Influence of truthful and misleading instructions on statistical learning across the autism spectrum

The effect of instructions on learning across the autism spectrum

Nikitas Angeletos Chrysaitis[1] and Peggy Seriès[1]*

[1]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Nikitas Angeletos Chrysaitis: s1791272@ed.ac.uk

Peggy Seriès: pseries@exseed.ed.ac.uk (* corresponding author)

Abstract

Bayesian studies of perception have documented how the brain learns the statistics of a new environment and uses them to interpret sensory information. Impairments in this process have been hypothesised to be central to autism spectrum disorders. However, very few such studies have differentiated between implicit and explicit learning. We manipulated the instructions given before a cue-stimulus association task to investigate their effects on statistical learning. The task was conducted online, in 335 participants with varying autistic traits. In the implicit condition, where no information was provided, participants acquired weak prior beliefs about the task regularities. Conversely, explicit information about the presence of regularities resulted in strong priors, correctly reflecting the task's statistics, regardless of the information's veracity. Autistic traits correlated with greater uncertainty and faster updating in the implicit condition, but no significant differences were found in the influence of priors. Our findings illuminate how instructions affect statistical learning and how these effects differ across the autism spectrum.

Author Summary

Perception is greatly influenced by the brain's prior knowledge of the environment, through a process called *Bayesian inference*. Recent theories of psychiatric disorders and particularly autism view them as impairments in this process. A crucial aspect of this process is how individuals form their knowledge of the environment. However, previous studies have not differentiated between learning that occurs when participants are aware of what they are learning and learning that happens implicitly. In the present study, we conducted an experiment with four conditions, each varying in terms of what participants were trying to learn and whether they were aware of the general form of the regularities. Our findings revealed that

participants form much stronger beliefs about the regularities when they are informed about their presence. Additionally, we discovered that participants with strong autistic traits tend to be slightly more uncertain in their beliefs and quicker to update them, but only when they are unaware of the presence of regularities.

Introduction

The information we get from our senses is frequently noisy and ambiguous. To overcome that, the brain combines sensory evidence with prior knowledge about the environment. This process of integrating sensory inputs with prior beliefs is formalised in Bayesian theories of perception and cognition [1]. According to Bayesian models, the brain maintains internal, probabilistic models of the environment. These are combined with sensory information that takes the role of the likelihood in Bayesian inference to generate posterior probability distributions over possible states of the world. The posteriors are then used to form what we end up perceiving and how we understand the world.

There are different ways that the brain could learns about the world. It could be that gradual experience with the regularities of the environment results in implicit prior beliefs. Alternatively, knowledge of the regularities could be explicitly communicated to the individual. This raises a few questions. Do priors that are based on explicit information differ from those that are formed implicitly? Does misleading information affect subsequent learning? And does that differ when there are no regularities in the environment versus when there are, but they differ from what you were told?

In this study, we set to answer these questions. Previous research on implicit and explicit learning has mostly focused on serial reaction time tasks and artificial grammar learning [2,3]. The results show evidence for two distinct processes [4,5], affecting different brain areas [6–

8]. These processes depend on different factors, with explicit learning being correlated with IQ [9,10] and the resulting knowledge decaying over time [11,12], while implicit learning is largely independent from intelligence and its effects being consolidated with the passage of time. Interestingly, when they are compared on the same task, alerting participants to the presence of regularities often leads to inhibition of implicit learning and worse participant performance [7,13–15]. However, to our knowledge, no studies have investigated implicit and explicit learning when environmental regularities are based on the frequency of the stimuli, instead of specific rules that they follow. Accordingly, no studies have investigated participant behaviour from a Bayesian viewpoint.

Our study is partially inspired by the Bayesian literature on autism spectrum disorder (ASD). Initially, such theories proposed that autistic individuals exhibit weaker prior influences compared to the general population [16–18]. While this remains the main hypothesis, in recent years the focus has shifted towards explanations that are centred on the learning of regularities and specifically the processing of environmental volatility [19,20]. Indeed, our systematic review found that ASD groups differed more often from neurotypical groups in tasks where priors were learned compared to tasks involving pre-existing priors [21]. Our review also showed that studies using implicitly learned priors tended to find weaker influences in autistic individuals, while this was less common when prior learning was explicit or when participants were directly informed about the regularities [21]. This seems to align with non-Bayesian studies showing intact explicit but occasionally impaired implicit learning in ASD [see 22 for a review]. Unfortunately, most ASD studies did not describe in detail what instructions were given to the participants, meaning it was sometimes unclear what components of the learning were implicit or explicit. Moreover, the two kinds of tasks commonly used very different experimental designs, making the comparison between them difficult.

In the present study, we used one design, manipulating the regularities and the instructions of the task independently across four conditions to assess their individual impact on learning and their interaction. The participants were asked to discriminate between left- and right-tilted Gabor patches that were preceded by an auditory cue. The cue-stimulus relationship varied across conditions, as did the participants' knowledge of it. We modelled the data with the Hierarchical Gaussian Filter (HGF), so that we could estimate the prior belief trajectories of the participants and compare them between conditions. We also collected responses on the autism spectrum quotient questionnaire (AQ) and investigated the potential relationships between these scores and participant performance on the task. We found that when participants were not given any information about the regularities, they developed very weak beliefs about them. In contrast, when their attention was drawn to the presence of regularities, their beliefs were much stronger, independently of the veracity of the information they received. We also found a negative relationship between AQ scores and the precision of participants' beliefs about the regularities in the implicit condition.

Methods

Participants

We recruited 456 participants from the crowdsourcing platform Prolific, with approximately half of them having reported that they have a diagnosis of ASD or that they feel they belong in the autism spectrum. All participants completed the AQ questionnaire and then one of the conditions of our task. All participants gave informed consent and received monetary compensation for participation. The study was approved by the University of Edinburgh School of Informatics Research Ethics Committee (Application number 175134). As a quality control,

we discarded the data of the participants who failed our attention checks, who did not respond in more than 3.5% of the trials, or who could not discriminate between high- and low-pitch tones. This was done before any behavioural or computational analysis. Surprisingly, more than one fourth of the participants were excluded via this process, leaving us with 335 participants across all conditions. The group that self-identified as part of the autism spectrum had fewer rejections than the other group (20% vs 32%). AQ scores did not differ between accepted and rejected participants (mean AQ 24.9 vs 24.4, $p = 0.47$).

