.35$) were not significant predictors.

A second regression model that also contained baseline RT as a predictor narrowly missed overall significance ($F(6,23)=2.41$, $P=0.07$), and communication scores were once again the only significant

predictor (t=-2.81, P=0.012). A third regression model, in a reduced sample size (see Online Methods), additionally included sensory sensitivity scores (ASQ) as a predictor of UE-E RT. This model was not significant (F(4,21)=1.28, P=0.32) and the only predictor approaching significance was, again, communication scores (t=-1.89, P=0.076).

Communication, as measured by the ADOS-2, weights predominantly on stereotyped and repetitive speech and conversational reciprocity which arguably necessitate reflexive behavioural responses to change. Future studies should examine the specificity of this link between general behavioural adaptations to learned expectations and communication abilities; especially as measured by different instruments.

Non-clinical replication

Finally, beyond the range of clinical phenotypes seen in people diagnosed with ASD, a wider continuum of social-communicative ability is expressed as autistic traits in the general population [32]. Encouragingly, the relationship between our behavioural measure of surprise (UE-E RT) and autistic tendency replicates in an independent non-clinical sample (N=57) of participants characterised according to expression of autistic traits (Figure S4). Not only does this bolster confidence in our clinical finding but additionally supports generalisation of this result to the broader autism spectrum in the wider population.

Responses to different stimulus types

Control analyses indicated that there were no group differences in response time or accuracy across the face and house stimuli (Figure S5).

Computational Modelling

To investigate learning about distinct kinds of uncertainty in ASD we adopted a participant-specific Bayesian model to track the role of uncertainty on behaviour (log RTs). In the Hierarchical Gaussian Filter (HGF) [11] beliefs are updated via prediction errors, with dynamic learning rates (α) at each level (i) influenced by uncertainty about the accuracy of current beliefs and environmental volatility (Figure 4a). In the version of the HGF used here (introduced in [33]) learning occurs simultaneously on three coupled levels of an uncertainty hierarchy (x_1, x_2, and x_3). Level 1 (x_1) addresses uncertainty about outcomes (face or house), level 2 (x_2) addresses uncertainty about probabilities (cue-outcome contingencies) and level 3 (x_3) addresses uncertainty about environmental change (volatility). See Online Methods and Table S2 for more model details.

Model validation

First, to ensure that the HGF performs well as a model to describe the behaviour of our participants, we fit three alternative learning models to the data and compared them to the HGF with random-effects Bayesian model selection (BMS). Relative to simple Reinforcement Learning (RL) models with fixed (RW) and dynamic (SK1) learning rates and a 2-level HGF in which volatility updates are eliminated, the three-level HGF was the best model for explaining the data by a considerable margin (see Online Methods, Figure S6). Importantly, BMS evaluates the relative plausibility of competing models in terms of their log-evidences which quantifies the trade-off between accuracy (fit) and complexity of a model and accounts for the fact that the observed variability in log-model evidences could be due to chance. Additionally, the 3-level HGF model simulations captured the principal group differences in the behavioural effect of expectation on RT (see Online Methods and Figure S7).

Predicting diagnostic status

A summary of group differences in each of the estimated model parameters is presented in Figure S8.

A binary logistic regression model predicting group status (ASD=1, NT=0), with all eight model parameters as predictors was significant (X^2= 26.83, P=.001) and prediction success overall was 81.6% (76% ASD, 88% NT), with cross-validated prediction success of 68%. The Wald statistic demonstrated that outcome uncertainty (β_2, P=0.043), phasic volatility (β_4, P=0.006) tonic volatility at the 3rd level (ω_3, P=0.024) and baseline log RT (β_0, P=0.007) made a significant contribution to prediction (Figure 4b).

Interestingly these significant predictors predominantly pertain to the third level of the HGF, i.e. learning about environmental volatility. ω_3 can be understood as capturing 'metavolatility' (i.e., the tonic volatility of the phasic volatility, with higher values in the ASD group implying a belief in a world where instability itself is instable (Figure S8). β_4 captures the modulation of log RT in response to phasic volatility, here smaller (negative) values in the NT group (Figure S8) implies that when beliefs about volatility increase, participants become more attentive and respond faster. In contrast, the larger (positive) values in the ASD group (Figure S8) indicate that increased beliefs about volatility leads to slower reaction time. In general, these findings point towards problems representing and responding to environmental change in ASD, specifically, an increased tendency to expect the unexpected.

Learning rate update in response to volatility

From the HGF we can infer the trial-wise rate of learning about two different sources of information: probabilistic outcomes (α_2) and also the rate of learning about environmental change (α_3). When the environment is volatile people should give more weight to recent sensory outcomes in building expectations about what they will see next (e.g. adopt a high α), in contrast they should give information from the distant past more weight when the environment is stable (e.g. adopt a low α) [8,9]. To test the hypothesis that individuals with ASD have problems flexibly updating their rate of learning (c.f. precision weighting) in response to environmental change we examined the change (Δ) in α_2 (probability) and α_3 (environment) when switching from stable (highlighted in violet on Figure 1) to volatile (highlighted in green on Figure 1) periods of the task. We compared the change in α_2 and α_3 between these two periods, across the groups. This analysis revealed a trend towards a main effect of group (F(1,47)=0.26, P=0.061), a significant main effect of α type (F(1,47)=6.07, P=0.017) and crucially an α type* group interaction (F(1,47)=9.80), P=0.003). Follow up independent-samples t-tests revealed that the ASD group did not update α_2 as much as NT adults (t(47)=-2.37, P=0.02) whereas they updated α_3 more than NT adults(t(47)=3.16, P=0.03; Figure 4c).

Average learning rates

To examine learning overall, we calculated average values for α_2 and α_3 for each participant. This analysis revealed no main effect of α type (F(1,47)=2.61, P=0.11), no main effect of group (F(1,47)=2.01, P=0.16, and no group* α type interaction (F(1,47)=2.54, P=0.12), suggesting that, in general, both groups were able to learn the this task equally well.

Predicting learning rate update from tonic volatility

Finally, since the HGF estimation does not fit α_2 and α_3 directly, we ran two linear regression models predicting $\Delta\alpha_2$ and $\Delta\alpha_3$ respectively to determine which of the two ω parameters drive these learning

rate differences. In each case the model was significant ($\Delta\alpha_2$: F(2,48)=68.94, P<0.001, R^2 =0.75; $\Delta\alpha_3$: F(2,48)=102.53, P<0.001, R^2 =0.82). The results indicate that $\Delta\alpha_2$ is positively predicted by ω_2 (t=2.72, P=0.009), suggesting that a tendency to believe *cue-outcome associations* are unstable is associated with a larger update in α_2 when switching from stable to volatile phases of the task. Interestingly, ω_3 negatively predicts $\Delta\alpha_2$ (t=-8.89, P>0.001), indicating that a tendency to believe *instability* is unstable drives a smaller update in α_2 in response to volatility. This fits with our finding that the ASD participants, who tend towards a smaller $\Delta\alpha_2$ (Figure 4c), show reduced behavioural 'surprise' (Figure 2b) and also larger 'metavolitility' estimates (Figure 4b). For the model predicting $\Delta\alpha_3$ both of the ω parameters were significant positive predictors (ω_2, t=7.88, P<0.001; ω_3, t=14.24, P<0.001). For the ASD participants, who show larger $\Delta\alpha_3$, this is consistent with a tendency towards beliefs in the instability of both cue-outcome associations and instability itself.

Pupillometry

Predictive coding descriptions of ASD depart from normative Bayesian theories in that they make explicit predictions about the neurobiological basis of precision; namely, the action of neuromodulators such as noradrenaline (NA) which control the gain on cortical responses (prediction errors) [3,4,6]. Raised NA signalling in ASD is suggested by elevated blood plasma levels [34] and increased arousal; i.e. heart rate variability [35], but no studies have examined phasic NA function in the context of learning about uncertainty in ASD. To do so we acquired concurrent pupillometry in a reduced subset of the sample (Online Methods). Phasic pupil response to surprising outcomes (ground truth contrast of UE-E trials) revealed a significant increase in pupil size in NT's (Figure 5a), consistent with many previous studies [25]. Convergent with the behavioural data (Figure 2b & d), the ASD group did not show this distinction between UE and E trials (Figure 5a). This pattern mirrors previous findings in the domains of electrophysiology (reduced mismatch negativity in ASD/smaller P300 [36,37] and BOLD imaging (reduced fMRI repetition suppression in ASD [21,23]) but now in the novel domain of pupillometry. However, this notion of surprise is conditioned upon adequate learning of the ground truth, and our computational analysis indicates that ASD and NT participants show a dissociation in how they estimate volatility and adapt their learning rates in response to the changeability of the environment (Figure 4b & c).

Computational pupillometry analysis

The HGF provides a nuanced and individualised trial-by-trial "learning fingerprint" and better characterises when participants were *actually* surprised as a function their personal learning process, namely 'high-level' precision-weighted prediction errors (PE's) about changes in cue-outcome contingency (ε_3). Here the learning rate α_3 depends on the precision weight on the PE; that is proportional to the update of environmental volatility (See Online Methods). As such ε_3 is a model-based measure of high-level surprise that is formally related to the dynamic learning about environmental change where we see group differences (Figure 4c). Applying multiple regression across every trial and every time point in the pupil time trace, we found a sustained positive relationship between pupil size and precision-weighted PE's (ε_3) in the ASD participants (Figure 5b), which significantly differed from the NT group and zero. Furthermore, these strong effects persisted when controlling for the UE-E ground truth contrast, trial-wise differences in fixation compliance, mean RT and outcome image type (face/house) all of which were included in the model as covariates (Online Methods). Additional analyses revealed that the volatility learning rate (α_3), and the probability learning rate (α_2) are not encoded in the pupil

response in either group (Figure S9). See Figure S10 for analysis examining the relationship between precision-weighted PEs in specific (volatile/stable) phases of the experiment.

Pupillometry control analyses

Given the possibility that people with ASD might look at face stimuli differently to people without ASD [38], stimulus duration was purposefully short (150ms) to prevent saccades. Nonetheless, to ensure that there was no difference between the groups in fixation compliance across the stimulus types (faces, houses) we conducted a repeated measures ANOVA on the mean absolute deviation (MAD) from fixation (in degrees of visual angle) across outcome image type (face, house) with a between subjects factor of group. All main effects and interactions in this analysis were non-significant (Table S1).

To examine group differences in tonic pupil size (thought to be a measure of general noradrenergic tone [39]) we compared the average of the z-scored pupil measurement across all trials with an independent samples t-test. This demonstrated no group differences in tonic pupil size in this sample ($t(23)=.36$, $P=0.72$).

Finally, control analyses revealed that there were no group differences in fixation compliance across conditions (Figure S11) or the relationship between pupil size and simple behaviour such as trial-wise RT (Figure S12). Raw pupil traces for each group can be seen in the Supplementary Results (Figure S13).

Discussion

In this study, behavioural (RT/error rates) and pupillometric results based on the experimental ground truth converge on the finding of reduced distinction between unexpected and expected outcomes in ASD (Figure 2, Figure 5a) which is consistent with many previous studies across a range of methods reporting reduced 'surprise' in ASD [21,23,36,37]. Crucially, however, this low-level notion of 'expected' and 'unexpected' trials assumes optimal or at least adequate learning of the ground truth. However, the statistical regularities that underlie our sensory world and shape our expectations are changeable and humans have to learn about different kinds of uncertainty in order to adaptively adjust the weighting of prior expectations and sensory inputs. Knowing whether to disregard an unexpected outcome or take it seriously (i.e. whether to adopt a high or low learning rate about cue-outcome probabilities (α_2)) depends on the precision of your beliefs about environmental change (i.e. whether you adopt a high or low learning rate about volatility (α_3)). The present data go beyond previous work by specifically demonstrating that over-estimating volatility in the face of environmental change – at the expense of learning about probabilistically aberrant events - characterises the behaviour of adults with ASD during perceptual inference (Figure 4c).

Furthermore, computational-pupillometry analyses indicate heightened encoding of trial-wise surprise in phasic noradrenergic responses in ASD (Figure 5b). Thus, under the assumption that pupil size is an index of NA release from the locus coruleus (LC) [40] these results are suggestive of raised phasic neuromodulatory signalling in ASD. NA is believed to change cortical gain in response to surprise, specifically; salient events indicating that global context has changed cf. "unexpected uncertainty" [15,25]. Here our computational-pupillometric analysis indicates a strong relationship between noradrenergic responsivity and precision-weighted prediction errors in ASD participants. Consistent with our other model-based results (Figure 4b & c), these findings again support over-reactivity to environmental change in ASD, but now in the context of physiological measures that index phasic neuromodulatory function. If the NA system is signalling more high-level surprise in ASD then this may imply atypical

cortical gain during sensory processing, resulting in a state where one is disproportionately receptive to sensory inputs. Aberrant phasic NA (c.f precision on prediction errors[4,6]) may alter the signal-to-noise ratio of cortical responses [41,42], broaden the tuning functions of sensory responses and, subsequently improve discrimination behaviour[43]. Thus aberrant NA function may offer a neurobiological perspective on the profile of sensory processing strengths and weakness experienced by people on the spectrum.

Importantly, these findings provide preliminary empirical evidence for neurobiologically informed Bayesian accounts of autism that emphasise aberrant representation of volatility and, consequently, *inappropriate* setting of gain (precision) on cortical responses (prediction errors) [4,6] under conditions of uncertainty. A recent pharmacological study employing the HGF, indicates that noradrenaline antagonism selectively impairs volatility learning[13], which accords with the raised pupillometric response to surprise about volatility reported in the adults with ASD here (Figure 5b). We hypothesise that the noradrenergic LC and its coupling with the anterior cingulate cortex (ACC)[8,10] ratifies estimated volatility, and that the downstream gain modulations act on the precision of cortical responses that are behaviourally relevant to the task at hand. Atypical social prediction error processing in the gyral surface of the ACC(g) has recently been shown in autism[44], but whether differences in processing in the ACC region extend to non-social tasks with explicit computational models and manipulations of volatility remains to be seen**.** Carefully designed neuroimaging and neuropharmacology studies will be necessary to link these (presumed) noradrenergic effects, and the mathematical anatomy of uncertainty[11], to hierarchical processing in the brain [12]. Additionally, although we emphasise the role of noradrenaline here, we also acknowledge the likely importance of its direct precursor, dopamine, and the complementary relationship with acetylcholine and the signalling of expected uncertainty [15]. All three of these neuromodulators are likely candidates in the neurobiological mechanisms underlying responses to environmental change in ASD.