Design

The experiment consisted of a cue-stimulus association task (Figure 1). In all conditions, the experiment began with instructions explaining that the goal of the task was discriminating between left- and right-tilted (Gabor) gratings. Then participants completed one training block of 64 trials that included only the visual stimuli. The task began at a very easy level, with the gradings being presented at a very high contrast and at 70° or 120°. During the training, the contrast changed based on a 2/1 perceptual staircase, while the tilt of the stimuli was gradually reduced to 93° and 87°, respectively. This was done so that participants would get gradually acclimated to the task. The contrast staircase also led to stimuli that were similarly hard to see across our sample, as we could not control the environment of online participants. In each trial, gratings were presented for 150ms and were immediately followed by a static target for 300ms to prevent the formation of afterimages. Then, participants had 1000ms to choose if the grating was tilted to the left or to the right, using the left and right arrow keys. The screen displayed two arrows at that time corresponding to the possible orientations. When participants made

their choice, the other arrow disappeared and the remaining arrow turned green or red for 300ms, depending on if they were correct or not.

Figure 1: Example stimulus (A) and structure of individual trial (B). *Stimuli in the experiment were presented in significantly lower contrast than this example.*

The trials of the main task kept the same structure with two changes. Firstly, the contrast of the grating started at the value set by the training and was updated using a 3/1 staircase. Secondly, participants were initially presented with a random pure tone of either 330hz or 660hz for 300ms before the presentation of the grating. The relationship between the auditory cues and the visual stimuli differed across conditions, as did the instructions given to the participants about them. The main task lasted for 4 blocks or 256 trials.

Participants were split into four conditions based on the information they received from the instructions (I) and the regularities (R). Receiving no explicit information about the regularities or the tone not being associated with the image was signified with the number 0. Otherwise,

the presence of information was signified with a + symbol, or with a – when instructions disagreed with the regularities.

- In the condition I^0R^+, the implicit condition, the tone predicted the tilt of the grating with 75% probability. The pairing of the tone to the tilt was randomised between participants. Participants were naïve about the association between tones and gratings (instructions: 'The purpose of the tone is to alert you to the presence of the grating and focus your attention on it.'). 119 participants took part in this condition.
- In the condition I^+R^0, the tone had no relationship with the grating (50%). Participants, though, were told that 'A low-pitch tone will be followed 75% of the time by a [right/left]-tilted grating. A high-pitch tone will be followed 75% of the time by a [left/right]-tilted grating'. Thus, participants were given explicit information about the regularities, but this information was misleading as the true regularities were completely random. 106 participants took part in this condition.
- The final two conditions were a combination of the above, with the tone predicting the tilt of the grating with 75% probability and the instructions being the same as in the second condition. In approximately half the participants (I^+R^+, n = 51) the instructions were telling the truth, similarly to explicit conditions in the literature. In the other half (I^-R^+, n = 59) the association mentioned in the instructions was the reverse to the true association. That is, if in truth the low tone was followed 75% of the time by a left-tilted grating, the instructions said that it was followed 75% of the time by a right-tilted one and vice versa. The purpose of these conditions was to see how explicit information interacted with the learning of the regularities, both when agreeing and when disagreeing. The I^-R^+ condition has some similarities with the I^+R^0 condition as in both the instructions were misleading. However, it is possible that learning the regularities

of an environment employs different mechanisms from learning that there are no regularities or learning to simply ignore the instructions.

We expected that in all conditions but the implicit one (I^0R^+), the instructions would create a prior belief in the participants that would bias their responses at the beginning of the experiment. In the I^+R^0 condition, participants would gradually learn that the information provided was false and they would end up with no biases. In the rest, participants would gradually learn the actual regularities. This would result in the I^+R^+ condition forming the strongest priors by the end of the experiment, followed by the I^0R^+ condition. Prior development in the I^-R^+ condition would be hindered by the instructions but would eventually update towards the regularities. We also hypothesised that participants with high AQ scores would develop weaker priors in the implicit, I^0R^+ condition, while they would show no differences in the I^+R^+ condition, as they would be aided by the instructions.

Computational Modelling

Besides simple statistical methods, we modelled the evolution of the participants' prior beliefs using the Hierarchical Gaussian Filter [HGF, 23]. This is a Bayesian model which tracks the evolution of beliefs over time. Its most common version consists of three levels. The first level (x_1) represents the probability of a left- or right-tilted grating for a specific trial, the second one (x_2) represents the association between the tones and the gratings over time, and the third one (x_3) the volatility of the environment, i.e. the rate of change of the association. The output of the HGF is x_1, which corresponds to the participant's prior belief after they hear the tone. We amended the base model to combine this probability with the visual information from the stimulus or 'likelihood' according to Bayes' rule to give the posterior probability for the stimulus. To calculate the likelihood we used a logistic sigmoid function on the log contrast,

likelihood = 0.5 + 0.5 / (1 + exp(log(contrast) – c_0)), with c_0 functioning essentially as a baseline contrast level that was fit to the data.

Given that in our experiment the regularities do not change over time, it is possible that the inclusion of the volatility level is unnecessary. We compared both 3-level and 2-level versions of the model, as well as versions with different sets of fixed and free parameters. The only parameters that were not part of the model comparison was the starting point of the second-level beliefs, which was set at 0 for the I^0R^+ and was a free parameter for the other conditions, and the standard deviation of the second level beliefs before the first trial. This choice was made because these were the direct objects of our study and as they would affect only the first few trials, it was possible that they would not survive model comparison. The different versions of the HGF, as well as a baseline Rescorla-Wagner model, were compared using Bayesian model selection [24] in each of the four conditions.

Results

Behavioural Analysis

The main expected influence of the prior was to bias participants towards the tilts that were associated with the tones. This would result in higher accuracy when tilts and tones were congruent and lower when they were not (Figure 2). By 'congruent', we refer to trials where the tone-tilt pairings correspond to the most likely pairing (appearing in 75% of the trials, in the I^0R^+, I^+R^+, and I^-R^+ conditions) or to the pairing participants were told would appear in 75% of the trials (in the I^+R^0 condition). Note that in the I^-R^+ condition we are defining 'congruent' according to the actual regularities and not the instructions. Specifically, we expected that in the I^0R^+ condition participants would start with no difference between congruent and incongruent trials and would gradually develop a bias during the experiment.

On the other hand, in the I^+R^0 condition, participants would start being strongly biased and then they would realise that there are no regularities and gradually reduce or even eliminate their bias. Then, the I^+R^+ condition would be a combination of the I^-R^+ and explicit ones. Participants would start biased and would remain at the same level throughout the experiment. Finally, in the I^-R^+ condition, they would start being biased in the opposite direction and then would gradually develop a bias in the correct direction, albeit never reaching the magnitude of the bias of the I^+R^0 or I^+R^+ conditions.

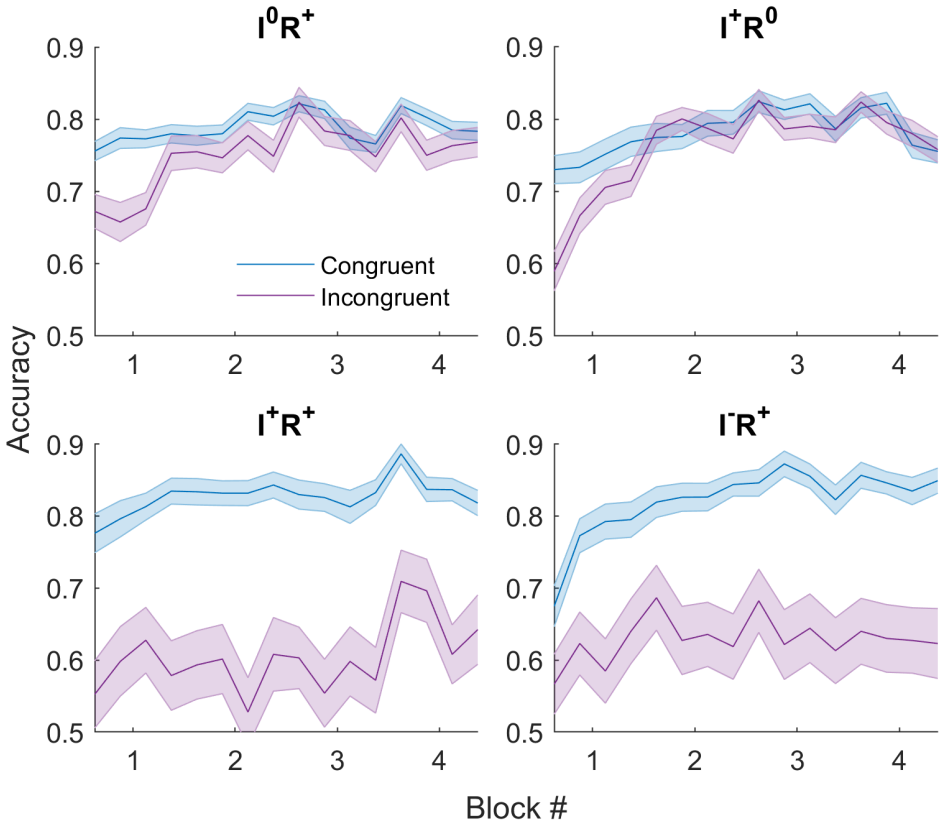


Figure 2: Accuracy comparison between congruent and incongruent trials in the four conditions. *Participants showed small accuracy differences in the implicit condition, large accuracy differences in the two conditions where explicit information was combined with regularities, and no differences in the I^+R^0 condition after the first block.*

Looking at the accuracy differences (Figure 3), we see that after the first block participants showed biases in the direction of the actual environmental regularities. They had significantly higher accuracy in the congruent trials than in the incongruent ones in the I^+R^0 (mean diff = 0.024, $t(118) = 2.73$, $p = 0.007$), I^+R^+ (mean diff = 0.226, $t(50) = 6.46$, $p = 10^{-8}$), and I^-R^+

conditions (mean diff = 0.204, $t(58) = 5.04$, $p = 10^{-6}$), but not in the I^+R^0 condition (mean diff = 0.004, $t(105) = 0.46$, $p = 0.65$). Biases in the I^+R^+ and I^-R^+ conditions were larger than those in the I^0R^+ condition ($t = 7.59$, $p = 10^{-12}$ and $t = 5.78$, $= 10^{-8}$, respectively), but they did not differ among each other ($t = 0.39$, $p = 0.69$). Reaction times did not significantly differ between the congruent and the incongruent trials in any condition, although there was a tendency for faster congruent reaction times in the I^+R^0 and I^+R^+ conditions (I^0R^+: $t(118) = -0.82$, $p = 0.42$; I^+R^0: $t(105) = -1.39$, $p = 0.17$; I^+R^+: $t(50) = -1.60$, $p = 0.12$; I^-R^+: $t(58) = -0.21$, $p = 0.84$).

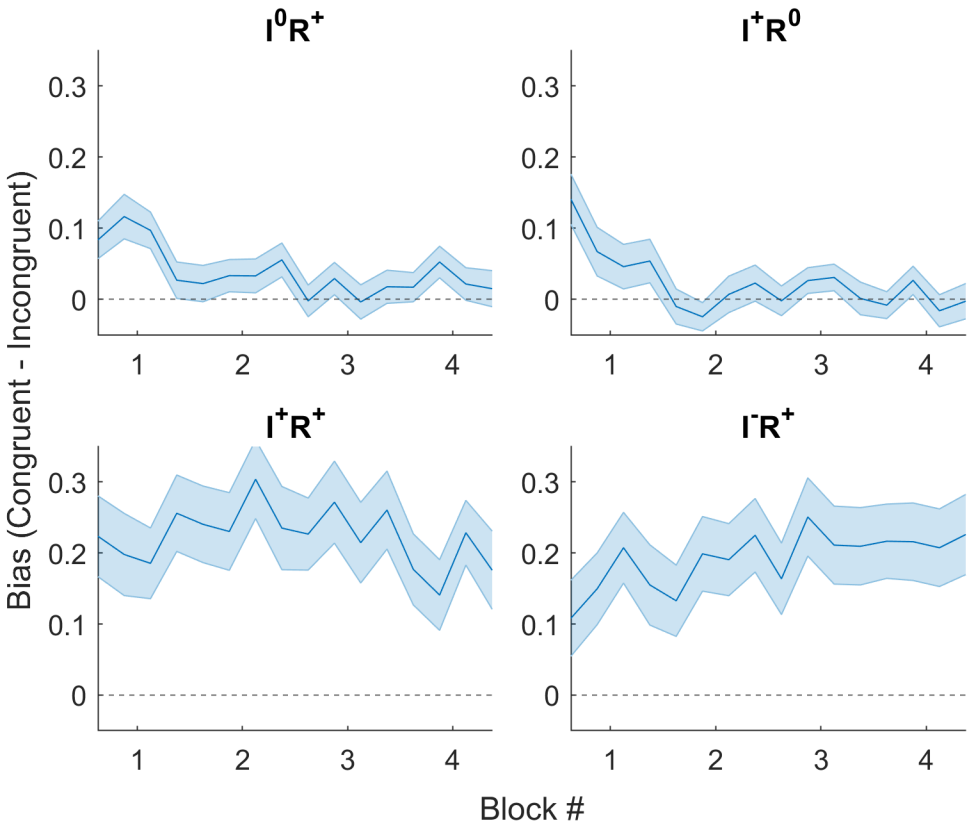


Figure 3: The evolution of bias over time in the four conditions.