From a Bayesian perspective, the simplest way in which persistent overweighting of all sensory inputs (relative to prior expectations) might occur would be a generally larger outcome α – reflecting *chronic* and inflexible overweighting of recent, relative to past, sensory history. Such an explanation is implied by conservative interpretations of non-hierarchical Bayesian accounts of ASD [2] and predictive processing accounts that emphasise “uniform” inflexibly high precision in sensory processing [3]. However, by logical extension, beyond a single ambiguous sensory event, all Bayesian accounts imply that dynamic learning about structural regularities (i.e. the formation of priors) is likely impaired in ASD [2–7]. Under the aberrant precision account of ASD it is problems with high-level learning about volatility, and the ratified neuromodulatory changes, that is hypothesised to underlie the difficulties faced by people on the spectrum [4–6]. It is for this reason that we designed a task to capture behaviour under orthogonal manipulations of expectations and sensory noise and built a model equipped with the ability to infer on learning about volatility.

The recent proposal that non-hierarchical, reinforcement learning (RL) models can speak directly to predictive coding theories of ASD[3] is perhaps too simplistic, not least because predictive coding is largely regarded as a neural process theory and therefore behavioural or modelling results in the absence of a proxy for brain function can only speak to such an account but are not truly able to test it**.** Motivated by these claims, a relatively recent study found no differences in learning rate malleability in autistic children during a reward learning task modelled with a delta learning rule [45]. Notably, however, there were no group differences in simple behaviour reported on this task. Here we made a specific behavioural prediction based on previous research (reduced ‘surprise’ in ASD) and a specific computational prediction to explain this behaviour (aberrant learning about volatility). We therefore designed a model complex enough to address simultaneous hierarchical learning. Using Bayesian model comparison we have shown (Figure S6) that the simplest learning model (similar to the one employed

previously[45]) performs poorest in explaining participant behaviour. Nonetheless, if it is the case that learning in the face of volatility is compromised in adults with autism (as reported here), but not children (as indicated previously[45]), this would be a significant discovery. It will be important for future studies to employ the same computational models and behavioural paradigms in adults and children to inform our understanding of how autism affects cognition across the lifespan, especially as some features of the disorder can become more severe with age [46]

Conclusion

The surprise experienced on finding a pineapple in your sock drawer depends on the strength of your prior expectation to see socks. The results of this study imply that adults with autism show a tendency to over-estimate the volatility of the sensory environment, at the expense of learning to build stable expectations that lead to adaptive surprise. In other words, adults with autism may be mildly surprised by the pineapple *and* the socks. Heightened encoding of prediction errors in pupil size measures is consistent with neurobiologically focused Bayesian accounts of autism, that emphasise neural gain impairments due to aberrant neuromodulatory funcion[4–6] . The distinct but complementary results provided by the ground truth and computational levels of analysis in our study underwrite the utility of computational approaches in better understanding neuro-developmental and psychiatric conditions with the aim of influencing clinical practice [47–49]. This study offers novel insight into the behavioural, algorithmic and physiological mechanisms that underlie learning about, and responses to, environmental change in ASD. Novel patterns of learning may emerge when the environment is more or less changeable, when expectations are formed explicitly or, or when outcomes are not incidental but instead tied to reward and/or social evaluation [10,50]. It will be important for future research to address these domains and test volatility learning in larger cohorts and people of different intellectual abilities across the autistic spectrum.

References

1. Kenny, L. *et al.* Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism* 1362361315588200 (2015).
2. Pellicano, E. & Burr, D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn. Sci.* **16,** 504–510 (2012).
3. Van de Cruys, S. *et al.* Precise minds in uncertain worlds: Predictive coding in autism. *Psychol. Rev.* **121,** 649 (2014).
4. Lawson, R. P., Rees, G. & Friston, K. J. An aberrant precision account of autism. *Front. Hum. Neurosci.* **8,** (2014).
5. Friston, K. J., Lawson, R. & Frith, C. D. On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends Cogn Sci* **17,** (2013).
6. Lawson, R. P., Friston, K. J. & Rees, G. A more precise look at context in autism. *Proc. Natl. Acad. Sci.* **112,** E5226–E5226 (2015).
7. Palmer, C. J., Lawson, R. P. & Hohwy, J. Bayesian Approaches to Autism: Towards Volatility, Action, and Behavior. *Psychol. Bull.* **143,** 521–542 (2017).
8. Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10,** 1214–1221 (2007).
9. Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X. & Bishop, S. J. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat. Neurosci.* (2015).
10. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative learning of social value. *Nature* **456,** 245–249 (2008).
11. Mathys, C. D. *et al.* Uncertainty in perception and the Hierarchical Gaussian Filter. *Front. Hum. Neurosci.* **8,** 825 (2014).
12. Iglesias, S. *et al.* Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* **80,** 519–530 (2013).
13. Marshall, L. *et al.* Pharmacological Fingerprints of Contextual Uncertainty. *PLOS Biol.* **14,** e1002575 (2016).

14. de Berker, A. O. *et al.* Computations of uncertainty mediate acute stress responses in humans. *Nat Commun* **7,** (2016).
15. Yu, A. & Dayan, P. Expected and unexpected uncertainty: ACh and NE in the neocortex. *Adv. Neural Inf. Process. Syst.* 173–180 (2003).
16. Berridge, C. W. & Waterhouse, B. D. The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain Res. Rev.* **42,** 33–84 (2003).
17. Hasselmo, M. E. & McGaughy, J. High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation. *Prog. Brain Res.* **145,** 207–231 (2004).
18. Kobayashi, M. *et al.* Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. *Eur. J. Neurosci.* **12,** 264–272 (2000).
19. Shepard, K. N., Liles, L. C., Weinshenker, D. & Liu, R. C. Norepinephrine Is Necessary for Experience-Dependent Plasticity in the Developing Mouse Auditory Cortex. *J. Neurosci.* **35,** 2432 (2015).
20. Lawson, R. P., Aylward, J., White, S. & Rees, G. A striking reduction of simple loudness adaptation in autism. *Sci. Rep.* **5,** 16157 (2015).
21. Ewbank, M. P. *et al.* Repetition Suppression in Ventral Visual Cortex Is Diminished as a Function of Increasing Autistic Traits. *Cereb. Cortex* bhu149 (2014).
22. Gomot, M. *et al.* Candidate electrophysiological endophenotypes of hyper-reactivity to change in autism. *J. Autism Dev. Disord.* **41,** 705–714 (2011).
23. Kleinhans, N. *et al.* Reduced neural habituation in the amygdala and social impairments in autism spectrum disorders. *Am. J. Psychiatry* **166,** 467–475 (2009).
24. den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J. & Stephan, K. E. Striatal prediction error modulates cortical coupling. *J. Neurosci.* **30,** 3210–3219 (2010).
25. Yu, A. J. Change is in the eye of the beholder. *Nat Neurosci* **15,** 933–935 (2012).
26. Solomon, M., Smith, A. C., Frank, M. J., Ly, S. & Carter, C. S. Probabilistic reinforcement learning in adults with autism spectrum disorders. *Autism Res.* **4,** 109–120 (2011).

27. South, M., Newton, T. & Chamberlain, P. D. Delayed reversal learning and association with repetitive behavior in autism spectrum disorders. *Autism Res.* **5,** 398–406 (2012).
28. Nemeth, D. *et al.* Learning in autism: implicitly superb. *PloS One* **5,** e11731 (2010).
29. Brown, J., Aczel, B., Jiménez, L., Kaufman, S. B. & Grant, K. P. Intact implicit learning in autism spectrum conditions. *Q. J. Exp. Psychol.* **63,** 1789–1812 (2010).
30. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: current issues and history. *Trends Cogn. Sci.* **20,** 260–281 (2016).
31. Wiecki, T. V., Sofer, I. & Frank, M. J. HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Front. Neuroinformatics* **7,** 14 (2013).
32. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *J. Autism Dev. Disord.* **31,** 5–17 (2001).
33. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5,** (2011).
34. Lam, K. S., Aman, M. G. & Arnold, L. E. Neurochemical correlates of autistic disorder: a review of the literature. *Res. Dev. Disabil.* **27,** 254–289 (2006).
35. Daluwatte, C. *et al.* Atypical pupillary light reflex and heart rate variability in children with autism spectrum disorder. *J. Autism Dev. Disord.* **43,** 1910–1925 (2013).
36. Courchesne, E., Kilman, B. A., Galambos, R. & Lincoln, A. J. Autism: processing of novel auditory information assessed by event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.* **59,** 238–248 (1984).
37. Jeste, S. S. *et al.* Electrophysiological evidence of heterogeneity in visual statistical learning in young children with ASD. *Dev. Sci.* **18,** 90–105 (2015).
38. Falck-Ytter, T. & von Hofsten, C. How special is social looking in ASD: a review. *Prog. Brain Res.* **189,** 209–222 (2010).

39. Aston-Jones, G. & Cohen, J. D. AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* **28,** 403–450 (2005).
40. Costa, V. D. & Rudebeck, P. H. More than Meets the Eye: the Relationship between Pupil Size and Locus Coeruleus Activity. *Neuron* **89,** 8–10 (2016).
41. Hasselmo, M. E., Linster, C., Patil, M., Ma, D. & Cekic, M. Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *J. Neurophysiol.* **77,** 3326–3339 (1997).
42. Hirata, A., Aguilar, J. & Castro-Alamancos, M. A. Noradrenergic activation amplifies bottom-up and top-down signal-to-noise ratios in sensory thalamus. *J. Neurosci.* **26,** 4426–4436 (2006).
43. Martins, A. R. O. & Froemke, R. C. Coordinated forms of noradrenergic plasticity in the locus coeruleus and primary auditory cortex. *Nat Neurosci* **18,** 1483–1492 (2015).
44. Balsters, J. H. *et al.* Disrupted prediction errors index social deficits in autism spectrum disorder. *Brain* **140,** 235–246 (2017).
45. Manning, C., Kilner, J. M., Neil, L., Karaminis, T. & Pellicano, E. Children on the autism spectrum update their behaviour in response to a volatile environment. *Dev. Sci.* **doi:10.1111/desc.12435,** (2016).
46. Happé, F. G. *et al.* Demographic and Cognitive Profile of Individuals Seeking a Diagnosis of Autism Spectrum Disorder in Adulthood. *J. Autism Dev. Disord.* **46,** 3469–3480 (2016).
47. Haker, H., Schneebeli, M. & Stephan, K. E. Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? *Front. Psychiatry* **7,** 107 (2016).
48. Corlett, P. R. & Fletcher, P. C. Computational psychiatry: a Rosetta Stone linking the brain to mental illness. *Lancet Psychiatry* **1,** 399–402 (2014).
49. Teufel, C. & Fletcher, P. C. The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain* (2016). doi:10.1093/brain/aww209
50. Sevgi, M., Diaconescu, A. O., Tittgemeyer, M. & Schilbach, L. Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. *Biol. Psychiatry* (2015).

Online Methods

Participants

29 adults with autism spectrum disorder (ASD) and 26 neurotypical volunteers (NTs) came to the UCL Institute of Cognitive Neuroscience as part of a testing day involving different researchers. Two adults with ASD did not complete this test owing to time constraints or an inability to tolerate the sounds and/or focus adequately on the test. Following data examination, participants with more than 20% overall errors or mean reaction times (RT's) > 2 standard deviations from their respective group mean RT were excluded from subsequent analysis to ensure the validity of the Bayesian modelling. This left 24 participants in the ASD group (18 males; mean age: 35.5, age range: 20-61) and 25 in the NT group (16 males; mean age: 36, age range: 19-62). The ASD and NT groups were matched on age ($t(47)=0.54$, $P=0.87$).

ASD participants had previously been diagnosed by an independent clinician, according to the DSM-IV[51] or ICD-10 criteria[52] [19 Asperger Syndrome, 3 Autistic Disorder, 1 High Functioning Autism, 1 Atypical Autism]. The Wechsler Adult Intelligence Scale (WAIS 3rd edition UK) had previously been administered to assess IQ [53] and participants were matched on full-scale scores (ASD mean: 117; range: 80-142; NT mean: 120, range: 99-145; $t(47)=-0.93$, $P=0.36$) The Autism Diagnostic Observation Schedule (2nd edition) [54] assessment was completed by a qualified administrator to assess symptom severity in the ASD participants. Mean ADOS total score was 9.9 (range 4-19). The mean scores for the communication and reciprocal social interaction sub scores were 3.3 (range: 0-7) and 6.6 (range 4-12), respectively.

An additional 57 NTs were studied as part of a replication of our key behavioural result (25 male, 32 female; mean age: 27.1, age range: 19–50) and additionally completed the Autism Spectrum Quotient (AQ) questionnaire; a 50-item self-report measure of autistic traits [32,55]. Mean AQ score was 18.43 (median: 17, range: 5-45). All participants had normal or corrected to normal vision and gave written informed consent. We performed a median split on the data such that participants were divided into high AQ (n=26) and low AQ (n=31) groups. AQ score was significantly higher in the in the high AQ group (mean=27, SD=6.4, range=18-45), relative to the low AQ group (mean=11.5, SD=3.4, range=5-17; $t(55) = 11.28$, $P<0.001$). The distribution of scores the low AQ group falls almost exclusively below the mean range of neurotypical scores reported in a recent meta-analysis of 73 studies administering the AQ [56]. Importantly there is considerable overlap between the scores in the high AQ group and the range reported, on average, in those with a diagnosis of ASD [56] – even though these participants do not present with any clinical need.