From Figures 2 and 3, it appears that participants were able to learn the regularities within the first block, at least partially, and they did not update their beliefs in the following trials. This was verified by fitting the participant biases to the block numbers by linear regression. No progression was shown in the I^0R^+ ($t = -0.47$, $p = 0.65$) and I^+R^0 conditions ($t = 0.34$, $p = 0.74$), while a weak negative and a weak positive relationship was found in the I^+R^+ ($t = -2.35$, $p =$

0.041) and $I^{-}R^{+}$ conditions ($t = 2.07$, $p = 0.065$), respectively. However, as the staircase which determined the contrast changed between the main experiment and the training, it could be that any apparent change in the bias is due to the contrast slightly increasing during the experiment (Figure 1 in S1 Supplementary Information). When the contrast is lower, the participants have to rely on their priors more and any potential biases are stronger. This would also explain why the bias in the $I^{+}R^{0}$ condition starts relatively high and then drops within the first block. To properly account for the effects of contrast we need to analyse the data using computational modelling.

Computational Analysis

Comparing different models (Section 2 in S1 Supplementary Information), we found that the best ones were very similar across conditions, being all versions of the HGF and having mostly the same free parameters. In all conditions, the winning model assumed that the environment stayed stable over time. That is, x_3 was set at $-\infty$ and the change in second-level beliefs resulted only from acquiring more information from the environment. The only difference between the models was the starting point of the second-level beliefs, which was set at 0 in the $I^{0}R^{+}$ condition. All of the winning models also had k_1 as a free parameter. k_1 mediates the relationship between second and first-level beliefs, with $x_1 = s(k_1 \cdot x_2)$ and $s(\bullet)$ being the logistic sigmoid. Essentially, it allows to differentiate between the brain's best estimate of the regularities (x_2) and what ends up influencing perceptual decisions (x_1). Additionally, high k_1 values result in higher precision for the second-level beliefs. One can think of k_1 as encoding the confidence participants have in their beliefs about the regularities. Low confidence means that they assign lower precision to them, leading to larger updates, and they use them less during perception.

We analysed the trajectories of estimated first-level priors used by the participants (x_1). The results confirm the behavioural findings (Figure 4). Participants initially formed a belief based on the instructions, so their prior at the first trial differed from the uniform prior 0.5 (I^+R^0: mean prior = 0.61, t = 4.3, $p = 10^{-5}$; I^+R^+: mean prior = 0.57, t = 2.7, p = 0.01; I^-R^+: mean prior = 0.44, t = –2.0, p = 0.052). But when that belief disagreed with the actual regularities, it was quickly amended (I^+R^+ vs I^-R^+, blocks 2-4: t = 1.1, p = 0.25). Again, we see that priors in both the I^+R^+ and the I^-R^+ conditions were much stronger than priors in the implicit condition (t = 7.9, $p = 10^{-13}$ and t = 5.2, $= 10^{-6}$).

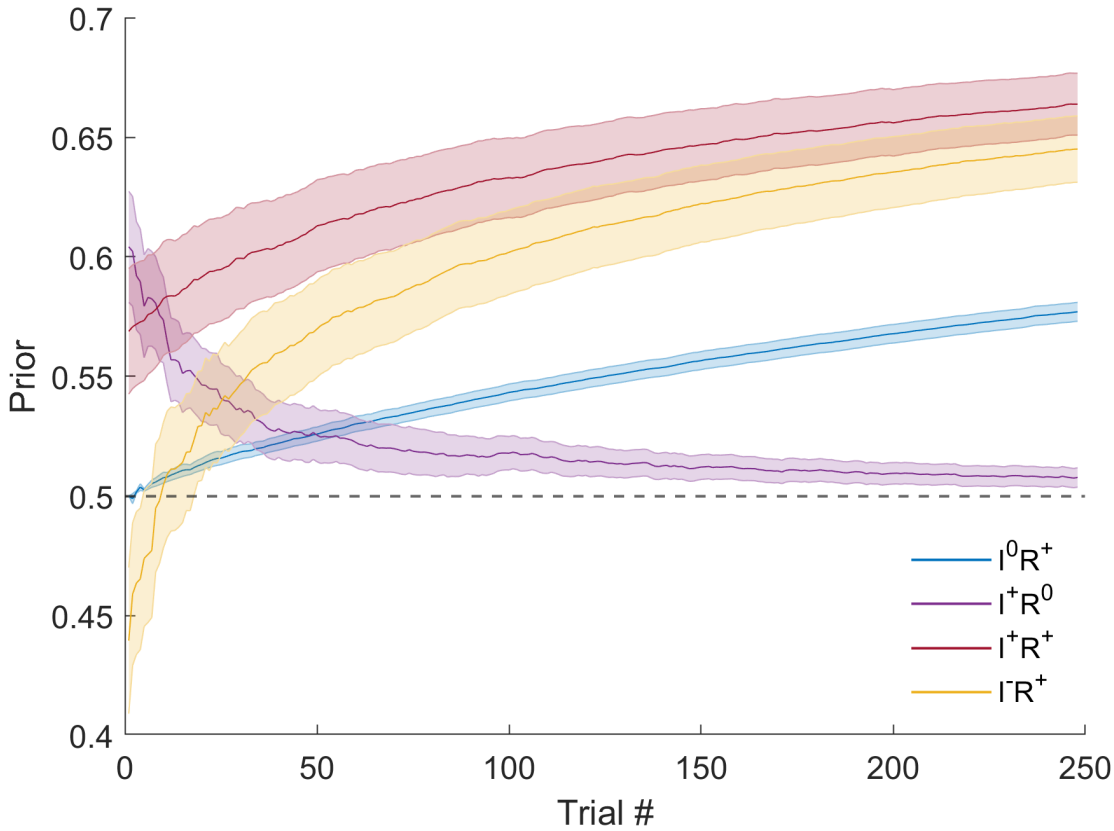


Figure 4: The trajectories of model-estimated first-level priors in the four conditions. *Priors were significantly affected by the instructions in the beginning of the task, but that was quickly amended based on the regularities in each condition. Priors in the conditions that combined explicit information and regularities were much stronger than those in the implicit condition.*

One difference with the behavioural findings is that the computationally estimated first-level prior in the I^+R^0 condition never reaches 0. Moreover, while the biases in Figure 3 remained relatively stable after the first block, Figure 4 shows the priors increasing in three out of four

conditions even in the final trials. Also note that the estimated priors start at different points among the explicit, I^+R^+, and I^-R^+ conditions, despite the instructions being identical. We suspect that these effects are artifacts cause by the structure of the HGF and do not reflect the participants' cognitive processes.