No randomisation was used to assign subjects or conditions. All participants provided written informed consent and were compensated financially for their time and travel expenses. The study was approved by the UCL Graduate School Ethics Committee (4357/001)

Stimuli

Auditory cues were either 330 Hz or 660 Hz pure tones generated in MATLAB R2012b (Mathworks, Ltd) and presented using the Cogent toolbox (http://www.vislab.ucl.ac.uk/cogent_graphics.php), via Sennheiser HD 201 headphones. Outcome images were either faces or houses. These stimuli were grayscale and comprised 6 different face identities (3 male, 3 female) or 6 different images of houses, masked by an ellipse and luminance matched using the SHINE toolbox [57]. Outcome images either had medium or high Gaussian noise added, with a mean of zero and a variance of 0.05 and 0.1 respectively,

using the image processing toolbox (MATLAB R2012b). Examples of no, medium and high noise face stimuli can be seen in Figure 1.

Procedure

Participants sat on a chair with their head in a chin rest at a viewing distance of 80 cm. An example trial can be seen in Figure 1. Each trial began with the 300 ms presentation of a pure tone that was either high or low in pitch. After 200 ms either a face or a house image was presented, that lasted for 150ms to prevent saccades. The participant's task was simply to respond to the image, indicating whether it was a face or a house (via left/right button press) and to be 'as fast and accurate as possible, trying to respond on every trial'. A variable response time of 1500-1800 ms followed the image; such that trials lasted 1950ms – 2250 ms. Participants were instructed that the tone preceding each image was probabilistically associated with the likelihood of seeing a face or house and that these probabilities would change across time. The probabilistic associations between the tones and the outcomes were either highly (p=0.84), weakly (p=0.16) or non-predictive (p=0.5) and changed pseudo randomly across trials in blocks of either 12, 36 or 72 trials (Figure 1). All participants completed 456 trials over 8 mini-blocks with optional periods of rest between.

To ensure that participants' responses were not biased by learned expectations about the relative frequencies of the visual stimuli, the task was designed such that the marginal probabilities of faces and houses were identical at any point in time (Figure 1) and each block contained equal numbers of randomly intermixed high and low tone trials. As employed in previous studies [12,24], this design ensured that the a priori probability of a face (or house) occurring was always 50% on any given trial, before the tone is presented. Thus, any expectations about the visual stimulus could depend only on the preceding tone. Additionally, and unique to this study, equal numbers of high, medium and no noise stimuli appeared in each of 12, 36 or 72 blocks of trials and across each cue type.

Data collection and analysis were not performed blind to the conditions of the experiment.

Pupillometry

To ensure fixation and measure neuromodulatory responses, gaze direction and pupil size were measured with an infrared eye tracker (Cambridge Research Systems) tracking the left eye at 200 Hz. Calibration of the eye tracker was unsuccessful in all participants wearing glasses and the eye tracker suffered a fatal technical failure before testing was completed, therefore eye tracking data are only available for 14 NT's and 11 ASD's.

Hierarchical Gaussian Filter

In the version of the HGF used here (introduced in [33]) learning occurs simultaneously on three coupled levels of an uncertainty hierarchy. The first level of the HGF (x_1) constitutes the outcome on any given trial (e.g. face or house), the second level (x_2) represents the probabilistic associations between the tones and the outcomes (e.g. the probability of seeing a house given that you've just heard a high tone), and the third level (x_3) quantifies the volatility of the probabilities (e.g. the changeability of the environment). On each trial, the model provides an estimate for each level, before the outcome is seen and the estimate updated accordingly. Predictions at each level are represented by a Gaussian distribution, described by its mean,$\hat{\mu}_i$ and variance, $\hat{\sigma}_i$. The variance $\hat{\sigma}_i$ represents the uncertainty of the estimate at each level. Updates of beliefs at each level occur via prediction errors that propagate upwards and are precision-weighted by the ratio of the uncertainty of the level that generated them to the uncertainty of the level being updated. The manipulation of perceptual noise (e.g. no, med, high) is captured trial-by-trial as a fixed parameter representing the variance of the noise on the inputs.

For each participant the perceptual model parameters, ω_2 and ω_3, learning rates, α_2 and α_3, and response model parameters ($\beta_{0,\ldots,4}$) were estimated from the trial wise log RT measures using variational Bayes as implemented in the HGF toolbox (http://www.translationalneuromodeling.org/tapas/). The ω's are the tonic log-volatilities at their respective levels, according to the generative model

$$x_1^{(t)} \sim \text{Bernoulli}\left(s\left(x_2^{(t)}\right)\right),$$

$$x_2^{(t)} \sim \mathcal{N}\left(x_2^{(t-1)}, \exp\left(x_3^{(t)} + \omega_2\right)\right),$$

$$x_3^{(t)} \sim \mathcal{N}\left(x_3^{(t-1)}, \exp(\omega_3)\right),$$

with $s\left(x_2^{(t)}\right) \coloneqq 1/\left(1 + \exp\left(-x_2^{(t)}\right)\right)$. This means that they determine the basic step size of the random walks in x_2 and x_3, without taking into account phasic modulation by higher levels of the hierarchy. The learning rate α_2 represents, trial-by-trial, the size of the update in μ_2 (i.e., the mean of the belief on x_2) relative to the size of the prediction error δ_1, as expressed in terms of the update in predicted outcome probabilities $\hat{\mu}_1$:

$$\alpha_2^{(t)} \coloneqq \frac{\hat{\mu}_1^{(t)} - \hat{\mu}_1^{(t-1)}}{\delta_1^{(t)}},$$

where $\hat{\mu}_1^{(t)} \coloneqq s\left(\mu_2^{(t-1)}\right)$. The learning rate α_3 is the equivalent quantity with respect to the size of the update in μ_3:

$$\alpha_3^{(t)} \coloneqq \frac{\mu_3^{(t)} - \mu_3^{(t-1)}}{\delta_2^{(t)}}.$$

Furthermore, α_3 is proportional to the precision-weight on the prediction error $\varepsilon_3^{(t)}$:

$$\alpha_3^{(t)} \propto \frac{\hat{\pi}_2^{(t)}}{\pi_3^{(t)}}.$$

Where $\pi_3^{(t)}$ is the posterior precision (inverse variance) at the third level and $\hat{\pi}_2^{(t)}$ is the precision (inverse variance) of the prediction at the second level. Accordingly $\varepsilon_3^{(t)}$, is the precision weighted-prediction error at the second level, which serves to update the estimate of log-volatility:

$$\varepsilon_3^{(t)} := \frac{\hat{\pi}_2^{(t)}}{\pi_3^{(t)}}\, \delta_2^{(t)}$$

More details can be found in the supplementary material to [6].

The βs are the coefficients of the response model, which describes how beliefs (i.e., the probability distributions on x_i as represented by their sufficient statistics μ_i and σ_i) are translated into log-reaction times. This is a straightforward linear model:

$$\log RT^{(t)} \sim \mathcal{N}\left(\beta_0 + \beta_1 \cdot \text{surprise}^{(t)} + \beta_2 \cdot \text{unc1}^{(t)} + \beta_3 \cdot \text{unc2}^{(t)} + \beta_4 \cdot \text{volatility}^{(t)}, \zeta\right),$$

with the independent variables defined as follows:

$$\text{surprise}^{(t)} \coloneqq \begin{cases} -\log_2\left(\hat{\mu}_1^{(t)}\right) & \text{if } u^{(t)} = 1 \\ -\log_2\left(1 - \hat{\mu}_1^{(t)}\right) & \text{if } u^{(t)} = 0 \end{cases}$$

$$\text{unc1}^{(t)} \coloneqq \hat{\sigma}_1^{(t)}$$

$$\text{unc2}^{(t)} \coloneqq s\left(\mu_2^{(t)}\right)\left(1 - s\left(\mu_2^{(t)}\right)\right)\sigma_2^{(t)}$$

$$\text{volatility}^{(t)} \coloneqq s\left(\mu_2^{(t)}\right)\left(1 - s\left(\mu_2^{(t)}\right)\right)\exp\left(\mu_3^{(t)}\right)$$

Here, $u^{(t)}$ is the outcome; $u^{(t)} = 1$ when the high tone cue is followed by a face or the low tone cue is followed by a house while $u^{(t)} = 0$ in the converse cases. Since $\hat{\mu}_1^{(t)}$ is the predicted probability of $u^{(t)} = 1$ (and $1 - \hat{\mu}_1^{(t)}$ of $u^{(t)} = 0$) this means that the first independent variable is the Shannon surprise associated with the outcome. Uncertainty at the outcome level (i.e, the first) is the variance $\hat{\sigma}_1^{(t)} = \hat{\mu}_1^{(t)}\left(1 - \hat{\mu}_1^{(t)}\right)$ of the Bernoulli distribution over predicted outcomes. This is the irreducible uncertainty associated with any kind of probabilistic prediction, referred to as risk in the economics literature. Uncertainty at the second level is the posterior variance σ_2 of the belief on x_2, expressed at the outcome level (hence the multiplication with the derivative of s taken at the current mean μ_2 of the belief on x_2; for details on this transformation to the first level, see the Supplementary Material to [12]. This is informational uncertainty, so called because it quantifies the lack of information about the quantity (here x_2) governing outcome probabilities. Volatility is the exponential of the phasic log-volatility μ_3, also expressed at the outcome level.

The choice of these models was hypothesis-driven. The reason for choosing the HGF as the learning model was twofold. First, because it reflects the hierarchical nature of changing environments in that it allows for volatility that is itself volatile, it allowed us to test the hypothesis that ASD participants differ from NT in the way they deal with a hierarchy of uncertainties and specifically address learning about volatility. The response model was chosen on the basis that log-RT's were approximately Gaussian

distributed and that a linear model allowed for the straightforward identification of the effects of all hypothesized modulating factors.

There were several reasons that we chose to fit reaction time over trial-wise errors. First, reaction times are a sensitive behavioural response measure which can take a range of values across trials, from fast to slow, and empirically have been shown to vary with the uncertainty of participant responses in both detection and discrimination experiments [58]. Second, reaction times were used previously where Bayesian learning models were applied to behavioural tasks very similar to ours, so modelling RT here increases comparability across studies[24,59]. Third, error rates are very low in this study (~3% overall), and any logistic model attempting to explain such a small incidence of states coded as 1 (relative to 0) would require more trials than we have in this study (increasing as a function of the explanatory variables in the model [60]). Fourth, (and most pragmatically) some participants didn't make any errors at all so modelling RT maximises the number of participants included in the analysis. Finally, the group*probability interaction for percent errors is not significant in our high and low AQ replication (supplemental results), and so in modelling RT we are modelling the most the effect most comparable across both experiments in this manuscript.

Sample Size

In our NT participants we sought a conceptual replication of Den Ouden et al. [24], albeit with a modified design. We calculated a minimum sample size *a priori* on the basis of the low probability minus high probability RT difference that they report (32 ms) and an assumed variance (actual SD not reported) of the same. This analysis indicated that we would need a minimum of 14 participants to achieve 95% power to detect a similar ($\alpha = 0.05$; 2-tailed) effect in the NT group. Given that initial effect sizes are often inflated [61] and that we sought power to detect a difference between two groups, we doubled this estimate and aimed to test ~28 participants in each group with some attrition expected.

As there is no prior precedent for detecting between-groups differences using this specific task, we additionally assessed the required sample size to detect a medium effect size for a between-subjects ANOVA with three levels and a between-subjects factor of group. This indicated that a total sample size of 48 participants would be necessary to have at least 90% power to detect an F-test effect size of 0.25.

For the pupil size regression, where it was not possible to calculate power a priori, the sample sizes and effect sizes (β's) reported for this particular analysis are in line with previous studies employing the same methods[9]. Post-hoc power calculations indicate that with 11 ASD participants included in the actual analysis, we had 86% power to detect the mean positive β (slope=0.72) that we observed in these participants ($\alpha = 0.05$; 2-tailed).

Statistics

Behavioural data

All statistical analysis of behavioral data were performed in MATLAB (Mathworks, Ltd.) and PASW Statistics 22 (SPSS inc./IBM). For the analysis of RT's, too fast and too slow (<100 or >1000ms) responses were excluded and, including missing responses, there was no significant difference between the groups in the overall percentage of missing data (1.9% ASD, 2.3% NT, $t(47)=0.45$, $p=0.65$). To maximize trial numbers per condition we collapsed across face/house trials and, for correct trials only, submitted RTs to a mixed ANOVA with within subject factors of expectedness (unexpected(UE),

neutral(N) and expected(E)) and stimulus noise (high(H), medium(M) and no (N)), and a between-subjects factor of group. We also quantified a behavioral measure of surprise, defined as the difference in RT between UE and E outcomes based on the ground truth, and compared this measure between the groups using independent-samples t-tests. An equivalent analysis was conducted for error rates and log transforms of both these measures. % errors were calculated for each condition separately. Data distributions were assumed to be normal but this was not formally tested. Where assumptions of heterogeneity of covariance were violated, degrees of freedom were corrected using the Greenhouse–Geisser approach.

Eye tracking data

All statistical analyses of eye tracking data were performed in MATLAB (Mathworks, Ltd.) Only trials in which 80% or more samples were successfully tracked were included in the analysis. There was no significant difference in the mean number of included trials between the groups (mean good trials ASD=298; NT=261; t(23)=0.803, *P*=0.43). For pupil data blinks were treated with linear interpolation and the resulting pupil traces were low-pass filtered and smoothed following the conventions outlined in [62]. To explore phasic pupil responses for correct trials traces were baseline corrected to the average response during the 500 ms preceding the outcome image. Tonic pupil responses were determined as the average of the z-scored pupil measurement across all trials. Z-scoring accounts for individual differences in baseline pupil size and has been employed previously in the literature [63,64]. Mean absolute deviation (MAD) from fixation (in degrees of visual angle) across groups and conditions was used to assess fixation compliance on each trial [65].

Regression analyses were conducted to examine the effects of surprise based on the ground truth and volatility surprise (ε_3: trial-wise precision-weighted prediction errors) on pupil dilation following outcome presentation. A similar approach has previously been employed in recent studies examining the relationship between pupil dilation and computational model parameters that vary across trials [9]. The post-outcome period for each trial was sampled using 370 5ms bins. Regression analyses were conducted for each individual time bin, with HGF estimates of precision-weighted prediction errors ($\boldsymbol{\varepsilon_3}$) and the 'ground truth' contrast of unexpected (1) minus expected (-1) included as regressors of interest; trial type (0=face, 1=house), fixation compliance (MAD), and RT for each trial were entered as control regressors. The resultant timeseries of beta-weights (e.g. multiple regression conducted at every time point) provide estimates of the effects of 'ground truth' surprise and volatility surprise on pupil dilation across all trials.