We also investigated the second-level learning rates of the participants across conditions (Figure 5), which are equal to variance of second-level beliefs. In all conditions, learning rates decreased during the task (1st block vs 4th block, $t > 12$, $p < 10^{-22}$), as participants got more information about their environment and formed more precise beliefs. The learning rates of the I^0R^+ condition were the highest ($t > 4.7$, $p < 10^{-5}$) and those of the I^+R^0 were the lowest ($t > 3.6$, $p < 0.001$), while the learning rates of the I^+R^+ condition were not significantly different to I^-R^+ ($t = 1.5$, $p = 0.14$).

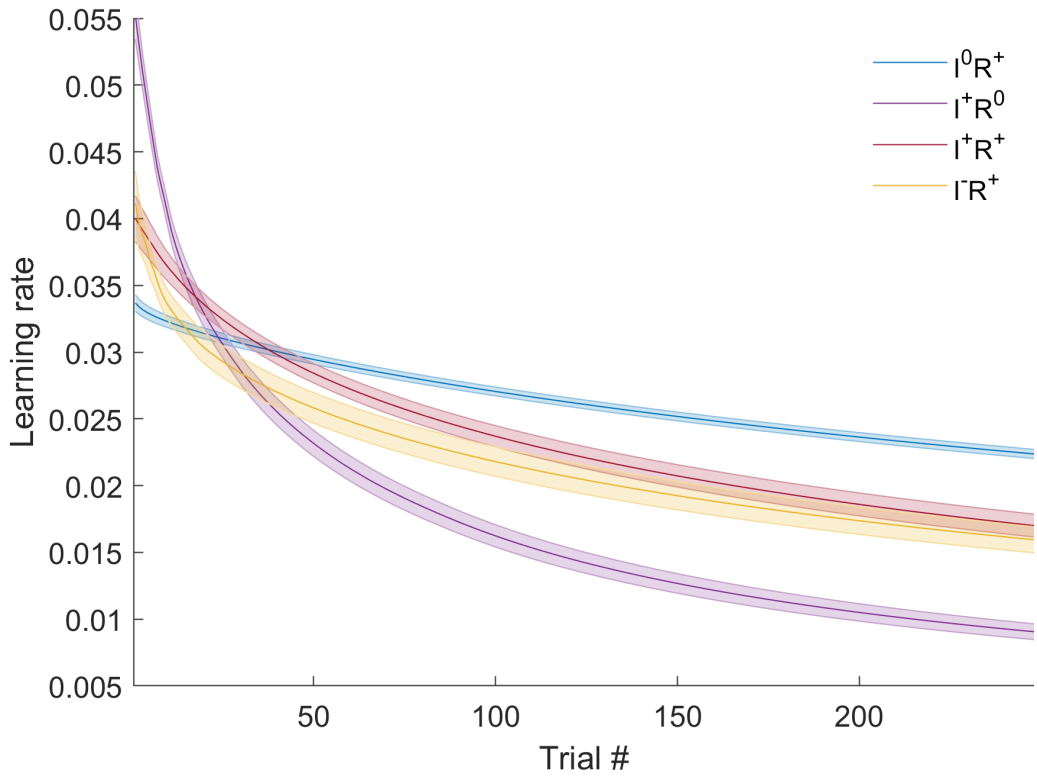
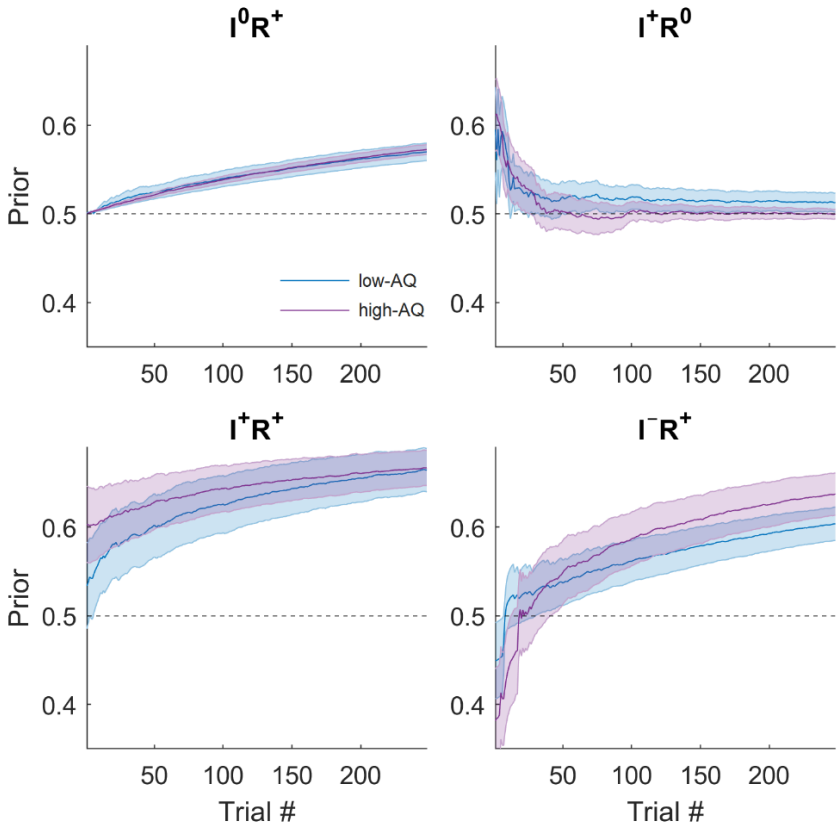


Figure 5: The evolution of second-level learning rates (variances) over time in the four conditions. *In all conditions the second-level priors become less uncertain during the task, leading to a decrease in learning rates.*

Autistic traits

To investigate the relationship between autism and prior beliefs, we looked at correlations between our results and AQ. Pearson correlations are very vulnerable to outliers. To avoid that we first selected all outliers following the interquartile range method. Then we winsorised the data, meaning that we equated the outliers with the most extreme non-outlier values [25]. This allowed us to keep some of the information in the outliers, while mitigating their outsized influence.