At the group level we then conducted t-tests for the positive or negative effect of the regressors of interest, and the independent-samples difference between groups, corrected for multiple comparisons using a cluster-based permutation approach at 2000 permutations (FWE alpha=0.05, 2-tailed) [66]. This allowed us to assess when our surprise metrics were significantly encoded in the pupil timeseries. This approach protects against false positives across correlated measurements (i.e. maximizes temporal sensitivity).

Learning rate data

To test the hypothesis that individuals with ASD have problems with flexibly updating their rate of learning (precision weighting) in response to environmental change we examined the change (Δ) in α_2 (probability) and α_3 (environment) when switching from stable to volatile periods of the task. For this we used the dynamic α trajectories estimated on the basis of all trials, but specifically interrogated a period

of 72 trials (highlighted in green on Figure 1) in which the probabilistic association between tones and outcomes remained fixed, followed by a period of 72 trials (highlighted in violet on Figure 1) in which the outcome probabilities switched three times. We compared the change in average α_2 and α_3 between these two periods, across the groups. Previous studies have examined how learning about how reward probability changes in response to volatility in typical volunteers [8,10] and also in aversive environments [9]. In these studies the participant's responses are fit with a simple delta learning rule, (cf. Rescorla-Wagner [67]) separately in volatile and stable task phases which annuls the elegance of the generative model approach by imposing knowledge of the task structure. In contrast, we fit subject RTs across all 456 trials using the HGF and the two learning rates (α_2 & α_3) dynamically vary as a function of each participant's inferred beliefs about cue-outcome informativeness and changes in these associations over time. While simpler models approximate participant's outcome probability estimates, assuming they are an 'ideal' Bayesian observer, the HGF addresses what kind of Bayesian observer each participant actually is, making is a more sensitive means of capturing individual differences in learning about uncertainty (see 'HGF model validation' section below and Figure 3C (inset) for comparisons between the HGF and simpler reinforcement learning models).

Bayesian Model Selection

To disambiguate alternative explanations (models) for the participants' behaviour, we used Bayesian model selection (BMS). BMS evaluates the relative plausibility of competing models in terms of their log-evidences which quantifies the trade-off between accuracy (fit) and complexity of a model. Here, we used a recently updated random effects BMS method to account for potential interindividual variability in our sample quantifying the *protected* posterior probabilities of four competing models [68]. Protected exceedance probabilities quantify the probability that any one model is more frequent than the others and also accounts for the fact that the observed variability in (log-) model evidences could be due to chance [68].

Regression analyses

To examine the relationship between the primary behavioural measure of surprise (UE-E RT) and autism symptom severity we conducted a multiple linear regression with ADOS-2 scores for communication and reciprocal social interaction, and IQ as predictors. A secondary regression model was also conducted in which an sensory sensitivity scores (as measured by the adult sensory questionnaire [69]) was also included as a predictor. Sensory scores were only available for 21/24 ASD participants, therefore this analysis was conducted on a reduced sample size. In response to a reviewer request we also conducted a third regression to predict UE-E RT that included baseline RT as an additional predictor. As both communication scores (r=-.421, P=0.04) and mean RT (r=-.341, P=0.017) correlate with UE-E RT difference in the ASD participants, we created centered versions of these variables and their interaction effect in the regression model.

To assess the validity of the HGF model parameters in predicting group status (ASD=1, NT=0) we conducted binary logistic regression (method=enter) using SPSS. The predictor variables in this analysis were the eight free parameters estimated by the HGF, namely the five response model betas ($\beta_{0\ldots4}$) plus decision noise (ζ) and the two omega parameters from the perceptual model(ω_2, ω_3). Additionally, we recreated this analysis in R and used the cv.glm function in the boot package to perform leave-one out cross validation.

Please see the *Life Sciences Reporting Summary* for more details about the methods in this manuscript.

Online Methods References

51. American Psychiatric Association. *Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000)*. (American Psychiatric Association, 2000).
52. Organization, W. H. International classification of diseases (ICD). (2012).
53. Wechsler, D. & Hsiao-pin, C. *WASI-II: Wechsler abbreviated scale of intelligence*. (Pearson, 2011).
54. Lord, C. *et al.* The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30,** 205–223 (2000).
55. Woodbury-Smith, M. R., Robinson, J., Wheelwright, S. & Baron-Cohen, S. Screening adults for Asperger syndrome using the AQ: A preliminary study of its diagnostic validity in clinical practice. *J. Autism Dev. Disord.* **35,** 331–335 (2005).
56. Ruzich, E. *et al.* Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Mol. Autism* **6,** 2 (2015).
57. Willenbockel, V. *et al.* Controlling low-level image properties: The SHINE toolbox. *Behav. Res. Methods* **42,** 671–684 (2010).
58. Bonnet, C., Ars, J. F. & Ferrer, S. E. Reaction times as a measure of uncertainty. *Psicothema* **20,** 43–48 (2008).
59. Vossel, S. *et al.* Spatial Attention, Precision, and Bayesian Inference: A Study of Saccadic Response Speed. *Cereb. Cortex* (2013). doi:10.1093/cercor/bhs418
60. Hosmer Jr, D. W. & Lemeshow, S. *Applied logistic regression*. (John Wiley & Sons, 2004).
61. Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19,** 640–648 (2008).
62. Jackson, I. & Sirois, S. Infant cognition: going full factorial with pupil dilation. *Dev. Sci.* **12,** 670–679 (2009).
63. Kang, O. & Wheatley, T. Pupil dilation patterns reflect the contents of consciousness. *Conscious. Cogn.* **35,** 128–135 (2015).
64. Knapen, T. *et al.* Cognitive and Ocular Factors Jointly Determine Pupil Responses under Equiluminance. *PloS One* **11,** e0155574 (2016).

65. Schwarzkopf, D. S., Anderson, E. J., de Haas, B., White, S. J. & Rees, G. Larger Extrastriate Population Receptive Fields in Autism Spectrum Disorders. *J. Neurosci.* **34,** 2713–2724 (2014).
66. Groppe, D. M., Urbach, T. P. & Kutas, M. Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology* **48,** 1711–1725 (2011).
67. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement. in *Classical Conditioning II: Current Research and Theory* 64–99 (Appleton-Century-Crofts, 1972).
68. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage* **84,** 971–985 (2014).
69. Kinnealey, M., Oliver, B. & Wilbarger, P. A phenomenological study of sensory defensiveness in adults. *Am. J. Occup. Ther.* **49,** 444–451 (1995).

Acknowledgments

This work was supported by the Wellcome Trust Senior Clinical Research Fellowship (100227: G.R). We thank all the participants who gave up their time to take part in this research and Michael Browning for helpful comments on an earlier poster presentation of this data.

Contributions

R.P.L conceived the study, collected and analysed the data. R.P.L. and C.M modelled the data. R.P.L, C.M and G.R wrote the manuscript.

Data availability

The data that support the findings of this study are available on reasonable request from the corresponding author in accordance with local ethics rules.

Code availability

We used the HGF toolbox (http://www.translationalneuromodeling.org/hgf-toolbox-v3-0/) for modelling learning. For pupillometry analysis we used modified versions of the freely available pre-processing code available here (http://www.tqmp.org/RegularArticles/vol10-2/p179/index.html) and analysis was conducted using code after the Mass Univariate Toolbox (http://openwetware.org/wiki/Mass_Univariate_ERP_Toolbox). Code to control the low level image properties of the stimuli used in this experiment is available here: http://www.mapageweb.umontreal.ca/gosselif/SHINE/.

Competing Financial Interests

The authors declare no competing financial interests.

Figure Legends

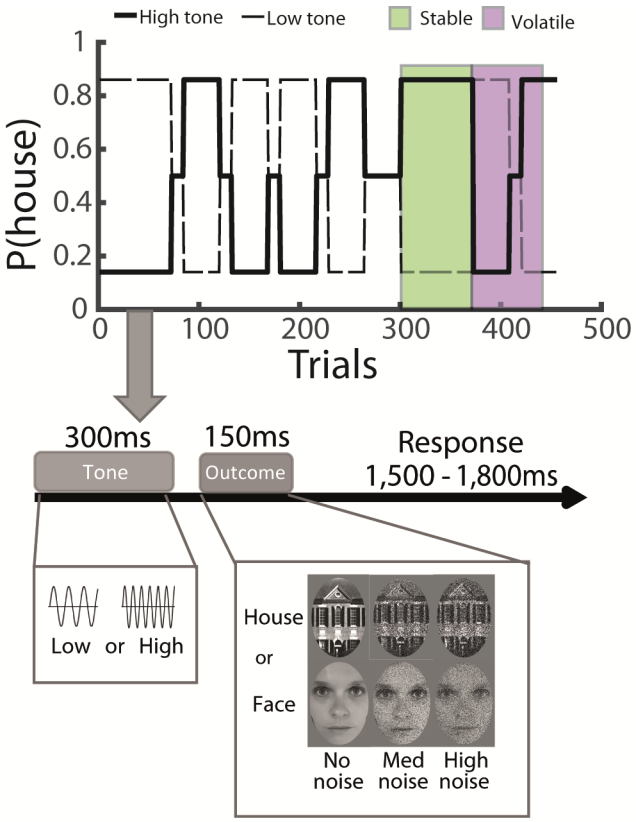
Figure 1: **Task structure** - Schematic of the task showing the volatile environmental structure (top) e.g. the probability of seeing a house (given the preceding high or low tone) across trials. Green area shows a "stable" period of 72 trials when the probabilities remained fixed and the violet area shows a "volatile" period of 72 trials where the outcome probabilities switched three times. A single trial is also seen (bottom) showing example stimuli.

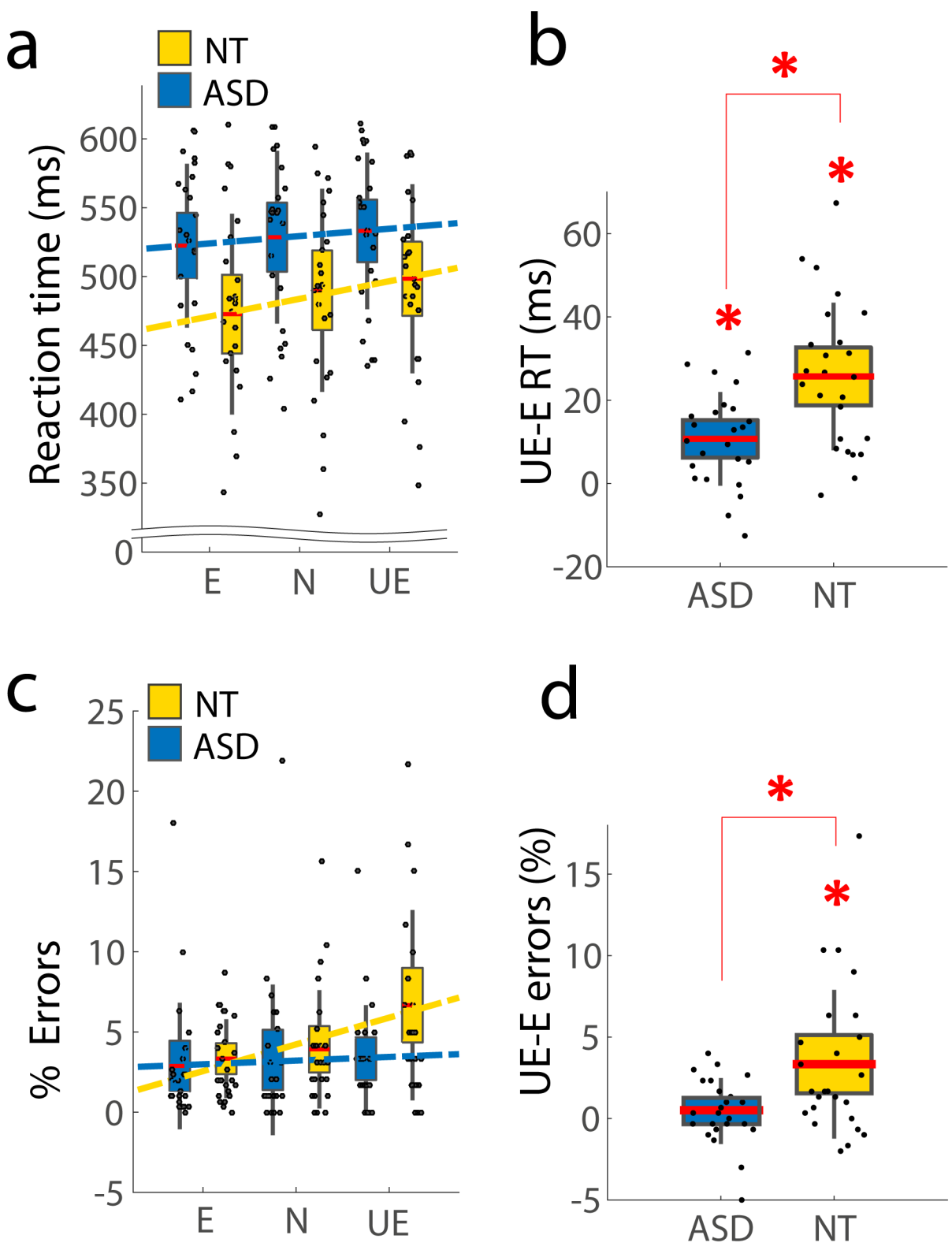
Figure 2: **Behavioural results based on the ground truth** – (a, b) reduced modulation of reaction time and (c, d) error rates as a function of expectation in ASD adults (n=24) relative to NT (n=25). Dotted lines show linear fits. ASD, autism spectrum disorder. NT, neurotypical. RT, reaction time. UE, unexpected. N, non-predictive. E, expected. Data points represent individual participants, red line shows the mean, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean for each condition and group. au, arbitrary units. Star indicates significance at P<0.05

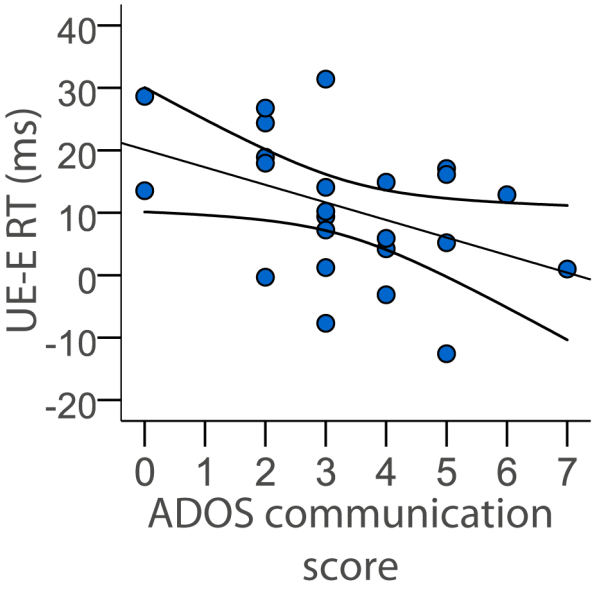
Figure 3: Relationship between behavioural surprise and symptoms - The magnitude of the reaction unexpected (UE) minus expected (E) reaction time (RT) effect is predicted by communication symptoms in the ASD group (Pearson correlation: r=-0.421, P=0.04) Data points show individual participants, n=24. ADOS, autism diagnostic observation schedule.

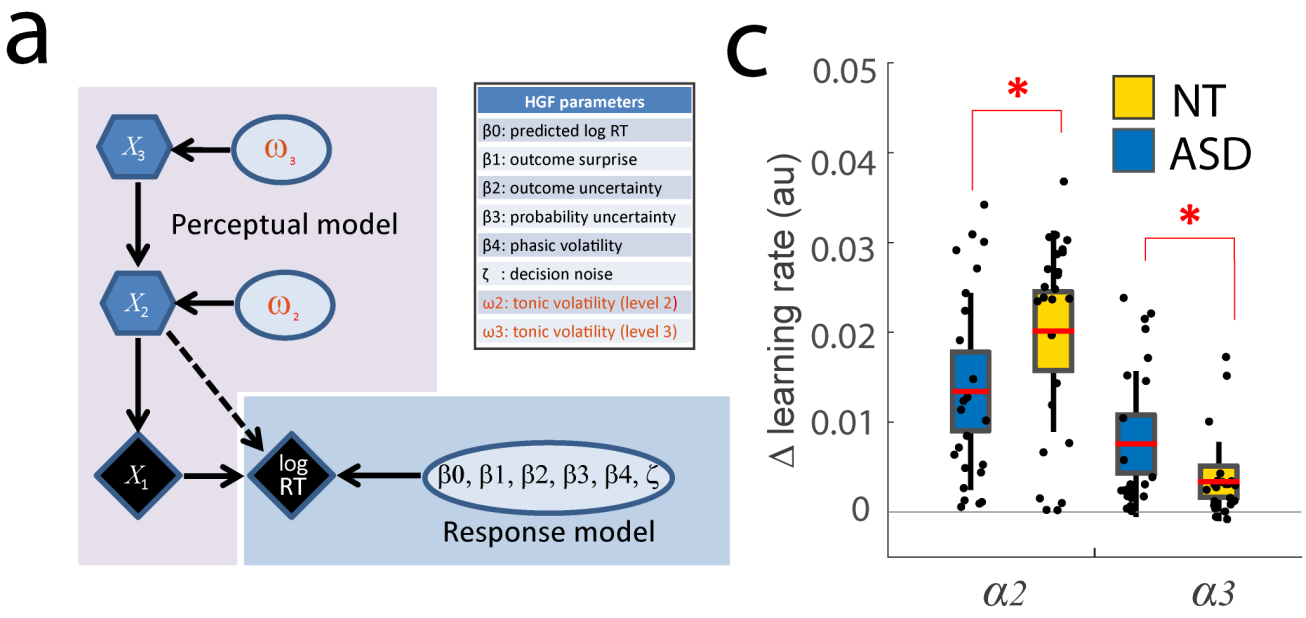
Figure 4: **Computational model details and results** - (a) Schematic depiction of the 3-level HGF. The perceptual model comprises three hierarchical states (x_1, x_2, and x_3). Participant specific free parameters (ovals) are estimated from individual log RT data - red parameters relate to the perceptual model whereas black parameters relate to the response model. Diamonds represent quantities that change over time (trials); hexagons, like diamonds, represent quantities that change in time but that additionally depend on their previous state in time in a Markovian fashion. See main text for more details. (b) Binary logistic regression – beta weights for each of the HGF free parameters showing the contribution of each to predicting group status (ASD, NT) across all participants (n=49). Significant predictors (P<0.05) are denoted with star. Error bars show SEM for the beta estimates. All parameters were included in the same model but $\omega's$ are plotted on a separate scale (in red). Group differences in the model parameters at the level of individual subjects can be seen in Figure S8. (c) Group differences in learning rate update (i.e. change from stable to volatile periods of the task). ASD participants (n=24) update α_2 less than NT participants (n=25), whereas they update α_3 more than NT participants. Data points represent individual participants, red line shows the mean, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean. ASD, autism spectrum disorder. NT, neurotypical

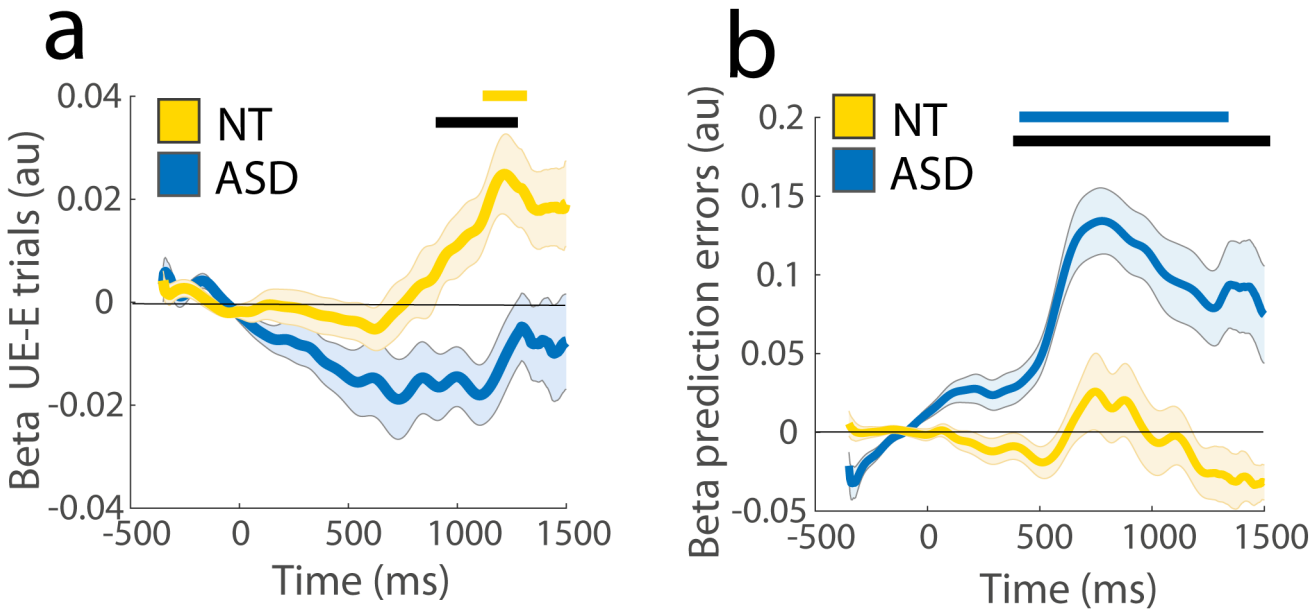
Figure 5: Pupillometry results – (a) solid yellow line shows cluster of time points where the UE-E group contrast was significantly positive in the NT participants; black solid line shows where NT's were significantly greater than ASD (2000 permutations; FWE alpha=0.05, 2-tailed). (b) Blue solid line indicates where ASD participants showed a significant pupil response to precision weighted prediction errors (ε_3), that is greater than zero and black solid line shows where this pupil response was significantly different from NT's (2000 permutations; FWE alpha=0.05, 2-tailed). NT, n=14, ASD, n=11. X-axis represents time since outcome. ASD, autism spectrum disorder. NT, neurotypical. UE, unexpected. N. E, expected. au, arbitrary units

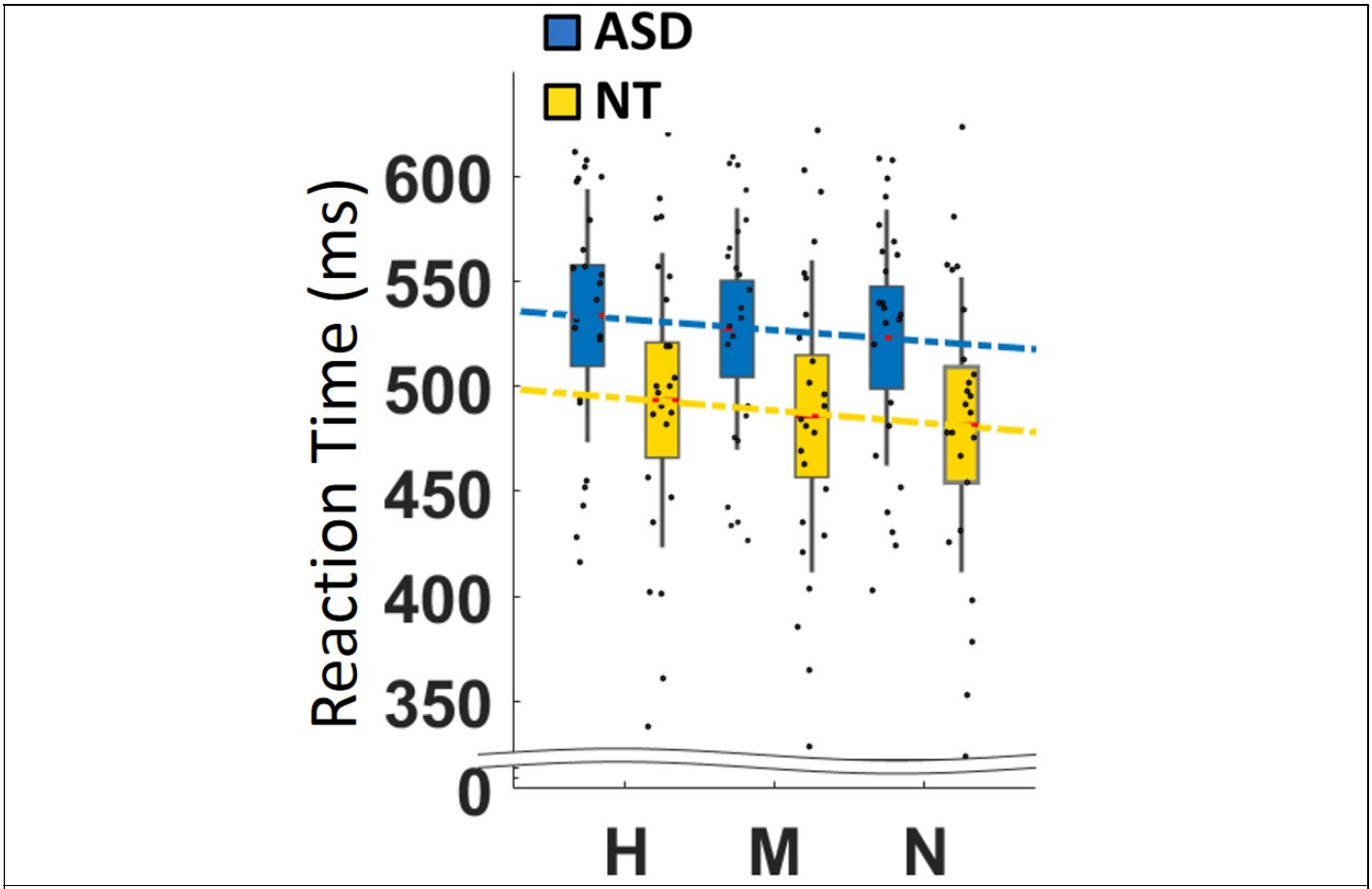

High tone
Low tone
Stable
Volatile
P(house)
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
Trials
300ms
Tone
150ms
Outcome
Response
1,500 - 1,800ms
Low or High
House
or
Face
No noise
Med noise
High noise


a
NT
ASD
Reaction time (ms)
600
550
500
450
400
350
0
E
N
UE
b
UE-E RT (ms)
60
40
20
0
-20
ASD
NT
c
NT
ASD
% Errors
25
20
15
10
5
0
-5
E
N
UE
d
UE-E errors (%)
15
10
5
0
-5
ASD
NT


40
30
20
10
0
-10
-20
UE-E RT (ms)
0
1
2
3
4
5
6
7
ADOS communication score


a
HGF parameters
β0: predicted log RT
β1: outcome surprise
β2: outcome uncertainty
β3: probability uncertainty
β4: phasic volatility
ζ : decision noise
ω2: tonic volatility (level 2)
ω3: tonic volatility (level 3)
Perceptual model
log RT
β0, β1, β2, β3, β4, ζ
Response model
c
NT
ASD
Δ learning rate (au)
0.05
0.04
0.03
0.02
0.01
0
α2
α3
b
Predicting diagnostic status
Beta weight (au)
60
40
20
0
-20
1.0
0.5
0
-0.5
β0
β1
β2
β3
β4
ζ
ω2
ω3