No relationship between AQ and biases, model-estimated first-level priors, likelihoods, or model parameters were significant (Figure 6 & Figure 7). Despite that, some measures showed non-significant tendencies for weak correlational relationships with AQ. We report them here as potential directions for future investigation. Specifically, biases in the final three blocks of the I^+R^0 condition (after they had stabilised) were marginally anticorrelated with AQ ($r = -0.17$, $p = 0.08$), a relationship that was more weakly present in the model-estimated priors ($r = -0.14$, $p = 0.15$). Parameter correlations showed a tendency for lower k_1 with higher AQ in the



I^+R^0 condition ($r = -0.18$, $p = 0.07$) and the I^+R^+ condition ($r = -0.2$, $p = 0.16$), but the reverse tendency in the I^-R^+ condition ($r = 0.18$, $p = 0.17$).

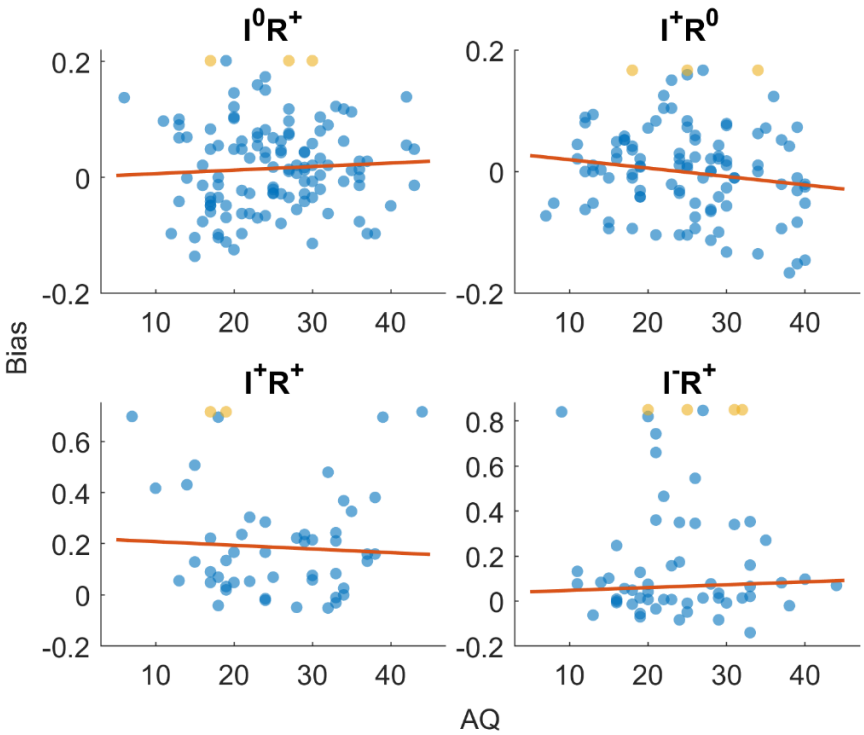


Figure 7: Correlations between autistic traits and response biases after the first block. *Orange dots correspond to winsorised outliers, and the red line to a robust linear regression fit. No correlations were significant.*

We also investigated the relationship between AQ scores and learning rates or prior variances (Figure 8). The resulting Pearson correlations showed a marginal effect in the I^+R^0 condition ($r = -0.19$, $p = 0.06$) and a significant positive correlation in the I^0R^+ condition ($r = 0.19$, $p = 0.03$). Note that this would not survive a correction for multiple comparisons. A comparison of implicit learning rates of the top versus the bottom AQ quartile in our data showed the same tendency, but was not significant ($t = 1.7$, $p = 0.09$).

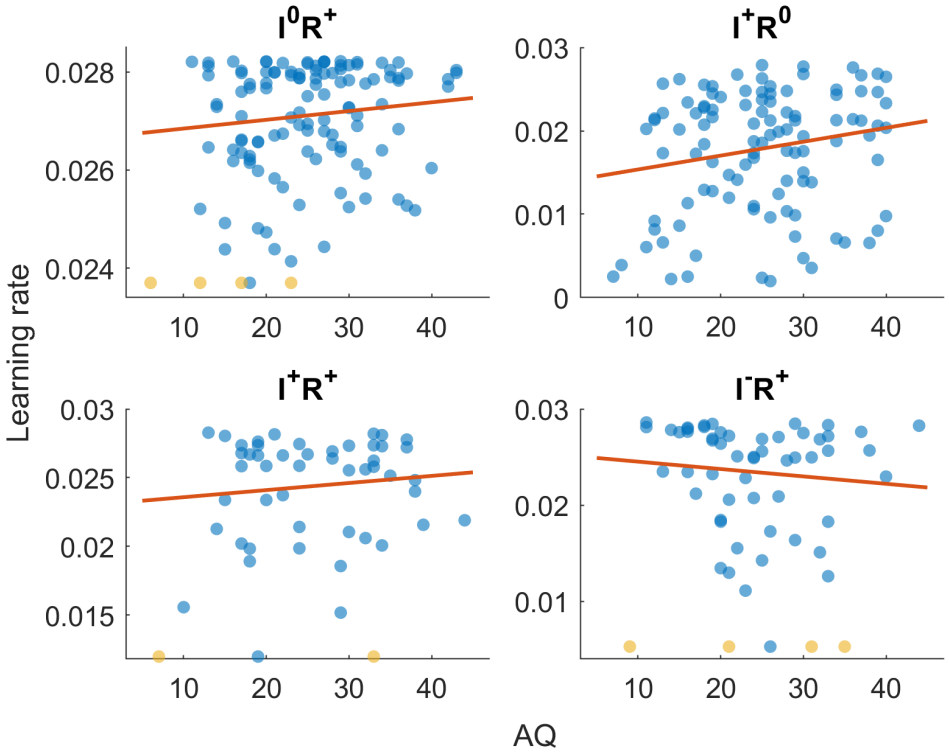


Figure 8: Correlations between AQ and average second-level learning rates (variances). *Orange dots correspond to winsorised outliers, and the red line to a robust linear regression fit. In the implicit condition the correlation was* $r = 0.19$, $p = 0.03$.

Discussion

Implicit and explicit learning

In this study, we explored how the learning of probabilistic associations between auditory cues and visual stimuli is impacted by the presence and the content of explicit instructions and how it is affected by autistic traits. We designed four experimental conditions that differed in the presence of associations and the veracity of the instructions given to participants. We expected participants to initially form a Bayesian prior based on the instructions of the task, which would then be updated following the statistics of the stimuli. Indeed, our results showed that participants were influenced by both the information explicitly provided to them and the regularities of the task.

Specifically, when participants were not alerted to the presence of any association between the cue and stimulus, they rapidly formed priors in line with the environmental regularities. However, through Bayesian modelling, we estimated these priors to be significantly weaker than the actual regularities (55% vs 75%). In another condition, where participants were falsely informed about the presence of an association, they initially showed a noticeable bias that diminished rapidly as they progressed in the task. Remarkably, in conditions where implicit and explicit information coexisted, participants developed more robust priors in the direction of the regularities (60%-65%), independently of the veracity of the instructions. This comes in contrast to our expectations, with misleading instructions having a positive effect on the participants' learning of the regularities. It also contrasts with previous work which has shown that searching for regularities inhibits the learning of environmental statistics [7,14,15].

Compared to previous studies, one difference with our task is that past work has mostly used serial reaction time and artificial grammar learning tasks. In both of these, the regularities that participants had to learn were rule-based, as opposed to our own task where regularities were frequency-based. For example, in some of these tasks, there would be a numeric sequence repeating (e.g., 1-1-3-1-4-2-2-4-1) and participants would have to respond as quickly as possible to each number by pressing the assorted key. In others, participants might have to learn that words with specific substrings are valid (e.g., VSXS and XXV) and others are not. In both of them, the regularities of the environment are far more precise (akin to rules) than a simple difference in the frequence of some stimuli or some probabilistic association between them. One would naively expect that being aware of the presence of regularities and consciously searching for them (explicit learning) is better suited for precise rules, while unconscious or implicit learning is better suited for general statistical patterns. However, this is the opposite pattern of what the results of this and past studies suggest. One possible explanation for the observed detrimental effect of instructions in the rule-based tasks could be that their exact

regularities are too complicated to be learned during a single experiment. At the same time, implicit learning might be able to extract some useful patterns, even if these are far from the exact rules. In the aforementioned example sequence, implicit learning might result in a general prior of the form ‘3s and 1s go together, 2s and 4s go together’. This obviously cannot explain the full sequence, but it is nonetheless useful. In contrast, participants who explicitly search for the exact form for the regularities might disregard incomplete rules, while dedicating many resources to discovering the exact pattern, leading to worse performance and no discernible benefits. In our tasks, on the other hand, explicit instructions would be perfectly able to provide an understanding of the true regularities, especially given that instructions explained the general nature of these patterns, if not their actual form.

We also compared the learning rates of participants across conditions. In the Hierarchical Gaussian Filter, the second level models the beliefs of the participants about the regularities. The learning rates of that level are equal to the variance of the beliefs, meaning that the more participants are uncertain about their priors the faster they update based on new information. Our results showed that the smallest learning rates after the first block were in the condition where misleading information was provided about the cue-stimulus association but no association was present, as participants quickly realised that there were no regularities. Conversely, when participants were given no explicit information, their uncertainty in their beliefs decreased significantly less with new stimuli, leading to the highest learning rates in the implicit condition after the first block. In the conditions where participants learned explicitly, they quickly formed beliefs about the association within the first block which then remained relatively stable, as their learning rates dropped.

Our results demonstrate that instructions can substantially influence participants' statistical learning, even when the information provided is inaccurate. Researchers should exercise caution when formulating task instructions, as merely directing participants' attention to

regularities can alter the nature of their responses. Future studies should clearly specify the type of learning they aim to evoke, whether implicit or explicit, as these processes likely rely on distinct mechanisms. Moreover, researchers should document the instructions given to participants with the same level of detail as the task design itself.

Autism

We investigated the relationship between autistic traits and task behaviour. Our hypothesis was that autism would be associated with weaker biases, primarily in the implicit condition. This was refuted by our results, as they showed no relationship between autistic traits and the magnitude of Bayesian priors or likelihoods in any condition. The only condition that showed a tendency for weaker biases was the one where no actual regularities were present. On a surface level these findings challenge common theories that suggest reduced biases in autism, attributed to either flatter priors or more precise likelihoods. However, a closer look reveals a more nuanced picture.

Our findings showed a significant positive correlation (before corrections for multiple comparisons) between autistic traits and learning rates in the implicit condition. In our version of the HGF the learning rate of a level is equivalent to that belief's variance. This means that prior beliefs about the regularities were more uncertain with stronger autistic traits, agreeing with the main hypothesis of flatter priors in autism [17]. It also offers additional support for the more recent theories of impaired volatility processing in autism [20] and specifically overestimating that volatility [19]. Believing that environmental volatility is high would lead to less certainty about current beliefs and therefore greater propensity to update them. Below we attempt to explain how flatter priors do not necessarily result in weaker response biases and

how both the theories of Pellicano & Burr and Palmer et al. might be correct when most studies show no bias differences in autism [21].

In tasks such as the one presented in this paper, prior beliefs about individual trials result from beliefs about task regularities but are distinct from them. Individual trial priors can be represented by a single number: after the tone is heard, how likely is it that the grating will be tilted to the left? Priors closer to 0 or 1 correspond to large certainty about what the stimulus would be (left or right). Priors at 0.5 are fully uncertain, not biasing responses towards any direction and completely amenable to new information. Prior beliefs about the task regularities, on the other hand, cannot be represented in the same way. These beliefs consist of *distributions* over these probabilities. In that case, a strong, certain prior is one that is more concentrated around a specific value, independently of what that value is. A completely uncertain prior would be one that is uniform over all possible values. Take the examples in Figure 9. There we see a strong prior that the experiment has no regularities and a weak prior that there is a 65% association between cue and stimulus. Despite the latter being a less certain belief about the regularities, it would result in a *stronger* prior for individual trials. This is because the precision of beliefs about regularities determines how these priors are updated, not their influence during perception. To give another example, a strong belief that a coin is fair would result in a very weak belief about the coin landing heads or tails in an individual flip.