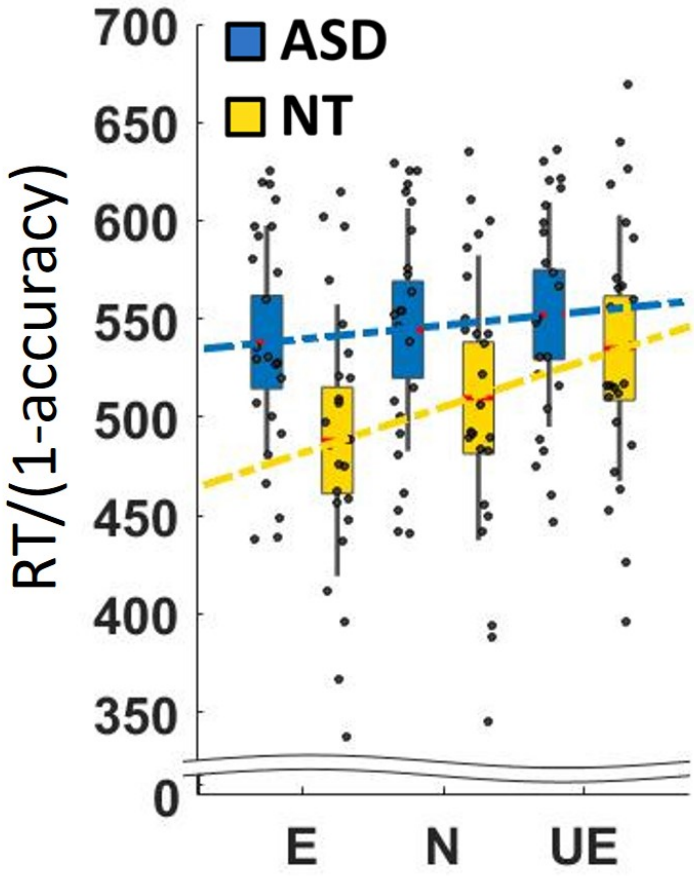

a
NT
ASD
Beta UE-E trials (au)
0.04
0.02
0
-0.02
-0.04
-500
0
500
1000
1500
Time (ms)
b
NT
ASD
Beta prediction errors (au)
0.2
0.15
0.1
0.05
0
-0.05
-500
0
500
1000
1500
Time (ms)



Supplementary Figure 1

Reaction time as a function of stimulus noise

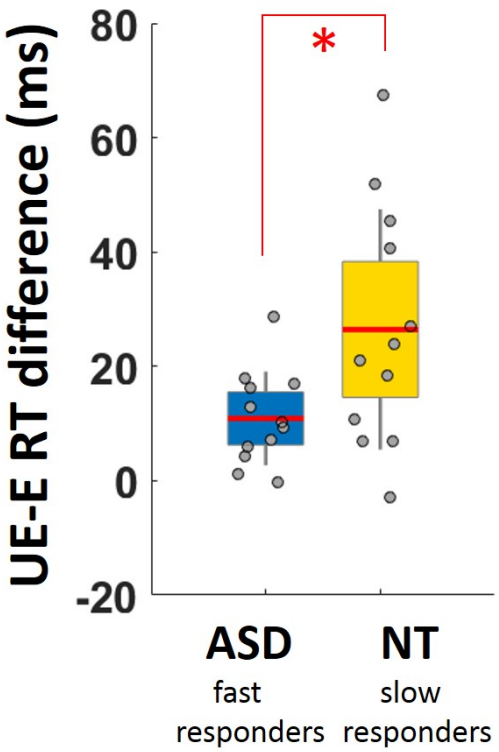
Collapsing across the three levels of expectedness, the non-significant stimulus noise*group interaction indicates that the linear relationship between noise and RT was equivalent in both groups (ASD, n=24; NT, n=25). See main text for supporting statistics. ASD, autism spectrum disorder. NT, neurotypical. RT, reaction time. Data points represent individual participants, red lines indicate the mean, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean for each condition and group. H=high noise, M=medium noise and N=no noise



Supplementary Figure 2

Inverse Efficiency Scores

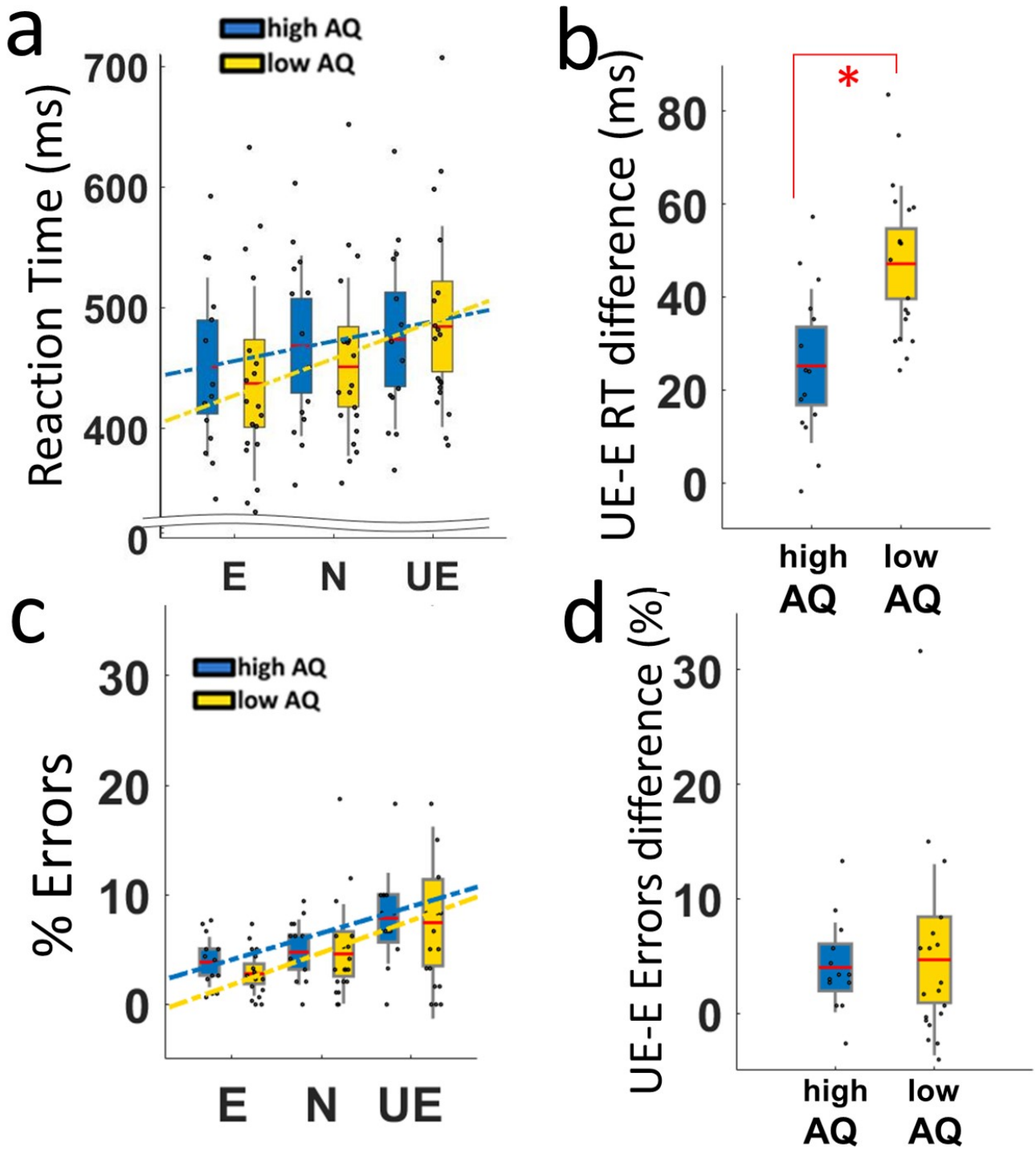
To ensure that the attenuated UE-E RT difference in the ASD participants was robust to correction accuracy, we calculated inverse efficiency scores (IES) as RT/(1-accuracy) for each condition. As for the analysis of RT and error rates alone, see main manuscript, there was a significant main effect of expectedness ($F(1.8,83.11)=34.24$, $P<0.001$) and noise ($F(2,94)=6.87$, $P=0.002$) and again only the expectedness*group interaction was significant in this analysis ($F(2,94)=9.98$, $P<0.001$). The noise*group ($F(2,94)=0.24$, $P=0.79$) and expectedness*noise*group interactions were not significant ($F(4,188)=2.2$, $P=0.07$). Thus, our primary reaction time finding is robust to correction condition-specific accuracy (ASD, n=24; NT, n=25). ASD, autism spectrum disorder. NT, neurotypical. RT, reaction time. Data points represent individual participants, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean for each condition and group. E=expected, N=neutral and UE=unexpected.



Supplementary Figure 3

Caution of responding control analysis

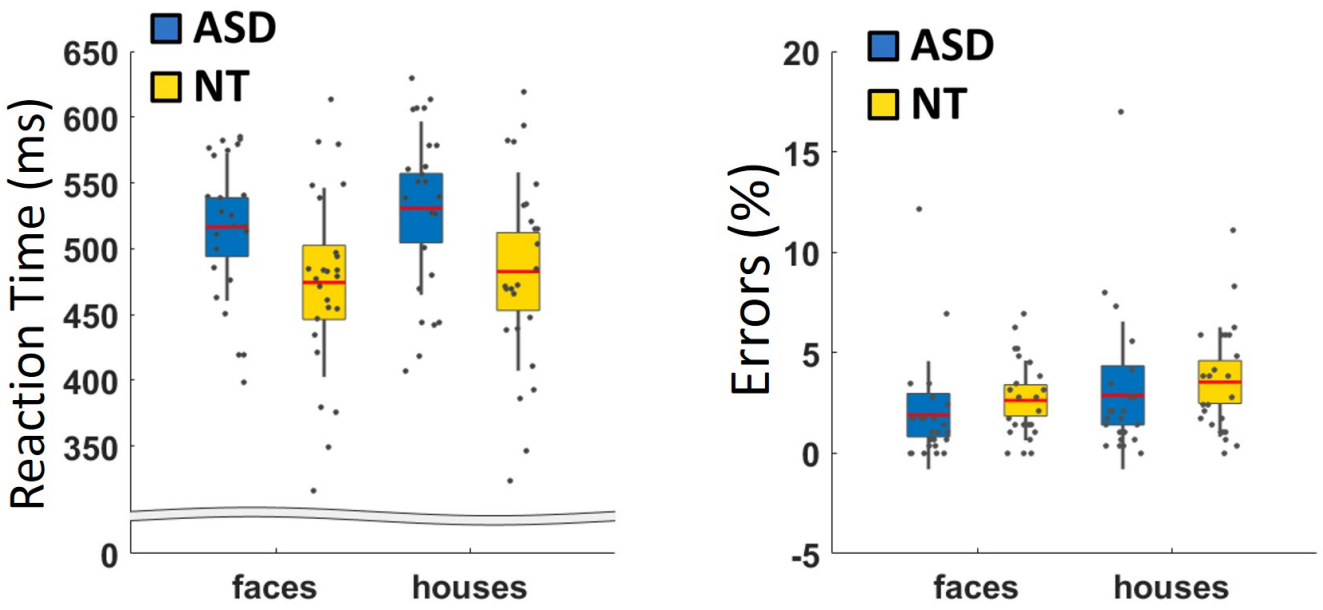
To exclude the possibility that our group difference in UE-E RT (i.e. reduced behavioural surprise in ASD) is explainable by increased response caution in the ASD participants we compared the 12 fastest responders from the ASD group (mean RT 418 ms) against the 12 slowest responders in the NT group (mean RT 540 ms) on the primary UE-E RT difference measure. Here the 12 fastest overall responding ASD participants are those who are most impulsive/least cautious in general responding (i.e. have the lowest response thresholds) whereas the 12 slowest overall NTs are the least impulsive/most cautious (i.e. have the highest response thresholds). Indeed, mean reaction time is significantly faster in this subgroup of ASD participants than in the subgroup of NTs ($t(22)=3.38$, $P=0.03$). Nonetheless, independent-samples t-tests revealed that the ASD participants (in this subset of fast general responders) still show significantly diminished behavioural surprise ($t(22)=2.39$, $P=0.026$) relative to NTs (in this subset of slow responders). ASD, autism spectrum disorder. NT, neurotypical. RT, reaction time. Data points represent individual participants, red lines indicate the mean, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean. Star indicates significance at $P<0.05$.



Supplementary Figure 4

Replication of behavioural result in a non-clinical sample

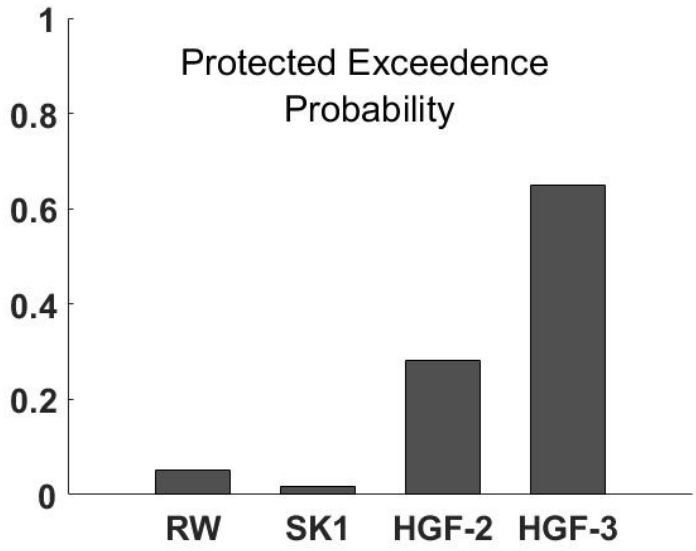
(a) The same task conducted in a sample of non-clinical volunteers characterised according to high or low autistic traits (AQ) replicates the interaction between expectedness (E=expected, N=neutral, UE=unexpected) and autistic tendency (high AQ, n=26; low AQ, n=31). There was a significant main effect of expectedness ($F(2,110)=69.46$, $P<0.001$) and, crucially, a significant expectedness*AQ group interaction ($F(2,110)=13.29$, $P<0.001$); suggesting that participants with high AQ scores show a reduced modulation of RT as a function of expectedness (e.g. reduced slope), relative to participants with low AQ scores. There was a main effect of noise ($F(2,110)=16.96$, $P<0.001$), and noise*group interaction ($F(2,110)=5.07$, $P=0.008$). No other linear interactions or main effects were significant (P's>0.2). (b) An independent samples t-test demonstrated that behavioural surprise was significantly attenuated in the high AQ group ($t(55)=4.32$, $P<0.001$). (c, d) Error rates were subject to the same analysis as above. There was a significant main effect of expectedness ($F(2,110)=19.89$, $P<0.001$) but the expectedness*AQ group interaction did not reach significance ($F(2,110)=.85$, $P=0.42$); suggesting that the main effect of expectedness on accuracy did not vary as a function of autistic traits. The main effect of noise narrowly missed significance ($F(2,110)=2.36$, $p=0.09$), but there was no noise*group interaction ($F(2,110)=1.67$, $P=0.1$). One low AQ participant showed relatively high % errors in the UE condition, but their overall errors were within reasonable limits and results are not changed if they are excluded. Compare with Figure 2a-d in the main text. Data points represent individual participants, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean. Star indicates significance at $P<0.05$