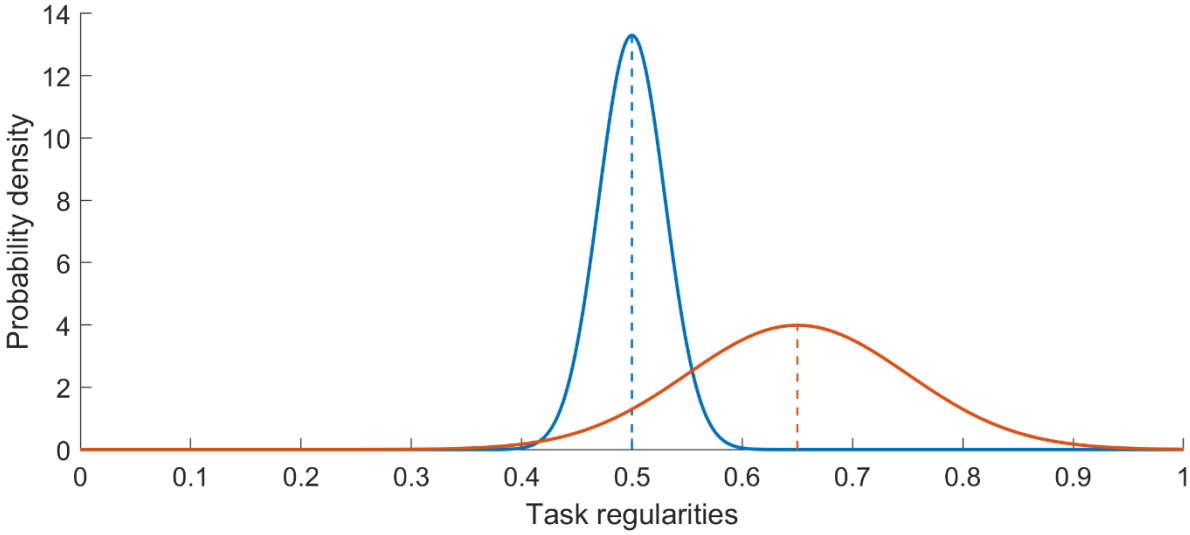


Figure 9: Examples of beliefs about task regularities. *The blue line corresponds a strong belief around 0.5, the red line corresponds to a weak belief around 0.65. Dashed lines show the mean of the distributions. In the Hierarchical Gaussian Filter these priors are encoded in the second-level of the model and their mean is used to form the first-level prior for the individual trials.*

In the HGF priors about individual trials are represented by the first level of the model, while prior beliefs about regularities are represented by the second level. Our results showed no correlation between autistic traits and first-level priors or the corresponding response biases, but offered evidence for weaker, flatter second-level priors. This supports the original hypothesis of Pellicano & Burr [17], but also supports the hypothesis of overestimated volatility in autism, as believing that the environment is volatile would decrease the certainty in one's internal model of the regularities and would result in faster updating. It also shows how simple measures of response biases commonly employed in behavioural experiments are unable to detect such differences [21].

Building on that, our results showed that this relationship goes away when participants are alerted about the presence of regularities. Therefore, it is possible that individuals with autism use explicit strategies to compensate for implicit learning differences. In the seminal paper by Lawson et al. [19], participants were informed about the presence of regularities and that these would change over time. Their results showed no relationship between diagnoses and the second-level volatility estimates of the participants (ω_2). However, participants were not

instructed that the rate of change would itself change during the task. And indeed the authors show differences across diagnostic groups in how the learning rates changed during the task and higher third-level volatility estimates (ω_3) in ASD.

References

1. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences. 2004 Dec;27(12):712–9.

2. Forkstam C, Petersson KM. Towards an explicit account of implicit learning: Current Opinion in Neurology. 2005 Aug;18(4):435–41.

3. VanPatten B, Smith M. Explicit and implicit learning in second language acquisition. Cambridge, United Kingdom New York, NY: Cambridge University Press; 2022. 49 p. (Elements in second language acquistion).

4. Foerde K, Knowlton BJ, Poldrack RA. Modulation of competing memory systems by distraction. Proc Natl Acad Sci USA. 2006 Aug;103(31):11778–83.

5. Maddox WT, Ashby FG. Dissociating explicit and procedural-learning based systems of perceptual category learning. Behavioural Processes. 2004 Jun;66(3):309–32.

6. Critchley H, Daly E, Phillips M, Brammer M, Bullmore E, Williams S, et al. Explicit and implicit neural mechanisms for processing of social information from facial expressions: A functional magnetic resonance imaging study. Hum Brain Mapp. 2000 Feb;9(2):93–105.

7. Fletcher PC, Zafiris O, Frith CD, Honey RAE, Corlett PR, Zilles K, et al. On the Benefits of not Trying: Brain Activity and Connectivity Reflecting the Interactions of Explicit and Implicit Sequence Learning. Cerebral Cortex. 2005 Jul 1;15(7):1002–15.

8. Hazeltine E. Attention and stimulus characteristics determine the locus of motor-sequence encoding. A PET study. Brain. 1997 Jan 1;120(1):123–40.

9. Gebauer GF, Mackintosh NJ. Psychometric intelligence dissociates implicit and explicit learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2007;33(1):34–54.

10. Reber AS, Walkenfeld FF, Hernstadt R. Implicit and explicit learning: Individual differences and IQ. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1991;17(5):888–96.

11. Liu H, Forest TA, Duncan K, Finn AS. What sticks after statistical learning: The persistence of implicit versus explicit memory traces. Cognition. 2023 Jul;236:105439.

12. Musen G, Treisman A. Implicit and explicit memory for visual patterns. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1990;16(1):127–37.

13. Caljouw SR, Veldkamp R, Lamoth CJC. Implicit and Explicit Learning of a Sequential Postural Weight-Shifting Task in Young and Older Adults. Front Psychol [Internet]. 2016 May 24 [cited 2024 Feb 12];7. Available from: http://journal.frontiersin.org/Article/10.3389/fpsyg.2016.00733/abstract

14. Reber AS. Implicit learning of synthetic languages: The role of instructional set. Journal of Experimental Psychology: Human Learning and Memory. 1976 Jan;2(1):88–94.

15. Song S, Howard JH, Howard DV. Implicit probabilistic sequence learning is independent of explicit awareness. Learn Mem. 2007 Mar;14(3):167–76.

16. Lawson RP, Rees G, Friston KJ. An aberrant precision account of autism. Front Hum Neurosci [Internet]. 2014 May 14 [cited 2020 Nov 27];8. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00302/abstract

17. Pellicano E, Burr D. When the world becomes 'too real': a Bayesian explanation of autistic perception. Trends in Cognitive Sciences. 2012 Oct;16(10):504–10.

18. Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, et al. Precise minds in uncertain worlds: Predictive coding in autism. Psychological Review. 2014;121(4):649–75.

19. Lawson RP, Mathys C, Rees G. Adults with autism overestimate the volatility of the sensory environment. Nat Neurosci. 2017 Sep;20(9):1293–9.

20. Palmer CJ, Lawson RP, Hohwy J. Bayesian approaches to autism: Towards volatility, action, and behavior. Psychological Bulletin. 2017;143(5):521–42.

21. Angeletos Chrysaitis N, Seriès P. 10 years of Bayesian theories of autism:a comprehensive review. Neuroscience & Biobehavioral Reviews. 2022 Dec;105022.

22. Kourkoulou A. Implicit learning of spatial context in adolescents and adults with autism spectrum disorder. Durham University; 2010.

23. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci [Internet]. 2014 Nov 19 [cited 2020 Jun 12];8. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00825/abstract

24. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. 2009;14.

25. Wilcox RR. Some results on a Winsorized correlation coefficient. Brit J Math & Statis. 1993 Nov;46(2):339–49.