Supplementary Figure 5

Responses to face and house stimuli

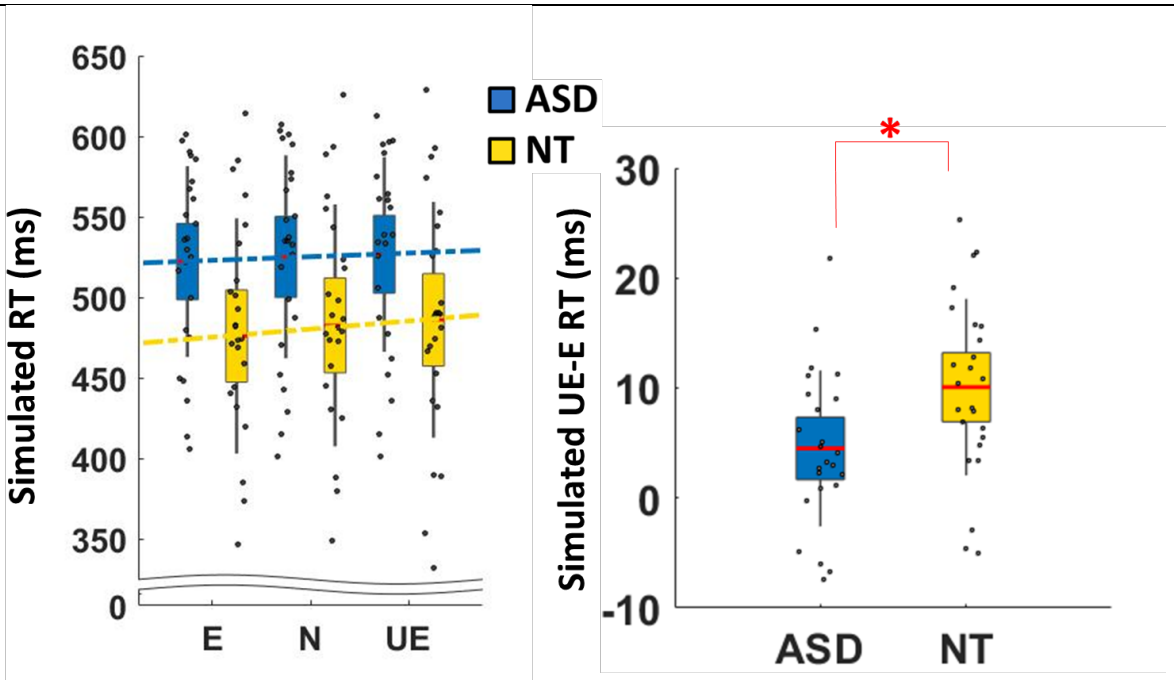
To confirm that there were no group differences in RTs or error rates in responding to the different outcome image types (faces, houses) we examined these responses in two separate repeated-measures ANOVAs with group (ASD, n=24; NT, n=25) as a between participants factor in each case. For reaction times there was a significant main effect of stimulus type, reflecting the fact that participants were in general slower to respond to house images over face images ($F(1,47)=16.52$, $P<0.001$). Additionally there was a main effect of group indicating that the ASD participants were generally slower to respond than the NT participants ($F(1,47)=5.54$, $P=0.023$) but crucially there was no interaction between stimulus type and group ($F(1,47)=1.23$, $P=0.2$). For error rates, participants generally made more errors on house trials (main effect of stimulus type: $F(1,47)=13.37$, $P=0.001$), but there was no group difference in errors overall (non-significant main effect of group: $F(1,47)=0.8$, $P=0.37$) and there was no stimulus type x group interaction ($F(1,47)=0.02$, $P=0.9$). ASD, autism spectrum disorder. NT, neurotypical. Data points represent individual participants, red lines indicate the mean, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean.



Supplementary Figure 6

Results of Bayesian model selection

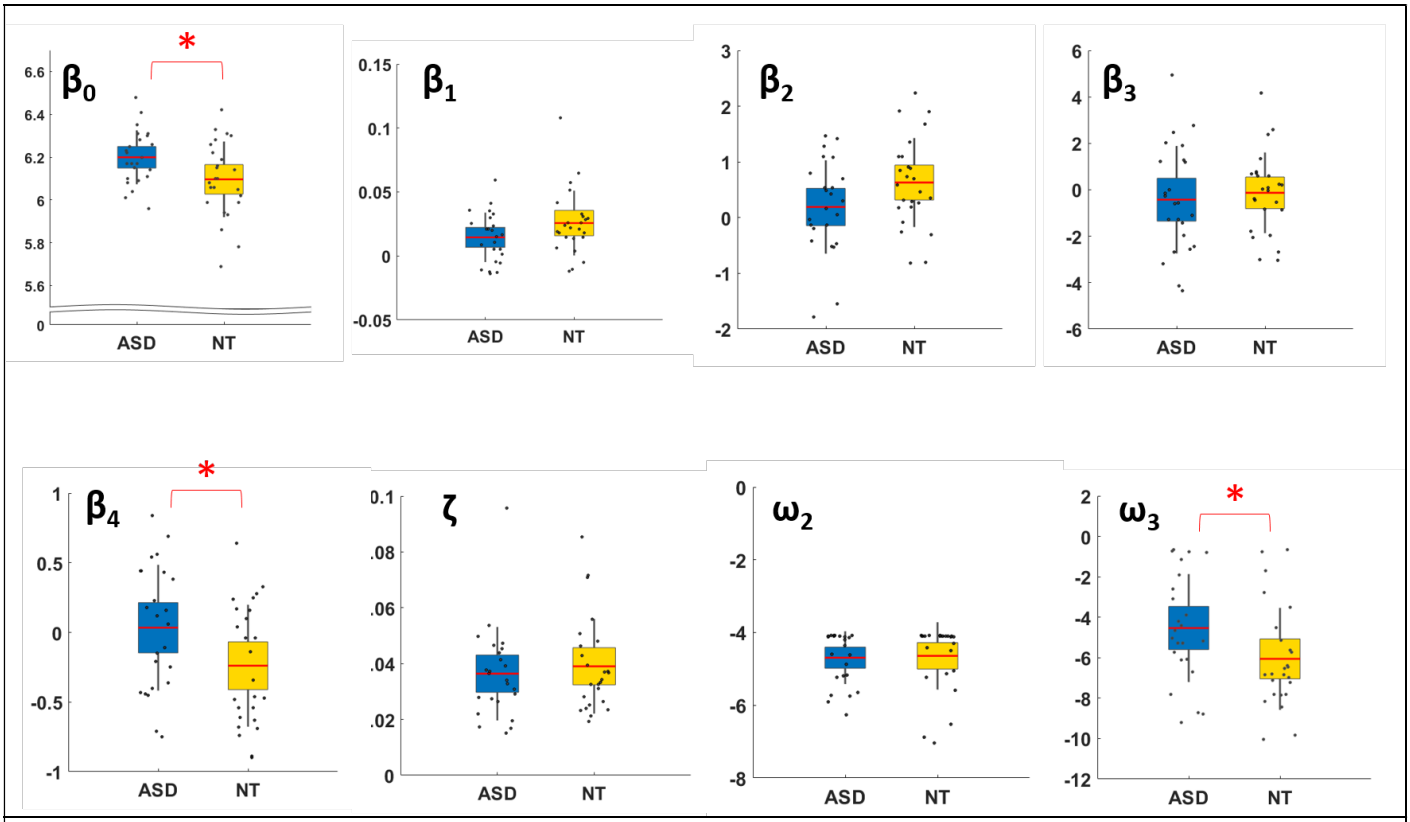
The protected exceedance probability from the Bayesian Model Selection (BMS) of log model evidences shows that the 3-level HGF (HGF-3) describes subject's behaviour better than alternative learning models (RW; Rescorla Wagner, SK1; Sutton K1, HGF-2; 2-level Hierarchical Gaussian Filter). See main text for details.



Supplementary Figure 7

Model simulated reaction times

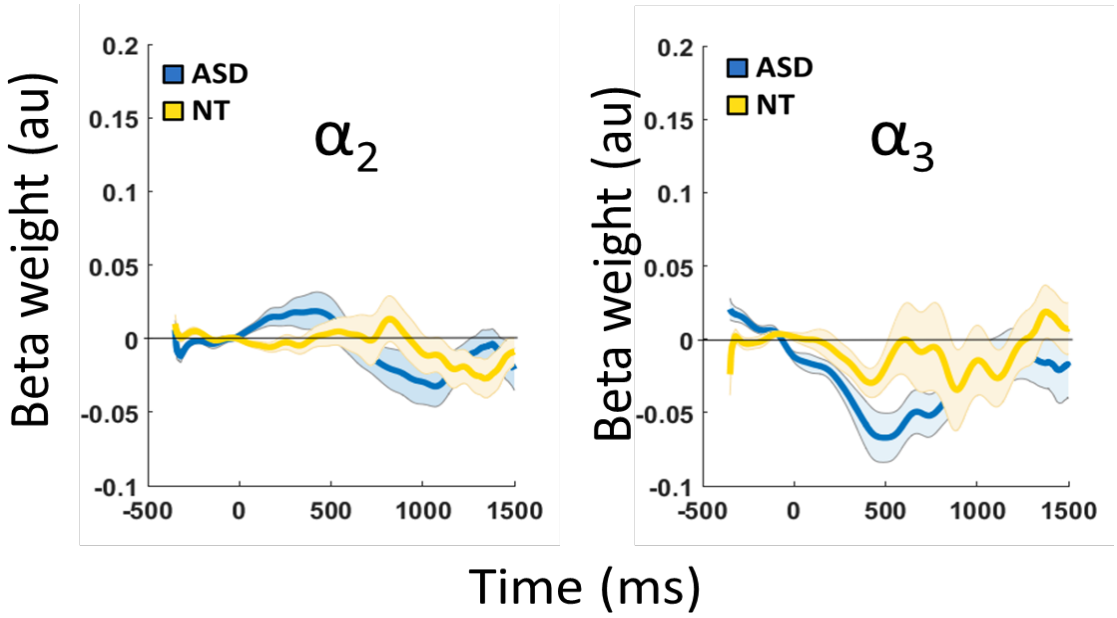
As an additional validation of the HGF model performance we simulated trial-wise RTs using the fitted perceptual and response model parameters from each of our 24 ASD and 25 NT participants. These simulations can recover the group differences in the main behavioural effect of expectation (compare to Figure 2a&b in the main manuscript). Statistical analysis of these model simulated RTs indicates a significant expectedness * group interaction ($F(1,94)=4.44$, $P=0.014$), and the simulated UE-E RT difference was significantly lower when simulated from the ASD parameters, relative to the NT parameters($t(47)=2.57$, $P=0.013$). ASD, autism spectrum disorder. NT, neurotypical. UE, unexpected. N, non-predictive. E, expected. Data points represent the mean of 32 simulations for each individual participant, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the group mean, red lines indicate the group mean. Star indicates significance at $P<0.05$



Supplementary Figure 8

Average HGF parameter estimates across groups

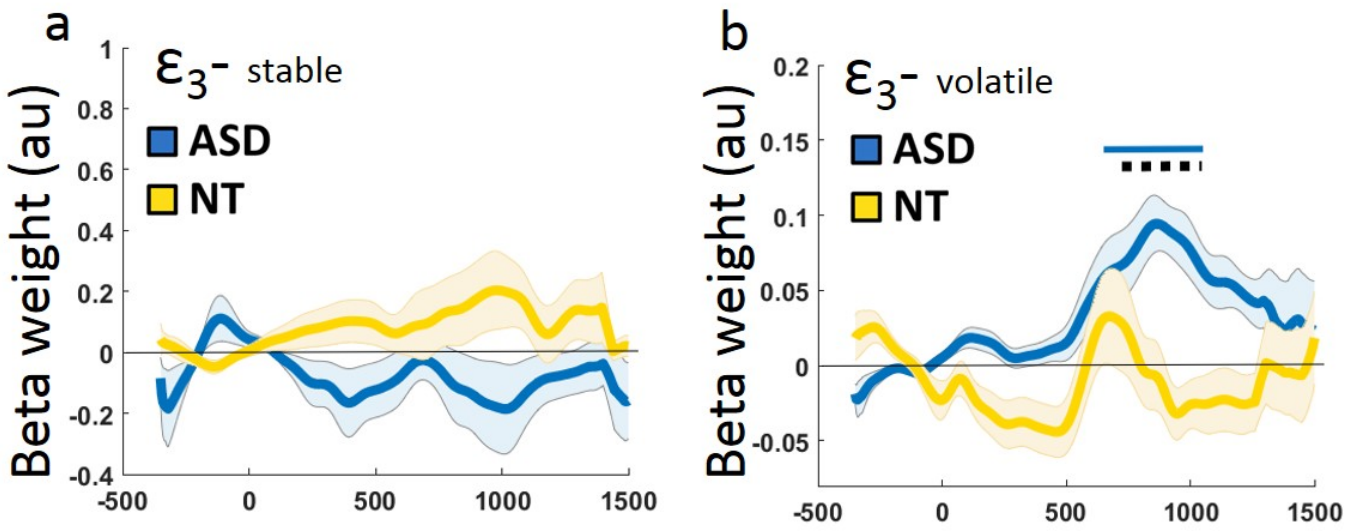
Individual participant parameter estimates for each of the free parameters estimated from the HGF, for both the ASD (n=24) and NT (n=25) groups. A statistically significant MANOVA effect indicated that the groups would differ on one or more of the estimated model parameters, Pillai's' Trace = .43, F(8, 40) = 3.81, P=0.002). Independent samples t-tests indicate a significant group difference in baseline log RT (β0; t(47) = 2.33, P=0.024), phasic volatility (β4; t(47) = 2.15, P=0.037) and tonic volatility at the third level (ω3; t(47)=2.10, P=0.045). Outcome surprise (β1; t(47) = -1.73, P=0.09) and outcome uncertainty (β2; t(47) = -1.87, P=0.06) narrowly missed significance. There were no group differences in probability uncertainty (β3; t(47) = -.51, P=0.61), decision noise (ζ; t(47) = -.55, P=0.59) or tonic volatility at the second level (ω2; -.21, P=0.84). See the main text and Figure 4b for a multiple linear regression analysis predicting group status from these same parameters. ASD, autism spectrum disorder. NT, neurotypical. Data points represent individual participants, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean, red lines indicate the group mean. Star indicates significance at P<0.05



Supplementary Figure 9

Pupil size and dynamic learning rates

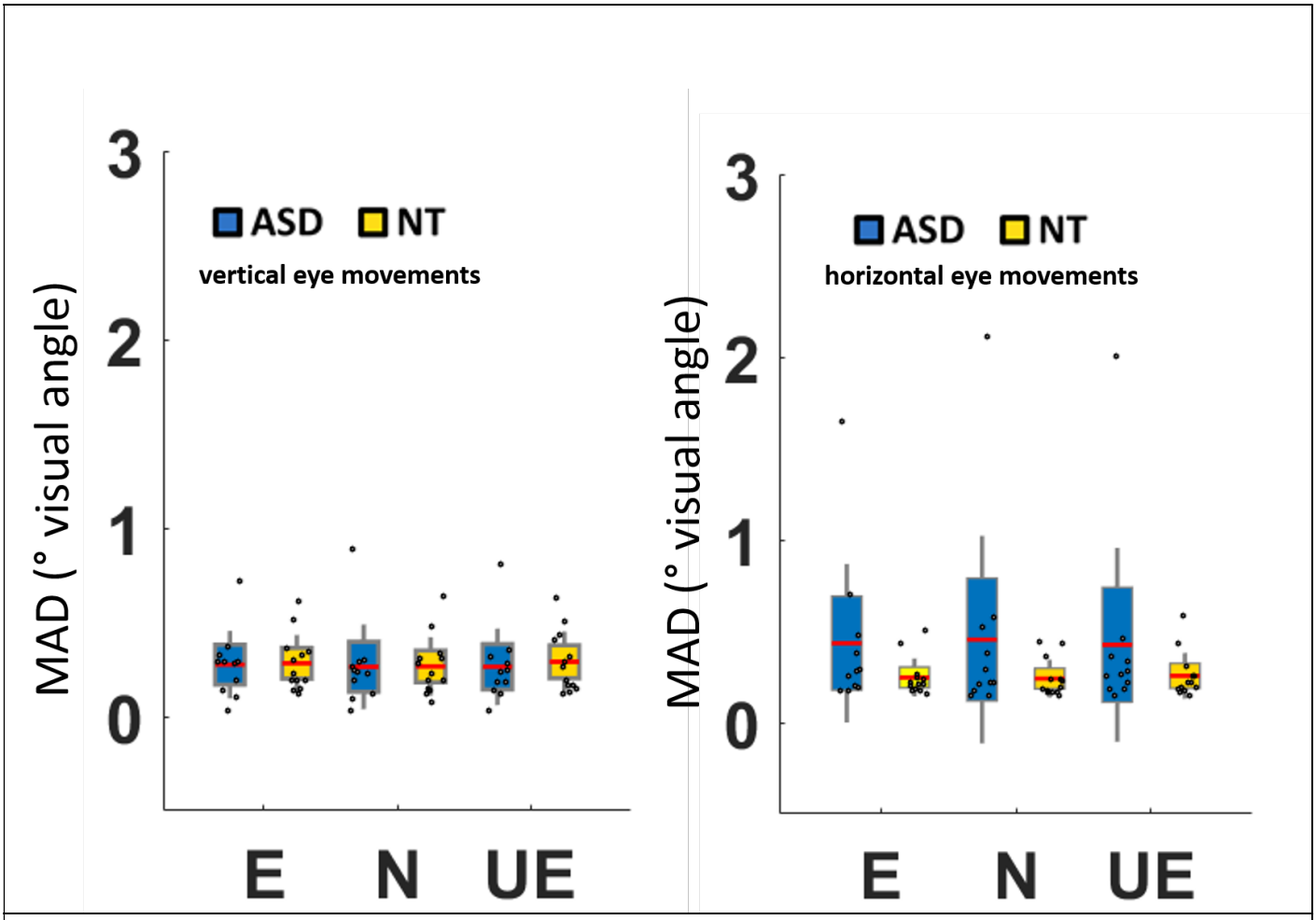
The analysis reported in the main text indicates a sustained positive relationship between pupil size and precision-weighted prediction errors (ε_3) in the ASD participants (Figure 5b). The precision weight (on the prediction error) is proportional to the update of environmental volatility and is formally related to dynamic trial-wise learning rate (α_3) This additional analysis indicates that the learning rates themselves (α_2 and α_3) do not have a significant influence on pupil dilation in either group. As for the results reported in the main text (see Online Methods) this regression analysis included, trial type (face, house), fixation compliance, mean RT and UE-E ground truth contrasts, as control regressors. Shaded regions represent standard error of the mean.



Supplementary Figure 10

Pupil size and precision-weighted PE's in stable and volatile task periods

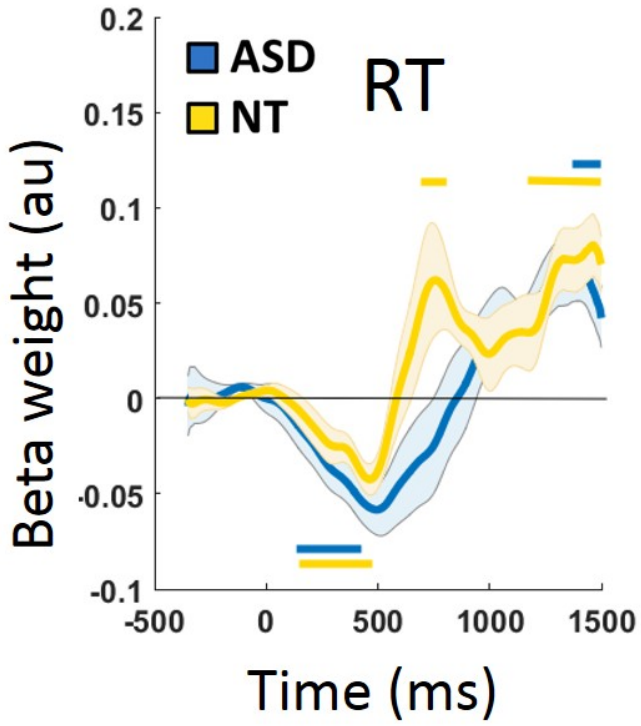
The pupil regression reported in the main text (Figure 5b) examined the relationship between precision-weighted prediction errors (ε_3; PE's) and pupil size across all trials in the experiment. A strength of this analysis is that it represents the pupil response when each participant was actually surprised, and does not impose knowledge of the task structure. Nonetheless, to examine the relationship between precision-weighted PE and pupil size in the volatile and stable periods of the task we conducted the same regression analysis (see main text and online methods) but separately for the 72 'stable' trials and '72' volatile trials (see Figure 1) towards the end of the experiment. (left) In the stable period there is no relationship between precision-weighted prediction errors and pupil size in either group or no differences between the groups. (right) The relationship between precision-weighted PE's and pupil size in the ASD participants (blue) is apparent 1000ms after the outcome appears in the volatile period of the task. Blue solid line shows where the ASD participants differ from zero and black dotted line shows where the ASD participants differed from the NT participants. Shaded region represents standard error of the mean. Consistent with the analysis of learning rates in the volatile and stable task periods (Figure 3c), this suggests that the ASD participants tend to show aberrant noradrenergic surprise about volatility, *in response to volatility* (e.g. over-updating learning about volatility and over-engaging noradrenergic responses to surprise about volatility, in the face of environmental volatility). However, we caution against the low trial numbers included in this analysis (72, vs a maximum of 456 in the analysis reported in the main text) and the fact that one control participant did not have enough good trials in the volatile period to be included in this analysis, so participant numbers are also reduced (ASD=11, NT=13).



Supplementary Figure 11

Fixation compliance across trial types

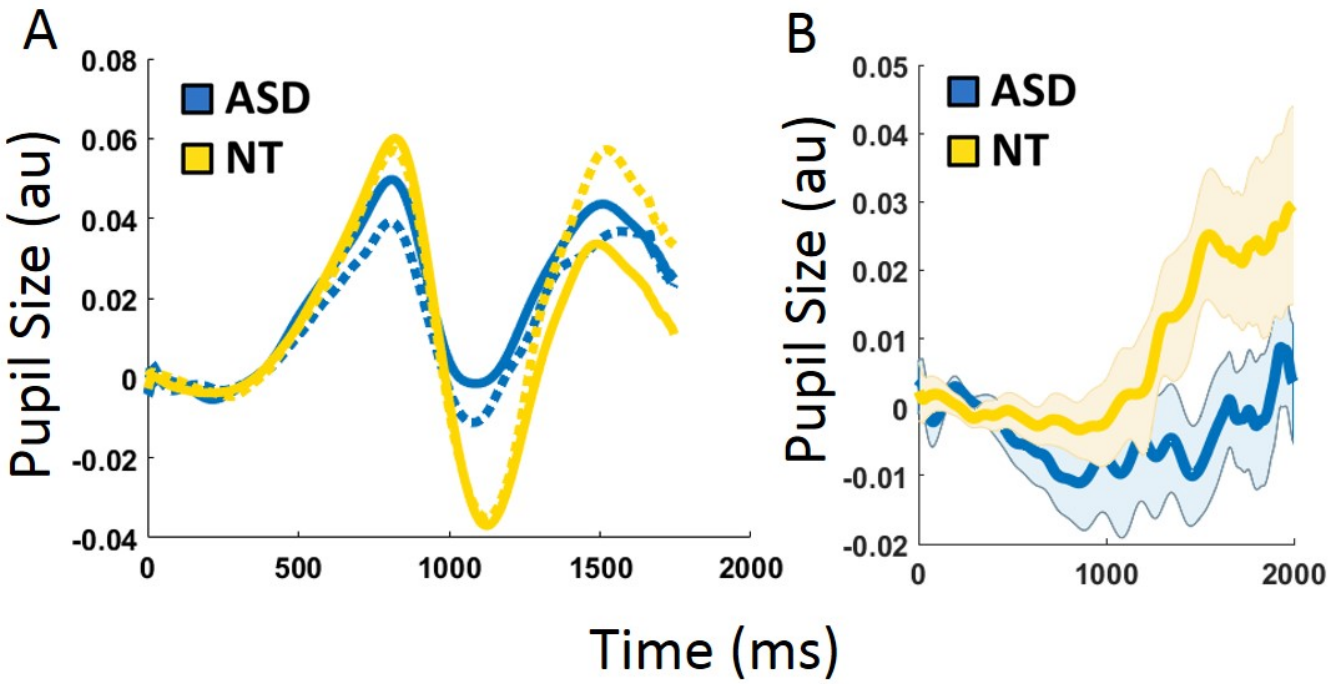
Mean absolute deviation (MAD) from fixation (in degrees of visual angle) across groups and conditions. Generally fixation compliance was very good, <1° of visual angle in both the vertical and horizontal axes. Stimulus duration was purposefully short to eliminate saccades. Crucially there is no systematic difference in fixation compliance that would impact on the pupillometry results, either across groups or conditions. One participant showed systematically larger deviation from fixation in the horizontal plane, though fixation was still good (below 2° visual angle) and not beyond the physical limits of the stimulus being presented. Importantly, trial-wise absolute deviation from fixation was included as a regressor of no interest in all pupillometry analyses reported in the main text and supplemental results, and thus our pupillometry analyses are corrected for eye movements. ASD, autism spectrum disorder. NT, neurotypical. UE, unexpected. N, non-predictive. E, expected. Data points represent individual participants, shaded regions and error bars show 95% confidence intervals and 1 standard deviation of the mean, red lines indicate the group mean.



Supplementary Figure 12

Pupil size and reaction time

The relationship between precision-weighted prediction errors and pupil size reported in the main text (Figure 5b), links pupil size to behaviour indirectly via the HGF model, since the precision-weighted prediction errors are estimated for each participant on the basis of their trial-wise RT. However, we conducted an additional regression analysis to investigate the relationship between basic behaviour (trial-wise RT) and pupil size directly. As RT increases, post-outcome pupil size shows an initial decrease from baseline followed by an increase towards the end of the trial. Crucially, there are no time points in which the relationship between RT and pupil size is significantly different between the ASD and NT groups. Notably, trial-wise RT is included as a control regressor in the results reported in the main text (Figure 5), and in the analyses reported above (Figure S9 & S10), so where there is a significant relationship between pupil size and the trial-wise model parameters this exists over and above any effect of RT on pupil size. Blue and yellow bars indicates where the relationship between RT and pupil size significantly differed from zero in the ASD and NT participants, respectively. As for the results reported in the main text (see online methods) this regression analysis included, trial type (face, house), fixation compliance and UE-E ground truth contrasts, as control regressors. Shaded region represents standard error of the mean.



Supplementary Figure 13

Raw pupil traces

(a) Raw mean pupil dilation in ASD participants (blue) and NT participants (yellow) separated into trials in which the outcome was unexpected (UE: dotted line) and trials where the outcome was expected (E: solid line) (a) The UE-E difference (i.e. ground truth 'surprise') in the ASD participants (blue) and NT participants (yellow). Equivalent to the UE-E contrast from the regression model presented in Figure 5a. Shaded region shows standard error of the mean.

Click inside this box and insert image for Supplementary Figure 14. For best results, use Insert menu to select a saved file; do not paste images. Source images must be JPEGs (no larger than 10 MB) saved in RBG color profile, at a resolution of 150–300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page.

Delete these instructions before inserting the image.

Supplementary Figure 14

Insert figure title here by deleting or overwriting this text; keep title to a single sentence; use Symbol font for symbols and Greek letters.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters.

Click inside this box and insert image for Supplementary Figure 15. For best results, use Insert menu to select a saved file; do not paste images. Source images must be JPEGs (no larger than 10 MB) saved in RBG color profile, at a resolution of 150–300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page.

Delete these instructions before inserting the image.

Supplementary Figure 15

Insert figure title here by deleting or overwriting this text; keep title to a single sentence; use Symbol font for symbols and Greek letters.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters.