THE UNIVERSITY *of* EDINBURGH

Examining and Extending Bayesian Theories of Autism

Nikitas Angeletos Chrysaitis

Doctor of Philosophy

Institute for Adaptive and Neural Computation

School of Informatics

University of Edinburgh

2025

Abstract

Recent approaches to understanding autism view it through the lens of the Bayesian brain framework. Within this framework, it is hypothesized that beliefs about the world are represented probabilistically in the brain and are updated using Bayesian statistics. Perception is understood a process of inferring the most likely cause of sensory input by combining sensory information with prior beliefs about the environment. Bayesian theories of autism propose that its heterogeneous symptomatology arises from an imbalance in the weighting or 'precision' of priors and likelihoods, favouring the latter. While this is a promising approach, the relevant research is highly diverse, occasionally resulting in contradictory findings, with some studies finding no evidence of such imbalances in autism and others showing more nuanced and complex results.

The aim of this work is twofold. First, we conducted a comprehensive examination of previous research, starting by conducting a systematic review of Bayesian studies of autism and autistic traits. The results were mixed, with a slight majority of studies finding no difference in the integration of Bayesian priors and likelihoods. Methodologically, many studies had low statistical power and inconsistent approaches. In a subsequent paper, we clarified the meaning of 'sensory precision', a central term in Bayesian theories of autism. Using a simple Bayesian perception model, we examined the role of two possible interpretations in the inferential process and their relevance to autism theories. We reanalysed data from an old study under this new light.

The second part of this work expands upon the main Bayesian theories of autism to better approximate the behaviours of autistic individuals. Informed by the literature review, we designed an experiment to compare implicit and explicit learning across autistic traits, manipulating the presence and content of instructions across conditions. We found that the presence of instructions exerted a significant influence on participant priors until the end of the task, while the effect of their content had only a weak, temporary influence. Our results showed no relationship between autistic traits and response biases, but offered hints of a weak correlation between autistic traits and participant uncertainty about implicit environmental regularities. In another study, we applied the circular inference model, a model of belief overconfidence, to autistic traits. Our findings did not reveal

any significant differences in circularity across the autism spectrum. To further explore this model in a more autism-relevant context, we designed a follow-up task that incorporated social elements. Surprisingly, the pilot results indicated that participants did not engage in Bayesian inference, instead they simply averaged the information obtained from different sources.

These results show that, while imbalance theories of autism are prominent within the field, their simplest form is weakly supported by experimental evidence. Both the literature review and our study on implicit and explicit learning highlighted the need for more nuanced approaches that focus on prior development. We discuss two such theories of volatility processing impairments in autism, which imply higher uncertainty in strong autistic traits, as hinted by our experimental findings. We also discuss methodological issues commonly present in the field and the need for thorough computational modelling.

Lay summary

Our brains understand the environment using both information from the senses and previous knowledge of environmental statistics. Both types of information are probabilistic in nature and are integrated according to the rules of probability. This means that they are weighted based on their reliability, with more certain or 'precise' information prioritised relative to noisier information. In recent years, this approach has been extensively applied to mental illness, and particularly in understanding autism spectrum disorders. Prominent current theories of autism hypothesise that at the core of the condition lies an imbalance in this weighting, where sensory information is given disproportionate influence relative to past knowledge.

In this thesis, we thoroughly examine past research and extend current theories of autism through novel experimental work. Our literature review reveals that a slight majority of studies find no difference in how autistic individuals and those with strong autistic traits integrate sensory information with past knowledge. However, differences are more frequently observed when individuals are presented with new environmental regularities which are implicitly learned during the experiments. We also identify various methodological issues across studies, such as small sample sizes and inconsistent computational approaches. In a separate article, we explore the concept of 'sensory precision' and its varying definitions across the literature. We propose new, clarifying terminology, examine its implications for the theories of autism, and offer recommendations for future research.

Building on the findings from the literature review, we develop new experiments that test more nuanced theories of autism. In one experiment, we investigate circular inference, a mathematical model of overconfidence, but find no relationship with autistic traits. In a follow-up pilot experiment, we show that participants do not always integrate information using the rules of probability, highlighting the importance of careful experiment design in the field. Finally, we design an experiment to investigate how individuals with varying levels of autistic traits are influenced by the presence of instructions during learning. Our results show that the mere presence of instructions significantly influenced behavior across our sample by directing attention to environmental regularities, while their specific content had only minor effects on participant behaviour. Surprisingly, We do not find a

relationship between autistic traits and response biases. However, we do observe hints of a weak correlation between autistic traits and uncertainty about implicit environmental statistics.

Our findings suggest that the simplest form of current autism theories lacks strong experimental support. They underscore the need for more nuanced approaches that consider how individuals develop their expectations about the world in stable and volatile environments, and how this interacts with their awareness of the environmental regularities. This work contributes to the field by providing both methodological insights and empirical results that can help refine computational accounts of autism.

Acknowledgements

First and foremost I would like to thank my supervisor, Peggy Seriés, for her guidance and mentorship throughout this process. I am also grateful to Renaud Jardri for his help and insights, especially during the early years of my PhD. I would also like to thank Sophie Denève and Povilas Karvelis for our fruitful collaborations, and Matthias Henning and Stephen Lawrie for their feedback during my annual reviews. I am especially thankful to Stelios Gkionis and Magda del Rio for their feedback on parts of this work, for helping me better understand the research at hand, and for patiently listening to my frustrations. Finally, I would like to thank Iro Karountzou, for her eternal support.

Table of Contents

Chapter 1

General Introduction

1.1 Perception as Bayesian Inference

The goal of perception is to create an accurate representation of our environment. This can be challenging if relying exclusively on sensory inputs. Our senses are imperfect and inherently noisy, and even high-fidelity sensory information is often ambiguous (e.g., Figure 1.1, left). To more accurately infer the state of the environment, the brain combines sensory information with internal models of the world (Frith and Dolan, 1997; Vetter and Newen, 2014). For example, a common facet of such a model is the belief that light comes from above (Brewster, 1826; Sun and Perona, 1998). This belief is frequently used to resolve ambiguities in convexity-concavity judgements, where different objects would create different shading patterns (e.g., Figure 1.1, middle & right). In fact, a great variety of visual illusions can be explained as influences of such beliefs, such as the hollow mask illusion (Gregory, 1970), which results from a belief that faces are convex, or the Kanizsa triangle (Kanizsa, 1987), which is the product of a belief that simpler shapes are more likely. The effects of internal models extend far beyond that. Even the simplest perceptual judgements are influenced by the brain's unconscious beliefs about the environment, such as orientation discrimination which is biased towards the cardinal orientation (Girshick et al., 2011) or speed estimation which is biased towards slower speeds (Stocker and Simoncelli, 2006).

This idea, that perception is inferring the state of the environment and uses previous

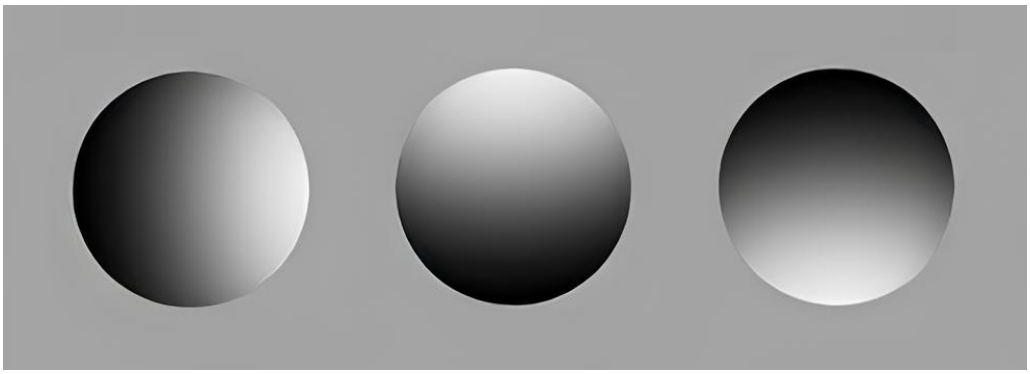

Figure 1.1: An illustration of the 'light from above' prior. The curvature of the left circle is unclear. Taking into account the fact that light usually comes from above can resolve this ambiguity, making the middle circle appear convex and the right one concave.

knowledge to achieve that, has been around at least since (von Helmholtz, 1866), but it rose to prominence with its modern formulation, the Bayesian Brain (Knill and Pouget, 2004). This hypothesises that the brain represents both the incoming sensory data and its beliefs about the world in a probabilistic manner and then combines them using Bayes' rule,

$$P(S \mid D) \propto P(D \mid S)P(S), \tag{1.1}$$

with S corresponding to a stimulus and D to the incoming data produced by that stimulus. $P(S)$ is the belief about the stimulus which originates in the brain's internal model of the world, called the prior in Bayesian terms. $P(D \mid S)$ represents the likelihood of the data for each possible stimulus value. $P(S \mid D)$ is the posterior, the updated belief based on the new data, based on which the percept is determined.

This formulation assumes not only that sensory inputs are combined with prior beliefs, but also that this process takes into account the uncertainty of each source of information. Priors and sensory inputs exert influence on the posterior proportional to the inverse of their uncertainty, the precision. This predicts that the percepts would be further away from the sensory inputs in situations where the individual holds strong beliefs about the environment or where the sensory information is noisy. Indeed, various studies have shown that the effects of the prior become stronger when sensory noise increases (e.g., Powell et al., 2012; Sotiropoulos et al., 2014) and that different sensory inputs are combined based on their precision (e.g., Ernst and Banks, 2002).

The Bayesian brain is a theory at the computational level of Marr's hierarchy (Marr, 1982). It focuses on what is computed – a Bayesian posterior – but it does not tell us how that computation might take place in a neural network such as the brain. The most prominent algorithmic counterpart to the Bayesian brain is Predictive Coding (Friston and Kiebel, 2009; Rao and Ballard, 1999; Aitchison and Lengyel, 2017). Predictive Coding conceptualises priors as predictions that the brain makes about the world, beginning at a high level in the cognitive hierarchy and being transmitted to the lower levels. Sensory information enters the hierarchy at the bottom level and is compared with predictions to calculating the prediction errors. These errors are then transmitted to the higher levels. At each level where these bottom-up and top-down streams meet, predictions are adjusted based on prediction errors; the higher the precision of the prediction error, the larger the adjustment. The Bayesian brain and predictive coding reconceptualise perception as a form of 'controlled hallucination' (Seth, 2019). Perception is no longer viewed as the direct result of sensory inputs, but as the product of the brain's beliefs or predictions about the environment, with sensory data being largely used to adjust these beliefs. In that way, perception stops being divorced from higher cognition, and is instead viewed as simply one of the processes that are used to model the environment. Bayesian principles of handling uncertainty and belief updating are used throughout these processes, from perception to categorisation, learning, and reasoning (Chater et al., 2010; Griffiths et al., 2008).

1.2 Bayesian theories of Autism

The Bayesian brain and predictive coding frameworks have been used extensively to understand and explain psychiatric disorders (Seriès, 2020). The most prominent example of the application of these theories is schizophrenia, which has been attributed to both an overweighting of sensory evidence at the expense of priors (Sterzer et al., 2018; Valton et al., 2017) and an overweighting of priors (Chambon et al., 2011; Powers et al., 2017). But Bayesian theories have been also proposed for depression (Huys et al., 2015), which is interpreted as a generalised pessimistic prior, anxiety disorder (Browning et al., 2015), which has been shown to correlate with impaired belief updating, and of course autism.

Autism is a neurodevelopmental condition that manifests in a wide variety of ways, making it challenging to explain based on one theory. Its symptoms are broad, including social difficulties, unusual sensory processing, and a preference for repetition (American Psychiatric Association, 2013; Cohen and Volkmar, 1997). In the social domain, autistic individuals often struggle to understand the nuances of the social context or other people's actions and motivations, potential leading to decreased interest in socialising and impairments in interpersonal communication (Baron-Cohen, 2000; Dawson et al., 1998). Sensory symptoms are usually divided into three categories: sensory overresponsivity, where individuals have strong negative reactions to specific sensory stimuli; sensory underresponsivity, where individuals have reduced sensitivity to some stimuli, such as pain; and sensory-seeking behaviours, where individuals are unusually preoccupied with specific stimuli (Hazen et al., 2014). Insistence on repetition could concern bodily movements, a strong need for routines and consistency, or restricted interests (Turner, 1999). On top of these symptom classes, other characteristics of autism include a heightened attention to detail (e.g., Baron-Cohen et al., 2009), executive function challenges (Pellicano, 2012), reduced susceptibility to some visual illusions (e.g., Happé, 1996), superior sound pitch memory (e.g., Stanutz et al., 2014), and better performance in visual search and discrimination tasks (e.g., O'riordan et al., 2001).

Due to this heterogeneity, previous theories of autism usually focused on one facet of the disorder. For example, Weak Central Coherence (Frith, 1990), one of the most prominent such theories views autism as an information processing style that prioritises details at the expense of a coherent whole. Autistic individuals, according to Frith, are biased towards featural or local information and have difficulties in global processing or the integration of high-order information. On the other side of the same coin, Enhanced Perceptual Functioning attributes autism to superior perceptual performance in low-level processing, another possible cause of the bias for local information (Mottron et al., 2006). Other theories of autism focus on different facets of the condition, such as hypersensitivities (Intense World Theory, Markram and Markram, 2010) or theory-of-mind deficits (Baron-Cohen et al., 1985).

The first Bayesian theory of autism was proposed by Elizabeth Pellicano and David Burr in 2012 (Pellicano and Burr, 2012). They hypothesised that autistic individuals

have flatter, less precise priors, which they termed 'hypo-priors'. In a reply to their article, Brock proposed an alternative theory in which likelihoods are more precise in autism (Brock, 2012). Two years later, Lawson et al. presented an analogous theory in predictive coding terms (Lawson et al., 2014). They suggested that the ratio of low-level to high-level precision is increased in autism. In other words, either predictions, situated high in the cognitive hierarchy, are less precise or prediction errors, coming from the lower levels, are more precise. They did not take a strong position on either of these possibilities, although they slightly leaned in favour of the latter. In the same year, Van de Cruys et al. also published a theory in the framework of predictive coding where they suggested that the precision of prediction errors is inflexibly high in autism (Van de Cruys et al., 2014).

All of these theories share a core prediction: that in autism, priors have weaker effects on perception. This seemingly explains a large part of autistic symptomatology. If priors can be thought of as a filter in perception, regularising sensory inputs towards what is expected from the environment, then weaker prior influences can explain sensory hypersensitivities and attention to detail. Less filtered sensory inputs are more extreme, more surprising, and they demand more attention from the individual. This could also potentially explain sensory seeking behaviours, where autistic individuals exhibit fascination with specific sensory stimuli. Weaker priors would be the reason for the social difficulties in autism, as social communication is often ambiguous making the reliance on previous knowledge paramount in navigating such situations. The insistence on repetition and routine could be seen as a coping strategy which avoids surprising sensory inputs by remaining in highly predictable environments. Lastly, weaker prior influences automatically explain the reduced susceptibility to visual illusions. The hierarchical nature of predictive coding also offers a way to view both top-down impairment theories (e.g., Weak Central Coherence) and bottom-up enhancement theories (e.g., Enhanced Perceptual Functioning) within the same framework. This allows for the formalisation of these theories, for example as reduced prior precision and increased sensory precision, respectively, and potentially provides the ability to experimentally distinguish between them.

In 2017, a new predictive coding theory of autism was published by Palmer et al. (2017)

They suggested that priors indeed exert less influence in autism, but that the reason for that was that autistic individuals overestimate the volatility of their environment. When the environment changes frequently, it makes sense to give less weight to prior expectations, as they were formed based on statistics that are no longer completely accurate. This theory potentially provides a more intuitive explanation of the insistence on repetition, as a way to minimise volatility.

Since 2012 there has been a multitude of studies investigating autism through a Bayesian or predictive coding lens. These came from different fields, including psychiatry, psychology, neuroscience, and informatics. They used simulations and both behavioural and neuroimaging experiments, focusing on auditory and visual perception, decision making, learning, and other modalities. The variety of the studies offers a significant amount of evidence for the effects of autism over a broad range of experimental designs. However, the same variety comes with a few important drawbacks. The differences between the approaches often do not allow for a direct comparison between the findings of the studies, making it difficult to draw overall conclusions. This is exacerbated by researchers of different fields using different language to understand and express their findings and drawing different conclusions.

1.3 Autism vs Autistic Traits

Historically, autism was viewed through a binary lens, with individuals classified as either autistic or not based on standardised clinical tools, primarily the Autism Diagnosis Observation Schedule (Lord et al., 2000) and the Autism Diagnostic Interview (Lord et al., 1994). However, contemporary approaches to psychopathology have shifted more towards dimensional frameworks (Krueger et al., 2018). This paradigm shift is reflected in autism research through the emergence of the Broader Autism Phenotype (BAP), which describes individuals exhibiting sub-clinical autistic characteristics (Sucksmith et al., 2011), and the increasing popularity of quantitative autistic trait measures (Mottron and Bzdok, 2020).

The assessment of non-clinical autistic traits in the general population typically relies on self-report questionnaires designed to yield a continuous measure of trait strength,

rather than categorical distinctions. The most prominent questionnaire among these is the Autism Spectrum Quotient (AQ; Baron-Cohen et al., 2001). This consists of five subscales: Social Skill, Attention Switching, Attention-to-Detail, Communication, and Imagination. The AQ was initially validated in individuals with high-functioning autism and has demonstrated satisfactory sensitivity and specificity in clinical populations (Woodbury-Smith et al., 2005; Austin, 2005).

The construct validity of autistic traits has faced scrutiny, as has the capacity of self-report questionnaires to accurately capture the autism spectrum (Abu-Akel et al., 2019; Sasson and Bottema-Beutel, 2022). Critics highlight limitations such as the differing interpretation of questionnaire items and distinct trait clustering patterns between autistic and non-autistic individuals (Gernsbacher et al., 2017; Mottron and Bzdok, 2020). In addition, a clinical diagnosis requires some form of functional impairment, which is not the case for continuous measures.

Nevertheless, behavioral research has provided substantial evidence supporting a dimensional approach to autism as a continuum (e.g., Constantino et al., 2004; Wiggins et al., 2012) that maintains stability throughout childhood development (Robinson et al., 2011). Heritability studies further support this, as non-diagnosed relatives of autistic individuals have stronger autistic traits (e.g., Constantino et al., 2006), which exist in a continuum in the general population (e.g., Constantino and Todd, 2003; Hoekstra et al., 2007). This aligns with current understanding of autism's genetic makeup, which is that rare de novo variants account for less than 10% of diagnostic variability, while common variants contribute more than five times that (Eyring and Geschwind, 2021).

The experimental studies in this thesis primarily use a dimensional approach, based on participant AQ scores. This offers practical advantages, as it allows for online data collection with large sample sizes, greater variety of experimental designs, and easier controlling of medications or comorbidities (Landry and Chouinard, 2019). Our analyses also included self-reported ASD diagnoses, to complement the autistic trait findings.

1.4 Aims and Outline of the Thesis

The aim of this thesis is twofold: to critically examine existing research and, based on that, to extend current Bayesian approaches to autism through novel experimental work. The thesis is structured into two corresponding parts, with chapters in each part arranged chronologically.

Part I comprises two chapters. **Chapter 2** presents a systematic review of the relevant literature from 2012 to 2021, including 83 experimental studies. In this chapter, we categorise and synthesise the finding of the studies, evaluate their statistical power and effect sizes, and assess potential publication bias. This work substantially informs our subsequent research. One insight from the review was that 'sensory precision', a central term in autism research, was often vaguely defined. **Chapter 3** addresses this issue through a theoretical investigation of sensory precision. We propose two specific definitions, demonstrate their effects in a simple Bayesian model, and discuss their implications for autism theories. This chapter also includes a re-evaluation of findings from an earlier study (Karvelis et al., 2018) under these new definitions and offers methodological recommendations for future research.

Part II consists of three chapters presenting original research. These were inspired by our Chapter 2 results which suggested that more nuanced theories of autism might be required to explain the full spectrum of experimental results. **Chapter 4** investigates circular inference across autistic traits through a behavioural experiment and corresponding computational modelling. Circular inference is an extension of the basic Bayesian model (Jardri and Denève, 2013; Jardri et al., 2017), based on the excitatory-inhibitory imbalances observed in schizophrenia, that are also present in autism. **Chapter 5** describes a follow-up study that used a social version of the Chapter 4 task, adapted from the work of Simonsen et al. (2021), which would potentially be more relevant to autism. Unexpectedly, our results revealed that pilot participants did not perform Bayesian inference when combining prior beliefs with new evidence. In light of these findings, **Chapter 6** focuses on a different insight from our systematic review: that autistic impairments are more prevalent in tasks involving prior development, particularly when participants are unaware of the environmental regularities. We present an audiovisual association experi-

ment that investigates implicit and explicit learning, and their differences across a broad range of autistic traits.

Part I

Examining

* * *

In the previous chapter, we introduced how Bayesian theories of perception and cognition have been applied to understand autism spectrum disorder. These theories propose that autism may arise from an imbalance between prior beliefs and sensory evidence, with autistic individuals relying more on sensory inputs at the expense of prior knowledge. This framework is intuitively appealing because it offers a potentially unified explanation for the diverse symptoms of autism through the single underlying mechanism of impaired Bayesian inference.

In the years since these theories were first proposed, the field has expanded considerably, with researchers investigating the hypotheses through a large variety of experimental designs and methodological approaches. This has made it challenging to maintain a comprehensive understanding of the evidence for and against these theories. Different studies have focused on different types of prior beliefs, used both behavioural and neuroimaging methods, and interpreted their results through slightly different theoretical lenses. Some examine pre-existing priors that participants already have before starting a task, while others examine new priors that are learned during experiments. Some have focused on simple perceptual priors, while others have investigated more complex social expectations. This diversity in approaches, while valuable for exploring different aspects of the theories, makes it difficult to draw clear conclusions about the validity of Bayesian accounts of autism.

The next two chapters critically examine the existing literature on Bayesian theories of autism, both the experimental evidence (**Chapter 2**) and the technical details of the Bayesian models (**Chapter 3**). Chapter 2, specifically, is a comprehensive review of all experimental Bayesian studies of autism and autistic traits published between 2012-2021. It categorises and synthesises the experimental findings based on the types of priors investigated, evaluates the statistical power and effect sizes of these studies, and assesses potential publication bias in the field. Additionally, it examines how different researchers have operationalised and tested Bayesian theories, revealing both commonalities and inconsistencies in their approaches.

This detailed examination of the literature not only helps evaluate the current state of

evidence for the Bayesian theories of autism we introduced in Chapter 1 but also informs the direction of future research, including our own experimental work presented in **Part II** of this thesis. It also compares the technical approaches of the studies and highlights areas of potential misunderstanding, something we expand on in **Chapter 3**.

Chapter 2

Systematic Review of the Bayesian ASD literature

This chapter consists of the postprint of a published paper: Angeletos Chrysaitis, N., & Seriès, P. (2023). 10 years of Bayesian theories of autism: A comprehensive review. Neuroscience & Biobehavioral Reviews, 145, 105022.

Chapter 2 Abstract

Ten years ago, Pellicano and Burr published one of the most influential articles in the study of autism spectrum disorders, linking them to aberrant Bayesian inference processes in the brain. In particular, they proposed that autistic individuals are less influenced by their brains' prior beliefs about the environment. In this systematic review, we investigate if this theory is supported by the experimental evidence. To that end, we collect all studies which included comparisons across diagnostic groups or autistic traits and categorise them based on the investigated priors. Our results are highly mixed, with a slight majority of studies finding no difference in the integration of Bayesian priors. We find that priors developed during the experiments exhibited reduced influences more frequently than priors acquired previously, with various studies providing evidence for learning differences between participant groups. Finally, we focus on the methodological and computational aspects of the included studies, showing low statistical power and often inconsistent approaches. Based on our findings, we propose guidelines for future research.

Keywords: autism, Bayesian brain, predictive coding, perception, learning

2.1 Introduction

Autism spectrum disorder (ASD) is a highly diverse neurodevelopmental condition. It is usually characterised by its social symptoms, such as theory of mind deficits, reduced motivation to socialise, or communication and social skills problems. Autistic people also experience a range of sensory atypicalities, including both hyper- and hypo-sensitivities to stimuli or a fascination with them, what has been termed 'sensory-seeking' behaviours. Moreover, they frequently exhibit repetitive behaviours ('stimming'), be it body movements or sound production, strong insistence on routine and sameness, narrow interests, increased attention to detail, and various executive dysfunctions (American Psychiatric Association, 2013). To further complicate this picture, studies have shown that autistic individuals are less susceptible to illusions (Happé, 1996), have better sound pitch memory than the general population (Heaton, 2003), and show superior performance in a variety of visual search and discrimination tasks (O'riordan, 2004). Given this heterogeneity, it is understandable that many attempts at explaining the condition have focused on individual symptoms (e.g. theory of mind deficits, Baron-Cohen et al., 1985), as it is obviously challenging for one theory to explain the full breadth of the autism spectrum. Intriguingly, however, Bayesian and predictive coding theories of perception and cognition attempt to do just that.

Both theories view perception as a combination of external information originating at the senses, and internal models, consisting of previous knowledge about the environment (Knill and Richards, 1996; Rao and Ballard, 1999). External information is considered to provide a bottom-up signal, 'moving' from lower levels in the cognitive hierarchy to higher ones, from the specific to the general. For example, bottom-up visual signals can propagate from the primary visual cortex, which has neurons that are sensitive to specific stimulus orientations, to the fusiform face area, which can recognise faces. Internal information on the other hand is assumed to start at a higher, more cognitive or abstract level, influencing perception in a top-down way. The two signals meet at every level of the hierarchy and are combined to form what we end up perceiving.

These theories have the potential to explain a multitude of perceptual phenomena, with one of the most characteristic examples being optical illusions. In the hollow-mask illusion, for instance, individuals are presented with the concave backside of a mask, but what they perceive instead is a normal, convex face (Papathomas, 2017). According to Bayesian and predictive coding theories, this happens because the brain has a quite certain internal model of what faces look like, i.e. that they are convex. Consequently, this model exerts a large influence on the final percept, overpowering the information coming from possibly ambiguous visual depth cues.

Bayesian theories claim that perception is inferring the most probable cause of the sensory inputs using Bayes' rule. The percept arises from a (posterior) probability distribution $P(cause \mid sensory\ inputs)$ which the brain computes (explicitly or implicitly) by multiplying the likelihood $P(sensory\ inputs \mid cause)$ and the prior $P(cause)$:

$$P(cause \mid sensory\ inputs) \propto P(sensory\ inputs \mid cause)\ P(cause),$$

The likelihood is thought to be conveyed by bottom-up signals and expresses how probable the sensory input is for each possible stimulus. Priors correspond to the brain's prior beliefs or expectations about the stimulus, before observing the sensory inputs, and are often thought to be conveyed by top-down signals[1].

Predictive coding, like the Bayesian brain theories, views perception as a combination of top-down and bottom-signals. Top-down signals represent predictions that the brain makes for what it is about to perceive, but here bottom-up signals correspond to the errors between the predictions and the sensory inputs. Predictions are adjusted based on these prediction errors, weighted by their relative certainty or 'precision' in technical terms. In the Bayesian brain, this corresponds to the reciprocal of the distribution variance. Both theories were initially formulated to explain perception specifically, but have since been broadened to become general frameworks of cognition and even motor control (e.g. Adams et al., 2013). The theories are not

[1] In reality, priors can be encoded in various ways, not all of them directly corresponding to neural signals (e.g., the selectivity of receptive fields can be considered a form of prior).

mutually exclusive, but neither do they necessarily coincide: the Bayesian approach focuses on what is being computed and could, theoretically, be implemented in various ways in the brain, while predictive coding focuses on a plausible neural algorithm, which could be implementing either Bayesian or non-Bayesian computations (Aitchison and Lengyel, 2017).

In 2012, Pellicano and Burr proposed that autism may result from weaker influences of prior beliefs in Bayesian brain computations, arising from impaired construction or improper integration of priors (Pellicano and Burr, 2012). In a reply to their article, Brock raised the alternative possibility that the autistic symptomatology could stem from stronger bottom-up (likelihood) influences instead, as this effect would be mathematically indistinguishable from weaker priors (Brock, 2012). Then, in 2014, both Lawson, Rees, and Friston, and Van de Cruys et al. reformulated the Bayesian theories using the hierarchical predictive coding framework. Lawson et al. (2014) proposed that autism arises from an imbalance between the precision of predictions and that of prediction errors leading to stronger influences of prediction errors. Van de Cruys et al. (2014) argued that the precision of prediction errors in autistic cognition is both higher and less flexible compared to neurotypicality. While the details of each specific theory differ, the general structure remains the same: in autistic perception and cognition, there is an imbalance between the influence of priors/predictions and likelihoods/prediction errors, with the brain relying more on bottom-up information, at the expense of top-down knowledge. This we will refer to as the *imbalance hypothesis* for the remainder of this review.

Since 2012, there have been multiple studies trying to confirm, disprove, or simply use these theories to interpret their results. However, their findings have been mixed and their approaches substantially varied, such that it is difficult for a researcher to have a thorough understanding of the literature. The main purpose of the current review is to present all the relevant evidence in favour of or against the imbalance hypothesis for autism, highlighting their commonalities and differences. To do that, we will systematically search for ASD studies that mention our frameworks of interest and categorise their findings both based on the experimental priors and based on their results relative to the imbalance hypothesis. We will also present other related findings within the Bayesian and predictive coding frameworks and propose directions for future research. In terms of methodology, we will examine the various approaches, focusing

especially on theory-driven computational models, and attempt to provide guidelines for the studies to come.

2.2 Methods

In order to perform a thorough and impartial review of the literature, we decided a priori on the form of our literature search and the inclusion/exclusion criteria. The reviewed articles obeyed the following inclusion criteria:

a. Articles should have been peer-reviewed.

b. Articles should be written in English.

c. Articles should have been published in the years 2012 to 2021.

d. Articles should be mentioning Bayesian or predictive coding theories in their title, abstract, or keywords.

e. Articles should report the findings of an experiment or a survey with autistic participants or with neurotypical participants that have undertaken the autism quotient (AQ) questionnaire (Baron-Cohen et al., 2001).

Criterion d does not require the actual words ‘Bayes’ or ‘predictive coding’ to appear in the abstract or keywords, as other phrases can unambiguously be referring to these frameworks. For example ‘interoceptive inference’ (Gu et al., 2015, keywords) clearly refers to an interoceptive version of Bayesian inference, and therefore fulfils the criterion. Similarly, Retzler et al. (2021) are directly referring to the Bayesian/predictive coding framework in their abstract when they write ‘Interpreting the world around us requires integrating incoming sensory signals with prior information’. On the contrary, a simple mention of expectations or predictions would not qualify as fulfilling the criterion.

The steps of our search process are presented below:

1. We searched for articles that had a word beginning with ‘autis-’ (e.g., autism, autistic) and either a word beginning with ‘Bayes’ (e.g., Bayesian) or one of the phrases

'predictive coding' and 'predictive processing' in their titles, abstracts, or keywords. The search focused on articles published from 2012 to 2020 and was conducted on four of the largest scientific databases: Web of Science, BASE (Bielefeld Academic Search Engine), Scopus, and PubMed.

2. We chose the articles that fulfilled criteria a-d. For criterion e, we temporarily suspended the requirement of performing an experiment, so that theoretical articles could be included in step 4.

3. We added 2 articles that were known to us and did not appear in the initial search (Skewes et al., 2015; Utzerath et al., 2018).

4. We searched each of the selected articles on scite (https://scite.ai/), a citation analysis tool, focusing only on citations labelled as 'supporting' or 'contrasting' by the platform.

5. We kept all articles (either from scite or from the initial search) that obeyed all the inclusion criteria.

This process was conducted during the first week of January 2022 and resulted in 85 articles. Two of those were discarded as they did not find a main effect (an effect of priors) at all in any participant group, meaning that interactions with autism could not be investigated (Schütz et al., 2021; Todorova et al., 2021). This left us with a pool of 83 articles. In some cases, whether the presented findings could adequately be explained by a Bayesian framework was debatable. However, as such judgements are partly subjective, we have decided not to exclude any articles based on our interpretation. Instead, we have simply included them in different sections (2.3.3 and 2.3.4). Moreover, as most studies did not differentiate between Bayesian and predictive coding theories, we will not be making such a distinction in the rest of the review. Instead, we will be referring to the terms 'prior', 'expectation', 'prediction', and 'top-down influence', as well as 'likelihood', 'prediction error', and 'bottom-up influence' relatively interchangeably.

We decided to group studies based on the types of priors that they investigated. Did they study the influence of a pre-existing prior, such as the fact that light comes from above us (a prior that impacts our interpretation of shading and shapes, e.g. Croydon et al., 2017), or did they require that participants develop the relevant prior during the task? This classification can be seen as roughly analogous to the one by Seriès & Seitz (2013), with 'structural priors' being

equivalent to pre-existing priors and 'contextual priors' to learned priors, although the categories do not completely overlap. We reasoned that such a categorization might provide some insight into the details of the imbalance hypothesis. If there are general imbalances in the use of priors in autism, then we would expect to see effects in both categories. However, if ASD influences the development of priors, then studies in the category of learned priors should exhibit clearer differences. We additionally separated the learned priors into 'implicit' and 'explicit' priors, based on how the relevant information was presented to the participants. Explicit priors are a result of information being given directly to the participants, while implicit priors are formed when the participants infer the statistics of the environment on their own. This is an important distinction, as explicit and implicit learning might be a result of different processes in the brain, and participants could employ conscious response strategies for the explicitly learned priors. Finally, pre-existing priors were separated according to whether they had a 'social' component or not (e.g, a prior for the recognition of faces, Gómez et al., 2014, vs the light-from-above prior we mentioned before). These studies were separated to explore specific differences in the social domain, known to be particularly affected in autism (we focus on pre-existing priors here as there were very few learned social priors). Studies that investigated multiple types of priors were included in each of the respective groups, their findings mentioned separately.

Furthermore, the results of the studies were categorised based on whether the evidence they provided was in favour or against the imbalance hypothesis. It is important to note that, in specific cases, the predictions of some Bayesian and predictive coding theories of autism can contradict each other or be more nuanced in their implications than the scope of this review. For example, results indicating flat likelihoods but even flatter priors in autism, would disagree with Brock's and Van de Cruys et al.'s formulation, who emphasised increased bottom-up precision, but would support Pellicano & Burr's and Lawson et al.'s conceptualization, as they focused on flatter priors and precision imbalances, respectively. To simplify matters, the main question of this review will strictly revolve around the imbalance hypothesis, i.e., the hypothesis that the ratio of prior precision to likelihood precision is *smaller* in autism. This formulation captures the common ground between the theories and allows us to clearly

categorise the results of the studies as i) supporting the presence of such an imbalance, ii) showing no difference in the precision ratio between diagnostic groups or across autistic traits, or even iii) finding evidence for the *reverse* imbalance, where the aforementioned ratio is larger in autism. Behavioural, computational, or neuroimaging evidence for a smaller prior-to-likelihood precision ratio or simply reduced bias based on prior beliefs will be considered as supporting the hypothesis. Some studies did not investigate the imbalance hypothesis, although their findings are relevant to the Bayesian and predictive coding hypotheses. Therefore, we included two additional categories to capture their findings, based on showing or not showing other differences between diagnostic groups or across autistic traits. All studies and their categorisation, along with the processing script can be found here: https://osf.io/3tvq2/

Our results are divided into subsections based on the main prior types (2.3.1 to 2.3.4), with further prior categorization within each subsection. These are followed by interim discussions, summarizing and interpreting the relevant findings. Then, a subsection with summary statistics is presented (2.3.5), which includes a table containing all studies and findings in our pool. The last subsection (2.3.6) examines the Bayesian modelling techniques utilised by the researchers in investigating the imbalance hypothesis.

2.3 Results

2.3.1 Pre-existing priors

By ‘pre-existing priors’ we mean knowledge or expectations that the participants would have, either consciously or unconsciously, before receiving any task-relevant information. Overall, such priors seem to not be greatly affected in ASD. For example, low-level priors regarding geometrical shapes and line orientations were similar across diagnostic groups and autistic traits (Noel et al., 2021; Tulver et al., 2019; Utzerath et al., 2019; Van de Cruys et al., 2021). Priors for the continuation of motion were also unaffected by autistic traits (Andermane et al., 2020; Tulver et al., 2019). On the other hand, there is mixed evidence for the known prior for slow motion (Sotiropoulos et al., 2014), which had reduced influence in high-AQ participants

(Powell et al., 2016), but increased influence in autistics' perception of self-motion (Noel et al., 2020). Importantly, the latter difference was attributed by the researchers to wider likelihoods instead of more precise priors in such participants and was no longer present once feedback was given to the participants.

Priors can also encode more complex environmental statistics. These were unaffected in ASD, as well. Expectations regarding the direction of light and the common colours of natural scenes (e.g., the sky being blue) did not differ between diagnostic groups (Croydon et al., 2017; Maule et al., 2018). Priors for the weight of an object, invoked by its size or apparent material, resulted in gaze differences when participants were surprised, but similar perception and lifting behaviour were observed between groups and across autistic traits (Arthur et al., 2020, 2019). However, when stimuli changed frequently between expected and unexpected weights, autistic individuals performed worse than neurotypical ones (Arthur et al., 2021). Moreover, hand claps with no sound led to stronger prediction errors in autistic participants as indicated by EEG components (van Laarhoven et al., 2020). Pre-existing expectations can also be invoked when the participants are the ones who produce the experimental stimuli (e.g., the expectation to hear one's voice when speaking). The influence of these priors was weaker in ASD when stimuli were auditory (Lin et al., 2015; van Laarhoven et al., 2019), but not when they were tactile (Finnemann et al., 2021). Specifically, when speaking, control participants relied more on predictions about their voice, while autistics relied more on auditory feedback (Lin et al., 2015).

Another kind of prior can be found in the process of cue integration (Ernst and Banks, 2002), as the belief that all sensory signals originate from the same source can be described as a coupling prior. This could be measured by studying how participants combine information from different stimuli. A weak coupling prior would result in participants mostly focusing on one source of information. Such priors were similar between diagnostic groups when combining depth and disparity information (Bedford et al., 2016), depth and occlusion cues (Smith et al., 2017), or visual and vestibular cues (Zaidel et al., 2015). Contrary to the imbalance hypothesis, though, autistics perceived auditory and visual cues as originating from the same source more frequently than controls, suggesting a stronger coupling prior (Noel et al., 2018). Differences also appeared between groups when information concerned one's body.

Proprioceptive signals were more integrated with visual and haptic information in low-AQ neurotypical participants compared to autistic or high-AQ ones (Palmer et al., 2015). Integration between the body's pain levels and visual cues also differed in ASD (Gu et al., 2015), although this finding could be interpreted both in terms of stronger bottom-up, exteroceptive signals and as resulting from a stronger coupling prior in autism.

Priors can also be based on social constructs. For example, in western cultures, consonant sounds are associated with more positive emotions compared to dissonant ones. Consistent with the imbalance hypothesis, this effect was stronger with lower autistic traits (Bravo et al., 2017). Semantic priming, where the meaning of a stimulus influences the perception of other stimuli, was also reduced in autistic participants (Grisoni et al., 2019). At the same time, other language priors, such as the effects of word knowledge or phoneme categories, were unaffected by autistic traits or diagnoses (Chiodo et al., 2019; Tulver et al., 2019).

Social priors might be especially relevant for autism, given the condition's symptoms. Findings in tasks involving such priors varied. Priors concerning the recognition of social images gave rise to mixed findings, with differing neuroimaging measures but similar behavioural performance in ASD (Gómez et al., 2014; Król and Król, 2019), and better performance with higher autistic traits (Tulver et al., 2019). Gaze perception was biased towards the forward direction or the more salient features of the environment independently of diagnoses or autistic traits (Pantelis and Kennedy, 2017; Pell et al., 2016). Recognizing biological motion also yielded mixed findings. High-AQ individuals showed weaker prior influences (van Boxtel and Lu, 2013), while autistics had similar general recognition, performing worse than controls only when aided by interpersonal, social hints (Chambon et al., 2017; von der Lühe et al., 2016; although see Amoruso et al., 2019 for a hint at a general impairment). Another study found increased prevalence of transgender and non-binary identities in ASD, interpreting it as a weaker prior about accepted gender norms (Walsh et al., 2018).

2.3.1.1 Pre-existing priors: Discussion

What conclusions can we draw from these studies? Findings are mixed, but it seems that studies investigating pre-existing priors do not support the existence of a general prior-to-likelihood

imbalance. This is not necessarily surprising. Weaker prior influences are hypothesised to result in autistic individuals getting overwhelmed by their environment (Pellicano and Burr, 2012). However, we know that not all environments cause distress in ASD, with new environments being far more overwhelming, something that is expressed as an insistence on repetition and routine. The reason for this might be that priors for familiar environments (i.e., pre-existing priors) are unimpaired.

One important observation is that tasks having a social component were almost evenly split between supporting and contrasting the imbalance hypothesis, while tasks using simpler perceptual priors usually revealed no difference across groups and traits. On one hand, this could be expected, given the communication and theory of mind impairments in autism. On the other, the main appeal of the Bayesian theories for ASD is that they attempt to explain all symptoms with the same mechanism, from low-level sensory sensitivities to high-level social deficits. It seems, based on this part of the review, that this framework might not apply equally well to all facets of the disorder.

2.3.2 Learned priors

Another way to test the prior-likelihood imbalance is to use tasks where participants form new priors during the experiment. This provides more control to the researchers, as they can directly manipulate the environmental statistics that influence the participants. However, it also introduces a confounder: the observed results could be a product of deficits in the development instead of in the use of priors.

One way to create a prior is to make participants familiar with a complex stimulus. For example, the recognition of a distorted image is enhanced in participants who have seen the original image beforehand. Such a prior was unaffected by autistic traits (Van de Cruys et al., 2018), although findings in ASD were mixed (Król and Król, 2019; Van de Cruys et al., 2018). Autistics were also more biased by their initial interpretation of a partial picture when they were shown the complete one (Lacroix et al., 2021). On the other hand, prior influences that

resulted from priming the participants with a stimulus or asking them to imagine it were not modulated by autistic traits (Andermane et al., 2020).

Expectations can also be developed in participants when some experimental stimuli are more frequent than others. Such frequency priors were found to have a weaker effect in individuals with autism in about half of the tasks. That was the case when participants were asked about the location of a visual stimulus (Ganglmayer et al., 2020), or in their EEG responses to 'odd one out' tonal stimuli (Goris et al., 2018; Grisoni et al., 2019), but not when the auditory stimuli differed in their rhythm (Knight et al., 2020). The expectation of repeated images also led to different fMRI measures across diagnostic groups (Utzerath et al., 2018). By contrast, frequency priors had similar influences in autistics and controls when recognizing agent actions (Chambon et al., 2017) or searching for specific line orientations (Van de Cruys et al., 2021). The frequencies of made-up categories resulted in equally strong biases between diagnostic groups when classifying auditory stimuli (Skewes and Gebauer, 2016), although they influenced high-AQ individuals less in the case of visual stimuli (Skewes et al., 2015). These two studies can also be interpreted as measuring the malleability of stimulus categories, an interpretation whose compatibility with the Bayesian approach is unclear. The rest of the autistic trait frequency findings are also mixed, with high-AQ individuals showing weaker integration of top-down and bottom-up signals in a recognition task (Coll et al., 2020) and having fewer hallucinations of frequent stimuli (Aru et al., 2018), but producing similar fMRI responses to more frequent stimuli repetitions compared to their low-AQ counterparts (Ewbank et al., 2016).

Frequency priors can also have an effect in experiments using continuous stimuli, where participants implicitly learn the complete stimulus distribution. This learning usually results in participants being biased towards the mean of the distribution when estimating the value or nature of a stimulus. The strength of these 'central tendencies' is inconsistently affected by autism. For example, some studies with visual stimuli found them weaker in ASD (Edey et al., 2019; also hinted at Sapey-Triomphe et al., 2021b), while an experiment with auditory stimuli showed no differences between groups (Edey et al., 2019). Three further auditory experiments showed weaker biases in ASD (Jaffe-Dax and Eigsti, 2020; Lieder et al., 2019), but with

computational modelling suggesting weaker biases towards recent trials only and no differences at the level of the stimulus distribution. Surprisingly, when reproducing time intervals, autistics had stronger central tendencies (Karaminis et al., 2016). However, the authors explained their results in terms of flatter priors (and even flatter likelihoods) in ASD, an interpretation that would agree with the specific theory of Pellicano & Burr (2012), but contradicts the imbalance hypothesis. There is also some evidence that central tendencies are less malleable in ASD depending on the stimulus distribution (Sapey-Triomphe et al., 2021b), and that such biases are no longer present in individuals who have lost the autism diagnosis (Jaffe-Dax and Eigsti, 2020).

Studies that aimed to create more complex expectations in the participants yielded mixed findings. When distractors followed a specific distribution, no performance differences were present between diagnostic groups, although the groups did display more nuanced differences in their behaviour (Allenmark et al., 2021). Bimodal distributions for moving stimuli led to high-AQ participants exhibiting weaker biases and fewer visual hallucinations, although this apparently resulted from higher sensory precision rather than flatter priors (Karvelis et al., 2018). At the same time, no differences were present between 3-year-olds with high and low likelihood of autism or between diagnostic groups when stimuli were sequential (Rybicki et al., 2021; Ward et al., 2021). Finally, difficulties with prior development might also be revealed when participants have to learn that their prior beliefs are no longer accurate and therefore should stop being influenced by them. Autistics were unable to do that when presented with changing stimulus orientation statistics, but instead continued to exhibit pre-existing biases (Noel et al., 2021). They also kept relying on a stimulus that had stopped being informative in the course of the experiment (Zaidel et al., 2015).

Another way to manipulate expectations during an experiment is to include implicit cues associated with the presented stimuli. Similar to frequency priors, these expectations had reduced effects in ASD or high-AQ individuals in approximately half of the experiments. Autistic participants mostly exhibited reduced influence of priors when auditory cues predicted a visual stimulus (Lawson et al., 2017; Sapey-Triomphe et al., 2021a, but see 2021c), but not when self-produced cues were associated with the timing of an auditory stimulus (Finnemann

et al., 2021). Additionally, environmental cues influenced neurotypical participants more than autistic ones when recognizing agent actions (Amoruso et al., 2019). A couple studies also looked at cue associations in autistic traits. One of them showed reduced influence of cues in shape recognition with higher AQ (Bianco et al., 2020). The same study also included a replication of the experiment of Amoruso et al (2019), but found no relationship between cue influence and AQ. An AQ replication of the Lawson et al. (2017) results within the same study was only partially successful.

When the relationship between the cue and the stimulus location was made explicit by the researchers, experiments yielded slightly weaker evidence to support the imbalance hypothesis. Explicit cues about the direction of visual stimuli showed comparable influences across autistic traits (Retzler et al., 2021), as did cues that did not reflect the actual environmental regularities in a binocular rivalry task (Andermane et al., 2020). Furthermore, the explicit presentation of prior information in a decision making task did not produce any difference across AQ or diagnostic groups (Angeletos Chrysaitis et al., 2021). Visual cues about the timing of an auditory stimulus also had similar behavioural effects in autistic and control participants, but resulted in different EEG responses (Beker et al., 2021). Moreover, sequences of stimuli explicitly associated with a target stimulus had greater behavioural and EEG effects in neurotypical participants relative to autistic ones (Thillay et al., 2016). Explicit information about the intention of an agent also affected autistics less (Hudson et al., 2021), as did explicit cues when tracking stimulus locations (Greene et al., 2019). Contrary to the imbalance hypothesis, though, autistic participants showed stronger fMRI responses and similar skin conductance when anticipating a painful stimulus based on an explicit cue (Gu et al., 2018).

A few studies attempted to assess priors directly, by asking participants to make predictions based on their past experiences before showing them any new stimuli. For example, a cue would be presented to the participants, who are asked to make a guess about the ensuing stimulus. After their response, the stimulus would be shown to them and the assumption is that they would update their priors about the cue-stimulus association. Such experiments have yielded mixed findings (Manning et al., 2017; Sapey-Triomphe et al., 2021a, 2021c; Sevgi et al., 2020). One of these studies showed that participants with stronger autistic traits processed

only social cues differently (Sevgi et al., 2020). Interestingly, social priors that were based on explicit information about another person had greater influence in autistic individuals, but those that were based on observation of their behaviour were similar across groups (Maurer et al., 2018). In simple perceptual tasks, the extrapolation of motion or accumulation of visual stimuli was similar between autistic and control participants (Tewolde et al., 2018), while autistics performed worse in synchronising their responses to auditory rhythmic stimuli and were slower to update when their tempo changed (Vishne et al., 2021). Semantic priors also had similar effects between diagnostic groups (Tewolde et al., 2018). Finally, the behaviour of high-AQ individuals was less flexible when estimating the effects of their actions in the experimental environment (Perrykkad et al., 2021).

2.3.2.1 Learned priors: Discussion

Interpreting the results in tasks with learned prior proves to be challenging. Findings regarding implicitly learned priors were evenly split between positive and null, while tasks providing explicit prior information led to a significantly reduced proportion of supporting findings. These observations suggest a possible impairment in implicit prior acquisition in autism, which would be mitigated by conscious strategies when priors are provided explicitly. However, this comes in contrast with direct research on learning in autism, which finds intact implicit but impaired explicit learning (Brown et al., 2010; Izadi-Najafabadi et al., 2015). Another caveat is that such paradigms introduce important variables that cannot be easily tracked, namely task instructions and performance feedback. If the explicit presentation of information aids autistic participants significantly during experiments, then the exact phrasing of the instructions or the feedback by the researchers may have important effects on their task behaviour.

Studies that attempted to assess prior development directly by asking participants to predict the next stimulus slightly supported the presence of differences between groups. Unfortunately, the link between such paradigms and Bayesian priors is not clear: Bayesian priors are often thought to be implicit, and it is unknown whether or in what context they can be accessed consciously to guide such judgements. Even if they could be accessed consciously, though, they might not be directly relevant to the imbalance hypothesis. Say that participants are asked

to predict a characteristic of a stimulus (e.g., its position) based on previous stimuli. The optimal response in the absence of no other information would be the mean of their prior beliefs. Therefore, if two participants had the same prior mean, but different precisions, it is possible that they would give identical responses. Therefore, the results of such experiments cannot reveal any information about prior precision and cannot speak for the imbalance hypothesis. On the other hand, they could point towards deficits in prior acquisition.

2.3.3 Adaptation

Besides the studies mentioned above, in our literature search we came across some experimental designs that cannot be directly classified into either pre-existing or learned priors. A large part of those studies focused on sensory adaptation, a phenomenon where the presentation of a stimulus biases subsequent percepts away from that stimulus and enhances the perception of infrequent stimuli, possibly via a reduction of neuronal sensitivity (Kohn, 2007). These repulsive effects are distinct from those explained by the conventional Bayesian framework, where priors aid in the processing of expected stimuli and perception is biased towards previous stimuli (attractive biases). However, we have included them here as findings of weaker sensory adaptation in autism are commonly attributed to the imbalance hypothesis. We have decided to also include mismatched negativity effects measured using auditory oddball tasks, as we believe that their interpretation in terms of sensory adaptation is more accurate than viewing them as violations of Bayesian expectations (Symonds et al., 2017).

Visual adaptation experiments didn't generally find evidence for the expected reduced adaptation in ASD. Colour adaptation was similar across diagnostic groups (Maule et al., 2018), as was adaptation in binocular rivalry and biological speed estimation across autistic traits (Andermane et al., 2020; van Boxtel and Lu, 2013). Interestingly, though, in the latter study, high-AQ participants showed weaker sensory adaptation than low-AQ ones when stimulus location changed, differentiating between local (same location, same stimulus) and global (different location, same stimulus) adaptation. Contrary to the previous results, adaptation to the number of stimuli on the screen was reduced in autistic children (Turi et al.,

2015), while neural habituation to faces was stronger in two-year-olds with high likelihood of autism (Ward et al., 2020).

Sound repetition led to increased signal-to-noise ratio in ASD (Font-Alaminos et al., 2020), while deviant auditory stimuli produced similar EEG responses across groups, both when participants were warned about them and when they were not (Gonzalez-Gadea et al., 2015; Goris et al., 2018). Audiovisual adaptation was significantly weaker in ASD (Turi et al., 2016), but was negatively correlated only with the 'attention to detail' AQ subscore and not the total AQ scores (Stevenson et al., 2017).

2.3.3.1 Adaptation: Discussion

In general, these findings do not show weaker sensory adaptation in ASD. However, adaptation effects are not trivially explained by the core mechanism of either the Bayesian or the predictive coding frameworks. That is because in such frameworks priors are expected to bias percepts towards, rather than away from their mean. There have been various theoretical attempts to explain adaptation as part of an extended Bayesian framework (Grzywacz and Balboa, 2002; Stevenson et al., 2010; Wei and Stocker, 2015), although they are outside of the scope of this review.

2.3.4 Other findings

Not all studies can be neatly fit into the categories of pre-existing priors, learned priors, or adaptation. Some studies did not specify the nature of assumed priors and, while they interpreted their results in terms of the imbalance hypothesis, how their findings could be formalised with the Bayesian or predictive coding interpretations was unclear to us. We have included them in this section for completeness and so as to be consistent with our review selection criteria.

Autistic participants were better at remembering exact colours, but performed similarly to controls when averaging them (Maule et al., 2017). They also showed atypical processing of colour dimensions in terms of their separability (Hadad et al., 2017). Autistic participants were

more hindered by visual noise (Zaidel et al., 2015). These studies focus on global vs local processing, which has been associated with the Bayesian framework, but it is not clear what previous knowledge or expectations of environmental regularities they are using.

Gaze differences were found between participants with both ASD and ADHD and controls when viewing social images, but not in participants with ASD only (Ioannou et al., 2020). When the task goal was predictably changed during the task, no differences were found across groups (Lacroix et al., 2021). Autistic participants paid less attention than controls to deviant auditory stimuli which were not specified in task instructions (Gonzalez-Gadea et al., 2015). A qualitative survey of autistic individuals also showed that they have an 'interrupted experience of time' (Vogel et al., 2019). Finally, two neuroimaging studies focused on predictive-coding-associated neural markers, but in a passive viewing task or in resting state, where the use of prior information was not obvious. These showed diminished predictable information in neural signals (Brodski-Guerniero et al., 2018), and impaired feedback connectivity and local visual processing in ASD (Seymour et al., 2019).

2.3.5 Summary statistics

In the following section, we present statistics based on the study findings most relevant to the imbalance hypothesis for each participant sample and prior type. For example, Pell et al. (2016) investigated differences between autistic and neurotypical groups and across autistic traits, using two distinct samples; Bianco et al. (2020) used two different tasks in the same sample; and Chambon et al. (2017) used one task in one sample, but investigated the effects of both pre-existing and learned priors. For each of these studies, we have included (at least) two findings in our analysis. The statistics are based on the *p*-values of the results being lower or higher than the 0.05 significance threshold (or an adjusted one in the case of multiple comparisons).

We decided against performing a meta-analysis for two reasons. First, in most studies, the effect sizes of interest are expressed in terms of F-values or related statistics or are completely absent. Including F-values to the calculations of a meta-analysis risks significantly

overestimating the aggregate effect size (Hullett and Levine, 2003). Avoiding that would require calculations with statistics that were not available for most such studies. Moreover, F-values by nature do not denote the direction of the presented effect. This is not a problem for findings below the significance threshold, as in this case the researchers perform additional tests to show this direction (e.g., weaker vs stronger biases in ASD). On the other hand, these tests are usually not performed for non-significant results, leaving us unable to know if the results should be counted as weak evidence for the imbalance hypothesis or the reverse imbalance. This means that a large portion if not most studies could not have been incorporated in the meta-analysis. The second reason is the heterogeneity present in our pool of studies. These included both neuroimaging and behavioural experiments, which can be further separated into fMRI, EEG, and MEG studies on one hand, and estimation, reaction time, eye-tracking, etc. on the other. In terms of sample, they looked both at differences between autistic and neurotypical participants and across autistic traits in the general population, both in children and adults. In terms of domain, they looked both at perception (visual and auditory) and at decision-making, among others. As for the priors investigated, we have already demonstrated how they can be split into various categories. It is therefore quite likely that any number produced by aggregating all these findings would be misleading (Feinstein, 1995; Greco et al., 2013; Murphy, 2017).

We recognise that determining which finding of a given study is most relevant to the imbalance hypothesis can never be fully objective and also that classifying and counting results based only on their *p*-values is highly imperfect. Nonetheless, we think that this approach offers a much-needed broad overview of the field. The complete list of main findings and their classification can be seen at the end of this subsection.

The 83 studies in our sample yielded a total of 123 findings relevant to the Bayesian or predictive coding explanations for autism. Of those, 92 were a result of experimental designs which clearly included both priors and likelihoods. The larger portion of that subset (51%) showed no statistically significant differences between groups or across autistic traits regarding the imbalance hypothesis. 37% showed evidence in the direction of the imbalance, 7% showed evidence for the reverse, 4% showed differences that have no clear interpretation in terms of

the imbalance hypothesis, and 1 finding showed no difference between groups with no clear imbalance interpretation. Focusing only on the designs that investigated the imbalance hypothesis, neuroimaging studies lead much less frequently to null findings than behavioural ones (37% vs 58%).

Figure 2.1 displays the distribution of results across prior categories which, although noisy, seems to agree with the conclusions drawn in our interim discussions. Findings related to implicitly learned priors are evenly split among supporting and opposing the presence of an imbalance, with findings related to pre-existing social priors showing slightly less support. On the other hand, designs which relied on pre-existing non-social and explicitly learned priors mostly do not support the imbalance hypothesis. 'Prediction' findings, which were based on direct prior estimations without the involvement of sensory evidence, also hint at the presence of prior development deficits in ASD. Adaptation findings show a few differences but, as we have already mentioned, their Bayesian interpretation is currently unclear.

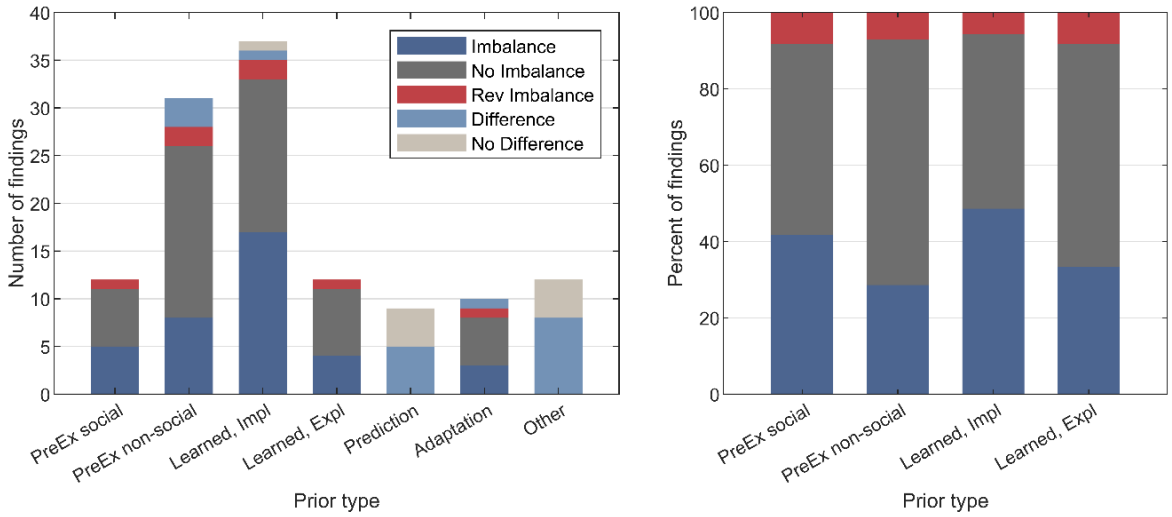

Figure 2.1 Distribution of imbalance-relevant findings across prior types and result categories. *The left panel includes all findings, while the right one presents only those that test the imbalance hypothesis. Findings in the three rightmost categories of the left panel are not interpreted as resulting from prior-likelihood combinations. 'Imbalance'-coded findings in the Adaptation category refer to reduced adaptation in ASD. The category of Reverse Imbalance refers to findings that found stronger prior influences in ASD. Difference and No Difference findings are not directly testing the imbalance hypothesis.*

Out of the 83 studies, 60 focused on ASD vs control group comparisons, 16 used AQ scores in the general population and 7 used both. Overall, autistic trait findings were more often null, with Imbalance/No Imbalance/Reverse findings being 10/18/1 in AQ and 26/28/5 in clinical group comparisons, when looking at pre-existing and learned priors. The median sample size was 44 participants for experiments focusing on autistic vs neurotypical participants, and 40 for those measuring autistic traits. The *p*-curves for the significant findings appear right-skewed (Figure 2.2), which suggests no strong selection effects (publication bias or *p*-hacking), which would result in left-skewed curves (Simonsohn et al., 2014). However, not all selection effects can be detected with this method (Bishop and Thompson, 2016) and the fact that the curves do not monotonically decrease as *p*-values increase is evidence for weak selection effects. We further used the *p*-curve app (https://www.p-curve.com/app4/, v4.06, Simonsohn et al., 2015) to estimate the statistical power of all studies that provided the necessary statistics (e.g. F-statistic; see https://osf.io/k6p5n for a complete list). This showed that both the set of all studies and the subset that investigated the imbalance hypothesis contain evidential value ($p < 0.001$). However, the estimated statistical power was a very weak 39% (90% CI [20%, 58%]) and 42% (90% CI [20%, 65%]), respectively. Statistical power for AQ findings was estimated at 61% (90% CI [26%, 85%]), while that for ASD was estimated at 32% (90% CI [13%, 55%]). Neuroimaging and behavioural studies did not differ, with respective powers at 42% and 41%.

For the effect sizes, we report partial eta squared (η_p^2) values, which correspond to the variance explained by the effect, when the effects of other predictors are partialled out. We made this choice because η_p^2, as opposed to Cohen's d or other statistics, is the reported effect size for most interaction effects or the only one that could be calculated with the provided statistics. It is also more comparable across studies, as many of the relevant effects in our pool are products of interactions and η^2 depends on the size of the main effect. In the case of only one predictor, η_p^2 is equal to η^2 or r^2. Out of the 123 main findings, 80 reported η_p^2 values or some other statistic which could be transformed to η_p^2. We calculated median effect sizes over all studies, irrespective of statistical significance, as using only significant results would overestimate the true effect size, especially in the presence of low statistical power (Gelman and Carlin, 2014).

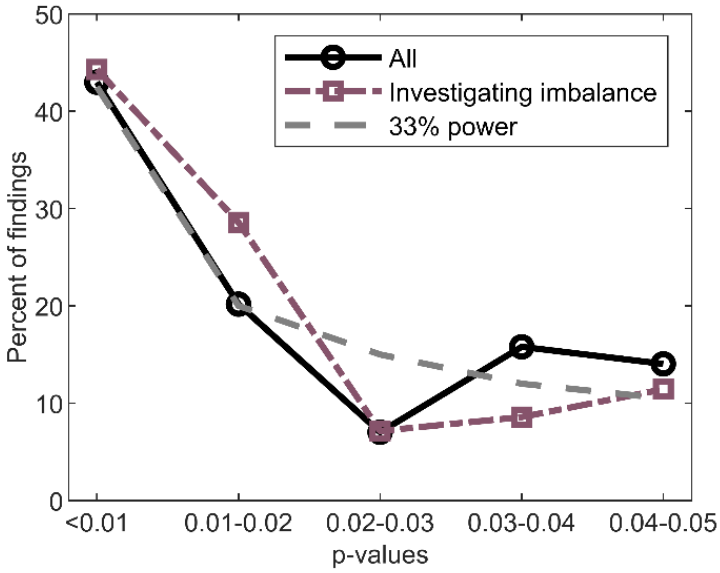

Figure 2.2 P-curve for all findings and for those investigating the imbalance hypothesis. *Only findings with reported p-values of less than 0.05 were included. Both curves are right-skewed, indicating no strong selection effects, which would result in left-skewed curves. Slightly higher percentages near p-values of 0.05 relative to smaller values are evidence of weak selection effects. The 33% power curve is provided here as an example of a curve when no selection effects are present.*

Median η_p^2 was 0.04 for studies that investigated the imbalance hypothesis and 0.05 overall. Studies that investigated the imbalance hypothesis in autistic individuals specifically had a median effect size of 0.05, while those that measured AQ scores in the general population ones had a medium effect size of only 0.02. Finally, studies investigating learned priors had a median partial eta squared of 0.05, while those investigating pre-existing priors a median of 0.02. In the case of one predictor, η_p^2 values of 0.01 are designated as small, values of 0.06 as medium, and values of 0.14 as large (Cohen, 1987). When multiple predictors are present, η_p^2 values might overestimate the true effect size, depending on the magnitude of the other predictors. This would mean that present effect sizes are small to medium. However, there are also other reasons the reported median effect sizes might be larger than the true effect size (Figure 2.3).

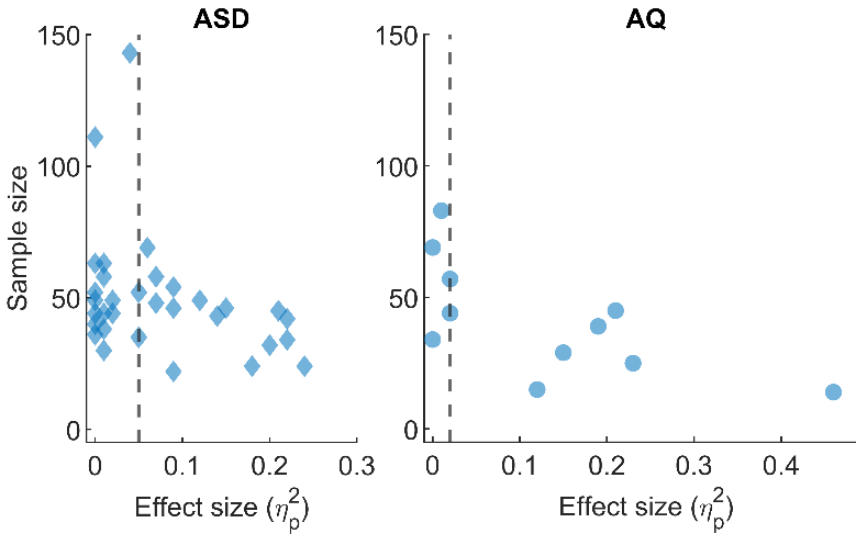

Figure 2.3 Sample size vs effect size for imbalance studies. *Dashed lines correspond to the median effect size. This plot is equivalent to a 'folded' funnel plot, as all the represented effect sizes can only be positive. This means that publication bias cannot be shown in the figure, contrary to a normal funnel plot. Nonetheless, the figure illustrates that higher sample sizes lead to smaller effect sizes. The folded funnel plots peaking slightly below the median effect sizes highlights that the medians are possibly overestimates of the true effect sizes.*

Partial eta squared takes non-negative values. The lack of information about the direction of the effects (e.g., stronger vs weaker biases in autism) automatically leads to overestimated median estimates. To understand that, imagine having three correlation coefficients: –0.2, 0, and 0.3, with a median of 0. If the effect sizes are presented in terms of r^2 or η_p^2, the median effect size between 0.04, 0, and 0.09 would be 0.04, while in reality it should be 0, as 0.04 corresponds to a negative correlation. This is something we could not avoid in our analysis, as many of the reported statistics included no information about the direction of the results. Publication bias in favour of positive results might also lead to the overestimation of the effect sizes. These considerations lead us to conclude that overall true effect sizes are small and in the case of pre-existing priors possibly close to zero. The full list of main finding effect sizes can be seen on Table 2.1.

Table 2.1 List of main findings and their classification.

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	*p*	Finding
van Boxtel & Lu, 2013	0	25	Point-light-walker distractors	PreEx, S	Resp	I	.23	.016	Better performance when distractors were biological
	0	30	Point-light-walker speed estimation	Adapt	Resp	I	.15	.036	Less adaptation when stimulus location changed
Gómez et al., 2014	10	14	Mooney faces viewing	PreEx, S	Image	I		*q*<.1	Reduced AIS in hippocampus
Gonzalez-Gadea et al., 2015	24	19	Auditory oddball	Adapt	Image	NI		>.05	No difference in MMN
				Other	Image	D		.044	Smaller P300 increase in unpredictable stimuli
Palmer et al., 2015	15	30*	Rubber hand illusion	PreEx NS	Resp	I	.21	.009	Weaker effect of RHI in kinematic measures
Turi et al., 2015	16	18	Numerosity estimation	Adapt	Resp	I		< .001	Less adaptation
Zaidel et al., 2015	14	22	Multisensory integration in self-motion	PreEx NS	Resp	NI			Intact multisensory integration
				Other	Resp	D		.004	More affected by visual noise
				Learned Imp	Resp	I			Less update after repeated pure visual noise exposure
Lin et al., 2015	11	13	Delayed auditory feedback in speech	PreEx NS	Resp	I	.18	.009	Stronger DAF effects
			Lombard effect	PreEx NS	Resp	I	.30	.006	Weaker Lombard effect
Gu et al., 2015	15	15	Pain discriminability in images	PreEx NS	Image	D		<.05	Increased AIC activation in empathetic pain
Skewes et al., 2015	0	29	Gabor orientation classification	Learned Imp	Resp	I	.15	.037	Less base rate bias
Pell et al., 2016	0	34	Gaze direction estimation	PreEx, S	Resp	NI	.00	.89	No difference in forward bias
	11	11	Gaze direction estimation	PreEx, S	Resp	NI	.09	.20	No difference in forward bias
von der Lühe et al., 2016	16	16	Point-light-walker classification with social information	PreEx, S	Resp	I	.20	.011	Less benefit of social information
			Point-light-walker recognition	PreEx, S	Resp	NI	.09	.10	No difference in recognition task
Powell et al., 2016	0	31	Aubert-Fleischl Phenomenon	PreEx NS	Resp	I			AQ + sensory noise model fitted best
	0	26	Contrast effect	PreEx NS	Resp	I			AQ + sensory noise model fitted best
Skewes & Gebauer, 2016	16	19	Sound location classification	Learned Imp	Resp	NI	.05	.22	No difference in base rate bias
Ewbank et al., 2016	0	29	Visual repetition suppression with unfamiliar faces	Learned Imp	Image	NI		>.16	No difference in the effect of block statistics
Bedford et al., 2016	16	14	Visual depth cue integration	PreEx NS	Resp	NI	.01	.56	No difference in cue integration
Turi et al., 2016	16	16	Audiovisual stimulus synchronicity	Adapt	Resp	I	.29	.001	Less adaptation

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	*p*	Finding
Karaminis et al., 2016	23	23	Time interval reproduction & discrimination	Learned Imp	Resp	R	.15	.007	More central tendency
Thillay et al., 2016	12	12	Predictable & unpredictable stimulus identification	Learned Exp	Image	I	.24	.015	Enhanced CNV to unpredictable stimuli
Manning et al., 2017	34	32	Non-stationary two-armed bandit	Predict	Resp	ND	.00	.98	No difference RL model learning rates
Chambon et al., 2017	18	20	Intention identification (video)	Learned Imp	Resp	NI	.01	.64	No difference in the overall effect of block statistics
				PreEx, S	Resp	NI	.08	.084	No difference in discriminability.
				PreEx, S	Resp	I	.19	.006	Smaller bias in favour of tit-for-tat
Croydon et al., 2017	18	18	Concavity-convexity judgements	PreEx NS	Resp	NI	.00	.99	No difference in light-from-above bias
Stevenson et al., 2017	0	54	Audiovisual stimulus synchronicity	Adapt	Resp	NI	.01	.53	No difference in adaptation
Pantelis & Kennedy, 2017	27	31	Gaze direction estimation with contextual cues	PreEx, S	Resp	NI	.01	.60	No difference in contextual cue weightings
Lawson et al., 2017	24	25	Probabilistic associative learning	Learned Imp	Resp	I	.12	.003	Less change in error rates with unexpected stimuli
	0	57	Probabilistic associative learning	Learned Imp	Resp	NI	.02	.42	No difference in error rates with changing expectancy
Hadad et al., 2017	19	17	Visual feature separation	Other	Resp	D	.15	.018	Atypical processing of integral and separable features
			Visual feature separation	Other	Resp	D	.13	.032	Atypical processing of integral and separable features
Bravo et al., 2017	0	39	Valence in sound	PreEx NS	Resp	I	.19	.006	Less positive ratings (valence) for consonant sounds
Smith et al., 2017	22	23	Depth judgement	PreEx NS	Resp	NI		>.2	No difference in use of occlusion information
Maule et al., 2017	16	16	Colour membership	Other	Resp	D	.28	.002	Better discriminability.
			Colour averaging	Other	Resp	ND	.02	.49	No difference in performance.
Karvelis et al., 2018	0	83	Motion direction estimation	Learned Imp	Resp	I		.011	Higher sensory precision
Tewolde et al., 2018	30	30	Visual motion & accumulation extrapolation	Predict	Resp	ND	.00	.62	No difference in prediction accuracy
			False memory task	Predict	Resp	ND	.04	.06	No difference in false memories
Maurer et al., 2018	17	25	Iterative trust game	Predict	Resp	D	.25	.001	Stronger effects of agent reputations
				Predict	Resp	ND	.01	.49	No difference in learned reciprocity

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	*p*	Finding
Van de Cruys et al., 2018	0	282	Mooney image recognition	Learned Exp	Resp	NI		.24	No difference in accuracy improvement
	23	24	Mooney image recognition	Learned Exp	Resp	NI		.82	No difference in accuracy improvement
Noel et al., 2018	38	33	Audiovisual simultaneity estimation	PreEx NS	Resp	R		.005	Stronger prior for one source
Goris et al., 2018	18	24	Auditory oddball	Adapt	Image	NI	.08	.080	No difference in overall MMN
				Learned Imp	Image	I	.14	.016	Less effect of frequency information
Brodski-Guerniero et al., 2018	19	19	Resting state	Other	Image	D		.031	Reduced average AIS
Aru et al., 2018	0	14	Auxiliary stimulus detection	Learned Imp	Resp	I	.46	.007	Lower hallucination intensity
	0	15	Auxiliary stimulus detection	Learned Imp	Resp	NI	.12	.22	No difference in hallucination intensity
Walsh et al., 2018	613	8064	Gender identity demographics	PreEx, S	Surv	I	.04	< .001	Higher percent of transgender and nonbinary identities
Gu et al., 2018	17	17	Pain anticipation	Learned Exp	Image	R	.22	.010	Higher rACC responses during anticipation
Maule et al., 2018	11	11	Colour adaptation in natural scenes & other images	Adapt	Resp	NI	.00	.78	No difference in overall adaptation
				PreEx NS	Resp	NI	.00	.81	No difference in the effect of natural scenes
Utzerath et al., 2018	22	22	Repetition suppression with images of objects	Learned Imp	Image	D	.09	.045	Different effects of expectations in V1 activity
Tulver et al., 2019	0	44	Illusory contours	PreEx NS	Resp	NI	.02	.40	No difference in illusion vividness
			Blur matching	PreEx NS	Resp	NI	.00	.75	No difference in the effect of word knowledge
			Mooney face recognition	PreEx, S	Resp	R	.16	.009	Better overall recognition
				PreEx, S	Resp	NI	.04	.17	No difference in the benefit of orientation
			Representational momentum	PreEx NS	Resp	NI	.00	.75	No difference in magnitude of displacement
van Laarhoven et al., 2019	30	30	Self-produced vs externally produced sounds	PreEx NS	Image	I		.03	Less N1 attenuation
Amoruso et al., 2019	24	24	Video action identification with contextual cues	Learned Imp	Resp	I	.07	.023	Less use of cue regularities
Vogel et al., 2019	26	0	Time experience questionnaire	Other	Surv	D			Interrupted experience of time
Greene et al., 2019	25	18	Anticipation of target location	Learned Exp	ET	I		.007	Fewer visits in unexpected cue-predicted locations
Seymour et al., 2019	18	18	Visual grating viewing	Other	Image	D		.032	Reduced V4toV1 feedback connectivity
Edey et al., 2019	24	25	Motor-auditory temporal synchronisation	Learned Imp	Resp	NI	.00	.66	No difference in central tendencies

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	*p*	Finding
	22	21	Motor-visual temporal synchronisation	Learned Imp	Resp	I	.14	.013	Less central tendency bias
Grisoni et al., 2019	20	22	Auditory semantic oddball	Learned Imp	Image	I	.22	.002	Less prediction potential during anticipation
				PreEx NS	Image	D	.07	.038	Less MMN reduction for 'talk' oddball in whistle standard
Chiodo et al., 2019	33	30	Categorical sound perception	PreEx NS	Resp	NI	.01	.89	No difference in categorisation
			Word and nonword recall	Other	Resp	ND	.06	.15	No difference in the effects of lexicality on recall
Król & Król, 2019	21	23	Repeated Mooney image recognition	PreEx NS	Resp	NI	.02	.36	No difference in overall recognition accuracy
				Learned Exp	ET	I		.02	Less decrease in number of fixations in 2nd presentation
Arthur et al., 2019	0	83	Material-weight illusion	PreEx NS	Resp	NI	.01	.34	No difference in weight ratings
Lieder et al., 2019	37	32	Sequential tone discrimination	Learned Imp	Resp	I	.06	.044	Less contraction towards recent trials
	16	23	Sequential tone discrimination	Learned Imp	Resp	I		.03	Less contraction towards recent trials
Utzerath et al., 2019	22	22	Kanizsa Triangle	PreEx NS	Image	NI		.14	No difference in the effect of the illusion in V1 activity
Ganglmayer et al., 2020	71	72	Stimulus location anticipation	Learned Imp	ET	I	.04	.013	Less frequent first anticipatory looks to target
van Laarhoven et al., 2020	29	29	Occasional missing sound from clapping video	PreEx NS	Image	I	.07	.042	Greater N1 in omissions
Ioannou et al., 2020	12	25	Social image viewing	Other	ET	ND		.62	No difference in face to face eye transitions
Sevgi et al., 2020	0	36	Non-stationary two-armed bandit with social cue	Predict, S	Resp	D	.23	.003	Less difference in card learning rate between conditions
Font-Alaminos et al., 2020	17	18	Repeated auditory tones	Adapt	Image	D	.22	.005	Larger increase in SNR with more repetitions
Coll et al., 2020	0	25	Probabilistic learning in stimulus discrimination	Learned Imp	Image	I		< .001	Smaller increase in IM SNR with more repetitions
Ward et al., 2020	32	24	Face habituation	Adapt	Image	R		.02	Increased habituation in P400
Knight et al., 2020	21	19	Temporal auditory oddball	Learned Imp	Image	NI	.00	.68	No difference in MMN across conditions
Arthur et al., 2020	29	82	Size-weight illusion	PreEx NS	Resp	NI	.00	.69	No difference in size-weight illusion magnitude
	0		Size-weight illusion	PreEx NS	Resp	NI	.02	.25	No difference in size-weight illusion magnitude
Jaffe-Dax & Eigsti, 2020	21	18	Sequential tone discrimination	Learned Imp	Resp	I		.048	Less contraction bias
Bianco et al., 2020	0	78	Video action identification with contextual cues	Learned Imp	Resp	NI		>.05	No effect of AQ subscores

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	p	Finding
			Video shape identification with contextual cues	Learned Imp	Resp	I		.019	Less effect of cues with increased attention switching
Noel et al., 2020	14	25	Virtual space navigation	PreEx NS	Resp	R		.01	Larger scaling of sensory uncertainty with velocity
Andermane et al., 2020	0	69	Binocular rivalry with rotating Gabor patches	PreEx NS	Resp	NI	.00	.98	No difference in rivalry bias
			BR with explicit nonpredictive cue	Learned Exp	Resp	NI	.00	.59	No difference in rivalry bias
			BR with adaptor	Adapt	Resp	NI	.03	.175	No difference in rivalry bias
			BR after imagining one of the stimuli	Learned Imp	Resp	NI	.04	.077	No difference in rivalry bias
			BR with prime	Learned Imp	Resp	NI	.00	.94	No difference in rivalry bias
Allenmark et al., 2021	22	22	Visual search	Learned Imp	RTs	NI	.00	.83	No difference in frequency effects in reaction times
Sapey-Triomphe et al., 2021	31	26	Size discrimination	Learned Imp	Resp	ND	.04	.13	No difference in fit coefficient change between conditions
Sapey-Triomphe et al., 2021b	29	25	Probabilistic associative learning	Learned Imp	Resp	I	.09	.032	Stronger effect of association in ambiguous trials
Lacroix et al. 2021	101	145	Emotional Shifting	Learned Imp, S	Resp	R		.018	Worse shifting performance
			Task Switching	Other	Resp	ND		.045	Removable switching performance interaction
Ward et al., 2021	23	21	Stimulus location detection	Learned Imp	RTs	NI			No difference in reaction times between conditions
Rybicki et al., 2021	28	35	Sequential cued tapping	Learned Imp	RTs	NI	.01	.65	No difference in reaction times between conditions
Finnemann et al., 2021	23	26	Force matching	PreEx NS	Resp	NI			No difference in self-produced force attenuation
	23	25	Intentional stimulus binding	Learned Imp	Resp	NI			No difference in estimate shift
Angeletos Chrysaitis et al. 2021	21	152*	Delayed cue integration	Learned Exp	Resp	NI		.08	No difference in prior weighting
		61		Learned Exp	Resp	NI		.64	No difference in prior weighting
Beker et al., 2021	31	21	Sound with or without isochronous cues	Learned Exp	RTs	NI	0	.68	No difference in entrainment between conditions
Sapey-Triomphe et al., 2021	26	26	Probabilistic associative learning	Learned Imp	Resp	NI	.05	.095	No difference in association effects in ambiguous trials
Hudson et al., 2021	23	23	Representational momentum with explicit cues	Learned Exp	Resp	I	.09	.044	Less change in displacement
Retzler et al., 2021	0	222	Incoherent motion direction estimation	Learned Exp	Resp	NI		.34	No difference in DDM start point change between conditions
Vishne et al., 2021	30	47	Motor-visual temporal	Predict	RTs	D		.033	Less correction of errors across trials

Article	ASD	NT	Task	Prior	Meas	Res	η_p^2	*p*	Finding
			synchronisation (isochronous)						
			Motor-visual temporal synchronisation (changing)	Predict	RTs	D		< .008	Less adaptation to changing tempos
Van de Cruys et al., 2021	24	25	Odd-one-out visual search	PreEx NS	RTs	NI	.02	.35	No performance difference between target orientations
				Learned Imp	RTs	NI	.00	.85	No difference in role reversal effects
Perrykkad et al., 2021	0	40	Judgement of agency	Predict	ET	D		.001	Less change in fixation error across changing variability
Arthur et al., 2021	26	54	Virtual reality racquetball	PreEx NS	Resp	D	.06	.027	Worse performance in volatile condition
Noel et al., 2021	17	25	Gabor orientation estimation	PreEx NS	Resp	NI		.23	Similar prior weights
				Learned Imp	Resp	I			Less change to prior weights between conditions

NT and ASD denote the respective sample sizes. AQ findings correspond to rows with ASD equal to 0 or NT including an asterisk. Possible prior types are pre-existing (PreEx) which is further divided into social (S) and non-social (NS), learned which is divided into implicit (Imp) and explicit (Exp), adaptation, prediction, and other. The measures used (Meas) are categorised into distribution of behavioural responses (Resp), neuroimaging (Image), reaction times (RTs), and eye-tracking (ET). The column 'Res' shows the classification of results in relation to the imbalance hypothesis as shown in Figure 2.1.

2.3.6 Bayesian modelling

Unfortunately (and perhaps surprisingly), only 13 studies in our pool used any sort of Bayesian, theory-driven computational methods. In this section we will present the details of their models and the results they yielded. Note that we refer specifically to the computational results, as the behavioural results of these studies have already been included in previous sections.

Zaidel et al. (2015) used sensory thresholds from single-cue conditions to predict Bayes-optimal multisensory integration thresholds. They find that experimentally observed thresholds in both diagnostic groups are similar to the theoretically expected ones. They also used simulations of Bayesian inference models with changing priors to illustrate how group differences can result from learning deficits in ASD. Karaminis et al. (2016) used the Weber

fractions estimated from a time discrimination task to calculate the participants' likelihood variances. These were combined with measures of bias in a time reproduction task to estimate prior precision, finding wide priors and even wider likelihoods in ASD. They also used simulations of a Bayesian model to argue that priors are wider than optimal in ASD. They do not account for possible prior influences in the discrimination task. Noel et al. (2021) calculated the predicted bias in an orientation estimation task as a function of the participants' Fisher information, a measure of the encoded information about the stimulus. Fisher information was modelled using two free parameters: one determining the total amount of information and one the resources allocated to the prior distribution versus a uniform stimulus distribution. They found similar initial allocation between autistic and neurotypical participants, something that changed throughout the experiment with only neurotypical participants updating their allocation of encoding resources to match the stimulus statistics. They also showed overall reduced Fisher information in ASD.

Powell et al. (2016) used two tasks in which participants had to compare the speed of two stimuli in each trial. In one of the tasks, the contrast varied across stimuli, while in the other, moving stimuli were presented either with a static or a moving point of reference. The researchers assumed that, under the imbalance hypothesis, the slow-speed prior variance would scale linearly with autistic traits ($\sigma_p^2 = k \cdot AQ$) in both tasks. They initially calculated the participants' perceptual thresholds (Δ) separately in two conditions (A, B) with different sensory noise (contrast or reference motion). Then, they estimated the point of subjective equality between conditions (V_A/V_B) from comparisons between them and fitted them using a Bayesian model in which prior widths were modulated by autistic traits:

$$\frac{V_A}{V_B} = \frac{k \cdot AQ + \Delta_A^2}{k \cdot AQ + \Delta_B^2},$$

with one free parameter, *k*. This model outperformed one based only on priors ($V_A/V_B = k \cdot AQ$) and one based only on sensory evidence ($V_A/V_B = k \cdot \Delta_A^2/\Delta_B^2$). This was the main finding of the study, showing flatter priors with higher AQ values.

Pantelis and Kennedy (2017) used a task in which participants had to estimate the direction of a person's gaze from a photograph. They superimposed a picture of a room on that photograph, expecting that environmental cues would bias participants. The researchers estimated the likelihood of gaze direction in trials with no environmental cues, and then estimated the influence of the cues in the other trials, using a Bayesian model. The model assumed that the probability of a perceived gaze direction given the presented gaze direction depends on the precomputed likelihood and the picture salience, weighted by a free parameter. Prior weightings did not differ between groups.

Karvelis et al. (2018), in a motion direction estimation task with a learned prior for the distribution of the motion directions, used a Bayesian model which assumed the general (bimodal) structure of the prior and had four free parameters: mode location, prior precision, likelihood/sensory precision, and a parameter expressing the probability of participants responding at random. This model was compared with other Bayesian and non-Bayesian models, which it outperformed. The results showed higher sensory precision with increased AQ, but no difference in prior variance. Noel et al. (2018) used a similar model in a simultaneity judgement task, with a greater number of parameters: prior strength, two likelihood precisions (one for each possible judgement), sensory noise, and two likelihood means. The values of half the parameters were fixed based on previous results, leaving the other half to be fit. They showed stronger coupling priors in autistic participants, but no differences in the other parameters. Noel et al. (2020) fit a Bayesian model with exponential slow-speed priors for angular and radial velocities, and likelihood functions with scaling variance based on noisy measurements of self-movement, resulting in four free parameters. This was compared with a model assuming flat priors but leaky integration of velocity measurements, which it outperformed. Results showed similar priors but increased radial likelihood scaling in ASD in the no-feedback condition.

To model the development of priors over time, Manning et al. (2017) used an ideal Bayesian learner which tracked the reward probabilities (r) and their volatility (v) over trials t, while also estimating the meta-volatility (k):

$$p(r_{t+1}|r_t, v) = \beta(r_t, e^v)$$

$$p(v_{t+1}|v_t, k) = N(v_i, e^k) \ ,$$

with β and N corresponding to the beta and the normal distributions. All parameters were estimated from the task information, so that the model had no free parameters. This was compared with three other, Bayes-suboptimal models, which it outperformed in both diagnostic groups. For the quantification of learning, the researchers used a simple reinforcement learning model, which showed no differences between the groups.

Angeletos Chrysaitis et al. (2021) used a circular inference model in a probabilistic decision-making task. The model combined the logit prior (L_p) and the logit likelihood of each trial (L_s) to predict participant logit confidence estimates (L_c) as follows:

$$L_c = F\big(L_p + F\big(a_p L_p, w_p\big), w_p\big) + F(L_s + F(a_s L_s, w_s), w_s) \ ,$$

$$\text{where } F(L, w) = \ln\left(\frac{we^L + 1 - w}{(1-w)e^L + w}\right).$$

The model had 4 free parameters: prior and likelihood weightings (w_p and w_s), and prior and likelihood reverberation or overcounting (a_p and a_s). It was compared with and outperformed an implementation of circular inference which included interference between the priors and likelihoods, a simpler weighted Bayes model, and a perfect Bayesian one. No statistically significant relationship was found between any of the parameters and diagnostic groups or autistic traits after accounting for multiple comparisons.

Three studies utilised Hierarchical Gaussian Filters (HGF) to model the Bayesian learning processes of the participants (Lawson et al., 2017; Sapey-Triomphe et al., 2021c; Sevgi et al., 2020). All of them used a three-level HGF, where the first level (x_1) corresponds to the object of each trial (i.e., what participants are asked to estimate/respond to), the second (x_2) to the coupling between contextual cues and objects, and the third to the volatility of the environment (x_3). Three-level HGFs are usually formalised within the following equations:

$$P(x_1^t = 1) = \frac{1}{1 + e^{-x_2^t}}$$

$$x_2^t \sim N\left(x_2^{t-1}, \mathrm{e}^{\kappa x_3^t + \omega}\right)$$

$$x_3^t \sim N(x_3^{t-1}, \theta)$$

The free parameters are κ and ω, which determine how volatility, x_3, influences coupling strength, and θ which represents meta-volatility or how much volatility changes over time.

Lawson et al. (2017) fixed $\kappa = 1$ at their implementation of the HGF, which was fitted on log reaction times via a response model, which had six free parameters (the HGF had two). This model was compared with two simple reinforcement learning models with fixed and dynamic learning rates and a two-level HGF, which it outperformed. Their results showed increased θ values, smaller changes in the 2nd level learning rate across conditions, and larger changes in the 3rd level learning rate in ASD. Sapey-Triomphe et al. (2021c) used the same HGF but probabilistically combined its output with the priming influence of the previous trial, the sensory memory of the previous ambiguous trial, and the sensory input. Eight models were compared with each other, based on all the possible combinations of coupling (HGF), priming, and sensory memory. All of them were fitted on both the predictions the participants made before seeing each stimulus and their subsequent perceptual responses. The winning model included only associations and priming and had four free parameters: ω, θ, and the combination weights (or precisions) of association and priming. The results showed smaller association precision in ASD. Sevgi et al. (2020), used a version of the HGF with fixed $\omega = 0$, which integrated a social and a non-social contextual cue. It had six free parameters: a κ and a θ for each cue, a weighting parameter for their combination, and a temperature parameter for the softmax response function. This was the winning model among seven others, both HGF and non-HGF variants, although none of the HGF ones had a non-fixed ω. The model was fitted separately on participant responses for the two conditions of the task. Results showed a two-way interaction effect between diagnostic group and experimental condition on social κ values, with low AQ participants changing their value more between conditions.

2.3.6.1 Bayesian modelling: Discussion

Theoretically, Bayesian models have the potential to resolve some of the contradictions observed in the literature, by revealing nuances that simple behavioural analysis is too crude to discover. In practice, though, their findings are equally mixed. Some show no imbalance evidence, while others do support the presence of an imbalance, with a few showing evidence for the reverse imbalance of strengthened priors in autism. Learning models mostly find evidence in favour of differences between groups, although these are not trivially interpretable in terms of support for the imbalance hypothesis.

A possible reason for the heterogeneity of these findings is the heterogeneity of the models themselves. All studies in our pool used different models. The computational methods ranged from simple simulations to multi-level Bayesian learners, from models with no free parameter to one with eight. Moreover, even similar models differed in various explicit and implicit ways. For example, while both Sevgi et al. (2020) and Lawson et al. (2017) used a three-level HGF, they each fixed a different parameter, leaving the motivation for their choice unclear. A more subtle difference can be found between the model of Karvelis et al. (2018), which assumes that the percept is the mean of the posterior distribution, and that of Karaminis et al. (2016), which randomly draws the percept from the posterior. Another modelling choice which was not explicitly stated in most articles revolves around the mean of the likelihood. For example, Pantelis and Kennedy (2017) model the task data using the same likelihood function for all trials with the same stimulus. The alternative would be to assume that, due to sensory noise, sensory inputs are drawn from a distribution centred on the stimulus, and therefore differ in each trial (e.g. Karvelis et al., 2018; Powell et al., 2016).

Finally, most studies incompletely satisfied the current recommended modelling guidelines (as stated, for example, by Wilson and Collins, 2019). Specifically, not all studies tested more than one model, few explored alternative explanations to the Bayesian framework, only one study performed model recovery (Angeletos Chrysaitis et al., 2021), only two performed parameter recovery (Angeletos Chrysaitis et al., 2021; Karvelis et al., 2018), and only few actually validated their models by comparing the models' behaviour to the experimentally observed

data. This is a serious limitation as these steps are necessary to understand how reliable modelling results are.

2.4 Discussion

In this comprehensive review, we looked at all studies that experimentally investigated Bayesian and predictive coding theories of autism spectrum disorder. We specifically focused on the imbalance hypothesis, which states that prior beliefs have a weaker influence in the perception and cognition of autistic individuals. This hypothesis describes the common structure between all relevant such theories (Pellicano and Burr, 2012; Brock, 2012; Lawson et al., 2014; Van de Cruys et al., 2014). We found a wide variety of innovative studies that explored this hypothesis using neuroimaging, behavioural, and computational means, across perceptual, decision-making, and social contexts.

Surprisingly, the majority of the findings did not support the presence of a prior-likelihood imbalance. This comes in contrast with the popularity of these theories (Haker et al., 2016; Palmer et al., 2017) and argues against a broad formulation of the imbalance hypothesis. However, more than a third of all studies presented evidence for reduced prior influences in ASD. Furthermore, our *p*-curve analysis found that study results contain evidential value, which shows that imbalances cannot be completely dismissed either. Effect sizes for the imbalance hypothesis were relatively small (median $\eta_p^2 = 0.04$). One possible reason for the mixed findings is that imbalances between priors and likelihoods could be produced by a variety of underlying mechanisms. Top-down signals could be underweighted or bottom-up ones overweighted in their integration. Alternatively, integration might be normal in ASD, but some priors could be less precise as a result of learning deficiencies. On the other hand, likelihoods might not be universally narrower in autistic individuals, but have less adaptable precision depending on the reliability of sensory information, as argued by Van de Cruys et al (2014).

We attempted to resolve some of the ambiguities by categorizing findings depending on the respective prior types. Specifically, priors which were developed during the experimental tasks were separated from those that the participants were expected to have acquired before the experiment, based on past experiences. Then, learned priors were split into explicit and implicit ones, depending on how the information was presented to the participants. Pre-existing priors were also split into social and non-social priors. We found that paradigms using implicitly learned priors resulted in equally many findings supporting the imbalance hypothesis and supporting no imbalance in ASD. This was the prior type that exhibited the strongest evidence in favour of the hypothesis, were closely followed by paradigms with pre-existing social priors. These results hint at an impairment in the way priors are developed in autism. If autistic individuals need more information or time relative to neurotypical participants to develop prior beliefs, differences between the groups would manifest more when the environmental regularities have to be extracted within a limited amount of time. Furthermore, it might be the case that social priors, which are more complex, are more difficult to develop and therefore learning impairments would affect them disproportionately. 'Prediction' tasks, where responses were presumably directly drawn from participant priors, also provided evidence for differences in prior acquisition. If this is the source of the precision imbalances in autism, future research will need to clarify whether the impairment lies in the learning of longer-term statistics or just use of recent trials (Lieder et al., 2019; but see Sapey-Triomphe et al., 2021c; Andermane et al., 2020 for no differences in priming).

We found that published behavioural studies lead more often to null findings than neuroimaging ones. A possible explanation is that neuroimaging methods do not only measure the influence of priors, but also differences in related, but distinct brain processes. For example, they might correspond to attentional shifts or updating processes that take place after the prior-likelihood integration. A diametrically opposite interpretation is that autistic participants utilise alternative strategies to compensate for reduced prior influences. In that case, their behavioural performance would be similar, but neuroimaging measures would reveal the underlying compensatory mechanisms. This interpretation might also explain why designs with explicitly learned priors showed frequent null results despite the aforementioned possible learning

impairments in autism. Studies that looked at AQ instead of ASD diagnoses also more frequently led to null results, which raises questions about the dimensional view of autism. It should be noted that studies were not always consistent in how they selected their neurotypical samples, e.g., if they clinically verified that neurotypical participants were not autistics or if they included AQ cut-offs.

A related theory to the imbalance hypothesis, that has seen some popularity in recent years, revolves around the updating of priors with new evidence and specifically the processing of environmental volatility (Palmer et al., 2017). This theory accepts the possibility of an imbalance between likelihood and prior precisions, but its focus is on how this imbalance affects the Bayesian updating of beliefs, instead of the inference about a specific stimulus. A seminal study supporting this approach has been the one by Lawson et al. (2017), which showed increased meta-volatility in ASD or how much the estimate of volatility changes over time. However, following studies with similar models have not reproduced this finding, instead finding complex interaction effects between conditions and groups (Sevgi et al., 2020) or no differences in the relevant model parameters (Sapey-Triomphe et al., 2021c). It should be noted, though, that no study has used the exact same model implementation. Non-modelling studies with similar focus have shown worse performance in volatile conditions in ASD (Arthur et al., 2021) and less change in eye-tracking measures following a change in variability (Perrykkad et al., 2021). It is unclear what these findings reveal about volatility processing in autism.

Another promising theory might be the specific formulation of the imbalance hypothesis called High, Inflexible Precision of Prediction Errors in Autism (Van de Cruys et al., 2014). HIPPEA proposes that likelihood precision in autism is not uniformly high. Instead, precision cannot flexibly change depending on the stimulus and the context, remaining high even when that is suboptimal. This could be a source of variability in the findings: differences would be observed only when neurotypical or low-AQ participants assign low enough precision to sensory information, and not otherwise. The theory also focuses on prior development, highlighting how assigning the same unduly high precision to all sensory inputs would lead to priors that are shaped by noise and are unlikely to be broadly useful. Some studies have attempted to

directly research this theory (e.g. Sapey-Triomphe et al., 2021b) with, once again, no clear conclusion.

On a methodological note, only 13 out of the 83 studies used Bayesian modelling to analyse their behavioural data. While computational models should theoretically enhance the clarity of the findings, those used in these studies were very heterogeneous, from simple simulations and group-level models to hierarchical learning models that describe individual differences. Taken as a whole, these studies yielded mixed findings, both supporting and contradicting the imbalance hypothesis. Due to the diversity of the models and the lack of within-study model comparisons, it is unclear if the differences in findings originated from the differences in the computational techniques or in the task designs.

A related systematic review published in 2021 also attempted to understand prediction impairments in ASD, although exclusively focusing in autistic individuals (Cannon et al., 2021). Their approach differed from ours in its theoretical basis, as it significantly relied on the Predictive Impairments in Autism hypothesis (Sinha et al., 2014) instead of the Bayesian formulations (Pellicano and Burr, 2012; Brock, 2012). It also included a completely different pool of articles, with 25 of their 47 studies not being included in our pool and 61 of our studies not included in theirs. The authors categorised the studies in 7 categories by ‘bases for prediction’, which would include both our broad prior categories and our within-section groupings (e.g., learned, explicit, social, frequency priors, cue associations). They also classified the studies into 11 categories by ‘evidence of prediction’, that is what type of response was measured by the researchers (e.g., specific neuroimaging measures, adaptation, reaction times). This greatly showcases the heterogeneity of the literature, something we have also grappled with. Their conclusions did partially reflect our own, finding impaired learning of predictive associations in ASD. One additional conclusion in that review was that autistic individuals differ in low-level predictive processing in the brain, largely drawn from habituation and adaptation studies.

2.4.1 Limitations of the literature

Despite the variety of the reviewed studies, we noticed some serious limitations in the presented approaches. Firstly, in this review we decided to group Bayesian and predictive coding approaches together. This choice was made because most studies do not differentiate between the two frameworks, although these do not necessarily coincide. A confusion between theoretical frameworks is also present in a broader way. There are multiple theories for autism that seem to share the same basic components. Weak Central Coherence (Frith, 1990), Enhanced Perceptual Functioning (Mottron et al., 2006), and in general local/global processing imbalances have the same basic structure as Bayesian or predictive coding theories, in that they all propose that autistic individuals overweight bottom-up, specific information relative to top-down, general information. Moreover, Predictive Impairments in Autism (Sinha et al., 2014) also proposes that such individuals have difficulties in learning the probabilistic associations of their environment. How could we distinguish between these theories? Are all of them formulations of the same fundamental mechanism and, if so, what is the added benefit of approaching this problem in Bayesian terms? It is very difficult to disentangle attenuated priors or overly precise likelihoods from other forms of top-down/bottom-up imbalances, something that gets significantly exacerbated by the fact that most of the reviewed studies did not offer any justification for their chosen interpretation. A follow-up question regards how complex or abstract prior beliefs are. Should we expect, for example, priors consisting of simple frequency information (e.g. Karvelis et al., 2018; Lieder et al., 2019) to arise from different mechanisms than coupling priors in cue integration (e.g. Bedford et al., 2016; Noel et al., 2018) or associations with contextual cues (e.g. Lawson et al., 2017)? Might that be the reason that social priors (e.g. Chambon et al., 2017; Walsh et al., 2018) seem to be more attenuated in autism, as they are presumably 'situated' higher in the cognitive hierarchy?

These ambiguities cannot be resolved without a clear framework for the hierarchy of priors in the brain and possibly its implementation in computational models. In spite of that, only few studies use theory-driven models and even fewer perform model comparisons, so as to make sure that their chosen model is the best description of the data. This is made worse by the fact

that models largely differ across studies, leaving the reader unable to compare their findings. It is also striking that the recommended guidelines for computational modelling (Wilson and Collins, 2019) are often incompletely followed, with parameter and model recovery being omitted from almost all of the studies.

A final, important issue is that most studies did not perform power analysis and used small sample sizes, resulting in an estimated statistical power of 42% for the studies that investigated the imbalance hypothesis and 39% overall. Unfortunately, low power is commonplace in the cognitive and biomedical sciences (Button et al., 2013; Dumas-Mallet et al., 2017; Szucs and Ioannidis, 2017). We found slightly higher power for imaging studies in our pool than related fields, but lower power for behavioural studies (Dumas-Mallet et al., 2017; Nord et al., 2017), presumably because of the low effect sizes in our pool. Low statistical power raises serious questions about the replicability of their findings, as it is one of the main factors underlying the replication crisis (Ioannidis, 2005). It also leads to an overestimation of effect sizes and adds noise to any conclusions drawn from the synthesis of study findings. No exact replication of the studies in our pool has been performed.

2.4.2 Limitations of our review

The greatest limitation of our review lies in the degrees of freedom we had during our analysis. As researchers were not usually clear in how they conceptualised the priors measured by their designs, we had to interpret and categorise them based on our own understanding. Similarly, we had to select the most relevant findings of each study for the 2.*3.5 Summary statistics* section. In both cases, we made our choice depending on our understanding of the experiment and the spirit of the analysis by the researchers. However, we were not blind to the study conclusions during this process, as these are mentioned in the study abstracts, something that theoretically could have introduced biases in our categorisation process. We were not able to have independent reviewers assessing each study, which would have minimised the possibility of bias on our part.

The conclusions of this review are drawn from an overview of the findings and the use of crude summary statistics. No meta-analysis was possible, given the statistics provided. Furthermore, the heterogeneity of the measures and techniques used in the literature to assess the prior-likelihood imbalances means that no single number could summarise the available evidence. Another limitation of this review was that we categorised the paradigms based on the development and the type of priors, but the choice of category for each paradigm was far from trivial. The imperfection of our categories might be an additional reason for the absence of a clear conclusion. Finally, given the scope of our literature search, various studies were not included in this review simply based on the choice of keywords and the interpretations provided by the researchers. This means that studies with possible Bayesian implications or even identical designs to our chosen studies were not included in this review (e.g. Solomon et al., 2011; Stevenson et al., 2014).

2.4.3 Conclusions and directions of future research

Many studies have investigated Bayesian and predictive coding theories in the 10 years since the seminal article of Pellicano and Burr in 2012. While the breadth of the research is impressive, synthesizing all the findings into a coherent whole has proven to be an almost impossible task. The results of the studies have been mixed and occasionally conflicting. In an attempt to make sense of the experimental evidence, we categorised studies based on how their priors were developed. This revealed a tendency towards priors learned throughout the experiment being impaired and pre-existing ones being intact. It also showed some differences when responses were drawn directly from priors. These results might suggest that rather than a general impairment in the integration of Bayesian priors, autism is characterised by a deficit in prior development. We also showed very small effect sizes, especially for pre-existing priors, and significantly low statistical power. These, combined with the variability of the findings, make any conclusions uncertain.

Based on our review we recommend that ensuing studies heed the following guidelines:

1. More clarity could be obtained in a study if researchers explain how they conceptualise priors/predictions and likelihoods/prediction errors in the context of the experiment, what the effects of their combination are expected to be, and how an imbalance would manifest in the data. Ideally this would be accompanied by diagrams and equations to describe the proposed mechanisms that underly the observed findings.
2. When possible, Bayesian/predictive coding computational models should be used to analyse the study findings.
    a. Comparing these models with non-Bayesian ones could be used to confirm the validity of the Bayesian approach, while comparison with other Bayesian models would support their exact choice of model structure.
    b. It is also important that model and parameter recovery is performed to verify the reliability of the model estimates, as well as a validation that the model is able to reproduce the relevant patterns in the data (Wilson and Collins, 2019).
3. Pre-registering of the hypothesis, experimental design, and analysis is a useful tool to constraint the degrees of freedom used in interpreting the results and minimise the bias towards positive findings.
4. Sample sizes should be increased, motivated by power analysis.
5. Studies that attempt to measure prior acquisition in conjunction with prior-likelihood imbalances throughout the experiment could clarify whether imbalance findings originate in learning or integration differences.

We believe that the inclusion of computational approaches to psychiatry can significantly aid our understanding of autism and other mental disorders. However, the field is currently young and relatively immature, which is evident both from the heterogeneity of the studies and the lack of formal structure in the analysis and interpretation of the results. Our suggestions could provide following studies with some necessary rigor, potentially aiding in future understanding of autism.

Chapter 2 Acknowledgements

The authors would like to thank Stelios Gkionis for his feedback and insights, without which this work might not have been possible, and Stephen Lawrie and Renaud Jardri for their suggestions and encouragement. Funding: This work was supported by the United Kingdom Research and Innovation [grant EP/S02431X/1], UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Declarations of interest: none. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

Chapter 2 References

Adams, R.A., Shipp, S., Friston, K.J., 2013. Predictions not commands: active inference in the motor system. Brain Struct. Funct. 218, 611–643. https://doi.org/10.1007/s00429-012-0475-5

Aitchison, L., Lengyel, M., 2017. With or without you: predictive coding and Bayesian inference in the brain. Curr. Opin. Neurobiol. 46, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Allenmark, F., Shi, Z., Pistorius, R.L., Theisinger, L.A., Koutsouleris, N., Falkai, P., Müller, H.J., Falter-Wagner, C.M., 2021. Acquisition and Use of 'Priors' in Autism: Typical in Deciding Where to Look, Atypical in Deciding What Is There. J. Autism Dev. Disord. 51, 3744–3758. https://doi.org/10.1007/s10803-020-04828-2

American Psychiatric Association (Ed.), 2013. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. ed. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Amoruso, L., Narzisi, A., Pinzino, M., Finisguerra, A., Billeci, L., Calderoni, S., Fabbro, F., Muratori, F., Volzone, A., Urgesi, C., 2019. Contextual priors do not modulate action prediction in children with autism. Proc. R. Soc. B Biol. Sci. 286, 20191319. https://doi.org/10.1098/rspb.2019.1319

Andermane, N., Bosten, J.M., Seth, A.K., Ward, J., 2020. Individual differences in the tendency to see the expected. Conscious. Cogn. 85, 102989. https://doi.org/10.1016/j.concog.2020.102989

Angeletos Chrysaitis, N., Jardri, R., Denève, S., Seriès, P., 2021. No increased circular inference in adults with high levels of autistic traits or autism. PLOS Comput. Biol. 17, e1009006. https://doi.org/10.1371/journal.pcbi.1009006

Arthur, T., Harris, D., Buckingham, G., Brosnan, M., Wilson, M., Williams, G., Vine, S., 2021. An examination of active inference in autistic adults using immersive virtual reality. Sci. Rep. 11, 20377. https://doi.org/10.1038/s41598-021-99864-y

Arthur, T., Vine, S., Brosnan, M., Buckingham, G., 2020. Predictive sensorimotor control in autism. Brain 143, 3151–3163. https://doi.org/10.1093/brain/awaa243

Arthur, T., Vine, S., Brosnan, M., Buckingham, G., 2019. Exploring how material cues drive sensorimotor prediction across different levels of autistic-like traits. Exp. Brain Res. 237, 2255–2267. https://doi.org/10.1007/s00221-019-05586-z

Aru, J., Tulver, K., Bachmann, T., 2018. It's all in your head: Expectations create illusory perception in a dual-task setup. Conscious. Cogn. 65, 197–208. https://doi.org/10.1016/j.concog.2018.09.001

Baron-Cohen, S., Leslie, A.M., Frith, U., 1985. Does the autistic child have a "theory of mind" ? Cognition 21, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E., 2001. The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. J. Autism Dev. Disord. 31, 5–17. https://doi.org/10.1023/A:1005653411471

Bedford, R., Pellicano, E., Mareschal, D., Nardini, M., 2016. Flexible integration of visual cues in adolescents with autism spectrum disorder: Flexible integration of visual cues in adolescents with ASD. Autism Res. 9, 272–281. https://doi.org/10.1002/aur.1509

Beker, S., Foxe, J.J., Molholm, S., 2021. Oscillatory entrainment mechanisms and anticipatory predictive processes in children with autism spectrum disorder. J. Neurophysiol. 126, 1783–1798. https://doi.org/10.1152/jn.00329.2021

Bianco, V., Finisguerra, A., Betti, S., D'Argenio, G., Urgesi, C., 2020. Autistic Traits Differently Account for Context-Based Predictions of Physical and Social Events. Brain Sci. 10, 418. https://doi.org/10.3390/brainsci10070418

Bishop, D.V.M., Thompson, P.A., 2016. Problems in using *p* -curve analysis and text-mining to detect rate of *p* -hacking and evidential value. PeerJ 4, e1715. https://doi.org/10.7717/peerj.1715

Bravo, F., Cross, I., Hawkins, S., Gonzalez, N., Docampo, J., Bruno, C., Stamatakis, E.A., 2017. Neural mechanisms underlying valence inferences to sound: The role of the right angular gyrus. Neuropsychologia 102, 144–162. https://doi.org/10.1016/j.neuropsychologia.2017.05.029

Brock, J., 2012. Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr. Trends Cogn. Sci. 16, 573–574. https://doi.org/10.1016/j.tics.2012.10.005

Brodski-Guerniero, A., Naumer, M.J., Moliadze, V., Chan, J., Althen, H., Ferreira-Santos, F., Lizier, J.T., Schlitt, S., Kitzerow, J., Schütz, M., Langer, A., Kaiser, J., Freitag, C.M., Wibral, M., 2018. Predictable information in neural signals during resting state is

reduced in autism spectrum disorder. Hum. Brain Mapp. 39, 3227–3240. https://doi.org/10.1002/hbm.24072

Brown, J., Aczel, B., Jiménez, L., Kaufman, S.B., Grant, K.P., 2010. Intact implicit learning in autism spectrum conditions. Q. J. Exp. Psychol. 63, 1789–1812. https://doi.org/10.1080/17470210903536910

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. https://doi.org/10.1038/nrn3475

Cannon, J., O'Brien, A.M., Bungert, L., Sinha, P., 2021. Prediction in Autism Spectrum Disorder: A Systematic Review of Empirical Evidence. Autism Res. 14, 604–630. https://doi.org/10.1002/aur.2482

Chambon, V., Farrer, C., Pacherie, E., Jacquet, P.O., Leboyer, M., Zalla, T., 2017. Reduced sensitivity to social priors during action prediction in adults with autism spectrum disorders. Cognition 160, 17–26. https://doi.org/10.1016/j.cognition.2016.12.005

Chiodo, L., Mottron, L., Majerus, S., 2019. Preservation of categorical perception for speech in autism with and without speech onset delay. Autism Res. 12, 1609–1622. https://doi.org/10.1002/aur.2134

Cohen, J., 1987. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Hillsdale, N.J.

Coll, M.-P., Whelan, E., Catmur, C., Bird, G., 2020. Autistic traits are associated with atypical precision-weighted integration of top-down and bottom-up neural signals. Cognition 199, 104236. https://doi.org/10.1016/j.cognition.2020.104236

Croydon, A., Karaminis, T., Neil, L., Burr, D., Pellicano, E., 2017. The light-from-above prior is intact in autistic children. J. Exp. Child Psychol. 161, 113–125. https://doi.org/10.1016/j.jecp.2017.04.005

Dumas-Mallet, E., Button, K.S., Boraud, T., Gonon, F., Munafò, M.R., 2017. Low statistical power in biomedical science: a review of three human research domains. R. Soc. Open Sci. 4, 160254. https://doi.org/10.1098/rsos.160254

Edey, R., Brewer, R., Bird, G., Press, C., 2019. Brief Report: Typical Auditory-Motor and Enhanced Visual-Motor Temporal Synchronization in Adults with Autism Spectrum Disorder. J. Autism Dev. Disord. 49, 788–793. https://doi.org/10.1007/s10803-018-3725-4

Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415, 429–433. https://doi.org/10.1038/415429a

Ewbank, M.P., von dem Hagen, E.A.H., Powell, T.E., Henson, R.N., Calder, A.J., 2016. The effect of perceptual expectation on repetition suppression to faces is not modulated by variation in autistic traits. Cortex 80, 51–60. https://doi.org/10.1016/j.cortex.2015.10.011

Feinstein, A.R., 1995. Meta-analysis: Statistical alchemy for the 21st century. J. Clin. Epidemiol. 48, 71–79. https://doi.org/10.1016/0895-4356(94)00110-C

Finnemann, J.J.S., Plaisted-Grant, K., Moore, J., Teufel, C., Fletcher, P.C., 2021. Low-level, prediction-based sensory and motor processes are unimpaired in Autism. Neuropsychologia 156, 107835. https://doi.org/10.1016/j.neuropsychologia.2021.107835

Font-Alaminos, M., Cornella, M., Costa-Faidella, J., Hervás, A., Leung, S., Rueda, I., Escera, C., 2020. Increased subcortical neural responses to repeating auditory stimulation in children with autism spectrum disorder. Biol. Psychol. 149, 107807. https://doi.org/10.1016/j.biopsycho.2019.107807

Frith, U., 1990. Autism: explaining the enigma, Repr. ed, Cognitive development. Blackwell, Oxford u.a.

Ganglmayer, K., Schuwerk, T., Sodian, B., Paulus, M., 2020. Do Children and Adults with Autism Spectrum Condition Anticipate Others' Actions as Goal-Directed? A Predictive Coding Perspective. J. Autism Dev. Disord. 50, 2077–2089. https://doi.org/10.1007/s10803-019-03964-8

Gelman, A., Carlin, J., 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect. Psychol. Sci. 9, 641–651. https://doi.org/10.1177/1745691614551642

Gómez, C., Lizier, J.T., Schaum, M., Wollstadt, P., Grützner, C., Uhlhaas, P., Freitag, C.M., Schlitt, S., Bölte, S., Hornero, R., Wibral, M., 2014. Reduced predictable information in brain signals in autism spectrum disorder. Front. Neuroinformatics 8. https://doi.org/10.3389/fninf.2014.00009

Gonzalez-Gadea, M.L., Chennu, S., Bekinschtein, T.A., Rattazzi, A., Beraudi, A., Tripicchio, P., Moyano, B., Soffita, Y., Steinberg, L., Adolfi, F., Sigman, M., Marino, J., Manes, F., Ibanez, A., 2015. Predictive coding in autism spectrum disorder and attention deficit hyperactivity disorder. J. Neurophysiol. 114, 2625–2636. https://doi.org/10.1152/jn.00543.2015

Goris, J., Braem, S., Nijhof, A.D., Rigoni, D., Deschrijver, E., Van de Cruys, S., Wiersema, J.R., Brass, M., 2018. Sensory Prediction Errors Are Less Modulated by Global Context in Autism Spectrum Disorder. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 3, 667–674. https://doi.org/10.1016/j.bpsc.2018.02.003

Greco, T., Zangrillo, A., Biondi-Zoccai, G., Landoni, G., 2013. Meta-analysis: pitfalls and hints. Heart Lung Vessels 5, 219–225.

Greene, R.K., Zheng, S., Kinard, J.L., Mosner, M.G., Wiesen, C.A., Kennedy, D.P., Dichter, G.S., 2019. Social and nonsocial visual prediction errors in autism spectrum disorder. Autism Res. 12, 878–883. https://doi.org/10.1002/aur.2090

Grisoni, L., Moseley, R.L., Motlagh, S., Kandia, D., Sener, N., Pulvermüller, F., Roepke, S., Mohr, B., 2019. Prediction and Mismatch Negativity Responses Reflect Impairments

in Action Semantic Processing in Adults With Autism Spectrum Disorders. Front. Hum. Neurosci. 13, 395. https://doi.org/10.3389/fnhum.2019.00395

Grzywacz, N.M., Balboa, R.M., 2002. A Bayesian Framework for Sensory Adaptation. Neural Comput. 14, 543–559. https://doi.org/10.1162/089976602317250898

Gu, X., Eilam-Stock, T., Zhou, T., Anagnostou, E., Kolevzon, A., Soorya, L., Hof, P.R., Friston, K.J., Fan, J., 2015. Autonomic and brain responses associated with empathy deficits in autism spectrum disorder. Hum. Brain Mapp. 36, 3323–3338. https://doi.org/10.1002/hbm.22840

Gu, X., Zhou, T.J., Anagnostou, E., Soorya, L., Kolevzon, A., Hof, P.R., Fan, J., 2018. Heightened brain response to pain anticipation in high-functioning adults with autism spectrum disorder. Eur. J. Neurosci. 47, 592–601. https://doi.org/10.1111/ejn.13598

Hadad, B.S., Goldstein, E.K., Russo, N.N., 2017. Atypical perception in autism: A failure of perceptual specialization? Autism Res. 10, 1510–1522. https://doi.org/10.1002/aur.1800

Haker, H., Schneebeli, M., Stephan, K.E., 2016. Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? Front. Psychiatry 7. https://doi.org/10.3389/fpsyt.2016.00107

Happé, F.G.E., 1996. Studying Weak Central Coherence at Low Levels: Children with Autism do not Succumb to Visual Illusions. A Research Note. J. Child Psychol. Psychiatry 37, 873–877. https://doi.org/10.1111/j.1469-7610.1996.tb01483.x

Heaton, P., 2003. Pitch memory, labelling and disembedding in autism: Pitch memory, labelling and disembedding in autism. J. Child Psychol. Psychiatry 44, 543–551. https://doi.org/10.1111/1469-7610.00143

Hudson, M., Nicholson, T., Kharko, A., McKenzie, R., Bach, P., 2021. Predictive action perception from explicit intention information in autism. Psychon. Bull. Rev. 28, 1556–1566. https://doi.org/10.3758/s13423-021-01941-w

Hullett, C.R., Levine, T.R., 2003. The Overestimation of Effect Sizes from F Values in Meta-Analysis: The Cause and a Solution. Commun. Monogr. 70, 1–1. https://doi.org/10.1080/03637750302475

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLoS Med. 2, e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannou, C., Seernani, D., Stefanou, M.E., Biscaldi-Schaefer, M., Tebartz Van Elst, L., Fleischhaker, C., Boccignone, G., Klein, C., 2020. Social Visual Perception Under the Eye of Bayesian Theories in Autism Spectrum Disorder Using Advanced Modeling of Spatial and Temporal Parameters. Front. Psychiatry 11, 585149. https://doi.org/10.3389/fpsyt.2020.585149

Izadi-Najafabadi, S., Mirzakhani-Araghi, N., Miri-Lavasani, N., Nejati, V., Pashazadeh-Azari, Z., 2015. Implicit and explicit motor learning: Application to children with

Autism Spectrum Disorder (ASD). Res. Dev. Disabil. 47, 284–296. https://doi.org/10.1016/j.ridd.2015.09.020

Jaffe-Dax, S., Eigsti, I.-M., 2020. Perceptual inference is impaired in individuals with ASD and intact in individuals who have lost the autism diagnosis. Sci. Rep. 10, 17085. https://doi.org/10.1038/s41598-020-72896-6

Karaminis, T., Cicchini, G.M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D., Pellicano, E., 2016. Central tendency effects in time interval reproduction in autism. Sci. Rep. 6, 28570. https://doi.org/10.1038/srep28570

Karvelis, P., Seitz, A.R., Lawrie, S.M., Seriès, P., 2018. Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. eLife 7, e34115. https://doi.org/10.7554/eLife.34115

Knight, E.J., Oakes, L., Hyman, S.L., Freedman, E.G., Foxe, J.J., 2020. Individuals With Autism Have No Detectable Deficit in Neural Markers of Prediction Error When Presented With Auditory Rhythms of Varied Temporal Complexity. Autism Res. 13, 2058–2072. https://doi.org/10.1002/aur.2362

Knill, D.C., Richards, W. (Eds.), 1996. Perception as Bayesian Inference, 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Kohn, A., 2007. Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. J. Neurophysiol. 97, 3155–3164. https://doi.org/10.1152/jn.00086.2007

Król, Magdalena, Król, Michał, 2019. The world as we know it and the world as it is: Eye-movement patterns reveal decreased use of prior knowledge in individuals with autism. Autism Res. 12, 1386–1398. https://doi.org/10.1002/aur.2133

Lacroix, A., Dutheil, F., Logemann, A., Cserjesi, R., Peyrin, C., Biro, B., Gomot, M., Mermillod, M., 2021. Flexibility in autism during unpredictable shifts of socio-emotional stimuli: Investigation of group and sex differences. Autism 136236132110627. https://doi.org/10.1177/13623613211062776

Lawson, R.P., Mathys, C., Rees, G., 2017. Adults with autism overestimate the volatility of the sensory environment. Nat. Neurosci. 20, 1293–1299. https://doi.org/10.1038/nn.4615

Lawson, R.P., Rees, G., Friston, K.J., 2014. An aberrant precision account of autism. Front. Hum. Neurosci. 8. https://doi.org/10.3389/fnhum.2014.00302

Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M., Ahissar, M., 2019. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. Nat. Neurosci. 22, 256–264. https://doi.org/10.1038/s41593-018-0308-9

Lin, I.-F., Mochida, T., Asada, K., Ayaya, S., Kumagaya, S.-I., Kato, M., 2015. Atypical delayed auditory feedback effect and Lombard effect on speech production in high-functioning adults with autism spectrum disorder. Front. Hum. Neurosci. 9. https://doi.org/10.3389/fnhum.2015.00510

Manning, C., Kilner, J., Neil, L., Karaminis, T., Pellicano, E., 2017. Children on the autism spectrum update their behaviour in response to a volatile environment. Dev. Sci. 20, e12435. https://doi.org/10.1111/desc.12435

Maule, J., Stanworth, K., Pellicano, E., Franklin, A., 2018. Color Afterimages in Autistic Adults. J. Autism Dev. Disord. 48, 1409–1421. https://doi.org/10.1007/s10803-016-2786-5

Maule, J., Stanworth, K., Pellicano, E., Franklin, A., 2017. Ensemble perception of color in autistic adults. Autism Res. 10, 839–851. https://doi.org/10.1002/aur.1725

Maurer, C., Chambon, V., Bourgeois-Gironde, S., Leboyer, M., Zalla, T., 2018. The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder. Cognition 172, 1–10. https://doi.org/10.1016/j.cognition.2017.11.007

Mottron, L., Dawson, M., Soulières, I., Hubert, B., Burack, J., 2006. Enhanced Perceptual Functioning in Autism: An Update, and Eight Principles of Autistic Perception. J. Autism Dev. Disord. 36, 27–43. https://doi.org/10.1007/s10803-005-0040-7

Murphy, K.R., 2017. What inferences can and cannot be made on the basis of meta-analysis? Hum. Resour. Manag. Rev. 27, 193–200. https://doi.org/10.1016/j.hrmr.2015.06.001

Noel, J.-P., Lakshminarasimhan, K.J., Park, H., Angelaki, D.E., 2020. Increased variability but intact integration during visual navigation in Autism Spectrum Disorder. Proc. Natl. Acad. Sci. 117, 11158–11166. https://doi.org/10.1073/pnas.2000216117

Noel, J.-P., Stevenson, R.A., Wallace, M.T., 2018. Atypical audiovisual temporal function in autism and schizophrenia: similar phenotype, different cause. Eur. J. Neurosci. 47, 1230–1241. https://doi.org/10.1111/ejn.13911

Noel, J.-P., Zhang, L.-Q., Stocker, A.A., Angelaki, D.E., 2021. Individuals with autism spectrum disorder have altered visual encoding capacity. PLOS Biol. 19, e3001215. https://doi.org/10.1371/journal.pbio.3001215

Nord, C.L., Valton, V., Wood, J., Roiser, J.P., 2017. Power-up: A Reanalysis of "Power Failure" in Neuroscience Using Mixture Modeling. J. Neurosci. 37, 8051–8061. https://doi.org/10.1523/JNEUROSCI.3592-16.2017

O'riordan, M.A., 2004. Superior Visual Search in Adults with Autism. Autism 8, 229–248. https://doi.org/10.1177/1362361304045219

Palmer, C.J., Lawson, R.P., Hohwy, J., 2017. Bayesian approaches to autism: Towards volatility, action, and behavior. Psychol. Bull. 143, 521–542. https://doi.org/10.1037/bul0000097

Palmer, C.J., Paton, B., Kirkovski, M., Enticott, P.G., Hohwy, J., 2015. Context sensitivity in action decreases along the autism spectrum: a predictive processing perspective. Proc. R. Soc. B Biol. Sci. 282, 20141557. https://doi.org/10.1098/rspb.2014.1557

Pantelis, P.C., Kennedy, D.P., 2017. Deconstructing atypical eye gaze perception in autism spectrum disorder. Sci. Rep. 7, 14990. https://doi.org/10.1038/s41598-017-14919-3

Papathomas, T.V., 2017. The Hollow-Mask Illusion and Variations. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199794607.003.0087

Pell, P.J., Mareschal, I., Calder, A.J., von dem Hagen, E.A.H., Clifford, C.W.G., Baron-Cohen, S., Ewbank, M.P., 2016. Intact priors for gaze direction in adults with high-functioning autism spectrum conditions. Mol. Autism 7, 25. https://doi.org/10.1186/s13229-016-0085-9

Pellicano, E., Burr, D., 2012. When the world becomes 'too real': a Bayesian explanation of autistic perception. Trends Cogn. Sci. 16, 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Perrykkad, K., Lawson, R.P., Jamadar, S., Hohwy, J., 2021. The effect of uncertainty on prediction error in the action perception loop. Cognition 210, 104598. https://doi.org/10.1016/j.cognition.2021.104598

Powell, G., Meredith, Z., McMillin, R., Freeman, T.C.A., 2016. Bayesian Models of Individual Differences: Combining Autistic Traits and Sensory Thresholds to Predict Motion Perception. Psychol. Sci. 27, 1562–1572. https://doi.org/10.1177/0956797616665351

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87. https://doi.org/10.1038/4580

Retzler, C., Boehm, U., Cai, J., Cochrane, A., Manning, C., 2021. Prior information use and response caution in perceptual decision-making: No evidence for a relationship with autistic-like traits. Q. J. Exp. Psychol. 74, 1953–1965. https://doi.org/10.1177/17470218211019939

Rybicki, A.J., Galea, J.M., Schuster, B.A., Hiles, C., Fabian, C., Cook, J.L., 2021. Intact predictive motor sequence learning in autism spectrum disorder. Sci. Rep. 11, 20693. https://doi.org/10.1038/s41598-021-00173-1

Sapey-Triomphe, L.-A., Temmerman, J., Puts, N.A.J., Wagemans, J., 2021a. Prediction learning in adults with autism and its molecular correlates. Mol. Autism 12, 64. https://doi.org/10.1186/s13229-021-00470-6

Sapey-Triomphe, L.-A., Timmermans, L., Wagemans, J., 2021b. Priors Bias Perceptual Decisions in Autism, But Are Less Flexibly Adjusted to the Context. Autism Res. 14, 1134–1146. https://doi.org/10.1002/aur.2452

Sapey-Triomphe, L.-A., Weilnhammer, V.A., Wagemans, J., 2021c. Associative learning under uncertainty in adults with autism: Intact learning of the cue-outcome contingency, but slower updating of priors. Autism 136236132110450. https://doi.org/10.1177/13623613211045026

Schütz, M., Boxhoorn, S., Mühlherr, A.M., Mössinger, H., Freitag, C.M., Luckhardt, C., 2021. Intention Attribution in Children and Adolescents with Autism Spectrum

Disorder: An EEG Study. J. Autism Dev. Disord. https://doi.org/10.1007/s10803-021-05358-1

Seriès, P., Seitz, A.R., 2013. Learning what to expect (in visual perception). Front. Hum. Neurosci. 7. https://doi.org/10.3389/fnhum.2013.00668

Sevgi, M., Diaconescu, A.O., Henco, L., Tittgemeyer, M., Schilbach, L., 2020. Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. Biol. Psychiatry 87, 185–193. https://doi.org/10.1016/j.biopsych.2019.09.032

Seymour, R.A., Rippon, G., Gooding-Williams, G., Schoffelen, J.M., Kessler, K., 2019. Dysregulated oscillatory connectivity in the visual system in autism spectrum disorder. Brain 142, 3294–3305. https://doi.org/10.1093/brain/awz214

Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: A key to the file-drawer. J. Exp. Psychol. Gen. 143, 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J.P., Nelson, L.D., 2015. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). J. Exp. Psychol. Gen. 144, 1146–1152. https://doi.org/10.1037/xge0000104

Sinha, P., Kjelgaard, M.M., Gandhi, T.K., Tsourides, K., Cardinaux, A.L., Pantazis, D., Diamond, S.P., Held, R.M., 2014. Autism as a disorder of prediction. Proc. Natl. Acad. Sci. 111, 15220–15225. https://doi.org/10.1073/pnas.1416797111

Skewes, J.C., Gebauer, L., 2016. Brief Report: Suboptimal Auditory Localization in Autism Spectrum Disorder: Support for the Bayesian Account of Sensory Symptoms. J. Autism Dev. Disord. 46, 2539–2547. https://doi.org/10.1007/s10803-016-2774-9

Skewes, J.C., Jegindø, E.-M., Gebauer, L., 2015. Perceptual inference and autistic traits. Autism 19, 301–307. https://doi.org/10.1177/1362361313519872

Smith, D., Ropar, D., Allen, H.A., 2017. The Integration of Occlusion and Disparity Information for Judging Depth in Autism Spectrum Disorder. J. Autism Dev. Disord. 47, 3112–3124. https://doi.org/10.1007/s10803-017-3234-x

Solomon, M., Smith, A.C., Frank, M.J., Ly, S., Carter, C.S., 2011. Probabilistic reinforcement learning in adults with autism spectrum disorders. Autism Res. 4, 109–120. https://doi.org/10.1002/aur.177

Sotiropoulos, G., Seitz, A.R., Seriès, P., 2014. Contrast dependency and prior expectations in human speed perception. Vision Res. 97, 16–23. https://doi.org/10.1016/j.visres.2014.01.012

Stevenson, I.H., Cronin, B., Sur, M., Kording, K.P., 2010. Sensory Adaptation and Short Term Plasticity as Bayesian Correction for a Changing Brain. PLoS ONE 5, e12436. https://doi.org/10.1371/journal.pone.0012436

Stevenson, R.A., Siemann, J.K., Woynaroski, T.G., Schneider, B.C., Eberly, H.E., Camarata, S.M., Wallace, M.T., 2014. Evidence for Diminished Multisensory Integration in

Autism Spectrum Disorders. J. Autism Dev. Disord. 44, 3161–3167. https://doi.org/10.1007/s10803-014-2179-6

Stevenson, R.A., Toulmin, J.K., Youm, A., Besney, R.M.A., Schulz, S.E., Barense, M.D., Ferber, S., 2017. Increases in the autistic trait of attention to detail are associated with decreased multisensory temporal adaptation. Sci. Rep. 7, 14354. https://doi.org/10.1038/s41598-017-14632-1

Symonds, R.M., Lee, W.W., Kohn, A., Schwartz, O., Witkowski, S., Sussman, E.S., 2017. Distinguishing Neural Adaptation and Predictive Coding Hypotheses in Auditory Change Detection. Brain Topogr. 30, 136–148. https://doi.org/10.1007/s10548-016-0529-8

Szucs, D., Ioannidis, J.P.A., 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLOS Biol. 15, e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tewolde, F.G., Bishop, D.V.M., Manning, C., 2018. Visual Motion Prediction and Verbal False Memory Performance in Autistic Children: Prediction and false memory in autism. Autism Res. 11, 509–518. https://doi.org/10.1002/aur.1915

Thillay, A., Lemaire, M., Roux, S., Houy-Durand, E., Barthélémy, C., Knight, R.T., Bidet-Caulet, A., Bonnet-Brilhault, F., 2016. Atypical Brain Mechanisms of Prediction According to Uncertainty in Autism. Front. Neurosci. 10. https://doi.org/10.3389/fnins.2016.00317

Todorova, G.K., Pollick, F.E., Muckli, L., 2021. Special treatment of prediction errors in autism spectrum disorder. Neuropsychologia 163, 108070. https://doi.org/10.1016/j.neuropsychologia.2021.108070

Tulver, K., Aru, J., Rutiku, R., Bachmann, T., 2019. Individual differences in the effects of priors on perception: A multi-paradigm approach. Cognition 187, 167–177. https://doi.org/10.1016/j.cognition.2019.03.008

Turi, M., Burr, D.C., Igliozzi, R., Aagten-Murphy, D., Muratori, F., Pellicano, E., 2015. Children with autism spectrum disorder show reduced adaptation to number. Proc. Natl. Acad. Sci. 112, 7868–7872. https://doi.org/10.1073/pnas.1504099112

Turi, M., Karaminis, T., Pellicano, E., Burr, D., 2016. No rapid audiovisual recalibration in adults on the autism spectrum. Sci. Rep. 6, 21756. https://doi.org/10.1038/srep21756

Utzerath, C., Schmits, I.C., Buitelaar, J., de Lange, F.P., 2018. Adolescents with autism show typical fMRI repetition suppression, but atypical surprise response. Cortex 109, 25–34. https://doi.org/10.1016/j.cortex.2018.08.019

Utzerath, C., Schmits, I.C., Kok, P., Buitelaar, J., de Lange, F.P., 2019. No evidence for altered up- and downregulation of brain activity in visual cortex during illusory shape perception in autism. Cortex 117, 247–256. https://doi.org/10.1016/j.cortex.2019.03.011

van Boxtel, J.J.A., Lu, H., 2013. Impaired Global, and Compensatory Local, Biological Motion Processing in People with High Levels of Autistic Traits. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00209

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., Wagemans, J., 2014. Precise minds in uncertain worlds: Predictive coding in autism. Psychol. Rev. 121, 649–675. https://doi.org/10.1037/a0037665

Van de Cruys, S., Lemmens, L., Sapey-Triomphe, L., Chetverikov, A., Noens, I., Wagemans, J., 2021. Structural and contextual priors affect visual search in children with and without autism. Autism Res. 14, 1484–1495. https://doi.org/10.1002/aur.2511

Van de Cruys, S., Vanmarcke, S., Van de Put, I., Wagemans, J., 2018. The Use of Prior Knowledge for Perceptual Inference Is Preserved in ASD. Clin. Psychol. Sci. 6, 382–393. https://doi.org/10.1177/2167702617740955

van Laarhoven, T., Stekelenburg, J.J., Eussen, M.L., Vroomen, J., 2020. Atypical visual-auditory predictive coding in autism spectrum disorder: Electrophysiological evidence from stimulus omissions. Autism 24, 1849–1859. https://doi.org/10.1177/1362361320926061

van Laarhoven, T., Stekelenburg, J.J., Eussen, M.L.J.M., Vroomen, J., 2019. Electrophysiological alterations in motor-auditory predictive coding in autism spectrum disorder. Autism Res. 12, 589–599. https://doi.org/10.1002/aur.2087

Vishne, G., Jacoby, N., Malinovitch, T., Epstein, T., Frenkel, O., Ahissar, M., 2021. Slow update of internal representations impedes synchronization in autism. Nat. Commun. 12, 5439. https://doi.org/10.1038/s41467-021-25740-y

Vogel, D., Falter-Wagner, C.M., Schoofs, T., Krämer, K., Kupke, C., Vogeley, K., 2019. Interrupted Time Experience in Autism Spectrum Disorder: Empirical Evidence from Content Analysis. J. Autism Dev. Disord. 49, 22–33. https://doi.org/10.1007/s10803-018-3771-y

von der Lühe, T., Manera, V., Barisic, I., Becchio, C., Vogeley, K., Schilbach, L., 2016. Interpersonal predictive coding, not action perception, is impaired in autism. Philos. Trans. R. Soc. B Biol. Sci. 371, 20150373. https://doi.org/10.1098/rstb.2015.0373

Walsh, R.J., Krabbendam, L., Dewinter, J., Begeer, S., 2018. Brief Report: Gender Identity Differences in Autistic Adults: Associations with Perceptual and Socio-cognitive Profiles. J. Autism Dev. Disord. 48, 4070–4078. https://doi.org/10.1007/s10803-018-3702-y

Ward, E.K., Braukmann, R., Buitelaar, J.K., Hunnius, S., 2020. No evidence for neural markers of gaze direction adaptation in 2-year-olds with high or low likelihood of autism. J. Abnorm. Psychol. 129, 612–623. https://doi.org/10.1037/abn0000518

Ward, E.K., Buitelaar, J.K., Hunnius, S., 2021. Implicit learning in 3-year-olds with high and low likelihood of autism shows no evidence of precision weighting differences. Dev. Sci. https://doi.org/10.1111/desc.13158

Wei, X.-X., Stocker, A.A., 2015. A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. Nat. Neurosci. 18, 1509–1517. https://doi.org/10.1038/nn.4105

Wilson, R.C., Collins, A.G., 2019. Ten simple rules for the computational modeling of behavioral data. eLife 8, e49547. https://doi.org/10.7554/eLife.49547

Zaidel, A., Goin-Kochel, R.P., Angelaki, D.E., 2015. Self-motion perception in autism is compromised by visual noise but integrated optimally across multiple senses. Proc. Natl. Acad. Sci. 112, 6461–6466. https://doi.org/10.1073/pnas.1506582112

* * *

In the previous chapter, we conducted a comprehensive review of experimental evidence of Bayesian theories of autism in the years 2012-2021. This review revealed a complex and occasionally contradicting list of findings. The overall evidence for reduced prior influences in autism was mixed, with the majority of studies not finding the expected differences between autistic and neurotypical individuals or across autistic traits. Importantly, our review highlighted some methodological issues in the field, including low statistical power, inconsistent approaches to modelling, and a lack of clarity in how different researchers conceptualised Bayesian processes.

One of the findings of the review was that computational studies appear to interpret sensory precision in two, distinct ways. Some focused on the actual reliability of sensory inputs, while others examined the brain's estimation of that reliability, but these interpretations were rarely made explicit. Moreover, the studies that focused on the actual reliability equated it with the estimated reliability in their models, something that was not the case in the models of the other studies. This difference and the lack of clear terminology in general could help explain some of the inconsistent computational findings we observed and makes it difficult to compare results across studies or precisely evaluate competing theories.

The next chapter addresses this gap by presenting a formal mathematical framework for the two types of sensory precision in Bayesian perception. We introduce clear definitions of the actual and estimated sensory precision, examine how they interact with prior beliefs, and explore their implications for theories of autism, both mathematically and through computational simulations. Our analysis reveals that some Bayesian theories of autism may be mathematically indistinguishable from each other, depending on the experimental design. We then apply these insights to re-analyse data from our previous work, demonstrating how specific tasks could help differentiate between the two precisions and advance our understanding of their role in autism.

Chapter 3

On Sensory Precision in Bayesian Models of Autism

This chapter consists of a manuscript to be submitted for publication: Angeletos Chrysaitis, N., Karvelis, P., & Seriès P. (in prep.) The Meaning of 'Sensory Precision' in Bayesian Perception and Its Importance for Theories of Autism

Chapter 3 Abstract

Bayesian theories of perception suggest that the brain continuously integrates prior beliefs about the environment with sensory information to make sense of uncertain or ambiguous inputs and increase average perceptual accuracy. The impact of prior beliefs is stronger when their reliability is high, but also when the reliability of the sensory information, known as 'sensory precision', is low. However, there often is a confusion as to what the term 'sensory precision' refers to: is it the actual reliability of the sensory input or the brain's estimate of that reliability? In this paper, we aim to clarify this confusion by introducing new terms for the two precisions and examining their individual effects in a basic model of perception.

Altered precision of sensory information or prior beliefs is often proposed to be at the origin of psychological disorders, such as autism or schizophrenia. Here, we apply the new terminology to Bayesian theories of autism and provide a brief overview of relevant modelling studies. We reanalyse data from a past study in this light (Karvelis et al., 2018) and show how its design holds promise in advancing our understanding of the Bayesian processes in autism and the general population.

Keywords: Bayesian inference, Predictive coding, Perception, Precision, Autism

3.1 Introduction

Imagine you are walking in the desert at night, and you see the silhouette of a four-legged animal that appears to be a bear. Most likely you would think that you are mistaken; as bears do not live in the desert, what you saw was probably a camel or another desert animal. This thought process makes sense. Given that our eyes are imperfect and sensory information is often ambiguous, relying on prior knowledge leads to better judgements than simply trusting our senses. Fortunately, this process is already implemented in the brain at a fundamental level. All perception is an amalgam of sensory information and the brain's prior beliefs about the environment. This is why you would probably never perceive a bear in that scenario to begin with, even if it was the closest match to what your eyes saw. Indeed, studies have repeatedly shown that the brain uses its knowledge of the environment during perception (e.g., Seriès & Seitz, 2013; Weiss et al., 2002).

On average, it is optimal to be influenced by prior knowledge, but the magnitude of that influence should not be constant across different scenarios. Instead, it should depend on two factors: the reliability of the sensory information and the certainty of prior knowledge. In the previous example, if it was daylight, it would make more sense to believe your eyes and ignore how unlikely it would be a priori to see a bear in the desert. Similarly, if what you saw looked like an antelope, then it is not as unlikely and you should not necessarily dismiss it, despite camels being more common in the desert. This weighting of information is also present in the brain, as shown through behavioural experiments. For example it has been demonstrated that the influence of visual and haptic evidence on the percept is proportional to their reliability (Ernst & Banks, 2002).

These are the core components of the Bayesian brain framework, namely that the brain infers the cause of sensory inputs, by combining them with prior beliefs about the environment, each weighted by its reliability (Knill & Pouget, 2004). In mathematical terms, the *posterior probability* about the cause (C) given the sensory inputs (I) is proportional to the product of the *prior* ($p(C)$) and the *likelihood* ($p(I/C)$) of the sensory inputs, weighted by their *precision*:

$$p(C|I) \propto p(I|C)\, p(C)\,. \tag{1}$$

That perception can be described in terms of such Bayesian computations has been thoroughly confirmed for different modalities over the years (Penny, 2012). Furthermore, researchers have developed methods to estimate the nature of the priors individual participants use in behavioural tasks (e.g., Stocker & Simoncelli, 2006) and have showed that these correspond well to the natural statistics of the environment (e.g., Girshick et al., 2011). However, the relationship between the likelihood function and sensory inputs has received less attention.

In their recent primer, Yon and Frith examined the role of precision in the Bayesian brain, both for prior distributions and for sensory inputs (Yon & Frith, 2021). They described how the brain uses beliefs about precision and how these beliefs can be inaccurate. They concluded, among other things, that the scientific community 'must get more precise about how precision works'. Here, we provide some of the requested precision about (sensory) precision. We introduce the terms 'actual sensory precision' and 'estimated sensory precision', the latter corresponding to what Yon and Frith call 'beliefs about precision', when applied specifically to sensory precision. We mathematically formalise the relevant concepts and examine some common assumptions in the modelling of the brain's likelihoods. We also discuss the implications of our formalisation for the Bayesian theories of autism.

3.2 A Simplified Bayesian Model of Perception

In this section, we will provide a detailed explanation of the mechanism of a basic Bayesian model of perception. This will give us the opportunity to analyse how its components impact perception and to clearly define the relevant terms. For simplicity, we will assume all the components involved, such as prior beliefs, can be well approximated by Gaussian distributions; a common assumption in modelling perceptual inference.

We will start with a stimulus or cause in the environment, represented by C. In this case we are concerned with a single characteristic of that stimulus, so C is a scalar that could represent, for example, the stimulus' position, velocity, or colour. The measurement of that characteristic is a noisy process due to our imperfect senses, the environmental conditions (e.g., poor

illumination), other characteristics of the stimulus (e.g. short duration, small size), and the attention paid to the stimulus. These can result in measurement error, where the sensory input (scalar *I*) does not match the actual cause. We can model this process as the sensory input being drawn from a distribution centred on *C*:

$$I \sim p_A(I|C) = N(C, 1/\pi_s^A) \tag{2}$$

The index *A* (for 'actual') signifies that these symbols refer to the actual process of drawing sensory inputs, which takes place independently of the observer's beliefs. π_s^A is the precision of that distribution, or the *actual sensory precision*, where precision is the inverse of variance, $\pi = 1/\sigma^2$. Low amounts of sensory noise would result in higher actual sensory precision and therefore lower measurement error or sensory inputs that are closer to their cause on average.

If there is no prior knowledge about the cause, the best estimate of the stimulus is the sensory input itself. But if the brain has formed a prior belief about the environment, then the optimal judgement should be influenced by both types of information. Bayesian theories hypothesise that the brain assigns a specific reliability to the sensory input in order to integrate it with its priors. More reliable inputs should have more sway in the perceptual process. Optimally, this reliability would be equal to the actual sensory precision. However, it is widely believed that the brain cannot directly access π_s^A, and must instead measure or estimate it. We will call the result of this measurement *estimated sensory precision* and symbolise it with π_s^M. Here, the index *M* (for 'model') indicates that π_s^M is part of the observer's model or beliefs about the world. Critically, actual and estimated precision might differ, or in Yon and Frith's terminology the brain might entertain false beliefs about sensory precision (Yon & Frith, 2021).

A way to represent the sensory input along with its reliability is the likelihood function:

$$p_M(I|C) = \mathcal{L}_M(C|I) = N(I, 1/\pi_s^M) \tag{3}$$

This shows the estimated probability of the current sensory input for each possible value of the stimulus. It also serves as an estimate for how likely each stimulus value is, before the information of the prior beliefs is considered. The stimulus is more likely to be close to *I*, with the probability diminishing as we get further away, based on the brain's estimate of the actual

sensory precision. The index *M* once again points to the fact that the likelihood is part of the brain's beliefs about the world – in this case it is based on a model of the draw of sensory inputs (see equation 2).

We have defined the sensory input and the likelihood as separate entities, i.e. a scalar measurement and the distribution centred on it, formed at different stages of the inferential process. This distinction is not meant to reflect the structure of sensory processing. For example, it is possible that the reliability of a sensory input is represented as part of the input itself, meaning that the sensory input is directly encoded by the brain as the likelihood distribution. Nonetheless, this choice does not fundamentally affect the structure of the Bayesian models or the presented results.

The brain's prior beliefs about *C* are also represented with a probability distribution. There is a stimulus value that the brain believes to be most likely *before* receiving any sensory information, which functions as the mean of the prior, μ_{pr}. Then, there is the certainty about this prior belief, π_{pr}, which leads us to the definition of the brain's prior:

$$p_M(C) = N\left(\mu_{pr}, 1/\pi_{pr}\right) \tag{4}$$

This is combined with the likelihood, according to Bayes' rule (equation 1). Because both distributions are Gaussian, as per our assumptions, the posterior distribution is Gaussian as well:

$$p_M(C|I) = N\left(\mu_{post}, 1/\pi_{post}\right) \tag{5}$$

$$\text{with } \pi_{post} = \pi_{pr} + \pi_s^M \tag{6}$$

$$\text{and } \mu_{post} = \frac{\pi_{pr} \cdot \mu_{pr} + \pi_s^M \cdot I}{\pi_{post}} . \tag{7}$$

This means that the brain's final belief about the stimulus characteristic is centred at a value that lies between its prior mean and the sensory input, with the exact position depending on their relative precisions. That is, the more reliable the information, the greater influence it has on the posterior mean. The precision of the posterior belief is equal to the sum of the individual

precisions; as the posterior combines both types of information, it is more reliable than either of them.

This highlights why both sensory precisions should be taken into account by a perceptual model. Actual sensory precision determines the relationship between stimuli and sensory inputs, while estimated sensory precision determines the weight of sensory inputs during inference. While estimated sensory precision is the brain's representation of actual sensory precision, the two might differ (Yon & Frith, 2021), and therefore a model that doesn't distinguish the two precisions would not be able to explain the full relationship between stimuli and percepts.

Finally, it is commonly hypothesised that one, singular estimate is chosen from the posterior distribution to form what we will end up perceiving. There are two different schools of thought for how this happens, two variants of Bayesian models: the *mean variant* and the *sampling variant* (e.g., Fiser et al., 2010; Knill & Pouget, 2004; Sanborn & Chater, 2016). The first variant assumes that the brain chooses the mean of the posterior ($\hat{C} = \mu_{post}$), as the mathematically optimal estimate of the cause, while the other suggests that the percept is sampled from the posterior distribution ($\hat{C} \sim p_M(C|I)$).[1] There is also the possibility of choosing the mode of the posterior as the percept, but for our Gaussian model this is equivalent to the mean variant. Figure 3.1 illustrates the complete process.

Translating all of this to our toy example: a camel in the desert (*C*) might appear to our eyes as the silhouette of a bear (*I*), due to low visibility during night-time (low π_s^A). The brain assigns a low weight to the sensory information (low π_s^M), due to its low reliability. Therefore, it is more influenced by its prior belief that camels are more frequent in the desert ($p_M(C)$) and ends up perceiving a camel ($\hat{C} \approx C$).

[1] Some sampling models use sampling in other steps of the inferential process, such as representing the prior distribution. They also might draw more than one sample from the posterior, in contrast to our simplified model.

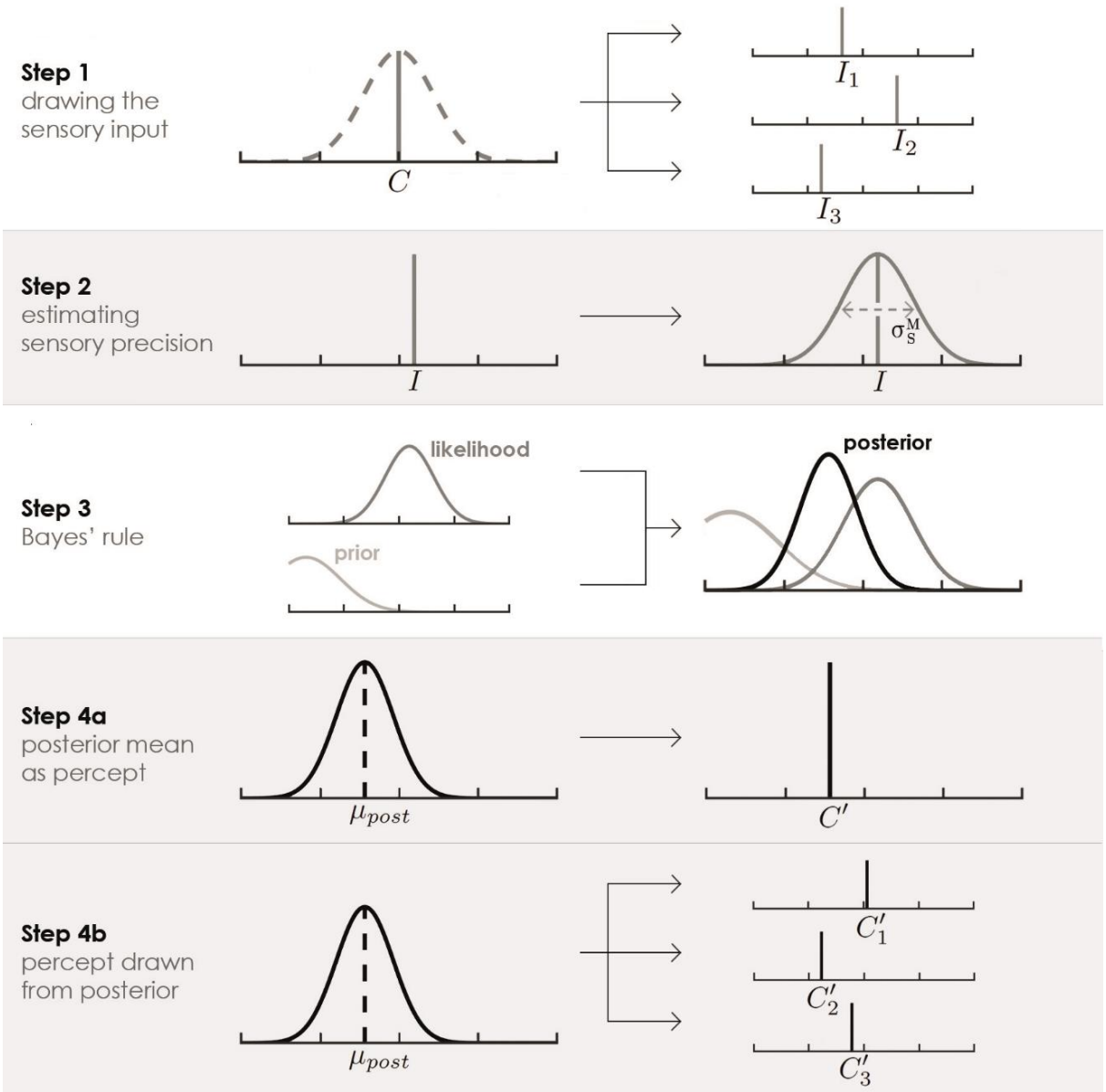

Figure 3.1: Schematic of Bayesian Inference in Perception. *The procedure is represented as 4 distinct steps for illustration purposes. In reality, steps might happen simultaneously and as part of the same process. Step 1: the sensory input is drawn from a distribution centred at the stimulus. Step 2: the actual sensory precision is estimated to create the likelihood function (σ_s^M is the standard deviation corresponding to the estimated sensory precision). Step 3: the likelihood is combined with the prior. Step 4: a percept is chosen, either (a) as the mean of the posterior or (b) by drawing from the posterior.*

In the sampling variant, posterior precision directly influences the variance of the percept. But, independently of the variant, posterior precision is thought to influence the observer's certainty about what they have perceived (Geurts et al., 2022; Pouget et al., 2016). It is assumed that, once again, the brain has to estimate the precision of these representations to produce the feeling of confidence (Yon & Frith, 2021). This is another case where the distinction between actual and estimated precision is useful, although in a domain that is not purely perceptual. Yon and Frith also argue that beliefs that are situated high in the cognitive hierarchy are based on diverse kinds of information and that this complexity leads to a higher discrepancy between estimated and actual precisions than in low-level perception.

In the following sections, we will focus on the implications of the model presented above. However, it is crucial to acknowledge the model grossly oversimplifies the complexity of the brain's functioning. For example, it is unlikely that the brain will receive only one 'piece of information' regarding a specific cause. In reality, multiple sensory inputs, whether they occur simultaneously or in temporal succession, will shape each percept. These influence perception one after the other, with the posterior of each inference becoming the prior for the next, as the brain eventually forms a precise belief about the stimulus. Additionally, the brain operates hierarchically, meaning that information about a cause also affects the observer's higher-level beliefs about the general statistics of the environment, how often these change, and so on. Conversely, the brain's higher-level beliefs about the environment influence the perception of specific features of that environment. In fact, this is where the prior in our example originated, as a higher-level belief about the desert and the animals that may inhabit it, rather than a belief about a specific animal.

This simplified model also does not account for the possibility of a changing environment. For instance, imagine we are interested in the colour of a stimulus we observe for a few seconds. What should we be inferring if the sensory inputs we receive are very different from each other over time? There are two possible explanations for this. One is that there is a lot of sensory noise, causing inaccurate sensory inputs, what is frequently termed 'expected uncertainty' (Yu & Dayan, 2005). In this case, the posterior mean should be their average and the posterior precision should be low, expressing the unreliability of the sensory information. This is what

would be calculated by the simplified model. The other explanation is that the stimulus is changing colour, an instance of 'unexpected uncertainty (Yu & Dayan, 2005). In that case, it would be optimal to be more influenced by the more recent sensory inputs, for example via increases in the estimated sensory precision. To choose between these alternatives and determine the magnitude of the belief updates, a more complex model would estimate the rate of change in the environment. The higher the rate of change, the less informative old estimates of the stimulus are, and therefore the larger the updates to the prior should be (Yon & Frith, 2021). It is commonly hypothesised that the neurotransmitter noradrenaline modulates belief updating under uncertainty (Yu & Dayan, 2005), possibly by encoding the rate of change in the environment. Indeed, a recent study found that suppressing noradrenaline resulted in larger prior influences and slower belief updating when the environment was volatile (Lawson et al., 2021; Yon, 2021).

3.3 The Parameters of Bayesian Perception

So far, we have mentioned four 'parameters' that could affect Bayesian perception of a given stimulus. In this section we examine the effects of each of them and how their values are determined. We will omit μ_{pr}, as its effects are easy to understand and have been described above. Instead, we will focus on the remaining three parameters: π_{pr}, π_s^M, π_s^A.

Prior precision corresponds to the brain's certainty about its prior beliefs. It is therefore understandable that its magnitude is analogous to the influence of the prior on the final percept. It also influences the brain's certainty about the percept. That is, higher π_{pr} leads to larger biases towards the prior mean, μ_{pr}, and higher π_{post}. Prior precision should ideally be a veridical expression of the regularities in the environment and the brain's knowledge of them (e.g., Girshick et al., 2011), with more predictable environments leading to higher precision. If the brain estimates that its priors are no longer accurate, it might be optimal to lower their precision to avoid costly biases and allow for rapid belief updating (Mathys et al., 2014).

The estimated sensory precision (the precision of the likelihood) determines the weight given to the sensory input in the inferential process. Accordingly, higher π_s^M reduces the influence of the prior, while also increasing posterior precision. As previously mentioned, the estimated sensory precision is the observer's belief about the informativeness of the sensory input and therefore it would optimally be equal to π_s^A. However, estimated and actual sensory precisions might differ, for the same reason the sensory input itself is not a perfect estimate of the stimulus. In the mean variant, unequal π_s^M and π_s^A would lead to suboptimal estimates, as the percept is insufficiently or overly influenced by the prior. In the sampling variant, the effects of $\pi_s^M \neq \pi_s^A$ depend on various factors, such as the magnitude of their difference and the number of samples used to form a percept. In both variants, higher estimated than actual sensory precision means that the observer is disproportionately confident in the information of the sensory inputs. This would result in overly precise posteriors, that is posteriors that are more concentrated around the mean than is optimal. In theory, the brain could introduce a discrepancy between actual and estimated sensory precisions 'on purpose'. For example, it could increase π_s^M under conditions of high environmental volatility to facilitate a change in inaccurate priors. However, this would create overconfident posteriors. Since posterior beliefs are used as priors in subsequent perceptual inferences, this would have the side-effect of creating overconfident priors.

Finally, the actual sensory precision directly corresponds to the reliability of the sensory input. Higher π_s^A values reduce the average distance between sensory input and stimulus, making the sensory input more informative. Therefore they lead to more accurate perception, regardless of the other parameters. This parameter is largely dependent on the properties of the physical stimulus itself and is not entirely under the control of the brain. However, it is possible that the brain could temporarily increase the actual sensory precision of a stimulus, by dedicating more perceptual resources to it, for example via the deployment of attention.

It should be noted that some theories link attention with the modulation of *estimated* (sensory) precision instead of the actual sensory precision (e.g. Friston, 2009; Hohwy, 2012). Although of clear theoretical interest, a mechanistic theory of attention is outside of the scope of this paper. Nevertheless, it is important to note here that a computational account of attention

should be consistent with the fact that attention can lead to more accurate perceptual judgements for the attended stimuli. We have shown that increasing estimated sensory precision is mathematically equivalent to decreasing prior precision in the mean variant, while in the sampling variant it would also result in a narrower posterior. These effects do not reliably improve perceptual accuracy, since accurate priors are beneficial to perception. Therefore, accounts where attention simply increases estimated precision appear incomplete, especially for spatial attention where estimated precision would be uniformly increased for a section of space. In our view, the precise computational underpinnings of attention remain an open question. Neither enhanced actual sensory precision nor enhanced estimated sensory precision alone can offer a sufficient explanation of all relevant behavioural findings, including both the increased accuracy and the attention-induced perceptual distortions (Carrasco & Barbot, 2019).

3.4 Estimating the Parameters from Behavioural Data

Bayesian models like the one we have discussed can be used to investigate various psychological phenomena, such as participant biases, the learning of environmental statistics, and the effects of attention. This is usually done by collecting participant responses in behavioural experiments and working backwards, using these data to infer the details of the brain's underlying process. One common approach is to fit computational models to behavioural data and extract parameter values that characterise the shape of priors and likelihoods (e.g., Chalk et al., 2010; Schmack et al., 2016; Stocker & Simoncelli, 2006). However, even for our simplified model, extracting these parameter values can be challenging and, in some cases, impossible.

Let's consider a task such as trying to infer the direction of motion of a cloud of dots (C), where some motion directions are presented in more trials than others (a unimodal version of the task in Chalk et al., 2010; Karvelis et al., 2018; and Valton et al., 2019). In such a task, it has been shown that the participants will implicitly develop an expectation for the most frequent directions. We describe this prior with a distribution with mean μ_{pr} and precision π_{pr}. When the dots have a low contrast, this prior will induce an attractive bias in the participants'

estimation of motion direction (Bias$[\hat{C} \mid C]$, see Figure 3.2) and a modulation of the variance of the estimation responses (Var$[\hat{C} \mid C]$). Our assumption of components being well-approximated by Gaussian distributions allows us to easily predict the bias and show that it is the same for both the mean and the sampling variants (see Appendix 3.A, Section 1):

$$\text{Bias}[\hat{C} \mid C] = \frac{\pi_{pr}(\mu_{pr} - C)}{\pi_{post}} . \tag{8}$$

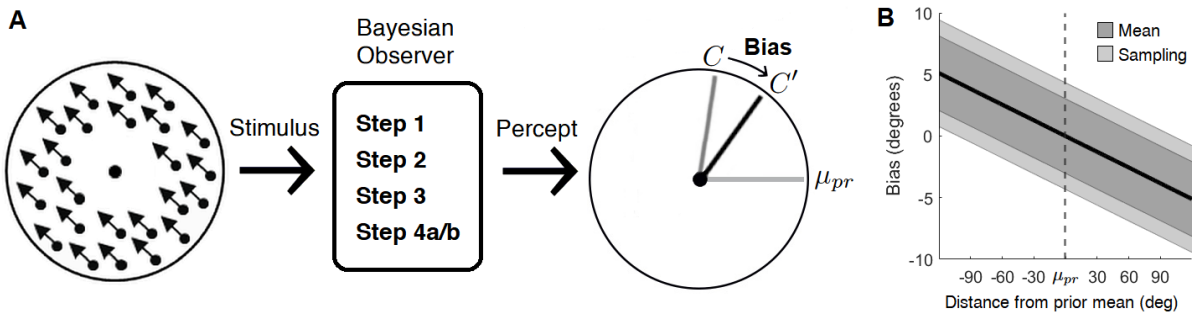

Figure 3.2: A generative model for the Moving Dots task (A) & illustration of the resulting bias (B). *A) A stimulus (C) is presented to the participant. Then, the sensory input is combined with the prior about the most frequent directions to form the posterior from which the percept (C') is chosen (see Figure 3.1 for the four steps). The participant reports a direction based on the percept, which on average shows an attractive bias towards the prior mean (μ_{pr}). B) Average biases are identical between the variants. The bias at the prior mean (μ_{pr}) is equal to 0. Biases for stimuli before the prior mean are positive, biases after are negative. Shading corresponds to the standard deviation of responses. Parameters used:* $\pi_{pr} = 0.0044$, $\pi_s^A = 0.1$, $\pi_s^M = 0.1$.

However, the variances of the motion direction responses differ depending on which generative model is used. In the mean variant, the response variance is caused by the sensory noise ($\sigma_A^2 = 1/\pi_s^A$). If two sensory inputs are the same, this would result in the same likelihood distribution, the same posterior, and the same percept as that is equal to the posterior mean. But due to sensory noise, inputs will slightly differ between trials, leading to different percepts. Priors reduce this variance, by biasing all percepts towards the prior mean. Mathematically, this is formalised as

$$\text{Var}_{\text{mean}}[\hat{C} \mid C] = \frac{(\pi_s^M)^2}{\pi_s^A \cdot \pi_{post}^2} . \tag{9}$$

In the sampling variant, response variance is caused both by the sensory noise and by the posterior variance, making the response variance always larger than that of the mean variant. The sensory noise causes the shift in the likelihood across trials and the posterior variance determines the spread of the sampled percepts. Priors reduce response variance as they both bias percepts towards one direction and increase posterior precision. In mathematical terms:

$$\mathrm{Var}_{\mathrm{sampling}}\left[\hat{C} \mid C\right] = \frac{(\pi_s^M)^2 + \pi_s^A \cdot \pi_{post}}{\pi_s^A \cdot \pi_{post}^2} \,. \tag{10}$$

These equations show how the assumptions we make about which generative model the brain uses influences our predictions of participants' responses in behavioural experiments. An important question then is: is it possible to infer all the parameters of such a model and which variant does the brain use?

From the participants' response bias, we can estimate both the mean of their prior (μ_{pr}) and the ratio between prior and likelihood precision (π_{pr}/π_s^M). Assuming the mean variant is correct also allows us to calculate the actual sensory precision using the response variance (π_s^A). However, the mean variant renders it impossible to independently estimate the individual values of prior (π_{pr}) and estimated sensory precision (π_s^M); something that holds true in all designs with Gaussian priors and potentially extends to other distributions as well. If the sampling variant is correct, no parameter besides the prior mean (μ_{pr}) can be inferred in such a task (Figure 3.3 & Appendix 3.B). Attempts to infer these values from behavioural data usually require additional assumptions about Bayesian processes. For example, assuming that the brain's estimation of sensory precision is accurate ($\pi_s^M = \pi_s^A$) allows us to calculate a value for prior precision.

Additionally, the mean and sampling variants are indistinguishable in this simple experimental design, which can be shown both mathematically (Appendix 3.A, Section 2) and through model recovery (Appendix 3.B). For model recovery, we simulated data using different combinations of parameter values for both variants and then used them to fit the same models to see if we could distinguish the variant from response data. Our results showed that both fitted variants achieve the exact same goodness of fit (measured by the likelihood of the data) for the vast

majority of simulated participants, independently of which variant generated the data. This means that the models cannot be differentiated via the most common approaches, such as likelihood ratios, BIC, or AIC scores. These results show the limitations of a particular type of experimental design, but they also illustrate the necessity of performing parameter and model recovery before any conclusions are drawn from the data.

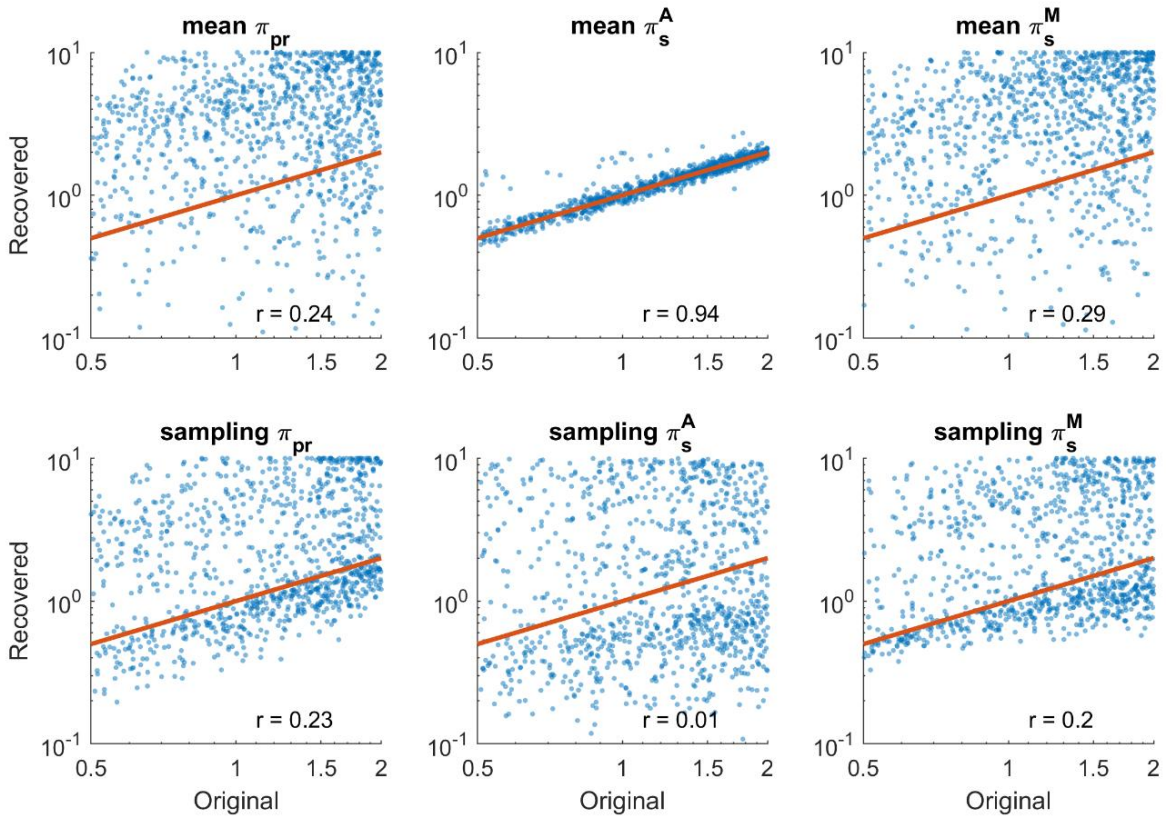

Figure 3.3: Parameter recovery in the mean and sampling variants. *We simulate data and then use the same models to estimate the original parameters (see Appendix 3.B for details). Blue dots correspond to simulated participants. The red lines correspond to perfect recovery and r to the Pearson correlation between original and recovered parameters. If the parameters could be estimated, the simulated and recovered parameters would be close to each other and the correlations between them high. The results show that estimation from behavioural data only works well for the actual sensory precision in the mean variant.*

3.5 The Bayesian Theories of Autism

One prominent application of the Bayesian framework has been in the study of autism, with most theories hypothesising reduced prior influences in autistic individuals (Brock, 2012; Lawson et al., 2014; Pellicano & Burr, 2012; Van de Cruys et al., 2014). However, the exact computational mechanisms behind these theories, have been subject to debate. Moreover, the language used in these theoretical articles has been often ambiguous, leaving them open to more than one interpretation. In this section, we attempt to clarify these theories, by rephrasing them as faithfully as possible using precise terminology. We will provide relevant quotes from the original articles to support our interpretation. In the rephrased theories, accented parameters represent hypothesised values in autism, while non-accented ones represent neurotypical values.

Pellicano and Burr (2012) initially suggested that priors are less precise in autism, that is $\pi'_{pr} < \pi_{pr}$.[2] Brock (2012) put forth a different explanation, where the actual sensory precision is increased instead.[3] This was based on Mottron et al.'s 'Enhanced Perceptual Functioning' hypothesis, which assumes better performance in low-level perception in autistic individuals. Brock's theory would be formalised as $\pi_s^{A'} > \pi_s^A$, with no explicit mention of estimated sensory precision, as his account does not differentiate between precisions. Given that the estimated sensory precision should be based on the actual sensory precision, if the relationship between π_s^A and π_s^M is not affected in autism this would result in $\pi_s^{M'} > \pi_s^M$.

Two additional theories with the same core structure have been proposed within the framework of predictive coding. According to Lawson et al. (2014), the estimated sensory precision is higher in autism relative to prior precision, $\pi_s^{A'} / \pi'_{pr} > \pi_s^A / \pi_{pr}$.[4] This could be either due to increased likelihood precision or decreased prior precision, but importantly no assumptions are made about the actual sensory precision, seemingly implying that this parameter remains

[2] E.g. "We suggest specifically that attenuated Bayesian priors – 'hypo-priors' – may be responsible for the unique perceptual experience of autistic people […]"
[3] E.g., "However, bottom-up accounts of enhanced autistic perception can also be formalized in Bayesian terms […]"
[4] E.g. "Here, we suggest that abnormalities in autistic perception, action and social interaction can be explained by an imbalance of the precision ascribed to sensory evidence relative to prior beliefs."

unaffected in autism. Van de Cruys et al. (2014) also hypothesised that the estimated sensory precision is higher in autism, but they also proposed that it is inflexible, i.e. that the autistic brain does not take into account the environmental volatility when setting the value of π_S^M.[5] This means that, under this hypothesis, autistic individuals do not update their beliefs less when they perceive the environment to be slow-changing. It is unclear if the estimated sensory precision is only inflexible in regard to environmental volatility or if it is inflexible in regard to the actual sensory precision as well.

The method most commonly used to experimentally investigate Bayesian theories in autism has been to assess the magnitude of bias in participant estimates caused by prior beliefs (Angeletos Chrysaitis & Seriès, 2022). However, this approach alone cannot distinguish between the aforementioned theories, as reduced biases could be caused either by flatter priors, as in the theory of Pellicano & Burr, or by higher actual or estimated sensory precision, as in the theories of Brock and Van de Cruys et al.

A more comprehensive approach involves response variance measurements and fitting computational models to compare the theories (e.g., Karvelis et al., 2018). Even in that case though, unless other assumptions are made, only the theory of Pellicano & Burr could be differentiated from the others in a simple task design with Gaussian priors. This is because it is the only theory that makes a specific prediction for response variance in ASD in both variants, namely that it is higher compared to neurotypical response variance. The accounts of Brock, Lawson et al., and Van de Cruys et al. could result in lower, higher, or the same response variance in ASD, depending on the variant and the specific values of the parameters (Appendix 3.A, Section 3).

In Table 3.1, we present the few studies of autism that used Bayesian computational models to test or compare the theories. These studies were selected from our systematic review (Angeletos Chrysaitis & Seriès, 2022), excluding those that estimated no parameters or did not focus in the integration of priors and sensory inputs during perception. We observe a variety

[5] E.g. "Low-level sensory prediction errors are generally set at a level of precision that is too high and independent of context."

of designs and modelling choices, yielding an analogous variety of results. Unfortunately, in most cases, the modelling assumptions were not explicitly stated in the studies and thus possible alternatives were not investigated. For example, most studies equated actual and estimated sensory precision, which was apparent only if one delved into the equations presented.

Table 3.1 Characteristics of Bayesian computational ASD studies.

Study	Variant	$\pi_s^M = \pi_s^A$	Sample	Findings	Notes
Karaminis et al. (2016)	Sampling	Y	ASD	lower π_{pr} lower π_s^A	π_s^A estimated from another task. Higher π_{pr}/π_s^A.
Powel et al. (2016)	Mean	Y	AQ	lower π_{pr} same π_s^A	Two levels of sensory noise; only plots for π_s^A.
Lawson et al. (2017)	n/a	Y	ASD	higher ω_3	Binary prior. ω_3 determines how much the volatility estimates change.
Pantelis and Kennedy (2017)	Sampling	No π_s^A	ASD	same π_{pr}	Non-Gaussian priors and likelihoods.
Karvelis et al. (2018)	Mean	Y	AQ	same π_{pr} higher π_s^A	Bimodal prior. Two, similar levels of sensory noise; one π_s^A value.
Noel et al. (2018)	Mean	N	ASD	higher π_{pr} same π_s^A	Binary prior. π_s^M was not fitted.
Noel et al. (2020)	Mean	Y	ASD	same π_{pr} lower π_s^A	Exponential prior.
Noel et al. (2021)	Mean	Y	ASD	same π_{pr}	Bimodal prior. ASD changed their π_{pr} less when experiment statistics changed.

The third column shows if models differentiated between estimated and actual sensory precision or if they used the same values. The Sample column shows if studies compared model parameters across diagnostic groups (ASD) or autistic traits as measured by the Autism Spectrum Quotient (AQ) questionnaire of Baron-Cohen et al. (2001). The Notes column highlights important details of the studies or how they differ from our simplified Gaussian model.

While all these studies progress our understanding of Bayesian inference in autism, the lack of comparisons between possible alternative explanations makes the interpretation of some results difficult. For example, findings of decreased prior precision could also be caused by increased estimated sensory precision. On the other hand, findings of increased actual sensory precision are less likely to be caused by other parameters (see the recoverability estimates in Figure 3.3).

Note that our investigations using a simple Bayesian model cannot be directly mapped to any of the aforementioned studies, as all included additional complications (e.g., binary or multiple tasks, learning models, etc.). One important way in which the present work might not be generalisable to other studies is our assumption of Gaussian distributions. We made this choice as it allowed us to significantly simplify the calculations and such distributions are commonly used in the field (e.g., Mathys et al., 2014; Noel et al., 2018). However, it might not hold true for the majority of the tasks. What we are trying to convey here, based on our prototypical model, is that the interpretation of such experiments is complicated by a significant number of problems. It is possible that some other types of priors would produce different response patterns which might circumvent such problems entirely. However, as we will show in the next sections, these issues extend to at least one more type of prior. Therefore, studies using this kind of approach should attempt to address these considerations and present evidence that their design or computational approach is able to overcome these problems, instead of a priori assuming that such issues would not be present.

Specifically, future Bayesian studies should compare models implementing the mean and sampling variants and show that their designs are able to distinguish between the two. If unable to do so, they should investigate interpretations of their findings under different sets of assumptions and be precise about the predictions made by their hypotheses. Importantly, any assumptions used should be clearly described as part of the study's methods. Furthermore, the broad use of simple, baseline models, such as the one we have formalised here, would provide a precise common language to describe the results of the studies and clarify the possible similarities and differences between them. More complicated models then could add nuance to the results and hint at possible underlying mechanisms, while their comparison with the

baseline would determine the extent of their explanatory power and function as an argument for their use.

3.6 A re-examination of the data of Karvelis et al. 2018

In the previous sections, we have shown two common assumptions that are made by Bayesian studies: equating actual and sensory precision and holding one of the variants as the correct one without comparing it with the alternative. These assumptions were made in our past work as well (Chalk et al., 2010; Karvelis et al., 2018; Valton et al., 2019), where all models followed the mean variant.

In this section, we revisit the data of Karvelis et al. (2018). In that study, we had explored how 91 healthy participants, scored for autistic and schizotypal traits, implicitly learned and combined priors with sensory information. We had found that autistic traits were associated with more veridical perception of the motion stimuli and weaker influence of expectations, as measured by weaker estimation biases, smaller estimation variance, and fewer detection false alarms. We also reported a significant correlation between AQ (Baron-Cohen et al., 2001) and sensory precision as defined by our Bayesian model. Here we re-analyse this data without the assumptions described above and interpret the findings using our new understanding of the theories of autism (Section 3.5).

3.6.1 Methodological analysis based on simulated data

Karvelis et al (2018) used the Moving Dots statistical learning task (Figure 3.2), but with two frequent directions of motion instead of a single direction as in our basic model (Section 3.4), which led participants to form a bimodal prior. Here, we focus on variants of the winning, Bayesian model of that study (see the modelling section of the Karvelis et al., 2018 methods). Similar to our methodology in Appendix 3.B, we simulate 500 participants with 175 responses each to explore parameter and model recovery for our experimental design. We use Pearson correlations between the simulated and recovered parameters to estimate their identifiability and we compare models with the Bayesian Information Criterion (BIC).

The winning model of the original study used a likelihood function of the form $p_M(I|C) = V(I, \pi_s)$, with $p_A(I|C) = V(C, \pi_s)$ and V being the von Mises distribution. This did not differentiate between π_s^A and π_s^M, instead representing both with π_s, the 'sensory precision'. It also assumed that percepts consist of the mean of the posterior distribution. Here, we investigate four different versions of the winning model: with both actual and estimated sensory precisions or with only one sensory precision, and using either the mean or the sampling variant to choose a percept. In keeping with the original study, for the precision correlations we are using the corresponding standard deviation instead ($\sigma = \sqrt{1/\pi}$).

Parameter recovery with one sensory precision is high for both prior and sensory precision in the mean (r = 0.87 & r = 0.81) and the sampling variant (r = 0.88 & r = 0.86). However, when the two sensory precisions are treated as distinct parameters, only the recovery for actual sensory precision in the mean variant remains high (mean variant: π_{pr} r = 0.41; π_s^A r = 0.83; π_s^M r = 0.3; sampling variant: π_{pr} r = 0.51; π_s^A r = 0.44; π_s^M r = 0.43), replicating the findings of our simplified model (Section 3.4). Note that, despite recovery being relatively low for the other parameters in this model (average r = 0.4), it is significantly higher than our results with the simpler experimental design and model illustrated in Figure 3.3 (average r = 0.19). This could be attributed to two possible causes. Firstly, it could be that using a bimodal prior increases the identifiability of the parameters. Secondly, in the Moving Dots task the stimuli are presented in very low contrast, leading participants to occasionally sample percepts directly from the prior. This was accounted for in our models and could have provided additional information about π_{pr}, which in turn would increase the identifiability of all parameters.

Interestingly, model recovery is much better than expected based on our basic model (Section 3.4). 66% of participants simulated using the mean variant under the $\pi_s^M = \pi_s^A$ hypothesis have better BIC scores when fitted with the mean variant compared to the sampling variant. Similarly, 74% of participants simulated by the sampling variant were better fitted by the sampling variant compared to the mean variant. This was also true when the sensory precisions were treated as distinct parameters by the models (mean variant: 68%, sampling variant: 61%). This, again, could be attributed to either the bimodal nature of the task or to the fact that some

percepts are drawn directly from the prior. In both cases it raises the possibility that the Moving Dots task in its current form could differentiate between the variants. However, for a sizeable portion of the simulated participants (57% for the mean variant and 62% for the sampling variant), the differences in BIC scores between the variants were very small, smaller than 1% of the BIC scores' standard deviation. This means that it is unclear whether differentiating between the two variants will robustly generalise when real participants are concerned, who do not behave in perfect accordance with any mathematical model. Moreover, our process was not able to discriminate between models that equate between actual and estimated sensory precision and those that do not.

3.6.2 Fitting the models to the Karvelis et al. (2018) data.

Despite these limitations, we decide to compare the mean and the sampling variant in the experimental data of Karvelis et al. (2018) using Bayesian model selection (Stephan et al., 2009). Doing so, we find that the mean variant fits our data significantly better (posterior model probability 97.6%). This result contradicts a few studies that have shown behavioural evidence for the sampling variant (e.g., Moreno-Bote et al., 2011; Sundareswara & Schrater, 2008). However, to our knowledge, this is the first study that directly compares computational models of the two variants using behavioural data.

In the original study, we had found reduced response variance in individuals with higher AQ, offering evidence against the hypothesis of Pellicano & Burr. We had also provided modelling evidence for higher sensory precision with stronger autistic traits, when the two precisions were forced to be equal. Here, we re-examine the correlations between parameters and AQ in a model without the $\pi_s^M = \pi_s^A$ assumption. We find evidence for a positive relationship between actual sensory precision and higher autistic traits (Figure 3.4), although the *p*-value was higher than the threshold of statistical significance (AQ vs σ_{sens}: $r = -0.21$, $p = 0.055$; see the original study for details). As actual sensory precision has high recovery in the mean variant, this relationship is meaningful. This offers weak evidence for the theory of Brock (2012). We find no relationship between AQ and estimated sensory precision or prior precision ($p > 0.53$).

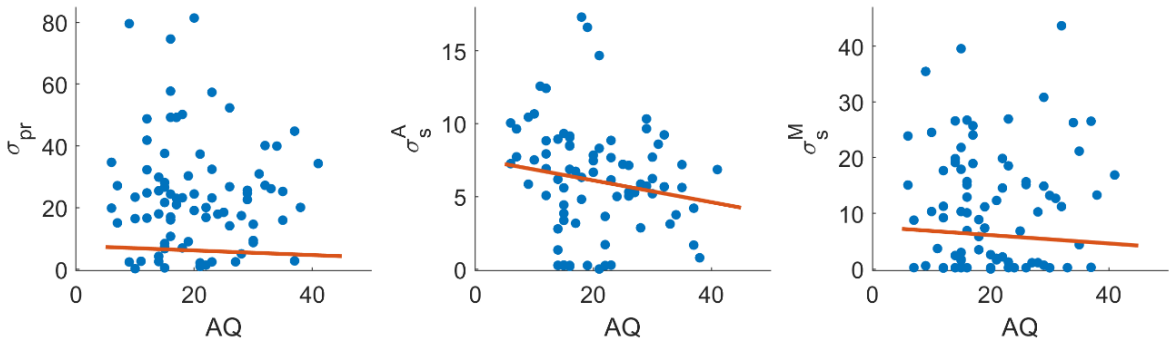

Figure 3.4: Model parameters and AQ scores. *The parameters are estimated using the mean variant of the winning Karvelis et al. (2018) model with distinct actual and estimated sensory precisions. The red line corresponds to the regression fit of the robustfit MATLAB function, which assigns less weight to outliers.*

As a whole, these results highlight the limitations present in our design if one would like to robustly disentangle between the mean and sampling variants or precisely estimate the prior and estimated sensory precision. With the caveats that we have mentioned, they nonetheless extend our previous findings, offering evidence for the mean variant and suggesting that stronger autistic traits are associated with higher *actual* sensory precision, i.e. higher reliability of the sensory inputs, rather than its estimation by the brain. This supports the theory of Brock (2012) and offers evidence against the prominent theories of Pellicano and Burr (2012), Lawson et al. (2014), and Van de Cruys et al. (2014). To our knowledge, this is the first study to disentangle between the different Bayesian theories of autism, advancing our understanding of the condition's underlying computational mechanisms.

3.7 Conclusions & Recommendations

The Bayesian brain framework has been suffering from a lack of clarity in its terminology, particular with regards to sensory precision. This issue is not unique to Bayesian models, as a similar phenomenon has been observed in the literature of sensory sensitivity differences in autism (He et al., 2023; Ward, 2019; Williams et al., 2021). In line with these efforts, we propose a new set of terms for sensory precision, aiming to facilitate better communication and understanding among researchers. To do that, we formalised a basic model of perception which

allowed us to clarify concepts that are present in every Bayesian model. We defined *actual sensory precision* as the reliability of sensory information, arising from the interaction of the stimulus characteristics and our senses. On the other hand, the *estimated sensory precision* is our brain's belief or estimate of the actual sensory precision and functions as the precision of the likelihood distribution, that is the weight given to the sensory information during the inferential process. We continued by illustrating the possible effects of these parameters. For instance, increasing actual sensory precision always improves perceptual accuracy, while larger estimated sensory precision than actual sensory precision decreases the influence of prior beliefs and leads to perceptual estimates that are suboptimal on average. Importantly, these effects depend on the way that the brain selects the percept from the posterior distribution. In the mean variant, the percept is the mean of the posterior, while in the sampling variant it is drawn from it.

By using Bayesian theories of autism as an example, we showed how imprecise terminology and implicit modelling assumptions can lead to multiple interpretations of both hypotheses and results. We presented evidence that the effects of different theories of autism cannot be easily distinguished from each other, as the two variants do not make distinct predictions in all tasks and their parameters are often not identifiable. We suggested potential directions for future research, which we followed while revisiting one of our studies. We showed that the task used in that study, called the Moving Dots task, could potentially distinguish between the mean and sampling variants and is able to provide some evidence for the autism theory of Brock. However, designing tasks that can accurately identify all Bayesian parameters remains an open problem.

Chapter 3 References

Angeletos Chrysaitis, N., & Seriès, P. (2022). 10 years of Bayesian theories of autism:a comprehensive review. *Neuroscience & Biobehavioral Reviews*, 105022. https://doi.org/10.1016/j.neubiorev.2022.105022

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: Comment on Pellicano and Burr. *Trends in Cognitive Sciences*, *16*(12), 573–574. https://doi.org/10.1016/j.tics.2012.10.005

Carrasco, M., & Barbot, A. (2019). Spatial attention alters visual appearance. *Current Opinion in Psychology*, *29*, 56–64. https://doi.org/10.1016/j.copsyc.2018.10.010

Chalk, M., Seitz, A. R., & Series, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2–2. https://doi.org/10.1167/10.8.2

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130. https://doi.org/10.1016/j.tics.2010.01.003

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Geurts, L. S., Cooke, J. R. H., van Bergen, R. S., & Jehee, J. F. M. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, *6*(2), 294–305. https://doi.org/10.1038/s41562-021-01247-w

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

He, J. L., Williams, Z. J., Harris, A., Powell, H., Schaaf, R., Tavassoli, T., & Puts, N. A. J. (2023). A working taxonomy for describing the sensory differences of autism. *Molecular Autism*, *14*(1), 15. https://doi.org/10.1186/s13229-022-00534-1

Hohwy, J. (2012). Attention and Conscious Perception in the Hypothesis Testing Brain. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00096

Karaminis, T., Cicchini, G. M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D., & Pellicano, E. (2016). Central tendency effects in time interval reproduction in autism. *Scientific Reports*, *6*(1), 28570. https://doi.org/10.1038/srep28570

Karvelis, P., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *eLife*, *7*, e34115. https://doi.org/10.7554/eLife.34115

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Lawson, R. P., Bisby, J., Nord, C. L., Burgess, N., & Rees, G. (2021). The Computational, Pharmacological, and Physiological Determinants of Sensory Learning under

Uncertainty. *Current Biology, 31*(1), 163-172.e4. https://doi.org/10.1016/j.cub.2020.10.043

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00302

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00825

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496. https://doi.org/10.1073/pnas.1101430108

Noel, J.-P., Lakshminarasimhan, K. J., Park, H., & Angelaki, D. E. (2020). Increased variability but intact integration during visual navigation in Autism Spectrum Disorder. *Proceedings of the National Academy of Sciences*, *117*(20), 11158–11166. https://doi.org/10.1073/pnas.2000216117

Noel, J.-P., Stevenson, R. A., & Wallace, M. T. (2018). Atypical audiovisual temporal function in autism and schizophrenia: Similar phenotype, different cause. *European Journal of Neuroscience*, *47*(10), 1230–1241. https://doi.org/10.1111/ejn.13911

Pantelis, P. C., & Kennedy, D. P. (2017). Deconstructing atypical eye gaze perception in autism spectrum disorder. *Scientific Reports*, *7*(1), 14990. https://doi.org/10.1038/s41598-017-14919-3

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Penny, W. (2012). Bayesian Models of Brain and Behaviour. *ISRN Biomathematics*, *2012*, 1–19. https://doi.org/10.5402/2012/785791

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Powell, G., Meredith, Z., McMillin, R., & Freeman, T. C. A. (2016). Bayesian Models of Individual Differences: Combining Autistic Traits and Sensory Thresholds to Predict Motion Perception. *Psychological Science*, *27*(12), 1562–1572. https://doi.org/10.1177/0956797616665351

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Schmack, K., Weilnhammer, V., Heinzle, J., Stephan, K. E., & Sterzer, P. (2016). Learning What to See in a Changing World. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00263

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00668

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). *Bayesian model selection for group studies*. 14.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. https://doi.org/10.1038/nn1669

Sundareswara, R., & Schrater, P. R. (2008). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, *8*(5), 12. https://doi.org/10.1167/8.5.12

Valton, V., Karvelis, P., Richards, K. L., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2019). Acquisition of visual priors and induced hallucinations in chronic schizophrenia. *Brain*, *142*(8), 2523–2537. https://doi.org/10.1093/brain/awz171

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675. https://doi.org/10.1037/a0037665

Ward, J. (2019). Individual differences in sensory sensitivity: A synthesizing framework and evidence from normal variation and developmental conditions. *Cognitive Neuroscience*, *10*(3), 139–157. https://doi.org/10.1080/17588928.2018.1557131

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn0602-858

Williams, Z. J., He, J. L., Cascio, C. J., & Woynaroski, T. G. (2021). A review of decreased sound tolerance in autism: Definitions, phenomenology, and potential mechanisms. *Neuroscience & Biobehavioral Reviews*, *121*, 1–17. https://doi.org/10.1016/j.neubiorev.2020.11.030

Yon, D. (2021). Prediction and Learning: Understanding Uncertainty. *Current Biology*, *31*(1), R23–R25. https://doi.org/10.1016/j.cub.2020.10.052

Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*(17), R1026–R1032. https://doi.org/10.1016/j.cub.2021.07.044

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Chapter 3 Supplementary Information

3.A Mathematical Appendix

3.A.1 The bias and variance of the mean and sampling variants

In this section, we will calculate the average bias and the variance of the percepts of a simple Bayesian model of perception. All distributions involved are assumed to be Gaussian, and are parameterised using precision (π) instead of variance, where $\pi = 1/\sigma^2$. The notation has been kept similar to the main text, with the exception of π_s^A and π_s^M which have been replaced with π_s (for 'sensory') and π_l (for 'likelihood'), respectively, for symbolic clarity. The symbol A for 'actual' denotes distributions or variables that are a product of the interaction of the observer's senses and the characteristics of the environment or the stimulus. M for 'model' is used for distributions or variables that are part of the observer's model of reality.

The stimulus or the cause of the sensory inputs is denoted as C. Then, due to sensory noise, the sensory input I can be modelled as drawn from a Gaussian distribution centered at C, with precision π_s: $I \sim p_A(I \mid C) = \mathcal{N}(C, \pi_s^{-1})$. The brain assigns a reliability π_l to that stimulus, where commonly $\pi_l \simeq \pi_s$, So the likelihood function is $\mathcal{L}(C \mid I) = p_M(I \mid C) = \mathcal{N}(I, \pi_l^{-1})$.

Then, if the prior is $p_M(C) = \mathcal{N}(\mu_{pr}, \pi_{pr}^{-1})$, the posterior according to Bayes' rule will be $p_M(C \mid I) = \mathcal{N}(\mu_{post}, \pi_{post}^{-1})$, with the posterior mean being $\mu_{post} = (\pi_{pr}\mu_{pr} + \pi_l I)/(\pi_{post})$ and the precision being $\pi_{post} = \pi_{pr} + \pi_l$.

A common assumption is that the brain chooses a single percept ($\hat{C}$) or estimate from the posterior. There are two proposed ways to do that: The *mean* variant ($\hat{C} = \mu_{post}$) and the *sampling* variant ($\hat{C} \sim p_M(C \mid I)$).

In the mean variant we have

$$\hat{C} = \mu_{post} = \frac{\pi_{pr}\mu_{pr} + \pi_l I}{\pi_{post}} = \frac{\pi_{pr}\mu_p r}{\pi_{post}} + \frac{\pi_l}{\pi_{post}} I \tag{1}$$

which is simply a linear transformation of the sensory inputs. Therefore:

$$\hat{C} \sim \mathcal{N}\left(\frac{\pi_{pr}\mu_{pr}}{\pi_{post}} + \frac{\pi_l}{\pi_{post}} C,\ \left(\frac{\pi_{post}^2}{\pi_l^2}\pi_s\right)^{-1}\right) \tag{2}$$

and thus we get the average bias and the variance of the percepts:

$$\text{Bias}(\hat{C}) = \text{E}(\hat{C}) - C = \frac{\pi_{pr}\mu_{pr}}{\pi_{post}} + \frac{\pi_l}{\pi_{post}} C - C = \frac{\pi_{pr}}{\pi_{post}}(\mu_{pr} - C) \tag{3}$$

$$\text{Var}(\hat{C}) = \frac{\pi_l^2}{\pi_{post}^2 \pi_s} \tag{4}$$

In the sampling variant, we have $\hat{C} \mid I \sim \mathcal{N}(\mu_{post}, \pi_{post}^{-1})$. From the law of total expectation and (1), we get $\text{E}(\hat{C}) = \text{E}[\text{E}(\hat{C} \mid I)] = \text{E}(\mu_{post})$. This means that

$$\text{E}(\hat{C}) = \text{E}\left(\frac{\pi_{pr}\mu_{pr}}{\pi_{post}} + \frac{\pi_l}{\pi_{post}} I\right) = \frac{\pi_{pr}\mu_{pr}}{\pi_{post}} + \frac{\pi_l}{\pi_{post}}\text{E}(I) = \frac{\pi_{pr}\mu_{pr}}{\pi_{post}} + \frac{\pi_l}{\pi_{post}} C \tag{5}$$

and therefore the bias is the same as with the mean variant, $\text{Bias}(\hat{C}) = (\mu_{pr} - C)\pi_{pr}/\pi_{post}$. For the variance, we use the law of total variance:

$$\text{Var}(\hat{C}) = \text{E}[\text{Var}(\hat{C} \mid I)] + \text{Var}[\text{E}(\hat{C} \mid I)] = \text{E}(\pi_{post}^{-1}) + \text{Var}(\mu_{post}) \tag{6}$$

But $\text{E}(\pi_{post}^{-1})$ doesn't vary with I, so $\text{E}(\pi_{post}^{-1}) = \pi_{post}^{-1}$. And $\text{Var}(\mu_{post})$ is known from (2). Therefore:

$$\text{Var}(\hat{C}) = \pi_{post}^{-1} + \frac{\pi_l^2}{\pi_{post}^2 \pi_s} = \frac{\pi_{post}\pi_s + \pi_l^2}{\pi_{post}^2 \pi_s} \tag{7}$$

Essentially the variance of the sampling model is the sum of the variance of the mean of the posterior (4) plus the variance of the posterior itself.

3.A.2 Inferring the parameters from bias and variance measurements

Commonly in behavioural experiments participants are presented with different stimuli (C) that all have the same level of sensory noise (π_s). Then, researchers calculate the response bias relative to each stimulus and the response variance. In this section, we will investigate which parameters can be estimated from such measurements and what kind of designs might offer more information.

Given that μ_{pr} is not affected by the change in stimulus value (C), having at least two different allows us to calculate both the ratio of prior to posterior precision and the mean of the prior using the formula for the bias (3), which is identical in both models. Let us define $\pi_{pr}/\pi_{post} = a$. Then, since $\pi_{post} = \pi_l + \pi_{pr}$, we can easily see that

$$\pi_l/\pi_{pr} = a^{-1} - 1 \tag{8}$$

The variance does not depend on the stimulus value and therefore remains constant across trials, $\mathrm{Var}(\hat{C}) = v$. Using the formula for variance in the mean variant (4), we can calculate the actual sensory precision

$$\pi_s = (1-a)^2/v \tag{9}$$

However, it is not possible to estimate the values of π_{pr} and π_l independently.

In the sampling variant, we get $v = \pi_{post}^{-1} + a/\pi_s$ from (7), which means that π_s cannot be estimated independently of π_{post}. Moreover, without any additional assumptions, the two variants cannot be differentiated from each other, as there is no prohibiting combination of bias and variance values. That is, fixing the ratio of prior to posterior precision at a does not restrict the possible variance values in any way.

One possible way to estimate the parameters in the sampling variant is to assume that the estimated sensory precision is proportional to the actual sensory precision, $\pi_l = \kappa\pi_s$, and use data from trials with two sensory noise levels (e.g. contrast levels), $\pi_{s,1}$ and $\pi_{s,2}$. From the proportionality assumption, the variance (7) becomes

$$v = \frac{\pi_{post} + \kappa\pi_l}{\pi_{post}^2} = \frac{\pi_{pr} + (\kappa+1)\pi_l}{(\pi_{pr} + \pi_l)^2} \tag{10}$$

but π_l can be expressed in terms of π_{pr} from (8) and therefore

$$v = \frac{((\kappa+1)(a^{-1}-1)+1)\pi_{pr}}{a^{-2}\pi_{pr}^2} = \frac{\kappa a - \kappa a^2 + a}{\pi_{pr}} \tag{11}$$

Using two sensory noise levels results in two values for variance. Given that π_{pr} does not change between sensory noise levels, taking the ratio of the variances gives us

$$\frac{v_1}{v_2} = \frac{\kappa a_1 - \kappa a_1^2 + a_1}{\kappa a_2 - \kappa a_2^2 + a_2} \implies \kappa = \left(\frac{v_1}{v_2}a_2 - a_1\right)/(a_1 - a_1^2 - \frac{v_1}{v_2}a_2 + \frac{v_1}{v_2}a_2^2) \tag{12}$$

with v_1, v_2, a_1, a_2 being known. Knowing κ, we can use (11) to calculate π_{pr}, then use (8) to calculate $\pi_{l,1}$ and $\pi_{l,2}$, and finally use our proportionality assumption to calculate $\pi_{s,1}$ and $\pi_{s,2}$.

From (4), (8), and the proportionality assumption we get that in the mean model $v_1/v_2 = \pi_{s,2}/\pi_{s,1} = \pi_{l,2}/\pi_{l,1} = (a_2^{-1}-1)/(a_1^{-1}-1)$, a relationship that does not hold true in the sampling model. In that way, we can differentiate between the two variants as well.

3.A.3 Bias and variance in the Bayesian theories of ASD

The three main theories of ASD can be formalised as follows (accented parameters correspond to autistic values; see main text, Section **3.5** for details):

I. $\pi'_{pr} < \pi_{pr}$

II. $\pi_l' > \pi_l$, with $\pi_s' = \pi_s$

III. $\pi_s' > \pi_s$, with the relationship between π_l and π_s remaining intact

A reminder here that the formula for the bias is identical between the variants, $Bias(\hat{C}) = \pi_{pr}/(\pi_{pr} + \pi_l)$. From that it is clear that all theories predict reduced bias in ASD.

In terms of variance it is clear that theory I predicts that it would be increased in ASD for both variants, as a decrease in prior precision would lead to a decrease in posterior precision, without affecting the remaining variables (4, 7). Theory II would result in increased variance for the mean model as well, as both π_l and π_{post} would get reduced by the same amount, but $\pi_{post} > \pi_l$. For the sampling model, the change in variance depends on the values of the other parameters. Specifically, if we take the derivative of the formula for variance (4), keeping in mind that $\pi_{post} = \pi_l + \pi_{pr}$, we get

$$\frac{\partial v}{\partial \pi_l} = \frac{(2\pi_{pr} - \pi_s)\pi_l - \pi_{pr}\pi_s}{\pi_s(\pi_l + \pi_{pr})^3} \tag{13}$$

Since all parameters are positive, the sign of the derivative depends only on the numerator. If $\pi_s > 2\pi_{pr}$, then the derivative is negative, leading to lower variance in ASD. Otherwise, the effects of the theory depend on the exact values of the parameters. If both π_l' and π_l are larger than $\pi_s\pi_{pr}/(2\pi_{pr} - \pi_s)$, variance in ASD is increased and if both are smaller it is reduced.

Similarly, in theory III, both higher and lower variance could theoretically be found in ASD. Even if we assume that the proportionality assumption holds, then in the mean variant

$$\frac{\partial v}{\partial \pi_s} = \frac{\kappa^2\pi_{pr} - \kappa^3\pi_s}{\pi_{post}^3} \tag{14}$$

Therefore if $\pi_s', \pi_s < \pi_r/\kappa$, variance would be lower in ASD, while if $\pi_s', \pi_s > \pi_r/\kappa$ variance would be higher.

In the sampling variant with the proportionality assumption, the derivative is

$$\frac{\partial v}{\partial \pi_s} = -\frac{\kappa(-\kappa\pi_{pr} + \pi_{pr} + \kappa^2\pi_s + \kappa\pi_s)}{\pi_{post}^3} \tag{15}$$

with a root of $\pi_s = (\kappa - 1)\pi_{pr}/(\kappa^2 + \kappa)$. Therefore, if both π_s' and π_s are smaller than the root, variance is lower in ASD, whereas if both are larger it would be higher.

3.B: Simulations Appendix

3.B.1 Simple Bayesian model

We run simulations using both the mean and the sampling variants of our simplified model with various parameter values to estimate parameter and model recoverability from behavioural responses. Specifically, we assumed that priors and likelihoods were gaussians and thus the posterior mean (μ_{post}) and precision (π_{post}) were determined as follows:

$$\mu_{post} = \frac{\pi_{pr} \cdot \mu_{pr} + \pi_s^M \cdot I}{\pi_{post}}$$

$$\pi_{post} = \pi_{pr} + \pi_s^M$$

with the sensory input *I* being drawn randomly from a distribution centred at the stimulus *C* with precision π_s^A (see the main text for details). Then, for the mean variant, we took the posterior mean as the response ($\hat{C}$), while for the sampling variant we drew the response from the posterior. In both cases it was assumed that motor noise is 0. Stimuli were drawn uniformly from the interval (0, 2) and μ_{pr} was set at 0 for all simulated participants. This was done to simplify the fitting process, but does not affect our conclusions as only the *distance* between stimulus and prior mean is relevant for the model.

We created 1000 simulated participants for each variant. Each participant had a random π_{pr} value, a random π_s^A value, and a random π_s^M value, all between 0.5 and 2, chosen with a uniform probability. Each participant 'saw' 1000 different stimuli between 0 and 2 and gave 1000 responses. From these responses, we estimated the simulated parameters using maximum likelihood estimation (MLE). MLE is considered substandard in real experimental settings, with few data which might not be normally distributed. However, in an ideal setting and with

enough data MLE is able to recover parameters that are recoverable, especially given the simplicity of our models. Moreover, in contrast with other fitting methods, MLE makes no assumption about the distribution of parameters.

To test parameter recoverability, we looked at Pearson correlations between simulated and recovered parameters, as well as scatterplots of the parameter values. Our results showed that only the actual sensory precision π_s^A could be recovered, and that only for the mean model (main text, Figure 3.3). For model recovery, we compared the log likelihood values between the models. Because the models have the same number of parameters, there is no difference between AIC and BIC values and the likelihood values can be compared directly (similarly to a likelihood ratio test). Our results showed that model log likelihoods are equal between the mean and sampling variants of the model for the vast majority of the participants, meaning that both variants can exactly equally well approximate the majority of the data. This means that the variant of the model cannot be recovered from simple behavioural data in such an experiment.

3.B.2 Moving Dots task

For the Moving Dots task and the corresponding models as presented in our past study (Karvelis et al., 2018; see Section 6 of the main text), we used a similar approach. We generated 175 responses for 500 different parameter combinations drawn uniformly from intervals that were based on the estimated values from the behavioural data ($\theta_{pr} \in [0, 64]$, $\sigma_{pr} \in [5, 55], \sigma_s^A, \sigma_s^M \in [3, 13], a \in [0, 0.2]$). While sampling the parameters independently risks oversampling the parameter space, using the parameter value combinations estimated by our models risks undersampling it, especially given the poor identifiability of the prior and estimated sensory precisions and the demonstrated relationship between them. We decided that the second risk was more relevant to our study. We used four versions of the original model to generate and fit the data ($\pi_s^M = \pi_s^A$ vs $\pi_s^M \neq \pi_s^A$, posterior mean variant vs posterior sampling variant). Model fit was quantified using the Bayesian Information Criterion.

Table 3.B1. Model recovery.

$$\pi_s^M = \pi_s^A$$

		Recovered Participants	
		Mean	**Sampling**
Simulated Participants	**Mean**	328	172
	Sampling	130	370

$$\pi_s^M \neq \pi_s^A$$

		Recovered participants	
		Mean	**Sampling**
Simulated Participants	**Mean**	338	162
	Sampling	193	307

Part II

Extending

* * *

After examining the theoretical foundations of Bayesian theories of autism in Part I, Part II represents a shift towards extending and refining these theories through novel experimental work that attempts to move beyond the simplest version of these theories. **Chapter 4** investigates a more sophisticated model of prior-likelihood interactions, **Chapter 5** attempts to expand upon this work using a social task, and **Chapter 6** focuses on an experiment that distinguishes between explicit and implicit learning of environmental regularities.

Our systematic review in Chapter 2 revealed that straightforward accounts of autism as a result of underweighted priors or overweighted sensory evidence are likely insufficient. While many studies found the expected differences between autistic and neurotypical individuals or across autistic traits, these differences were often inconsistent and difficult to reconcile with broad Bayesian theories. This suggested that the underlying mechanisms might be more complex and that correspondingly more nuanced approaches would be needed, both in terms of computational modeling and potentially experimental design. These insights appeared early in our review process, leading us to investigate circular inference in parallel with completing the systematic review. Therefore, the work of Chapter 4 was performed simultaneously with parts of Chapter 2, but is presented here, out of chronological order, as it fits thematically with Part II alongside other experimental work that takes a more nuanced approach to understanding autism.

Circular inference is a model that goes beyond simple imbalances between priors and likelihoods, formalising a more complex interaction between them. It proposes that excitation-inhibition imbalances in the brain can cause bottom-up or top-down signals to be overcounted in Bayesian inference, leading to perceptual or cognitive overconfidence. Our decision to investigate circular inference was influenced by its success in explaining aspects of schizophrenia, another condition characterised by excitation-inhibition imbalances. Autism and schizophrenia share some neurobiological characteristics, but they also have partially overlapping Bayesian theories, as some explanations of schizophrenia rely on overweighting likelihoods as the underlying mechanism. Given these similarities, circular inference appeared to offer a promising avenue for developing a more sophisticated

computational account of autism.

The timing of this work coincided with the COVID-19 pandemic, which necessitated a turn to online experimentation. While this presented certain challenges, it also offered an opportunity to recruit a larger sample than would typically be possible in laboratory studies. This led us to adopt a dimensional approach, examining circular inference across the spectrum of autistic traits in the general population, which we complemented by a categorical comparison between individuals who self-reported autism diagnoses online and those that did not identify as part of the autism spectrum.

Chapter 4

Circular Inferences in Individuals with Autistic Traits

This chapter includes the postprint and the supplementary information of a published paper: Angeletos Chrysaitis, N., Jardri, R., Denève, S., & Seriès, P. (2021). No increased circular inference in adults with high levels of autistic traits or autism. PLoS computational biology, 17(9), e1009006.

Chapter 4 Abstract

Autism spectrum disorders have been proposed to arise from impairments in the probabilistic integration of prior knowledge with sensory inputs. Circular inference is one such possible impairment, in which excitation-to-inhibition imbalances in the cerebral cortex cause the reverberation and amplification of prior beliefs and sensory information. Recent empirical work has associated circular inference with the clinical dimensions of schizophrenia. Inhibition impairments have also been observed in autism, suggesting that signal reverberation might be present in that condition as well. In this study, we collected data from 21 participants with self-reported diagnoses of autism spectrum disorders and 155 participants with a broad range of autistic traits in an online probabilistic decision-making task (the fisher task). We used previously established Bayesian models to investigate possible associations between autistic traits or autism and circular inference. There was no correlation between prior or likelihood reverberation and autistic traits across the whole sample. Similarly, no differences in any of the circular inference model parameters were found between autistic participants and those with no diagnosis. Furthermore, participants incorporated information from both priors and likelihoods in their decisions, with no relationship between their weights and psychiatric traits, contrary to what common theories for both autism and schizophrenia would suggest. These findings suggest that there is no increased signal reverberation in autism, despite the known presence of excitation-to-inhibition imbalances. They can be used to further contrast and refine the Bayesian theories of schizophrenia and autism, revealing a divergence in the computational mechanisms underlying the two conditions.

Keywords: circular inference, Bayesian Brain, autism, schizophrenia, computational psychiatry, excitation-to-inhibition ratio

4.1 Introduction

Autism spectrum disorder (ASD) and schizophrenia (SCZ) are two heterogeneous mental disorders with a complicated relationship (Joyce & Roiser, 2007; Masi et al., 2017). While the term 'autism' was initially used to refer to one of schizophrenia's symptoms (Kuhn & Cahn, 2004), the two disorders have since been considered as separate conditions and have been studied as such by most researchers. Despite that, numerous links have been observed between them, from behavioural and neurophysiological similarities in social cognition impairments (Couture et al., 2010; Pinkham et al., 2008), to immune (Patterson, 2009) or intestinal (Cade et al., 2000) dysregulation and genetic overlap (Carroll & Owen, 2009), among others. Such findings suggest that the relationship between schizophrenia and ASD should be more thoroughly explored, within a framework that is able to handle and explain their differences (B. H. King & Lord, 2011; Rapoport et al., 2009).

In Bayesian theories of perception and cognition, the brain is viewed as constantly making probabilistic calculations in order to infer the true state of the environment. The information coming from sensory inputs is captured by the likelihood function and is combined with prior beliefs about the environment, in a process akin to Bayesian inference (Knill & Pouget, 2004). This framework has been widely adopted in both ASD and SCZ research, with a frequently proposed hypothesis for both disorders being that sensory inputs are overweighted relative to prior beliefs (Lawson et al., 2014; Palmer et al., 2017; Pellicano & Burr, 2012; Sterzer et al., 2018; Valton et al., 2017) (see (Cassidy et al., 2018; Chambon et al., 2011; Powers et al., 2017) for an alternative SCZ hypothesis). In schizophrenia, this theory attempts to explain the tendency of patients to jump to conclusions (Speechley, 2010) and their partial immunity to perceptual illusions (D. J. King et al., 2017), with hallucinations and delusions being interpreted as the formation of bizarre beliefs to account for strange, hypersalient sensory data (Fletcher & Frith, 2009). Intriguingly, the hypothesis of overweighted sensory information is also suggested to account for most of ASD's symptoms, such as sociocognitive impairments, attention to detail, sensory hypersensitivity, and decreased susceptibility to illusions (Palmer et al., 2017). The similarity of the proposed theories for autism and schizophrenia is surprising

given their distinct symptomatology. However, very few Bayesian studies have examined both conditions using the same experimental or computational paradigm, which would be crucial for understanding their relationship and mechanisms of action.

In 2013, Jardri and Denève proposed a new computational explanation for schizophrenia, called *Circular Inference* (Jardri & Denève, 2013), motivated by an attempt to understand the potential consequences of the increased excitation-to-inhibition (E/I) ratio that is associated with the condition (Grent-'t-Jong et al., 2018; O'Donnell, 2011). Using hierarchical network simulations, they showed that inhibitory impairments in the cortex might lead to sensory evidence or prior beliefs being reverberated throughout the network that the brain uses to represent the environment, overwhelming the inferential process. Sensory input reverberation could cause the reported 'jumping to conclusions' bias in schizophrenia, where patients get overconfident in their beliefs based on relatively little evidence (Evans et al., 2015). The positive symptoms, then, can be seen as an extension of the same process, where hallucinations and delusions are produced by misplaced certainty in noisy perceptual and other non-sensory information, respectively.

Jardri et al. supported this hypothesis with behavioural evidence from 25 SCZ patients and 25 controls (Jardri et al., 2017), using a probabilistic variant of the beads task (Ross et al., 2015), called the *fisher task*. In the fisher task, subjects are asked to estimate the chance that a red fish caught by a fisherman came from one of two lakes, while being presented with the lake preferences of the fisherman and the proportions of red fish in each lake (Figure 4.1). The researchers interpreted the preferences (which were presented first) as a Bayesian prior and the fish proportions as the sensory evidence. They showed that all participants exhibited signs of signal reverberation. Importantly, they found that sensory evidence was reverberated more in patients, with the magnitude of reverberation being correlated with their positive symptoms. A following study confirmed these findings, utilising a social version of the beads task in a sample of 35 patients with schizophrenia or schizoaffective disorder and 40 controls (Simonsen et al., 2021). The researchers found that the circular inference model fitted best the participants' behaviour, with increased sensory reverberation in patients. They also presented strong

evidence for an association between that reverberation and various clinical features in patients (e.g., delusions, anhedonia-asociality).

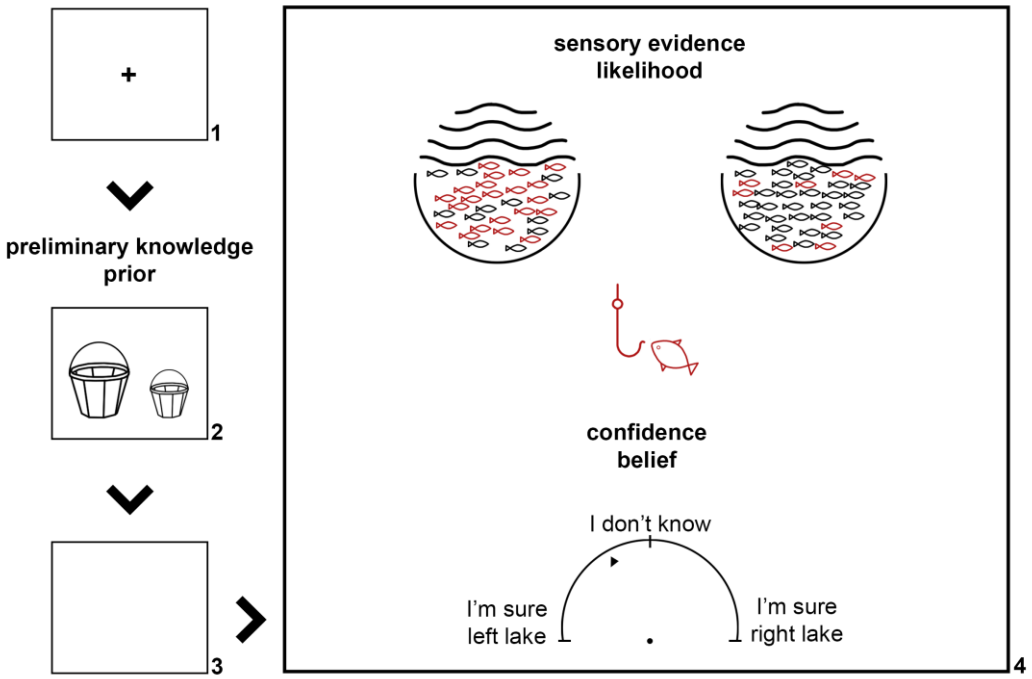

Figure 4.1 An outline of the four stages of the fisher task. *1) The fixation cross is presented; 2) participants are shown the preference of the fisherman, visualised as two baskets of varying sizes, one for each lake; 3) a blank screen is presented; 4) participants are shown the fish proportions and are asked to make a confidence estimate about the lake of origin of the fish (Adapted from (Jardri et al., 2017)).*

Impaired inhibition has been strongly associated with autism (Brown et al., 2005; Casanova et al., 2002; Gogolla et al., 2009; Kana et al., 2007; Rubenstein & Merzenich, 2003). A question that arises naturally is, therefore, whether circular inferences are present in ASD, and whether they would then be of the same nature as in schizophrenia (e.g., sensory vs prior reverberation). In the present study, we aimed to assess cue integration across a sample with a broad range of autistic traits, which also included some autistic participants ('autistic' is the preferred term by people on the autism spectrum (Kenny et al., 2016)). This allowed us to investigate signal reverberation within a dimensional as well as a more traditional, categorical view of autism (American Psychiatric Association, 2013; H. Kim et al., 2019; Robinson et al., 2011; Wiggins et al., 2012). To achieve that, we utilised an online version of the fisher task, and both circular inference and more traditional Bayesian models. This provided us with an opportunity to

explore the influences of ASD in probabilistic decision-making, while also allowing an additional, qualitative comparison with past SCZ findings.

4.2 Methods and Materials

4.2.1 Ethics Statement

The present study was approved by the University of Edinburgh, School of Informatics Ethics committee (RT number 29368).

4.2.2 Sample

We recruited 204 naive adults; 61 voluntarily via our social media networks and 143 with fixed monetary compensation via the Prolific recruiting platform (Palan & Schitter, 2018). All participants had normal or corrected-to-normal vision and were not taking any psychotropic medication. 28 subjects were excluded for providing low quality data (Section 4.A3 in Supplementary Information). The final sample included 102 male and 71 female participants, with a median age of 26.6 years. The study was conducted online. Participants were presented with detailed information about the study and had to click a button to indicate consent for the experiment to start.

Half of the Prolific subsample was selected to have a self-reported diagnosis of ASD or to identify as part of the autism spectrum (Section 4.A1 in Supplementary Information), with 21 subjects having a diagnosis in the final sample. All participants filled in the Autism Spectrum Quotient (AQ) questionnaire (Baron-Cohen et al., 2001) and the 21-item Peters et al. Delusions Inventory (PDI) (Peters et al., 2004). The final sample showed indeed stronger autistic traits (mean 22.9, SD 6.5) than what is usually found in the general population (mean 16.9, SD 5.6) (Ruzich et al., 2015), but no difference in delusional ideation (mean 6.1, SD 3.1 vs mean 6.7, SD 4.4) (Peters et al., 2004). Interestingly, the participants with the ASD diagnoses had AQ scores on the low-end (mean 28.0, SD 8.0) compared to those reported in the literature for

autistic individuals (mean 35.2, SD 6.3) (Ruzich et al., 2015). Statistical power for our tests could not be calculated, as model parameters were not verifiably following any known distribution. However, the strength of Jardri et al.'s findings (Jardri et al., 2017) suggests that comparable effects would reach statistical significance in our larger sample, according to an exploratory analysis (Section 4.A4 in Supplementary Information).

4.2.3 Procedure

The task was kept as similar to the original fisher task (Jardri et al., 2017) as possible. The participants were shown a fisherman having caught a red fish and were asked which of two lakes the fish was caught from. To make this decision, they were presented with two kinds of information in each trial: 1) the preferences of the fisherman for each of the lakes, visualised as two baskets of varying sizes (prior); 2) the proportions of red versus black fish in each lake, visualised as 100 fish in two lake drawings (sensory evidence or likelihood). Subjects were instructed to gauge their confidence and respond using a continuous semi-circular scale, ranging from 'I'm sure LEFT LAKE' to 'I'm sure RIGHT LAKE', with 'I don't know' in the middle. Confidence estimates were interpreted probabilistically, in a continuous manner, with a click on the left edge of the scale corresponding to a probability of 1 for the fish originating in the left lake (0 for the right) and vice versa.

Trials were structured as follows (Figure 4.1): Initially, a fixation cross was shown for 800ms, followed by the two baskets for 1000ms, and a blank screen lasting 50ms. Then, the lake drawings, the fisherman with the red fish, and the scale appeared on the screen until the subject gave a response. Participants were presented with detailed instructions which they could view many times before proceeding to the task. The instructions made clear that participants should respond 'as fast and as accurately as possible'. After the instructions, subjects completed 11 training trials with easy stimulus combinations to acclimate themselves with the task.

Due to concerns about participants' potential distractibility in an online environment if the task was too long, we reduced the number of trials to 130 (Section 4.A2 in Supplementary Information). The trials appeared in a random order, with lake drawings being different for

every trial. Every 22 trials, the participants were prompted to take a break, which they could end with the press of a button. Lakes had 9 possible ratios of red to black fish, while baskets appeared in 9 possible sizes, both corresponding to the probabilities 0.1 to 0.9. In all trials, likelihoods and priors were complimentary (e.g., if the left fish proportions were 0.3, the right would be 0.7). Therefore, probabilities mentioned in the text refer to the left lake, as the probabilities for the right can be immediately inferred.

4.2.4 Model-free analysis

Linear mixed-effects models (LMEs) were used to verify that participants combined the information of both baskets and fish ratios when making their decisions and to investigate any possible interactions with autistic traits. We chose the absolute confidence of the participants as the response variable (|c – 0.5|, with c being the participant confidence estimate). We modelled the following as fixed effects with repeated measures across the subjects in all LMEs: i) the absolute likelihood (|likelihood – 0.5|); ii) the prior congruency, that is how much the prior agreed with the likelihood (|prior – 0.5| * sgn[(prior – 0.5)(likelihood – 0.5)]); iii) the reaction times, which were used to investigate the possibility of a speed-accuracy trade-off. All LMEs also included the two-way interaction between i and ii, with the participants being treated as a random factor. We analysed our results with 5 different LME variants. The first one, LME_core only used the aforementioned components. LME_AQ expanded upon LME_core by including a fixed effect for AQ and the two- and three-way interactions of AQ with i and ii. LME_PDI was the same as LME_AQ but with the PDI scores instead of the AQ. Then, LME_full, used both AQ and PDI and their interactions with i and ii, but no interactions between them. Finally, LME_rtInteract expanded upon the LME_full to include interactions between AQ or PDI and reaction times. Full specification of the models in Wilkinson notation can be found in Section 4.B1 in Supplementary Information.

4.2.5 Bayesian models

Data were fitted with four Bayesian models: Simple Bayes (SB), Weighted Bayes (WB), and two variants of the circular inference model: Circular Inference – Interference (CII) and

Circular Inference – No Interference (CINI). Originally (Jardri & Denève, 2013), the inferential processes expressed by these models were simulated in a hierarchical network, where priors corresponded to top-down signals and likelihoods to bottom-up ones. In the current study, we followed Jardri et al. in fitting simplified models, that capture the network effects with significantly fewer free parameters (Jardri et al., 2017).

SB combines the two sources of information using Bayes' theorem. This is expressed in logits as

$$L_c = L_p + L_s \,, \tag{4.1}$$

with subscript *p* corresponding to trial prior, *s* to sensory evidence, and *c* to the confidence estimate, while *L* denotes the respective logit.

WB expands upon SB:

$$L_c = F\big(L_p, w_p\big) + F(L_s, w_s) \;\; , \tag{4.2}$$

where *F* is the sigmoid function

$$F(L, w) = \ln\left(\frac{we^L + 1 - w}{(1-w)e^L + w}\right) , \tag{4.3}$$

allowing for the underweighting of priors or likelihoods. Each weight *w* determines the influence of the corresponding signal to the confidence estimate. This depends on how the reliability of that signal is estimated by each participant.

CII has the form:

$$L_c = F\big(L_p + I, w_p\big) + F(L_s + I, w_s) \,, \tag{4.4}$$

$$I = F\big(a_p L_p, w_p\big) + F(a_s L_s, w_s) \,, \tag{4.5}$$

where top-down and bottom-up signals get reverberated, interfering with one another, and end up corrupting prior beliefs and sensory evidence by the same amount, *I*. Parameters a_p and a_s

affect the number of times the respective information is overcounted, expressing the signals' reverberation.

CINI is similar to CII, but it assumes that both signals get reverberated or overcounted separately and are only combined at the end of the process:

$$L_c = F\big(L_p + F\big(a_p L_p, w_p\big), w_p\big) + F(L_s + F(a_s L_s, w_s), w_s)\,. \tag{4.6}$$

SB has 0 free parameters, WB 2, and both CII and CINI have the same 4. True parameter ranges were [0.5, 1] for the weights (w) and [0, 60] for the reverberation parameters (a); however, these were rescaled to [0, 1] so that they could be easily compared with those reported by Jardri et al. 60 is an arbitrary upper limit, that however is high enough for our purposes, as no parameter approached it (max non-rescaled CINI a = 29.02). In the rest of this article, we will be referring exclusively to the rescaled parameters, however the word 'rescaled' will be omitted for conciseness. A (rescaled) weight value of w = 0 shows no influence of the corresponding signal, while both w = 1 make WB equivalent to SB and both a = 0 make CII and CINI equivalent to WB. The difference between CII and CINI is subtle, but important. In CINI, the sensory and prior signals are combined linearly, while in CII, one signal's influence on the model estimate depends on the strength of the other, due to the interference between them. Figure 4.2 illustrates the behavioural patterns predicted by the different models.

We followed Jardri et al., assuming Gaussian noise in the logit model estimates (L_c), and therefore fitted models via least squares, which is equivalent to maximum likelihood estimation in that case. Model comparison was performed using an approximation of the Bayesian information criterion (BIC) for normally distributed errors,

$$\mathrm{BIC} = n\ln(\sigma^2) + k\ln(n)\,, \tag{4.7}$$

where n is the number of datapoints, k the number of free parameters, and σ the model's mean squared error. To choose a model across all subjects, we followed the random-effects Bayesian model selection (Stephan et al., 2009), implemented in the SPM12 (*SPM12 Toolbox* -

Statistical Parametric Mapping, 2020). Group-level BIC (Li et al., 2008), a fixed-effects approach, produced similar results.

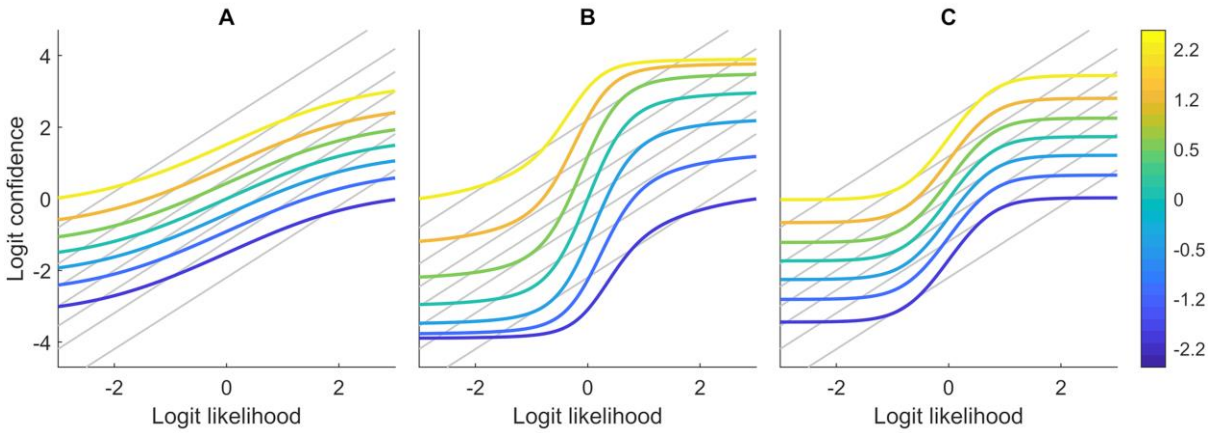

Figure 4.2 Illustration of WB (A), CII (B), and CINI(C) behaviour. *The graphs show how logit model confidence estimates change as a function of logit likelihood (fish proportions). Different colours represent different prior values (basket size) and grey lines represent the SB model predictions. The SB model simply combines the information of the two signals by adding their logits. WB can underweight either or both signals, while, in addition to that, the circular inference models allow for signal overcounting. In CII, the contribution of the likelihood on the confidence estimate depends on the prior value and vice versa. In contrast with that, in the CINI model, each source of information affects the confidence independently, and therefore the graph lines are completely parallel to each other. Parameter values were the same for all models (a_p=0.02, a_s=0.05, w_p=0.8, w_s=0.06).*

4.2.6 Statistical analysis and validity of results

We investigated the hypothesis of an association between autism and circular inference (H_1) in three ways: 1) correlations between model parameters and total AQ scores; 2) differences between the low- and high-AQ groups, defined as participants in the top and bottom 15% of the sample (AQ ≥ 30, n = 29, M/F: 13/14 vs AQ ≤ 16, n = 30, M/F: 15/15); 3) differences between subjects with an ASD diagnosis and those without, who also did not identify as part of the autism spectrum (ASD, n = 21, M/F: 13/8 vs ND, n = 61, M/F: 39/22; answers 1, 2 vs 5 in Section 4.A1 in Supplementary Information). The nonparametric measures of Kendall rank correlation coefficient and Mann-Whitney U test were chosen, as model parameters were not normally distributed (*Shapiro-Wilk test*; $p \leq 0.0068$) and there is no reason to expect a linear

relationship between them and psychiatric traits. The common language effect size statistic (*f*) was reported for the Mann-Whitney U (McGraw & Wong, 1992). All analyses were performed in MATLAB R2020a.

To quantify the evidence for the null hypothesis (H_0) in favour of the alternative one (H_1), we calculated the Bayes factors 01 (BF_{01}) for each of our tests. $1 < BF_{01} \leq 3$ constitutes weak evidence in favour of H_0, $3 < BF_{01} \leq 20$ positive evidence, and $BF_{01} > 20$ strong (Raftery, 1995). Note that $BF_{10} = 1/BF_{01}$. Bayes factors were calculated in JASP 0.14, using the default priors (*JASP*, 2020). To verify the fitting and model selection processes, we performed parameter and model recovery on CINI and CII using the current set with the 130 trials, as SB and WB scored very poorly in model comparisons.

Both models showed moderate recovery for the reverberation parameters (CINI a_p, $r = 0.54$; a_s, $r = 0.58$; CII a_p, $r = 0.54$; a_s, $r = 0.71$), although this was partly due to Pearson's correlation sensitivity to outliers (Y. Kim et al., 2015) (for details see Section 4.B3 in Supplementary Information). The models exhibited excellent recovery for the weight parameters (CINI w_p, $r = 0.96$; w_s, $r = 0.91$; CII w_p, $r = 0.94$; w_s, $r = 0.93$). They also showed no correlation between different parameters (Table 4.B3 and 4.B4 in Supplementary Information). Model recovery was good for both models, with approximately 80% of the simulated participants being better fitted by their generating model (Table 4.1). Experiment code, models, and anonymised data can be found at https://osf.io/yqug2/.

Table 4.1 Confusion matrix for model recovery.

		Recovered	
		CINI	CII
Simulated	CINI	799	201
	CII	185	815

Perfect model recovery would result in 1000 participants in the (CINI, CINI) and (CII, CII) cells, and 0 in the rest.

4.3 Results

4.3.1 Model-free findings

Participant responses adapted to changes in both priors and likelihoods, showing that they took both sources of information into account to make their confidence estimate (Figure 4.3). Despite that, their behaviour was not strictly Bayesian. A change from 0.5 to 0.4 or 0.6 in either prior or likelihood corresponded to a disproportionally large shift in the average response, indicative of signal reverberation.

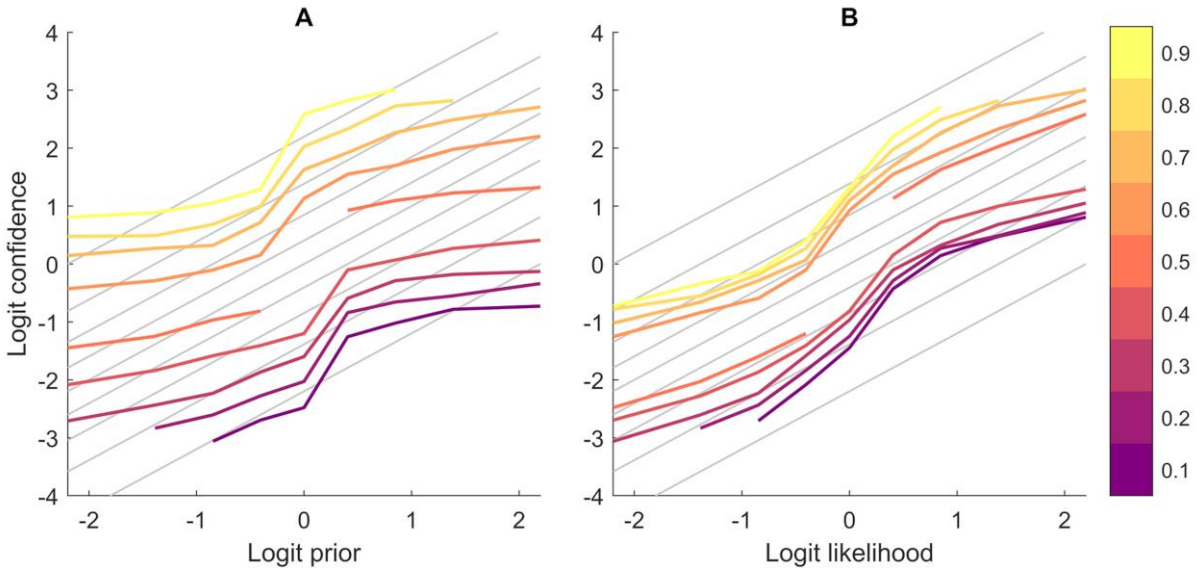

Figure 4.3 Average logit confidence estimates for all participants as a function of priors (A) and likelihoods (B). *Logit confidence estimates for the left lake increase following an increase in either prior probability for the left lake (baskets) or likelihood (fish ratios), showing that participants incorporate both information sources in their decision-making. However, their behaviour is far from strictly Bayesian, as evidenced by the differences between coloured and grey lines (SB confidence estimates). Different colours correspond to different likelihood (probability) values in the left graph and different prior (probability) values in the right.*

Among the linear mixed-effects models, the one which achieved the smallest BIC was LME_core (ΔBIC: LME_PDI, 17; LME_AQ, 35; LME_full, 51; LME_rtInteract, 69). All models confirmed the influence of both absolute likelihood (e.g., LME_core: $t = 44.50$, $p < 10^{-323}$) and prior congruency (e.g., LME_core: $t = 24.63$, $p = 10^{-132}$), as well as the interaction of

the two components (e.g., LME_core: t = 25.20, p = 10^{-138}). Despite the LME_core being the best model, both LME_PDI and LME_full showed significant association between absolute confidence and non-clinical delusional beliefs (PDI) (e.g., LME_PDI results: t = 2.08, p = 0.037) and an interaction between absolute likelihood and PDI (e.g., LME_PDI results: t = 2.31, p = 0.021). However, neither the influence of autistic traits (AQ) or its interactions with model components were significant in the LME_AQ and LME_full models. Reaction times showed a negative relationship with absolute confidence in all models (e.g., LME_core: t = –17.01, p = 10^{-64}), which is presumably a result of participants taking more time to respond when they are uncertain (Bonnet & Ars, n.d.). Importantly though, the LME_rtInteract achieved the worst BIC score, with no interaction between psychiatric traits and reaction times (LME_rtInteract: PDI t = 1.29, p = 0.20; AQ t = 0.72, p = 0.47). This suggests that any possible relationship between AQ or PDI and participant behaviour is not a result of differences in time management. The full LME results can be found in Section 4.B1 in Supplementary Information.

4.3.2 Model-based findings

Both random- and fixed-effects model comparisons showed that Circular Inference – No Interference was the best fitting model, followed by Circular Inference – Interference (Figure 4.4). Since model fit plots showed that both CINI and CII fit the data relatively well (Figure 4.D1 in Supplementary Information), for the sake of completeness, we conducted the same analysis with parameters from both models. Results from CINI are reported below, while those from CII can be found in Supplementary Information (Section 4.D2).

There was no evidence of a relationship between prior or likelihood reverberation and total AQ scores (Table 4.2). The only correlation that reached an (uncorrected) *p*-value of lower than 0.05 was a negative correlation between AQ and the CINI prior weight (τ = –0.12, p = 0.02, BF_{10} = 1.57), but this did not survive adjusting for multiple comparisons (Benjamini & Hochberg, 1995). Furthermore, the low- and the high-AQ groups behaved in a similar way (Figure 4.5), and the comparison between the parameters of high- and low-AQ groups did not reveal any difference, neither did the comparison between the ASD participants and those with

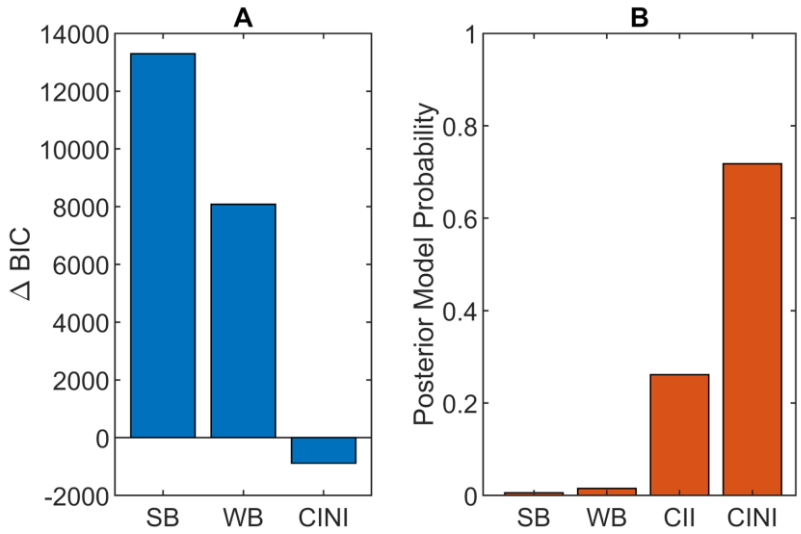

Figure 4.4 Results of fixed (A) and random (B) model comparisons. *(A) Group-level ΔBIC is defined as the sum of individual participant BIC scores for each model minus the sum for CII, used as a baseline, as it was the winning model in the Jardri et al. study (Jardri et al., 2017). The lower the BIC the better the model, with differences of more than 20 between BIC scores considered very strong evidence (Raftery, 1995). ΔBIC for CII is by definition 0. (B) Posterior model probabilities calculated using Bayesian model selection (Stephan et al., 2009). Both measures clearly show that Circular Inference models better account for the data, with CINI being a slightly better fit than CII.*

no diagnosis (ND) (Figure 4.6 & Table 4.3). Since it is possible that ND subjects with high autistic traits have an undiagnosed autism spectrum disorder, we performed an additional comparison between the ASD group and the subgroup of ND participants with weak autistic traits (AQ $\leq$ 17, n = 21, M/F: 13/8). No difference between these groups was found (Section 4.D1 in Supplementary Information). A weak positive correlation was found between PDI and the likelihood weight (τ = 0.13, p = 0.02, BF_{10} = 2.08; Table 4.2), that again is not significant when corrected. No relationship was present between psychiatric traits or diagnoses and CII parameters (Section 4.D2 in Supplementary Information).

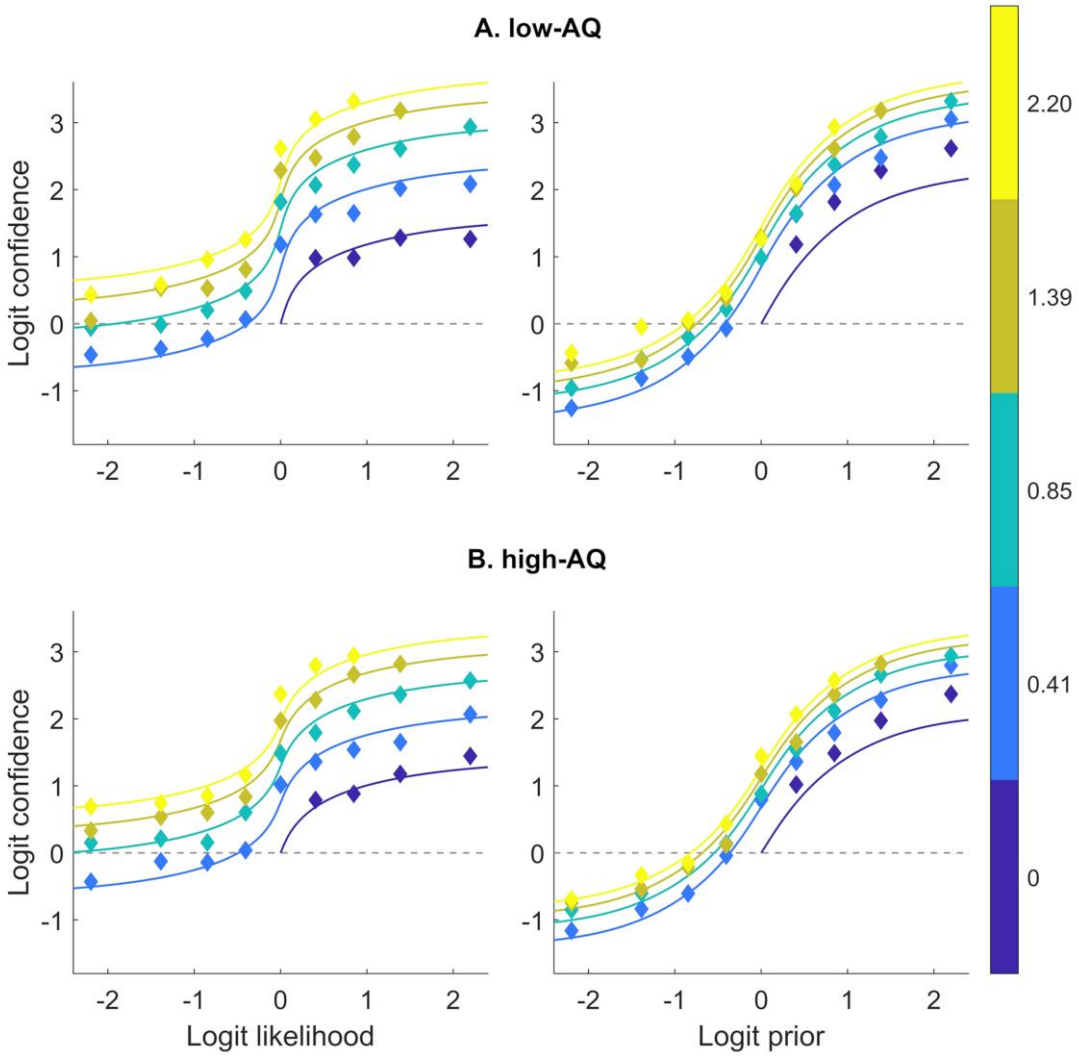

Figure 4.5 Participant confidence estimates and CINI model fits for the low-AQ (A) and the high-AQ (B) groups. *Model and participant logit confidence as function of logit likelihoods and priors. Coloured lines represent model predictions and rhombuses the participant confidence estimates. Different colours represent logit likelihood in A and logit prior values in B and are equivalent to probabilities of 0.5 to 0.9. Since both the task and the CINI model structure are symmetrical around 0 logit confidence (0.5 probability), participant estimates have been averaged between symmetric trials to reduce noise (e.g., a trial with a logit prior of –1 and a logit likelihood of 2 is symmetrical to one with a logit prior of 1 and a logit likelihood of –2).*

Table 4.2 Kendall rank correlations between CINI parameters and psychiatric traits.

	AQ			**PDI**		
CINI params	τ	p	BF_{01}	T	p	BF_{01}
a_p	0.04	0.5	7.98	–0.04	0.48	7.76
a_s	–0.02	0.65	9.11	0.01	0.85	9.94
w_p	–0.12	0.02	0.64	0.07	0.20	4.14
w_s	–0.02	0.69	9.35	0.13	0.02	0.48

Total AQ scores and Y/N PDI scores were used for the correlations. τ signifies the correlation coefficient. p-values are not adjusted for multiple comparisons. BF_{01} stands for the Bayes factor 01, with higher values corresponding to stronger evidence for the null hypothesis.

Table 4.3 Mann-Whitney U test results between the CINI parameters of the ASD and ND groups and the low-AQ and high-AQ groups.

	ND vs ASD			**low-AQ vs high-AQ**		
CINI params	f	p	BF_{01}	F	p	BF_{01}
a_p	0.55	0.50	3.82	0.54	0.63	3.60
a_s	0.45	0.46	3.15	0.47	0.70	3.88
w_p	0.47	0.64	3.85	0.40	0.19	1.96
w_s	0.53	0.71	3.71	0.42	0.31	2.17

Total AQ scores were used for the comparisons. f signifies the common language effect size, with larger f values corresponding to larger parameter values for the ASD and the high-AQ groups. An f of 0.5 corresponds to no differences. p-values are not adjusted for multiple comparisons. BF_{01} stands for the Bayes factor 01, with higher values corresponding to stronger evidence for the null hypothesis.

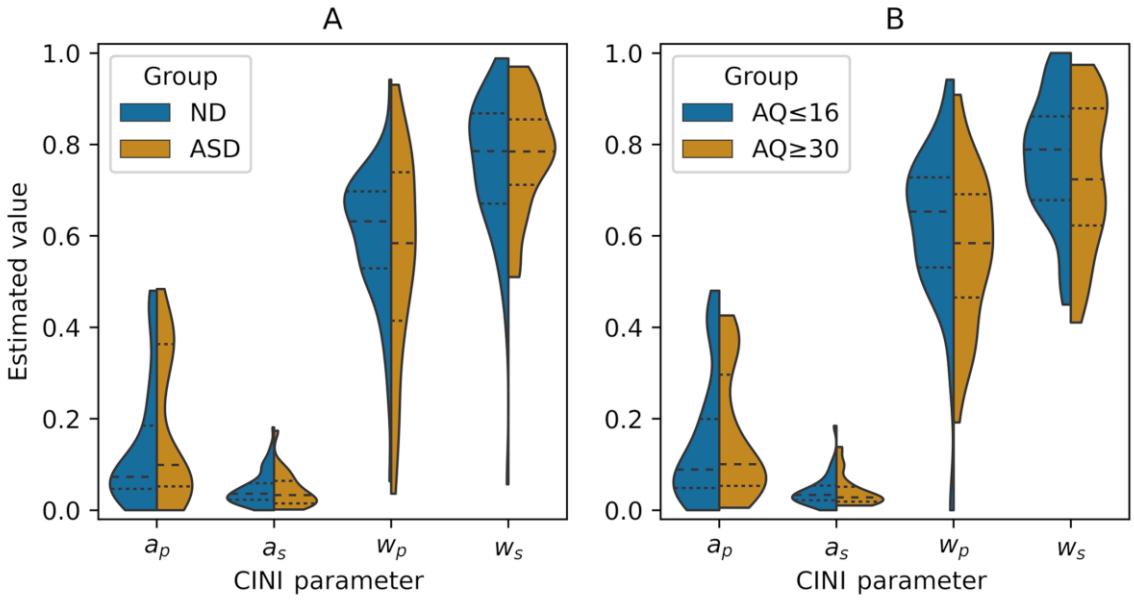

Figure 4.6 CINI parameter values of ND vs ASD groups (A) and low-AQ vs high-AQ groups (B). *Violin plots show the density of estimated parameters over the possible values, relative to the subgroup size. Dashed lines in the middle represent the median, while dotted ones represent the top and bottom quartiles in each group. No differences are observed between groups.*

4.4 Discussion

In the present study, we investigated the relationship between circular inference and autistic traits or autism. Circular inference is an impairment in Bayesian hierarchical networks where top-down or bottom-up signals get reverberated throughout the network, becoming significantly amplified (Jardri & Denève, 2013). We hypothesised that stronger autistic traits and ASD diagnoses would be associated with stronger reverberation of priors or sensory evidence. We used the fisher task, a probabilistic decision-making task that had been used previously with patients with schizophrenia (Jardri et al., 2017). To our knowledge, this is the first study to explore signal reverberation in ASD. Our analysis showed that the circular inference models perform best across the whole sample, similarly to previous results (Jardri et al., 2017; Simonsen et al., 2021). However, our hypothesis was refuted. Specifically, no correlation was found between autistic traits and either reverberation parameter. Similarly, there were no differences in these parameters between the groups with the strongest and

weakest autistic traits, and no differences between the autistic subjects and those with no self-reported diagnosis.

Circular inference attempts to model the effects of increased excitation-to-inhibition ratio, a phenomenon which has been strongly associated with schizophrenia (Grent-'t-Jong et al., 2018; O'Donnell, 2011). Indeed, Jardri et al. found clear experimental evidence for stronger likelihood reverberation in SCZ patients, using the fisher task (Jardri et al., 2017). On that account, the absence of any difference between our participant groups is surprising, given the observed inhibitory impairments in ASD (Gogolla et al., 2009; Kana et al., 2007) and the commonalities between autism and schizophrenia regarding E/I imbalances (Davenport et al., 2019; Foss-Feig et al., 2017). Moreover, prominent computational explanations for the two conditions suggest similar Bayesian impairments between them. Specifically, it has been proposed that an imbalance of likelihoods to priors, in favour of the former, lies at the heart of both ASD and SCZ (Lawson et al., 2014; Palmer et al., 2017; Pellicano & Burr, 2012; Sterzer et al., 2018; Valton et al., 2017). This seems to be contradicted by our findings, which showed no increase in reverberation along the autism spectrum, despite the presence of such an increase in schizophrenia. This is further exhibited in a qualitative comparison between the conditions, which showed higher likelihood reverberation in SCZ (Figure 4.E1 in Supplementary Information). A divergence in the Bayesian mechanisms of the two conditions has also been observed by Karvelis et al., which showed an association between autistic traits and increased sensory precision, but no discernible imbalance in schizotypy, in a statistical learning task (Karvelis et al., 2018). A partial divergence was also found by Noel et al., in an audio-visual synchrony task, where patients with schizophrenia showed increased unreliability in sensory representations, in addition to differences in their priors, which they shared with the autistic participants (Noel et al., 2018). No other studies are known to us that compare ASD and SCZ using the same tasks and Bayesian models, despite the commonalities between their computational explanations (for reviews see (Palmer et al., 2017; Valton et al., 2017)).

In agreement with the findings of Jardri et al. (Jardri et al., 2017), we found compelling evidence for signal reverberation across our sample. Interestingly though, one variant of the model, Circular Inference – No Interference (CINI), was a better fit for our data compared to

the other variant, Circular Inference – Interference (CII), contrary to the Jardri et al. study. Additional analysis of the Jardri et al. dataset revealed that this is partially because we used fewer trials than in the original study (Table 4.E3 in Supplementary Information). Furthermore, even in the original dataset, CII was dominant mostly in the SCZ subsample, while it performed equally well with CINI in controls. The dominance of CINI across our sample (Table 4.E3 in Supplementary Information) is an indication that prior beliefs and sensory evidence are reverberated, even in healthy participants. While we can only speculate about the possible neurobiological underpinnings of our circular inference models, we proposed that reverberation arises due to an increased E/I ratio, based on the network model of Jardri & Denève (Jardri & Denève, 2013). If that is the case, CINI might correspond to a weak or localised E/I imbalance, affecting the signals only separately. In schizophrenia, then, this imbalance would be larger and extend across the cognitive hierarchy, which would lead to interference between priors and likelihoods, making CII the better fit for these participants.

Surprisingly, we found no evidence for an association between prior or likelihood weights and ASD diagnoses or AQ scores. This result is seemingly in contrast with previous studies showing an overweighting of likelihoods relative to priors in autistic individuals or those with stronger autistic traits (e.g., (Karaminis et al., 2016; Karvelis et al., 2018; Król & Król, 2019; Powell et al., 2016)). However, these effects have been demonstrated exclusively in perceptual tasks, with the rare study of Bayesian decision-making in ASD showing no such imbalance (Lu et al., 2019; Manning et al., 2017). Another important difference is that in most of the literature, participants have to learn prior beliefs based on the observed statistics, while in our study they are explicitly presented by the size of the baskets. It is possible that the cause of the prior-likelihood imbalance found in the literature lies in impaired prior acquisition, rather than in the relative weighting of the prior per se.

Our analysis revealed a slight increase of absolute confidence with stronger non-clinical delusional beliefs (PDI), but no association between PDI and any reverberation parameter. This confirms the Jardri et al. findings of no such relationship in healthy subjects, although only 8 participants had scores above the clinical PDI mean of 11.9 (Peters et al., 2004). This result would deserve further investigation with a more thorough assessment of schizotypy, so as to

assess how it can fit with the dimensional view of schizophrenia (Nelson et al., 2013). Interestingly, PDI scores showed a significant interaction with likelihoods in the linear mixed-effects models and a slight correlation with the likelihood weights. While the latter result was not significant when adjusted for multiple comparisons, both of them agree with the prominent theory of overweighted likelihoods in schizophrenia (e.g., (Sterzer et al., 2018)).

4.4.1 Limitations and future work

Through our recruitment methods, we had aimed to recruit participants with a broad range of autistic traits. However, the resulting variance of AQ in our sample (SD 6.5) was only marginally higher than what is found in the general population (SD 5.6, (Ruzich et al., 2015)). Moreover, only 4 participants had an AQ score of more than 1 SD below the neurotypical mean of 16.9 (Ruzich et al., 2015) and only 5 participants had an AQ above the clinical mean of 35.2 (Ruzich et al., 2015). A wider range of autistic traits would be useful in investigating Bayesian impairments that might be associated with the extremes of the AQ distribution. Moreover, the diagnoses of our participants in the Prolific subsample were self-reported. What those diagnoses were based on, and when they were delivered is uncertain, which could explain the atypical AQ scores of the ASD group. Our findings will need to be confirmed in a sample verified by a mental health professional, especially as the criteria for an ASD diagnosis have largely changed between versions of the Diagnostic and Statistical Manual of Mental Disorders (Young & Rodi, 2014). Such a study would also benefit from cognitive measures, to ensure that perceptual or verbal reasoning abilities do not constitute a confounder for any differences between the groups. Another limitation that also nuances the comparison with previous investigations in schizophrenia (Jardri et al., 2017) concerns the fact that our experiment took place online. The lack of a lab-controlled environment could have substantially affected the quality of the collected data. Adding to that is the absence of in-person communication between participants and researchers, so the instructions of the task could have been clearly conveyed and possible questions answered. Such effects were visible in our dataset by the large portion of subjects that were excluded (≈14%).

In the fisher task, the baskets are presented before the lakes. This means that participants might simply display a recency bias, where the most recent evidence is overweighted. Under the Bayesian framework, the earlier evidence should create a prior belief in the participants, which is then combined with and updated by following evidence. Therefore, a recency bias is indistinguishable from an overweighting of sensory evidence. A possible issue, though, is that behavioural differences might be related to differences in the working memory of the participants. This could be especially important, since working memory is impaired in both ASD and schizophrenia (Forbes et al., 2009; Wang et al., 2017). However, Jardri et al. measured working memory performance in their sample and showed that it is correlated only with the prior weights, but not with the reverberation parameters. This would need to be validated in further studies, but we therefore expect that our findings regarding circular inference in autism should be robust to potential differences in working memory.

As with other findings relating behaviour to Bayesian inference impairments, it will be important to assess how our findings can be generalised to other tasks or modalities. Circular inference is formalised within a hierarchical Bayesian framework of cognitive processing. This framework assumes that priors express the (top-down) influences of the more abstract representations of the environment to the less abstract ones, while likelihoods encode the reverse (bottom-up) influences (Denève & Jardri, 2016). It is difficult to verify that the information presented in the current task (baskets and fish proportions) is encoded by subjects in the expected way – that is, that the preferences of the fisherman correspond to more abstract or contextual information and the fish proportions to more sensory. If these stimuli were processed by the participants as being in the same conceptual level, the task structure would be more akin to a delayed cue integration task (Ernst & Banks, 2002). Additionally, it is possible that the basket size is treated by some participants as a qualitative variable, leading them to disregard the exact difference in size, something that would appear as prior overcounting in the models (although see Section 4.D4 in Supplementary Information). We believe that these concerns do not invalidate our results, but further research would be needed to understand how delayed cue integration tasks or qualitative information fit within the circular inference framework.

Future research should replicate both ASD and SCZ findings in other tasks, involving different cognitive modalities. The social beads task of Simonsen et al. (Simonsen et al., 2021), for example, might be well suited for the investigation of signal reverberation in ASD, given the condition's impairments. Perceptual tasks, on the other hand, would avoid conscious strategies that are especially prevalent in decision-making, focusing instead on more fundamental computations in the brain and connecting circular inference with the rest of the Bayesian literature. Equally important is clarifying the connection between reverberation and neurophysiological measures, with a focus on the spatial patterns of E/I imbalances across brain areas. Differences in such patterns could explain why computational (Palmer et al., 2017; Valton et al., 2017) and neurobiological (Couture et al., 2010; B. H. King & Lord, 2011) theories of ASD and SCZ partially overlap, while their phenotypic expressions differ (Foss-Feig et al., 2017).

Chapter 4 References

American Psychiatric Association (Ed.). (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bonnet, C., & Ars, J. F. (n.d.). Reaction times as a measure of uncertainty. 7.

Brown, C., Gruber, T., Boucher, J., Rippon, G., & Brock, J. (2005). Gamma Abnormalities During Perception of Illusory Figures in Autism. *Cortex*, *41*(3), 364–376. https://doi.org/10.1016/S0010-9452(08)70273-9

Cade, R., Privette, M., Fregly, M., Rowland, N., Sun, Z., Zele, V., Wagemaker, H., & Edelstein, C. (2000). Autism and Schizophrenia: Intestinal Disorders. *Nutritional Neuroscience*, *3*(1), 57–72. https://doi.org/10.1080/1028415X.2000.11747303

Carroll, L. S., & Owen, M. J. (2009). Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Medicine*, *1*(10), 102. https://doi.org/10.1186/gm102

Casanova, M. F., Buxhoeveden, D. P., Switala, A. E., & Roy, E. (2002). Minicolumnar pathology in autism. *Neurology*, *58*(3), 428–432. https://doi.org/10.1212/WNL.58.3.428

Cassidy, C. M., Balsam, P. D., Weinstein, J. J., Rosengard, R. J., Slifstein, M., Daw, N. D., Abi-Dargham, A., & Horga, G. (2018). A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine. *Current Biology*, *28*(4), 503-514.e4. https://doi.org/10.1016/j.cub.2017.12.059

Chambon, V., Pacherie, E., Barbalat, G., Jacquet, P., Franck, N., & Farrer, C. (2011). Mentalizing under influence: Abnormal dependence on prior expectations in patients with schizophrenia. *Brain*, *134*(12), 3728–3741. https://doi.org/10.1093/brain/awr306

Couture, S. M., Penn, D. L., Losh, M., Adolphs, R., Hurley, R., & Piven, J. (2010). Comparison of social cognitive functioning in schizophrenia and high functioning autism: More convergence than divergence. *Psychological Medicine*, *40*(4), 569–579. https://doi.org/10.1017/S003329170999078X

Davenport, E. C., Szulc, B. R., Drew, J., Taylor, J., Morgan, T., Higgs, N. F., López-Doménech, G., & Kittler, J. T. (2019). Autism and Schizophrenia-Associated CYFIP1 Regulates the Balance of Synaptic Excitation and Inhibition. *Cell Reports*, *26*(8), 2037-2051.e6. https://doi.org/10.1016/j.celrep.2019.01.092

Denève, S., & Jardri, R. (2016). Circular inference: Mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences*, *11*, 40–48. https://doi.org/10.1016/j.cobeha.2016.04.001

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Evans, S., Averbeck, B., & Furl, N. (2015). Jumping to conclusions in schizophrenia. *Neuropsychiatric Disease and Treatment*, 1615. https://doi.org/10.2147/NDT.S56870

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Forbes, N. F., Carrick, L. A., McIntosh, A. M., & Lawrie, S. M. (2009). Working memory in schizophrenia: A meta-analysis. *Psychological Medicine*, *39*(6), 889–905. https://doi.org/10.1017/S0033291708004558

Foss-Feig, J. H., Adkinson, B. D., Ji, J. L., Yang, G., Srihari, V. H., McPartland, J. C., Krystal, J. H., Murray, J. D., & Anticevic, A. (2017). Searching for Cross-Diagnostic Convergence: Neural Mechanisms Governing Excitation and Inhibition Balance in Schizophrenia and Autism Spectrum Disorders. *Biological Psychiatry*, *81*(10), 848–861. https://doi.org/10.1016/j.biopsych.2017.03.005

Gogolla, N., LeBlanc, J. J., Quast, K. B., Südhof, T. C., Fagiolini, M., & Hensch, T. K. (2009). Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *Journal of Neurodevelopmental Disorders*, *1*(2), 172–181. https://doi.org/10.1007/s11689-009-9023-x

Grent-’t-Jong, T., Gross, J., Goense, J., Wibral, M., Gajwani, R., Gumley, A. I., Lawrie, S. M., Schwannauer, M., Schultze-Lutter, F., Navarro Schröder, T., Koethe, D., Leweke, F. M., Singer, W., & Uhlhaas, P. J. (2018). Resting-state gamma-band power alterations in schizophrenia reveal E/I-balance abnormalities across illness-stages. *eLife*, *7*, e37799. https://doi.org/10.7554/eLife.37799

Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241. https://doi.org/10.1093/brain/awt257

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*(1), 14218. https://doi.org/10.1038/ncomms14218

JASP (0.14). (2020). [Computer software]. JASP Team. https://jasp-stats.org/

Joyce, E. M., & Roiser, J. P. (2007). Cognitive heterogeneity in schizophrenia: *Current Opinion in Psychiatry*, *20*(3), 268–272. https://doi.org/10.1097/YCO.0b013e3280ba4975

Kana, R. K., Keller, T. A., Minshew, N. J., & Just, M. A. (2007). Inhibitory Control in High-Functioning Autism: Decreased Activation and Underconnectivity in Inhibition Networks. *Biological Psychiatry*, *62*(3), 198–206. https://doi.org/10.1016/j.biopsych.2006.08.004

Karaminis, T., Cicchini, G. M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D., & Pellicano, E. (2016). Central tendency effects in time interval reproduction in autism. *Scientific Reports*, *6*(1), 28570. https://doi.org/10.1038/srep28570

Karvelis, P., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *eLife*, *7*, e34115. https://doi.org/10.7554/eLife.34115

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, *20*(4), 442–462. https://doi.org/10.1177/1362361315588200

Kim, H., Keifer, C., Rodriguez-Seijas, C., Eaton, N., Lerner, M., & Gadow, K. (2019). Quantifying the Optimal Structure of the Autism Phenotype: A Comprehensive Comparison of Dimensional, Categorical, and Hybrid Models. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(9), 876-886.e2. https://doi.org/10.1016/j.jaac.2018.09.431

Kim, Y., Kim, T.-H., & Ergün, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, *13*, 243–257. https://doi.org/10.1016/j.frl.2014.12.005

King, B. H., & Lord, C. (2011). Is schizophrenia on the autism spectrum? *Brain Research*, *1380*, 34–41. https://doi.org/10.1016/j.brainres.2010.11.031

King, D. J., Hodgekins, J., Chouinard, P. A., Chouinard, V.-A., & Sperandio, I. (2017). A review of abnormalities in the perception of visual illusions in schizophrenia. *Psychonomic Bulletin & Review*, *24*(3), 734–751. https://doi.org/10.3758/s13423-016-1168-5

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Król, M., & Król, M. (2019). The world as we know it and the world as it is: Eye-movement patterns reveal decreased use of prior knowledge in individuals with autism. *Autism Research*, *12*(9), 1386–1398. https://doi.org/10.1002/aur.2133

Kuhn, R., & Cahn, C. H. (2004). Eugen Bleuler's Concepts of Psychopathology. *History of Psychiatry*, *15*(3), 361–366. https://doi.org/10.1177/0957154X04044603

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00302

Li, J., Wang, Z. J., Palmer, S. J., & McKeown, M. J. (2008). Dynamic Bayesian network modeling of fMRI: A comparison of group-analysis methods. *NeuroImage*, *41*(2), 398–407. https://doi.org/10.1016/j.neuroimage.2008.01.068

Lu, H., Yi, L., & Zhang, H. (2019). Autistic traits influence the strategic diversity of information sampling: Insights from two-stage decision models. *PLOS Computational Biology*, *15*(12), e1006964. https://doi.org/10.1371/journal.pcbi.1006964

Manning, C., Kilner, J., Neil, L., Karaminis, T., & Pellicano, E. (2017). Children on the autism spectrum update their behaviour in response to a volatile environment. *Developmental Science*, *20*(5), e12435. https://doi.org/10.1111/desc.12435

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neuroscience Bulletin*, *33*(2), 183–193. https://doi.org/10.1007/s12264-017-0100-y

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

Nelson, M. T., Seal, M. L., Pantelis, C., & Phillips, L. J. (2013). Evidence of a dimensional relationship between schizotypy and schizophrenia: A systematic review. *Neuroscience & Biobehavioral Reviews*, *37*(3), 317–327. https://doi.org/10.1016/j.neubiorev.2013.01.004

Noel, J.-P., Stevenson, R. A., & Wallace, M. T. (2018). Atypical audiovisual temporal function in autism and schizophrenia: Similar phenotype, different cause. *European Journal of Neuroscience*, *47*(10), 1230–1241. https://doi.org/10.1111/ejn.13911

O'Donnell, P. (2011). Adolescent Onset of Cortical Disinhibition in Schizophrenia: Insights From Animal Models. *Schizophrenia Bulletin*, *37*(3), 484–492. https://doi.org/10.1093/schbul/sbr028

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, *143*(5), 521–542. https://doi.org/10.1037/bul0000097

Patterson, P. H. (2009). Immune involvement in schizophrenia and autism: Etiology, pathology and animal models. *Behavioural Brain Research*, 9.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Peters, E., Joseph, S., Day, S., & Qarety, P. (2004). Measuring Delusional Ideation: The 21-Item Peters et aL Delusions Inventory (PDI). *Schizophrenia Bulletin*, *30*(4), 18.

Pinkham, A. E., Hopfinger, J. B., Pelphrey, K. A., Piven, J., & Penn, D. L. (2008). Neural bases for impaired social cognition in schizophrenia and autism spectrum disorders. *Schizophrenia Research*, *99*(1–3), 164–175. https://doi.org/10.1016/j.schres.2007.10.024

Powell, G., Meredith, Z., McMillin, R., & Freeman, T. C. A. (2016). Bayesian Models of Individual Differences: Combining Autistic Traits and Sensory Thresholds to Predict Motion Perception. *Psychological Science*, *27*(12), 1562–1572. https://doi.org/10.1177/0956797616665351

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Rapoport, J., Chavez, A., Greenstein, D., Addington, A., & Gogtay, N. (2009). Autism Spectrum Disorders and Childhood-Onset Schizophrenia: Clinical and Biological Contributions to a Relation Revisited. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(1), 10–18. https://doi.org/10.1097/CHI.0b013e31818b1c63

Robinson, E. B., Munir, K., Munafò, M. R., Hughes, M., McCormick, M. C., & Koenen, K. C. (2011). Stability of Autistic Traits in the General Population: Further Evidence for a Continuum of Impairment. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(4), 376–384. https://doi.org/10.1016/j.jaac.2011.01.005

Ross, R. M., McKay, R., Coltheart, M., & Langdon, R. (2015). Jumping to Conclusions About the Beads Task? A Meta-analysis of Delusional Ideation and Data-Gathering. *Schizophrenia Bulletin*, *41*(5), 1183–1191. https://doi.org/10.1093/schbul/sbu187

Rubenstein, J. L. R., & Merzenich, M. M. (2003). Model of autism: Increased ratio of excitation/inhibition in key neural systems. *Genes, Brain and Behavior*, *2*(5), 255–267. https://doi.org/10.1034/j.1601-183X.2003.00037.x

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). 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. *Molecular Autism*, *6*(1), 2. https://doi.org/10.1186/2040-2392-6-2

Simonsen, A., Fusaroli, R., Petersen, M. L., Vermillet, A.-Q., Bliksted, V., Mors, O., Roepstorff, A., & Campbell-Meiklejohn, D. (2021). Taking others into account: Combining directly experienced and indirect information in schizophrenia. *Brain*. https://doi.org/10.1093/brain/awab065

Speechley, W. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience*, *35*(1), 7–17. https://doi.org/10.1503/jpn.090025

SPM12 Toolbox—Statistical Parametric Mapping (Version SPM12). (2020). [Computer software]. Wellcome Centre for Human Neuroimaging. https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). *Bayesian model selection for group studies*. 14.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Valton, V., Romaniuk, L., Douglas Steele, J., Lawrie, S., & Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews*, *83*, 631–646. https://doi.org/10.1016/j.neubiorev.2017.08.022

Wang, Y., Zhang, Y., Liu, L., Cui, J., Wang, J., Shum, D. H. K., van Amelsvoort, T., & Chan, R. C. K. (2017). A Meta-Analysis of Working Memory Impairments in Autism Spectrum Disorders. *Neuropsychology Review*, *27*(1), 46–61. https://doi.org/10.1007/s11065-016-9336-y

Wiggins, L. D., Robins, D. L., Adamson, L. B., Bakeman, R., & Henrich, C. C. (2012). Support for a Dimensional View of Autism Spectrum Disorders in Toddlers. *Journal of Autism and Developmental Disorders*, *42*(2), 191–200. https://doi.org/10.1007/s10803-011-1230-0

Young, R. L., & Rodi, M. L. (2014). Redefining Autism Spectrum Disorder Using DSM-5: The Implications of the Proposed DSM-5 Criteria for Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *44*(4), 758–765. https://doi.org/10.1007/s10803-013-1927-3

Chapter 4 Supplementary Information

4.A Data collection and processing

4.A.1 Experiment details and pre-screening

The platform Prolific was chosen for its higher data quality compared to the alternatives (Peer et al., 2017). Both Prolific and social media participants filled in the questionnaires before they participated in the behavioural task. We included two attention checks in the AQ questionnaire (i.e., questions that straightforwardly asked for a specific response) to serve as validity checks for data quality. The task was implemented on the PsychoPy 2020.1.3 Builder, automatically translated to PsychoJS, and hosted on Pavlovia (Peirce & MacAskill, 2018), as this method achieves high temporal accuracy (Bridges et al., 2020).

Prolific asks its participants various questions, which we used to pre-screen our participants. We required all participants to have answered positively to the question ‘Do you have normal or corrected-to-normal vision? (i.e., You can see colour normally, and if you need glasses, you are wearing them or contact lenses)’ and negatively to the question ‘Are you currently taking any medication to treat symptoms of depression, anxiety or low-mood (e.g., SSRIs)?’.

Moreover, we used the question ‘Have you received a formal clinical diagnosis of autistic spectrum disorder, made by a psychiatrist, psychologist, or other qualified medical specialist? This includes Asperger's syndrome, Autistic Disorder, High Functioning Autism or Pervasive Developmental Disorder’ to ensure that we would get a broad enough range for autistic traits. The question had the following possible answers:

1. Yes – as a child
2. Yes – as an adult
3. I am in the process of receiving a diagnosis
4. No – but I identify as being on the autistic spectrum
5. No
6. Don't know / rather not say

We recruited half of our Prolific participants from those that had chosen options 1-4, and half from those that had chosen options 5 or 6. All participants had taken part in at least 3 more studies, with an approval rate of at least 98% (meaning that their submissions in previous studies have been approved 98% of the time).

4.A.2 Trial set

Trials where priors and likelihoods were both very high or very low were not included in our trial set, as participants would presumably respond to the extremes of the scale, not providing much information about the integration of priors and likelihoods. Instead, we focused on trials with stimuli closer to a probability of 0.5 (equal basket sizes or fish proportions), so that prior and likelihood overcounting would be more apparent. Trials with stimuli equal to 0.5 were avoided, as, independently of parameter values, stimuli equal to 0.5 do not influence the model estimates. The resulting trial set is presented in Table 4.A1.

Table 4.A1 Counts of prior-likelihood combinations in the trial set.

		Likelihood values								
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Prior values	**0.1**	–	–	2	2	1	2	2	1	1
	0.2	–	1	2	2	1	2	2	1	1
	0.3	2	2	2	2	2	2	2	2	2
	0.4	2	2	2	2	2	2	2	2	2
	0.5	1	1	2	2	–	2	2	1	1
	0.6	2	2	2	2	2	2	2	2	2
	0.7	2	2	2	2	2	2	2	2	2
	0.8	1	1	2	2	1	2	2	1	–
	0.9	1	1	2	2	1	2	2	–	–

4.A.3 Data cleaning

Crowdsourcing data for behavioural experiments online carries many risks in terms of data quality. Participants might misunderstand the instructions, be prone to external distractions, view the experiment only as a source of income so that they aim to finish it as fast as possible at the expense of quality, or even use automated programs (bots) to complete it (Chmielewski & Kucker, 2020). To filter out these risks, various measures of data quality have to be implemented. In this experiment, as detailed below, we used four criteria to help us distinguish between high- and low-quality responses: the attention checks, the responses to the more 'certain' trials of the task, the distributions of task responses, and the average time per response. All of them were chosen and implemented before the data were analysed in any other way.

Specifically, for the attention checks, all participants who failed at least one were immediately disqualified (n = 9). 'Certain' trials are those with very high (> 0.94) or very low (< 0.06) Bayesian posteriors, i.e., those with a (*prior, likelihood*) combination of (0.7, 0.9), (0.9, 0.7), (0.8, 0.8), (0.3, 0.1), (0.1, 0.3), and (0.2, 0.2). There was a total of 16 such trials, including the last 6 training ones. We discarded all participants who clicked on the opposite side of the scale relative to the posterior in more than 3 out of these 16 trials (n = 3), as they could not have been using the information provided to make a choice. Then, we disqualified the participants who mostly clicked outside of the scale (n = 1), and those who had clustered responses (n = 9), as they seemed to have discretized the scale instead of treating it as a continuous measure of confidence (Figure 4.A1). Finally, we discarded data from participants who had an average response time of more than 5*s* (n = 6), because it might signify alternative response strategies compared to the other participants, given that we had instructed them to answer 'as fast and precise as possible'. Data were discarded slightly more often from the ASD population (Table 4.A2). The final sample showed good reliability for the total AQ (*Cronbach's a* = 0.76) and slightly low reliability for the PDI Y/N (*Cronbach's a* = 0.67).

Table 4.A2 Number of discarded participants for each Prolific ASD category.

ASD category	Total sample	Discarded	Proportion
1	19	3	0.16
2	7	2	0.29
3	5	1	0.2
4	39	5	0.13
5	68	7	0.1
6	2	0	0
– (social media)	61	7	0.11

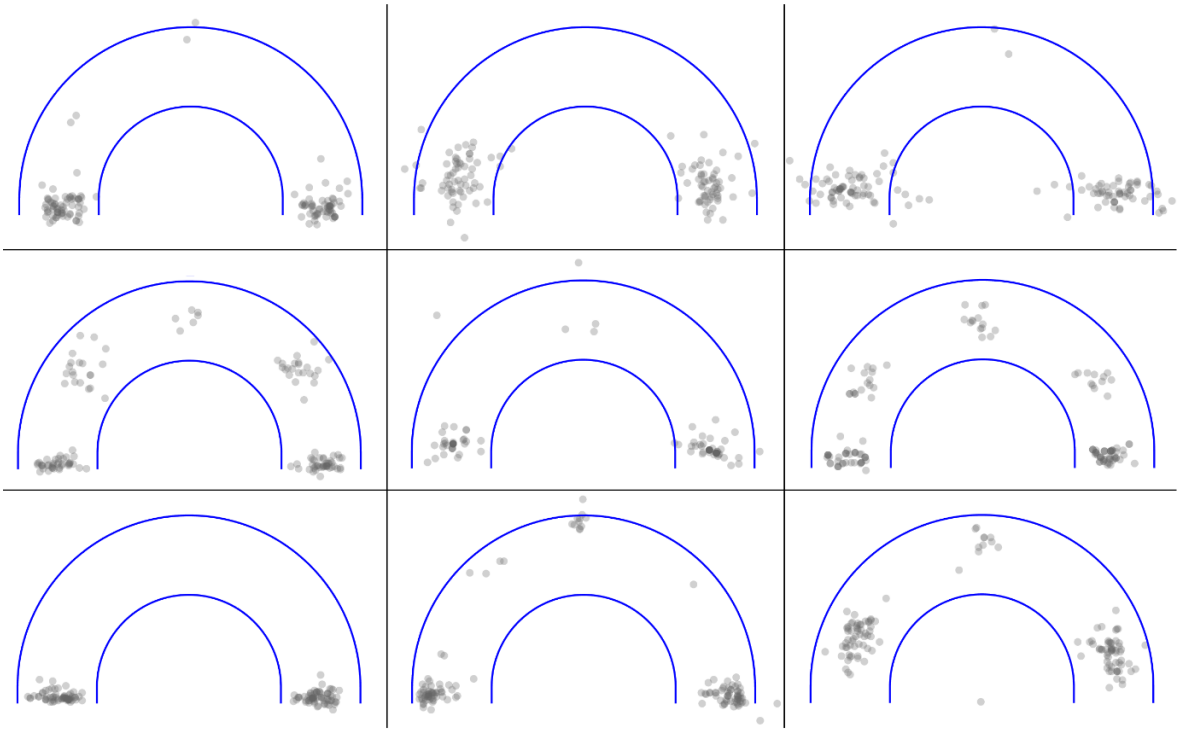

Figure 4.A1 The 9 datasets that got discarded because of response clustering. *Blue lines represent the edges of the response scale and grey dots the locations of the mouse clicks by the participants. The presented participants seem to have misinterpreted the instructions of the task, answering in a discrete instead of a continuous way, in 2, 3, or 5 distinct places.*

4.A.4 Statistical Power

Jardri et al. found a Pearson's correlation of 0.59 between sensory evidence reverberation and non-clinical delusionary beliefs in their whole sample, as well as a correlation of 0.45 between sensory evidence reverberation and psychotic symptoms in their SCZ subsample. With a significance level of $a = 0.05$ and our sample size $n = 176$, a Kendall's correlation coefficient of only $\tau = 0.2$ yields a statistical power of 98%, as calculated based on hypothesis testing (May & Looney, 2020). It is important to note that all power calculations for Kendall's τ assume bivariate normality, which is not verified in our dataset. Therefore, this value should be understood only as a weak indication of the actual power of our correlation tests. Similarly, the statistical power for the group comparisons cannot be calculated without any assumptions for the underlying distribution. However, group sizes in the present study are comparable to those used by Jardri et al., who showed very clear effects.

4.B Modelling details

4.B.1 Linear mixed-effects models

The linear mixed-effects models (LMEs) that were used in the current study are shown below in Wilkinson-Rogers notation. `absCnf` stands for absolute confidence, `absLl` for absolute likelihood, `prCng` for prior congruency, `RT` for the reaction times, and `ID` for the participant ID. For an analysis of the role of each model component, please see the corresponding Methods and Materials section in the main text of Chapter 4.

LME_core: `absCnf ~ absLl*prCng + RT + (1|ID)`

LME_AQ: `absCnf ~ absLl*prCng*AQ + RT + (1|ID)`

LME_PDI: `absCnf ~ absLl*prCng*PDI + RT + (1|ID)`

LME_full: `absCnf ~ absLl*prCng*AQ + absLl*prCng*PDI + RT + (1|ID)`

LME_rtInteract: `absCnf ~ absLl*prCng*AQ + absLl*prCng*PDI + RT*AQ + RT*PDI + (1|ID)`

The best BIC score was achieved by LME_core (Figure 4.B1), showing that AQ and PDI scores offered relatively little information. Despite that, the full results for all models can be seen in Table 4.B1.

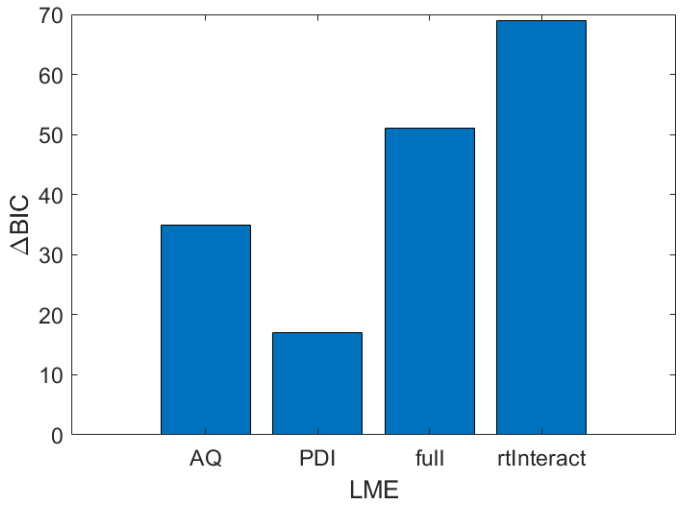

Figure 4.B1 Results of model comparisons. *ΔBIC is the difference between the BIC score of each model and that of LME_core. Lower BIC scores are better, with differences of 20 or more being considered very strong evidence (Raftery, 1995).*

Table 4.B1 Results of linear mixed-effects models.

Model	Component	t	p	Component	t	p
LME_core	absLl	44.50	$<10^{-323}$	absLl:prCng	25.20	10^{-138}
	prCng	24.63	10^{-132}	RT	–17.01	10^{-64}
LME_AQ	absLl	11.04	10^{-28}	AQ	–1.72	0.09
	prCng	8.09	10^{-16}	absLl:AQ	1.23	0.22
	absLl:prCng	5.46	10^{-8}	prCng:AQ	–1.39	0.16
	RT	–17.03	10^{-64}	absLl:prCng:AQ	1.52	0.13
LME_PDI	absLl	18.19	10^{-73}	PDI	2.08	0.04
	prCng	10.83	10^{-27}	absLl:PDI	2.31	0.02
	absLl:prCng	12.84	10^{-37}	prCng:PDI	0.40	0.69
	RT	–17.00	10^{-64}	absLl:prCng:PDI	–1.55	0.12
LME_full	absLl	9.03	10^{-19}	absLl:AQ	1.08	0.28
	prCng	7.10	10^{-12}	prCng:AQ	–1.42	0.15
	absLl:prCng	5.65	10^{-8}	absLl:prCng:AQ	1.63	0.10
	RT	–17.01	10^{-64}	absLl:PDI	2.23	0.03
	AQ	–1.91	0.06	prCng:PDI	0.50	0.62
	PDI	2.23	0.03	absLl:prCng:PDI	–1.66	0.1
LME_rtInteract	absLl	8.90	10^{-18}	PDI	1.88	0.06
	prCng	7.06	10^{-12}	absLl:PDI	2.32	0.02
	absLl:prCng	5.55	10^{-8}	prCng:PDI	0.55	0.58
	AQ	–2.03	0.04	absLl:prCng:PDI	–1.58	0.11
	absLl:AQ	1.14	0.25	RT	–5.08	10^{-7}
	prCng:AQ	–1.40	0.16	RT:AQ	0.72	0.47
	absLl:prCng:AQ	1.68	0.09	RT:PDI	1.29	0.20

Model components with $p < 0.05$ are shaded.

4.B.2 Bayesian models

All the following procedures were identical to those reported by Jardri et al. (Jardri et al., 2017). Participant and model confidence estimates were restricted to the range [0.01, 0.99], to avoid numerical issues in both the reporting of participant responses and the model fitting. Moreover, trials where participants did not click on or very close to the scale were not included in our analysis (see the code at https://osf.io/yqug2/ for the exact criterion).

During model fitting, we applied a small L2 regularization penalty on the reverberation parameters (a) to prevent 'degenerate' solutions with weights close to 0 and artificially large a. This penalty was equal to $0.00005(a_p^2 + a_s^2)$. Model predictions were almost completely insensitive to this added cost. Due to the regularization, reverberation parameters rarely exceeded a value of 0.5, as the minuscule benefit to the model predictions (Figure 4.B2) was outweighed by the L2 penalty. Model fitting was carried out by minimizing mean squared error, which is equivalent to least squares. This was chosen so that the L2 term would not change depending on the number of trials included for each participant.

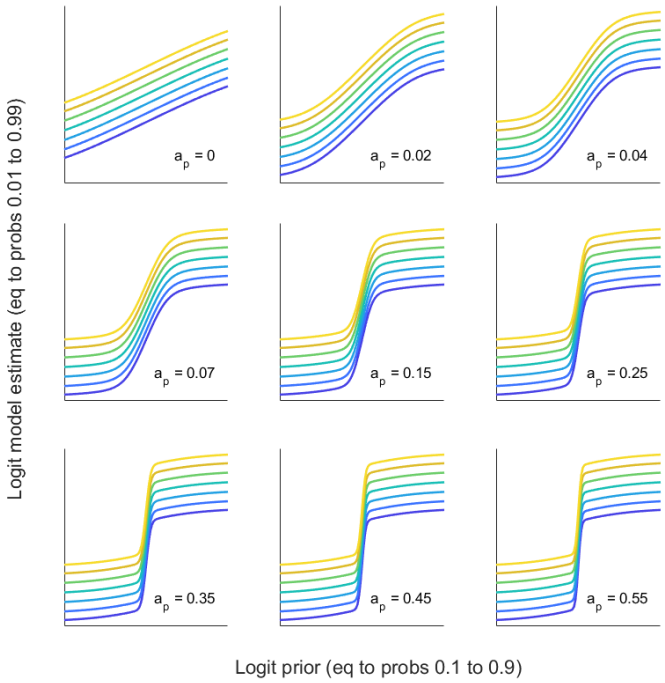

Figure 4.B2 CINI predictions with varying prior reverberation. *Coloured lines correspond to different likelihood values (probabilities 0.1 to 0.9). Changes from $a_p = 0$ (top left) to $a_p = 0.15$ (middle) greatly affect observed model predictions, while those from $a_p = 0.15$ to $a_p = 0.55$ (bottom right) have almost no perceptible effect, despite the larger difference between parameter values.*

4.B.3 Parameter Recovery

First, for each selected model, we created 1000 simulated participants by drawing values randomly from the set of estimated parameters and adding a small uniform noise term drawn from [–0.025, 0.025]. Each simulated subject was also assigned an error variance, calculated from a random participant's mean squared error around the model estimate, L_c. Then, we generated 130 trial responses for each simulated subject using a Gaussian distribution with the assigned variance, centred on the model response. Finally, both models were fitted on the simulated responses, and the recovered parameters were compared with the original ones using Pearson's product-moment correlation coefficient. We also tested for correlated parameters, by calculating the Kendall rank correlation between the recovered parameter values. Model recovery was performed by fitting the simulated responses of all participants with both models and calculating the confusion matrix.

Most parameters were recovered close to their original values for both CINI (Figure 4.B3) and CII (Figure 4.B4). However, reverberation parameters were much more likely to be badly recovered if they had values greater than 0.15, leading to a clustering pattern (top-left plot of Figures 4.B3 & 4.B4). In this range, the data can be roughly approximated by treating prior or likelihood information as binary (e.g., 'left basket larger' vs 'right basket larger'), with very little dependency on the exact size of the baskets or the exact fish ratios (Figure 4.B2). However, this was a minority of the estimated values.

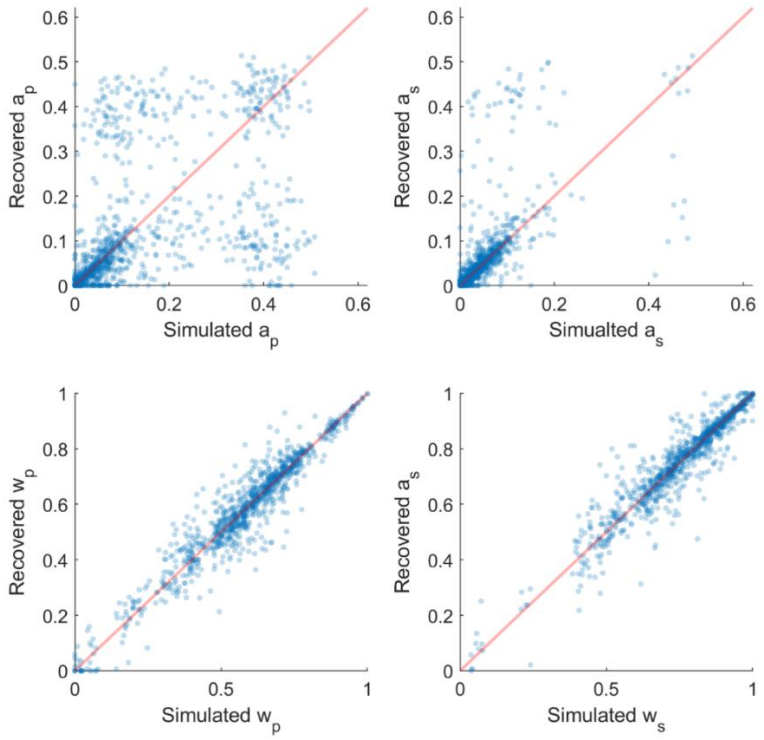

Figure 4.B3 Recovered vs simulated parameters of CINI model. *Each dot represents one simulated participant. The red line represents perfect recovery.*

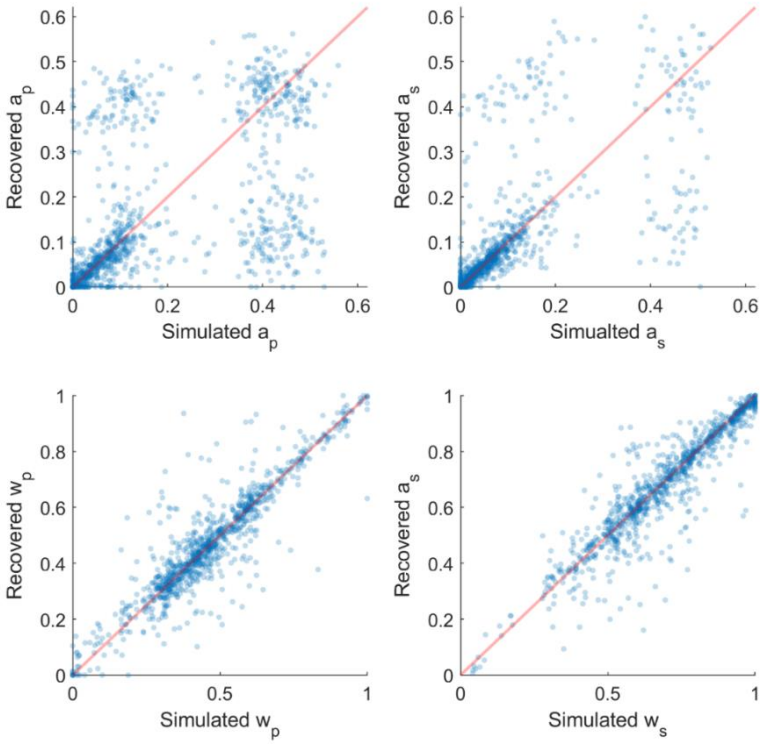

Figure 4.B4 Recovered vs simulated parameters of CII model. *Each dot represents one simulated participant. The red line represents perfect recovery.*

The clustering pattern can also be observed when parameter recovery used the set of 200 trials of Jardri et al. (Jardri et al., 2017) (Figure 4.B5), with minimal differences in the recovery correlations between the two sets (Table 4.B2).

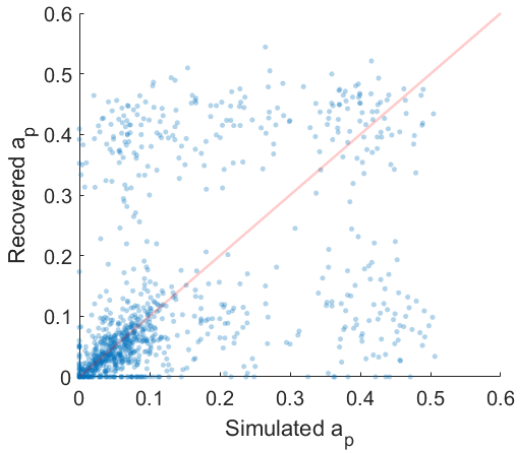

Figure 4.B5 Parameter recovery results for the CINI prior reverberation over 200 trials. *The red line corresponds to perfect recovery.*

Table 4.B2 Pearson correlations between simulated and recovered parameters in the trial set of Jardri et al. and the current study.

	CINI		CII	
Trials	130	200	130	200
a_p	0.54	0.46	0.54	0.51
a_s	0.58	0.54	0.71	0.60
w_p	0.96	0.95	0.94	0.92
w_s	0.91	0.94	0.93	0.93

Minimal correlations were observed between the recovered parameters of both the CINI (Table 4.B3) and CII model (Table 4.B4). The same pattern of minimal correlations was also observed between the simulated and recovered values across parameters (all $|\tau| < 0.1$).

Table 4.B3 Kendall rank correlations between recovered CINI parameters.

a_p	1			
a_s	–0.012	1		
w_p	–0.008	–0.022	1	
w_s	0.016	–0.043	–0.041	1
	a_p	a_s	w_p	w_s

Table 4.B4 Kendall rank correlations between recovered CII parameters.

a_p	1			
a_s	–0.019	1		
w_p	0.040	–0.037	1	
w_s	0.007	–0.020	–0.083	1
	a_p	a_s	w_p	w_s

4.C Visualizations

4.C.1 Model parameters and correlations with AQ

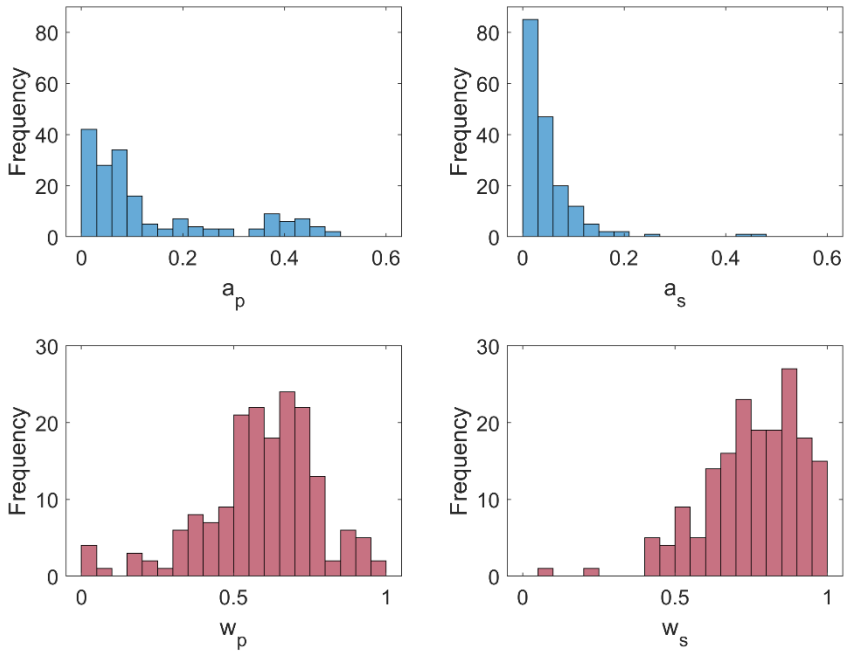

Figure 4.C1 Histogram of CINI parameter values estimated during model fitting. *Notice that axis limits differ between rows.*

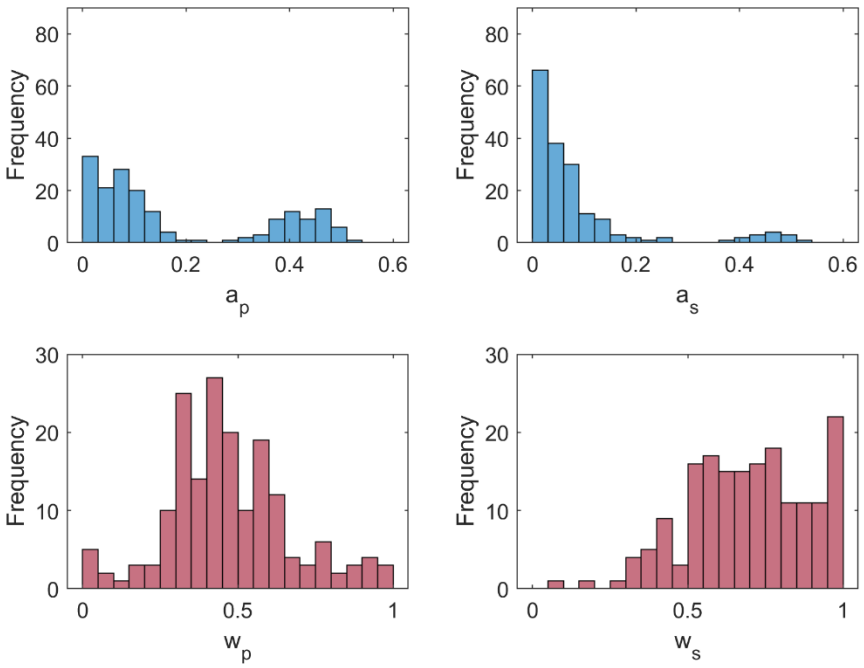

Figure 4.C2 Histogram of CII parameter values estimated during model fitting. *Notice that axis limits differ between rows.*

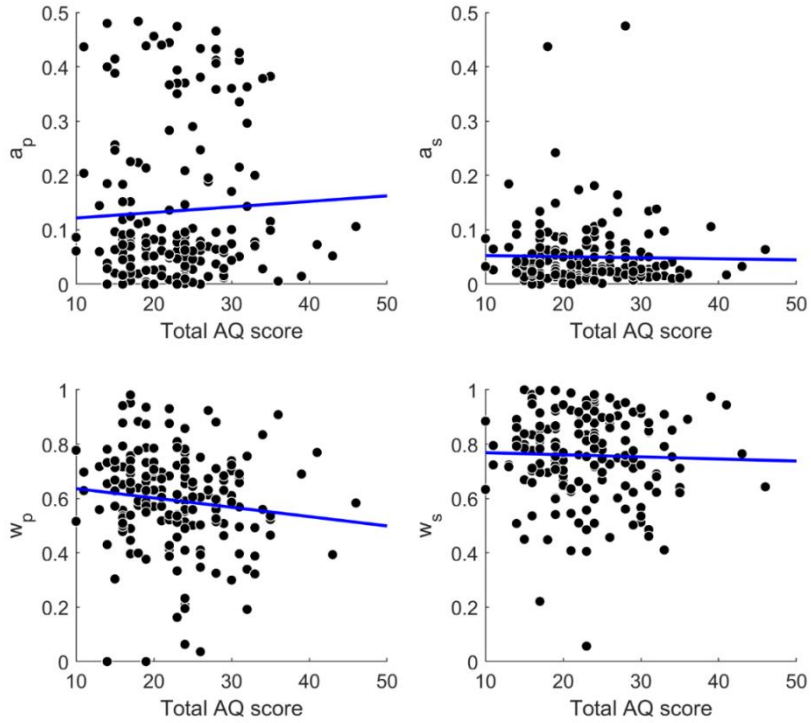

Figure 4.C3 CINI model fit parameters vs participant AQ. *The blue line represents the least squares fit. All of the presented correlations had an uncorrected p-value above 0.05, besides w_p (τ = –0.12, p = 0.02), that however would not survive correction for multiple comparisons.*

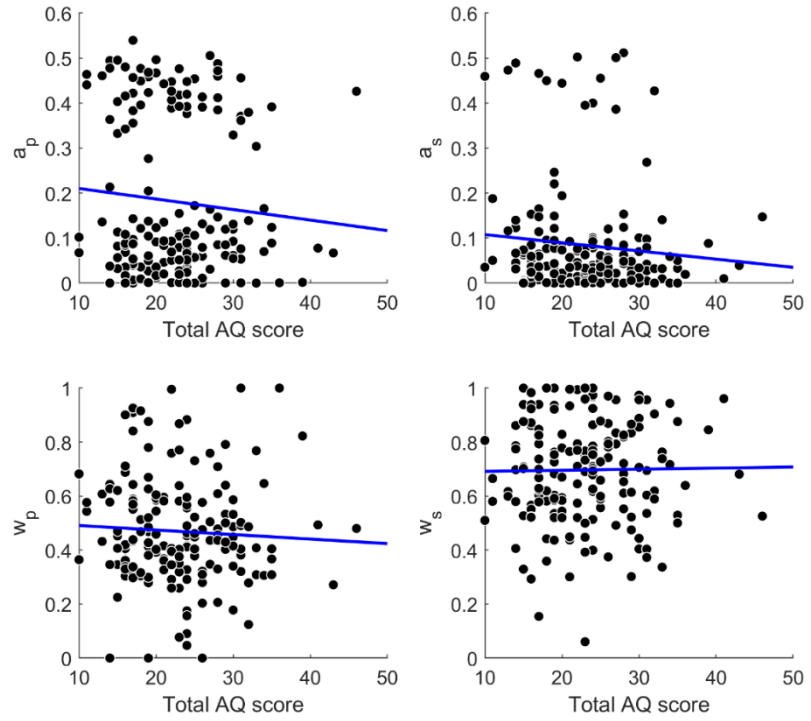

Figure 4.C4 CII model fit parameters vs participant AQ. *The blue line represents the least squares fit. All of the presented correlations had an uncorrected p-value above 0.05.*

4.D Additional tests

4.D.1 ASD vs ND with low AQ

To avoid comparing autistic participants to participants with a potential undiagnosed autism spectrum disorder, we performed an additional comparison with a limited ND group, which contained only participants with AQ scores lower than or equal to the reported mean value of the general population (Ruzich et al., 2015) (AQ ≤ 17, n = 21). The results showed no differences between the groups (Table 4.D1).

Table 4.D1 Mann-Whitney U test results between the CINI model fit parameters of the ASD and limited-ND participant groups.

	limited-ND vs ASD		
CINI params	f	p	BF_{01}
$\boldsymbol{a_p}$	0.51	0.90	3.26
$\boldsymbol{a_s}$	0.44	0.50	2.52
$\boldsymbol{w_p}$	0.46	0.63	2.88
$\boldsymbol{w_s}$	0.57	0.44	2.61

f signifies the common language effect size, with larger f values corresponding to larger parameter values for the ASD group, relative to the limited-ND. p-values are uncorrected for multiple comparisons.

4.D.2 Circular Inference – Interference

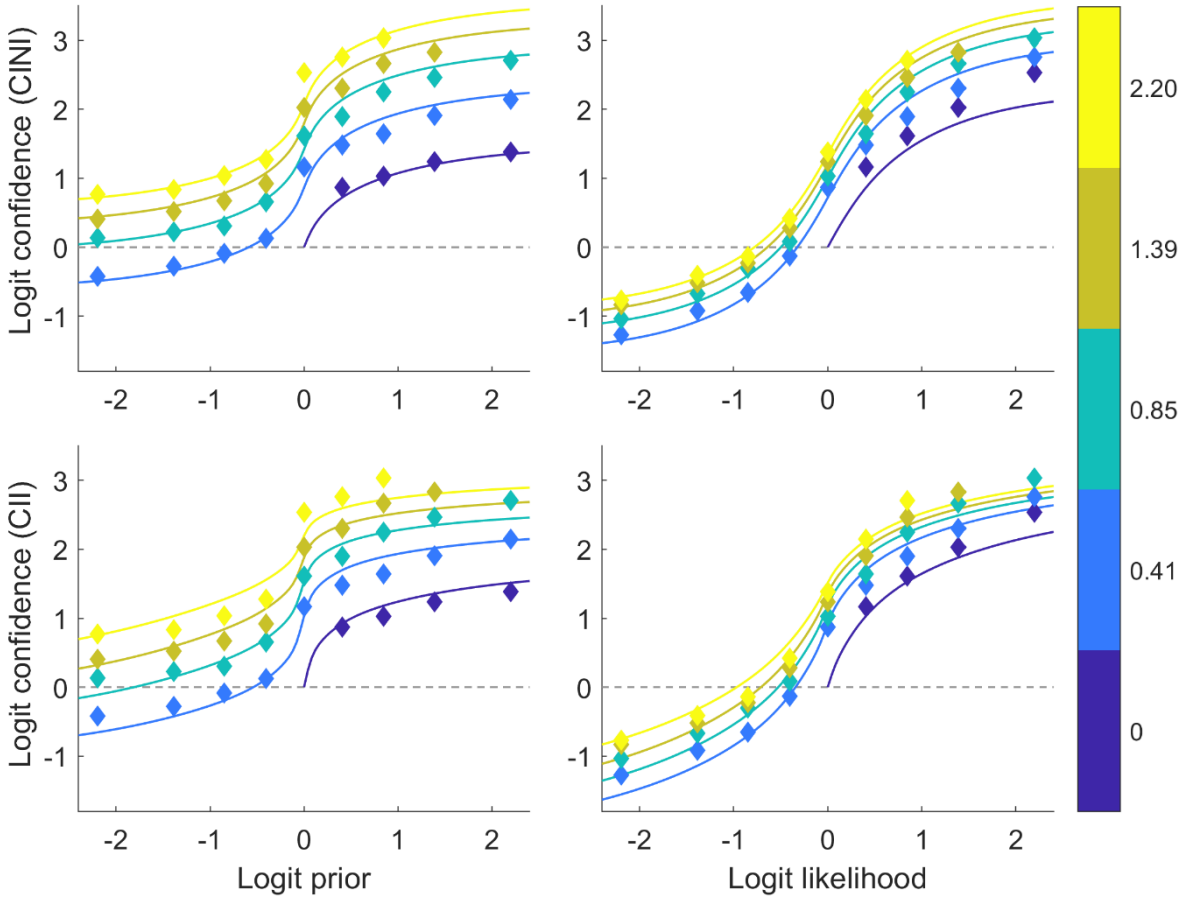

Figure 4.D1 CINI (top) and CII (bottom) model fit vs participant logit confidence estimates. *Model and participant logit confidence as function of logit likelihoods and priors. Coloured lines represent model predictions and rhombuses the subject confidence estimates. Different colours represent logit likelihood values in the left graph and logit prior values in the right and are equivalent to probabilities 0.5 to 0.9. Since both the task and the two model structures are symmetrical around 0 logit confidence (0.5 probability), participant estimates have been averaged between symmetric trials to reduce noise (e.g., a trial with a logit prior of –1 and a logit likelihood of 2 would have been symmetrical to one with a logit prior of 1 and a logit likelihood of –2).*

Kendall correlations showed no association between AQ and any CII parameters (Table 4.D2). Moreover, no differences were found in model parameters between the ND and ASD groups (Table 4.D3, Figure 4.D4). Pearson correlations between CII and CINI parameters were high (a_p, r = 0.75; a_s, r = 0.76; w_p, r = 0.82; w_s, r = 0.83).

Table 4.D2 Kendall rank correlations between CII parameters and psychiatric traits.

	AQ			PDI		
CII params	τ	p	BF_{01}	τ	p	BF_{01}
a_p	–0.05	0.38	6.84	–0.05	0.34	6.19
a_s	–0.08	0.14	3.23	0.00	0.94	10.1
w_p	–0.07	0.18	3.95	0.08	0.14	3.19
w_s	0.01	0.81	9.84	0.09	0.09	2.16

Total AQ scores and Y/N PDI scores were used for the correlations. τ signifies the correlation coefficient. p-values are presented without any correction for multiple comparisons.

Table 4.D3 Mann-Whitney U test results between the CII parameters of the ASD and ND groups and the low-AQ and high-AQ groups.

	ND vs ASD			high-AQ vs low-AQ		
CII params	f	p	BF_{01}	f	p	BF_{01}
a_p	0.55	0.50	3.82	0.44	0.40	2.37
a_s	0.45	0.46	3.15	0.40	0.18	1.81
w_p	0.47	0.64	3.85	0.43	0.36	3.14
w_s	0.53	0.71	3.71	0.49	0.93	3.77

Total AQ scores were used for the comparisons. f signifies the common language effect size, with larger f values corresponding to larger parameter values for the ASD and the high-AQ groups, relative to the others. p-values are uncorrected for multiple comparisons.

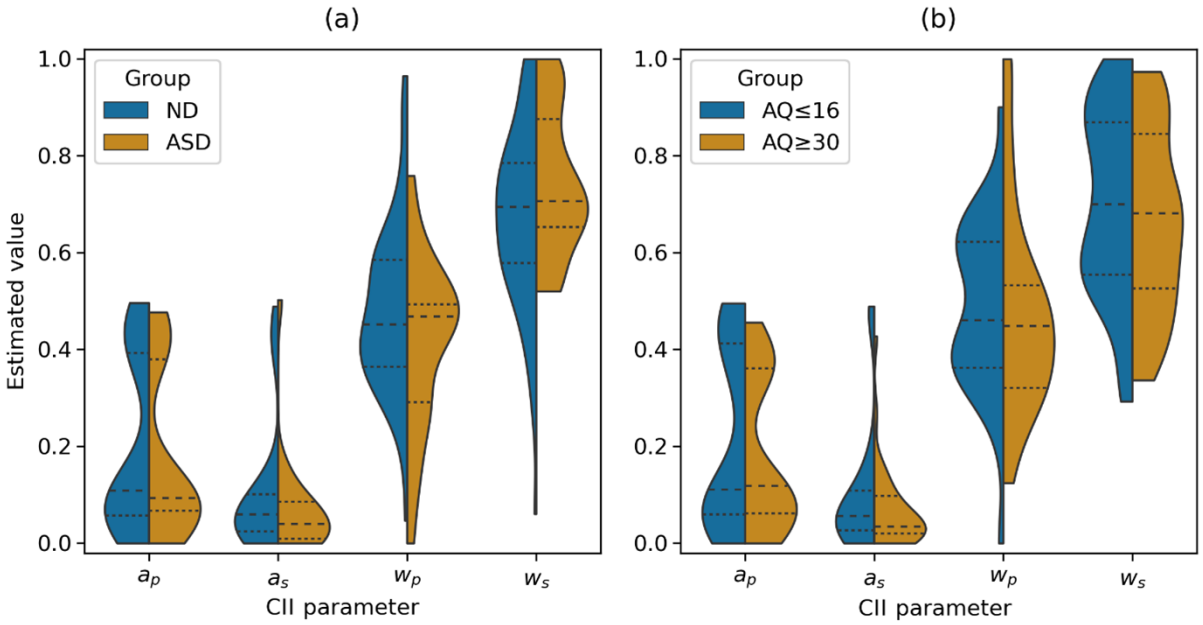

Figure 4.D4 CII parameter values of ND vs ASD groups (a) and low-AQ vs high-AQ groups (b). *Violin plots show the density of estimated parameters over the possible values, relative to the group size. Dashed lines represent the median, while dotted ones represent the top and bottom quartiles in each group. No differences are observed between groups.*

4.D.3 AQ – Likert scoring

Scoring the AQ questionnaire on a Likert scale instead of the usual binary one (scores of 1-4 for each question, instead of 0 and 1), increased the questionnaire's reliability (*Cronbach's a* = 0.81), but did not affect the observed relationships with CINI model fit parameters. The Kendall correlation with the largest magnitude and the smallest corresponding *p*-value remained the one with the prior weights (τ = –0.12, p = 0.02). As with the correlations reported in the main text, this relationship is non-significant when corrected for multiple comparisons. Moreover, once again, no comparisons were significant between the CINI model fit parameters of the new low- (Likert AQ ≤ 105, N = 28) and the new high- (Likert AQ ≥ 133, N = 29) AQ participant groups (e.g., w_p, uncorrected p = 0.08).

4.D.4 Alternative interpretations of basket sizes

Since the basket information is less quantitative than the fish ratios, we investigated whether it is interpreted by the participants as intended. For that, we tested four additional models that were based on CINI but accounted for different mappings between basket sizes and probabilities. The first model simply included a linear rescaling of the intended probabilities:

$$q_i = k(p_i - 0.5) + 0.5, \qquad i \in \{l, r\}, \qquad k \in (0, 1)\,, \tag{4.D1}$$

with *p* corresponding to the intended probabilities, *q* to the perceived ones, *i* to the left or the right basket, and *k* to the rescaling factor. The second model allowed for the exponential rescaling of the probabilities:

$$q_i = \frac{p_i^k}{p_l^k + p_r^k}, \qquad i \in \{l, r\}, \qquad k \in (0, +\infty)\,, \tag{4.D2}$$

The third was based on the Weber-Frechner law, under which the perceived sizes are proportional to the logarithm of the actual sizes (Frechner, 1860):

$$q_i = \frac{\ln\left(\frac{p_i}{p_0}\right)}{\ln\left(\frac{p_l}{p_0}\right) + \ln\left(\frac{p_r}{p_0}\right)}\,, \qquad i \in \{l, r\}, \qquad p_0 \in (0, 0.1)\,, \tag{4.D3}$$

where *p*0 is the probability corresponding to the just noticeable basket size, assuming that the participants can see all the presented baskets. Finally, the fourth model treated basket sizes as providing only binary information (‘left basket larger’ vs ‘right basket larger’), which nudged the fish frequency-based estimates by a constant amount. Then the logit confidence becomes:

$$L_c = F(L_s, w_s) \pm A, \qquad A \in (-\infty, +\infty)\,, \tag{4.D4}$$

with A being the constant amount, positive when the left basket was larger than the right, negative in the opposite case, and equal to 0 when sizes were equal with each other.

The first three models had 5 parameters each, while the last one had only 3. We compared these models with the original CINI. Both fixed and random effects pairwise comparisons showed clear superiority for CINI, with group ΔBICs > 715 and posterior model probabilities for CINI > 0.72. This indicates that participants interpreted the basket sizes in the intended way of probabilities being analogous to basket size.

4.E Comparisons between Jardri et al.'s and this study's datasets

To verify that the change of trial set did not have a strong influence on the parameter estimation, we compared the parameter values estimated from the original Jardri et al. dataset (Jardri et al., 2017), with a subset of that dataset, restricted on the trials of the present study. Pearson's correlations showed minimal changes between parameters estimated from 200 and 130 trials (Table 4.E1).

Table 4.E1 Pearson's correlations between the parameters estimated from 200 and 130 trials of the Jardri et al. dataset.

CINI params	R	CII params	R
a_p	0.86	a_p	0.82
a_s	0.99	a_s	0.97
w_p	0.98	w_p	0.91
w_s	0.97	w_s	0.91

Moreover, we compared the parameter values estimated from the trial subset of the Jardri et al. study to our study's parameter values (Table 4.E2). Interestingly, Mann-Whitney U tests showed increased weights in our sample (w_p, $p = 0.004$; w_s, $p = 0.002$).

Table 4.E2 Means and standard deviations estimated from 130 trials of the Jardri et al. dataset compared to the current study.

	Jardri et al. CTL		Jardri et al. SCZ		This study	
CII params	μ	SD	μ	SD	μ	SD
a_p	0.196	0.176	0.133	0.194	0.180	0.174
a_s	0.041	0.035	0.250	0.216	0.084	0.120
w_p	0.361	0.107	0.232	0.242	0.470	0.199
w_s	0.580	0.158	0.775	0.139	0.697	0.194

It is important to remember that a quantitative comparison with the findings of Jardri et al. is limited by the fact that in the current study the diagnoses are self-reported, the number of trials reduced, and the study took place online, as opposed to a lab environment. Nonetheless, a qualitative comparison between our ASD group and the SCZ group of Jardri et al. shows a clear difference in sensory reverberation, with higher values in the SCZ patients, especially in the CII model (Figure 4.E1).

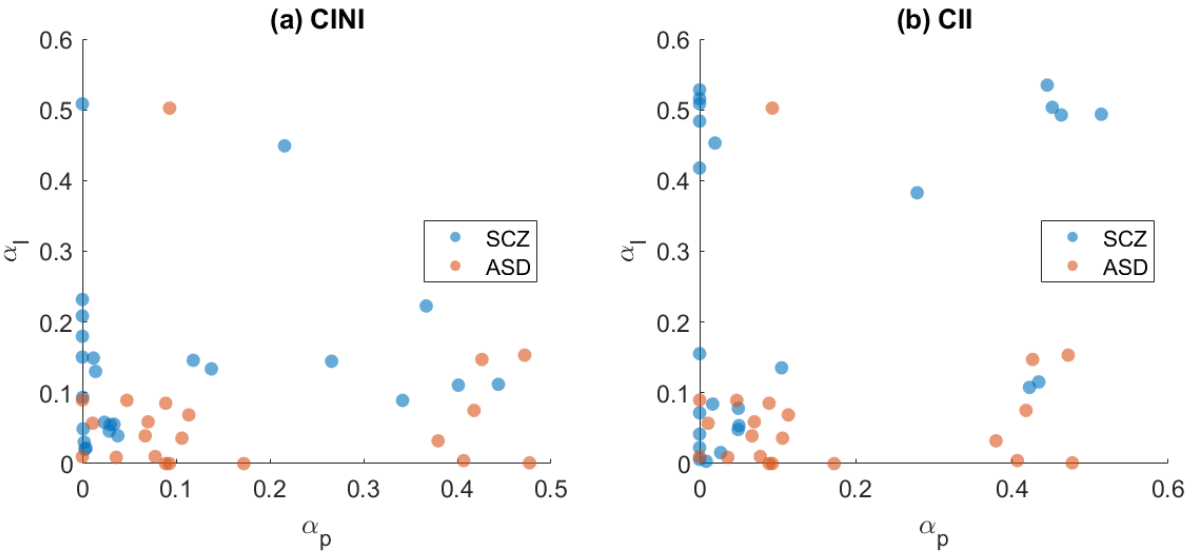

Figure 4.E1 Reverberation parameters of the current study's ASD sample and Jardri et al.'s SCZ sample. *Each dot corresponds to one participant. SCZ parameter values were estimated using 130 trials of the Jardri et al. dataset.*

We also used the Jardri et al. dataset to more deeply investigate the comparisons between CII and CINI goodness of fit, in relation to trial set changes, parameter values, and psychiatric diagnoses and traits. The results from the Jardri et al. dataset showed that CII lost ground to CINI in the 130-trial set, and that patients are much better fitted by CII than CINI, while controls were slightly better fit by CINI (Table 4.E3). Our study showed no difference between the ND and ASD groups, as both were clearly dominated by CINI.

Table 4.E3 Fixed and random effects model comparisons in both studies.

Trialset	**ΔBIC**	***Pr*(CII)**	**ΔBIC**	***Pr*(CII)**	**ΔBIC**	***Pr*(CII)**
Jardri et al.	Whole sample		CTL		SCZ	
200	192	0.61	–9	0.48	201	0.83
130	–27	0.53	–104	0.39	77	0.76
This study	Whole sample		ND		ASD	
130	–916	0.27	–291	0.31	–89	0.29

ΔBIC is the sum of individual CINI BIC scores minus individual CII BIC scores. Pr(CII) is the posterior model probability for CII, with Pr(CINI) = 1 – Pr(CII). CTL stands for control participants.

One possible explanation could be that the dominance of CII is associated with the size of the reverberation parameters. We tested that by looking at the Kendall correlations of CII reverberation parameters with the individual *Δ*BIC = BIC(CINI) – BIC(CII). Interestingly, the results showed that this correlation only existed for the prior reverberation parameter and not the likelihood, and it appeared almost exclusively in patients with schizophrenia and not controls (Table 4.E4).

Table 4.E4 Kendall rank correlations between CII reverberation parameters and ΔBIC scores.

	Jardri et al. CTL		**Jardri et al. SCZ**		**This study**	
CII params	τ	p	τ	p	τ	p
$\boldsymbol{a_p}$**, 200**	0.31	0.03	0.63	< 0.001	–	–
$\boldsymbol{a_s}$**, 200**	0.22	0.13	0.19	0.19	–	–
$\boldsymbol{a_p}$**, 130**	0.08	0.59	0.53	< 0.001	–0.13	0.01
$\boldsymbol{a_s}$**, 130**	–0.13	0.39	0.17	0.24	–0.02	0.70

Overall, these results show that the dominance of CINI in our study, does not contradict the findings of Jardri et al. On the contrary, the results suggest that SCZ patients exhibit joint signal reverberation, in contrast with controls who independently overcount sensory or prior information.

Chapter 4.S1 References

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Frechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel.

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*(1), 14218. https://doi.org/10.1038/ncomms14218

May, J. O., & Looney, S. W. (2020). *Sample Size Charts for Spearman and Kendall Coefficients*. *11*, 7.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Sage.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). 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. *Molecular Autism*, *6*(1), 2. https://doi.org/10.1186/2040-2392-6-2

Chapter 5

Pilot of a Social Decision-Making Task

5.1 Motivation and Design

As discussed in Chapter 4, circular inference is a theory for schizophrenia developed by Jardri and Denève (2013), inspired by the excitation to inhibition imbalances in the brains of patients. According to that theory, inferential feedback loops resulting from an increased excitation-to-inhibition ratio lead to overcounting of sensory evidence and overconfident beliefs in schizophrenic individuals. Experimental evidence for this theory was provided by Jardri et al. (2017) using the fisher task, a version of the beads task, where they showed that patients with schizophrenia exhibited increased likelihood overcounting relative to healthy controls. Autism shares a lot of similarities with schizophrenia, such as social impairments (e.g., Couture et al., 2010) or genetic risk factors (e.g., Carroll and Owen, 2009), which also include the increased excitation-to-inhibition ratio (e.g., Kana et al., 2007; Gogolla et al., 2009; Christ et al., 2011).

In Chapter 4, we used the fisher task to investigate circular inference in autism in a sample with a broad spectrum of autistic traits and individuals with self-reported autism diagnoses. Our findings showed no differences in circularity across the autism spectrum. One potential explanation for this result is that such impairments in autism are somewhat domain-specific. Autism is partially a social disorder, to a higher degree than schizophre-

nia, as reflected in the conditions' diagnostic criteria. Indeed, in our review in Chapter 2 we found that differences in Bayesian inference in autism and across autistic traits were more common in social compared to non-social tasks. A second potential reason for these results is that, as we highlighted in our discussion, this experiment could be interpreted more as a cue integration task, rather than an integration of priors and sensory evidence. This is because priors and likelihoods were simply separated by a short time interval, meaning that participants might had not formed a belief based on the first stimulus before they saw the second. Instead it is possible that they had kept the corresponding stimulus in mind and then used both stimuli concurrently to form a belief. For that reason, we decided to develop a different design which would be more likely to be interpreted in the intended way for the participants.

A recent study by Simonsen et al. (2021) used a version of the beads task which included social elements to measure circular inference in schizophrenia (Figure 5.1). In this social beads task, the participants were presented with a jar of obscured contents. They knew that the jar contained some number of green and red beads but not their number or ratio of colours. Their task was to guess the colour of the next bead, based on two sources of information. Firstly, the faces of four other people, the 'agents', were shown besides the jar. Each of them was shown to draw some beads from the jar and then provide their estimate for the colour of the next bead, expressed with 'high' or 'low' confidence. Then, the participant themselves drew some beads. By combining the agents' judgements and their own draw, the participants made a binary judgement about the colour of the next bead. This version of the beads task could be better suited for investigating circular inference in autism, as social difficulties are one of the most common symptoms of the condition. Moreover, the task design appeared to be more intuitive and better suited to probabilistic thinking, compared to the fisher task. However, we encountered a few issues with its structure as published by Simonsen et al. (2021), which motivated us to modify it for our study.

Firstly, similar to the fisher task, the task could also be more akin to a cue integration task rather than a combination of priors and sensory evidence. Despite the agents' guesses first appearing on the screen before the participant's draw, the fact that both sources of information are present on the screen at the same time might result in the participant not

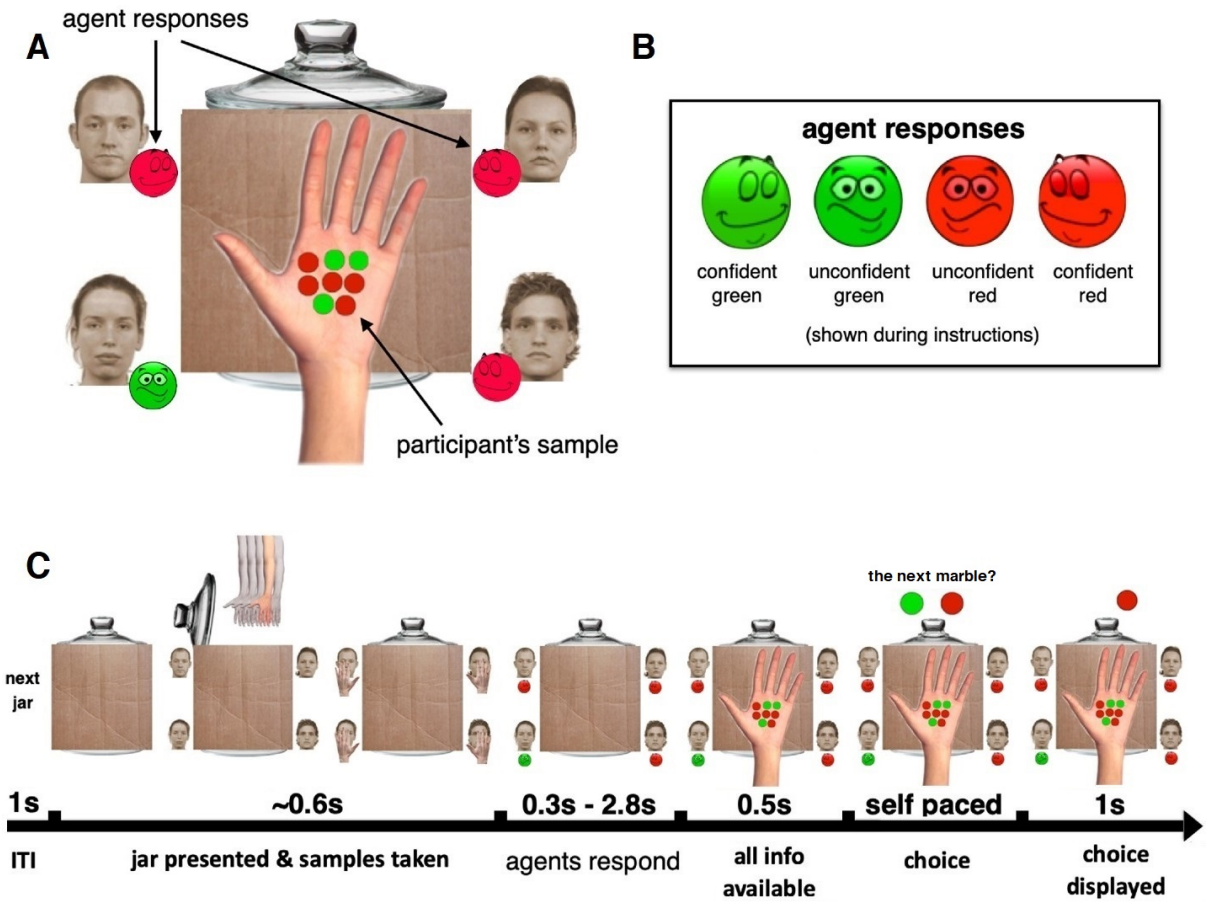

Figure 5.1: An illustration of the social beads task [adapted from (Simonsen et al., 2021)]. A) The information presented to the participant during their response. B) An explanation of the agents' judgements, as explained in the instructions of the task. C) The timeline of a single trial.

forming a prior belief in the meantime. Secondly, the binary response of the participant makes measuring overconfidence more difficult, something which is central to the task. If participants do not follow a probability matching strategy, but instead always select the option with the highest probability, then overconfidence would have no influence on the judgement. A probability of 60% for the next bead being red would lead to the participants choosing 'red' 100% of the time. A third issue is that participants have to account for both the nature of the jar and the randomness of the next bead drawn from that jar, complicating the interpretation of their responses. This happens because during the task, participants get information about the nature of the jar: both the bead draw and the responses of the agents make some red-to-green ratios in the jar more probable than others. But if, for example, a participant is extremely confident that the jar is

majority-red with 60 red beads and 40 blue ones, this would make them only 60% certain that the next bead would be red. The final issue is that the 'high' and 'low' confidence with which agents provide their guesses are not automatically mappable to probabilities. This results in one extra parameter that has to be estimated by the computational models and which adds additional complexity to parameter identifiability.

For these reasons we decided to modify the task. The new version of the task and the accompanying code was initially designed by Matthew Whelan. We changed the goal of the participants to guessing the jar the beads were drawn from instead of the next bead, so that we could circumvent the probabilistic nature of the draw itself. During instructions, participants were shown the two possible jars that the beads could be drawn from, a majority-red one (60-40) and a majority-blue one (40-60), and they were informed about their exact contents. Then, during the task, participants were shown a random jar of the two being picked, without its contents being revealed. The fact that the contents of the jars were explicitly determined allowed us to calculate the exact evidence a specific bead draw would provide for each jar. We also reduced the number of agents to one, to make the task simpler. That agent drew some beads from the jar and gave a certainty judgement about its nature using a slider, i.e. how likely was it that the jar was the majority-red vs the majority-blue one. This made their judgement clear and removed the need to interpret what 'high' or 'low' confidence could mean. Then, the participant drew some beads themselves and they had to combine the information from their draw with the belief of the agent to give a new probabilistic estimate, using the same slider. The starting point for the slider was the agent's judgement to reinforce the idea that participants should start with that as their prior belief and adjust it based on the new evidence provided by their draw. Before being able to use the slider, participants were asked which jar they deemed more probable. This was done to remind them that they were trying to choose between two jars, as opposed to choosing the jar ratio out of all possible red:blue ratios. The final version of the task can be seen in Figure 5.2.

We made sure that the instructions were very clear about the nature of the task. All the steps were followed by examples with accompanying images of beads or possible responses on the slider bar. Participants were told exactly what each end of the slider bar corresponded to and were given examples of certain and uncertain responses. We

also provided an example of the agent's response and what that could mean or how they could take the information into account. The task was preceded by four training trials in which participants could get familiarised with the task. In these trials, the draw of the agent was visible along with their judgement, so that they could see that the agent offered different information from their own draw, that the agent is reliable, and to give them a further indication of how the slider is interpreted.

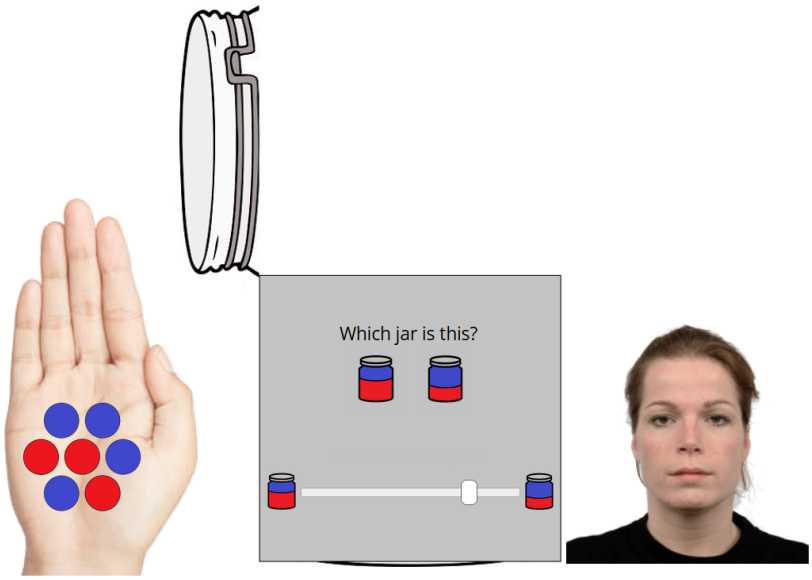

Figure 5.2: The new response screen. Participants saw their own bead draw on the left, the covered jar in the middle, and the agent's face on the right. The slider bar presented them with the judgement of the agent, which appeared on the screen before the bead draw. Participants had to first select which jar they thought was the most probable out of the two options and then adjust the slider to state their certainty.

The task lasted for 134 trials. The prior and likelihood values and their combinations were determined so that they maximised the parameter recovery without presenting the agent as unreliable. If priors disagreed too much with the sensory evidence, then it would be reasonable for participants to conclude that the agent's judgements cannot be trusted. On the other hand, if they were very similar then model fitting would not be able to disentangle their influences. For that reason we restricted the distance between while choosing the exact values to maximise parameter recovery. The resulting prior and likelihood values differed by 22.3% on average, with the maximum difference being 55%.

We planned to run the task on a participant pool with broad autistic traits and a substantial proportion of ASD diagnoses. We hypothesized that (1) participants would probabilistically integrate new evidence with their prior beliefs, but (2) this process would differ in those with strong autistic traits or diagnoses. Specifically, this could manifest as (2a) a reduced influence of priors or (2b) increased circularity, potentially both. We also considered dividing participants into two conditions, with the second one consisting of a non-social version of the task, where the prior would be based on the relative frequency of majority-red and majority-blue jars. In that case, we expected that (3) any differences observed between participants with strong and weak autistic traits in the social condition would be less pronounced in the non-social condition.

5.2 Results and Discussion

We initially piloted the social version of the experiment online with 12 participants, selected via word of mouth. The participants were naive as to the purposes of the experiment. Before the task, they completed the autistic quotient questionnaire (AQ; Baron-Cohen et al., 2001) and the 21-item Peters et al. Delusions Inventory (PDI; Peters et al., 2004), designed to quantify the strength of autistic traits and delusional ideation in the general population. After the task they completed an anonymous feedback questionnaire about their understanding of the task and the strategies they followed. All participants reported that the task was easy to understand and the instructions clear. Moreover, 10 out of the 12 participants reported that they trusted the agent and that they took their responses into account when making their judgement. Of the 2 that did not trust the agent, one said that they were nonetheless influenced by them. Indeed, fitting the data with a linear fixed-effects model showed that priors strongly influenced responses in the expected direction ($slopes > 0.36,\ p < 10^{-13}$) for all participants besides these two ($slope = 0.068,\ p = 0.22$ and $slope = 0.07,\ p = 0.025$). Likelihoods had a significant effect on all participants ($slopes > 0.38,\ p < 0.003$).

However, upon further analysis of our data, we discovered an unexpected result. According to Bayes' rule, when both the prior and the likelihood point to the same possibility (e.g., both give > 0.5 probability that beads are drawn from the majority-red jar), the

posterior should be more extreme than both of them. This holds true regardless of the specific value of the prior or the likelihood, as long as they agree, because more pieces of evidence increase the certainty of the corresponding possibility. The collected data revealed that participants frequently deviated from this pattern. Specifically, out of the 12 participants, 5 followed this rule less than 50% of the time. In trials with agreeing priors and likelihoods the median participant gave more extreme responses than both the prior and the likelihood on only 56% of them. This holds true even if the 2 participants that were not very influenced by the prior are excluded, with 56% becoming 59%. No correlations with AQ or PDI were observed ($p > 0.3$).

This pattern could result from a strategy more akin to averaging between priors and likelihoods. However, other explanations are possible as well. One possibility is that even participants that were influenced by the agent did not completely trust them, leading them to underweight the prior. This could result in responses between priors and likelihoods when priors are more extreme than likelihoods, even when following Bayesian principles. A follow-up analysis disproved this possibility, by showing that focusing only on trials where priors were less extreme than likelihoods did not significantly affect the median (59%). Another possibility is that participants were wrongly quantifying the likelihood of the drawn beads. However, plotting the data revealed that the likelihood had the expected influence on the participant responses when priors and likelihoods disagreed (Figure 5.3).

To confirm this, we included an additional question at the end of the task (Figure 5.4) and contacted past participants to interview them about their strategies in the task. Although there was a feedback question about task strategy, it did not cover this specific possibility and participants mostly did not provide any clarifying information. We managed to interview 7 of the 12 participants. Out of them, only 2 reported strategies similar to Bayes' rule when responding. The other 5 employed strategies more akin to a weighted average between priors and likelihoods. Even after being explained the Bayesian strategy, 3 of these participants found it counterintuitive and could not understand why it would be the preferred one mathematically. 4 new participants completed the revised version of the task with the additional final question. All of them responded between the likelihood and the prior, in accordance with an averaging strategy.

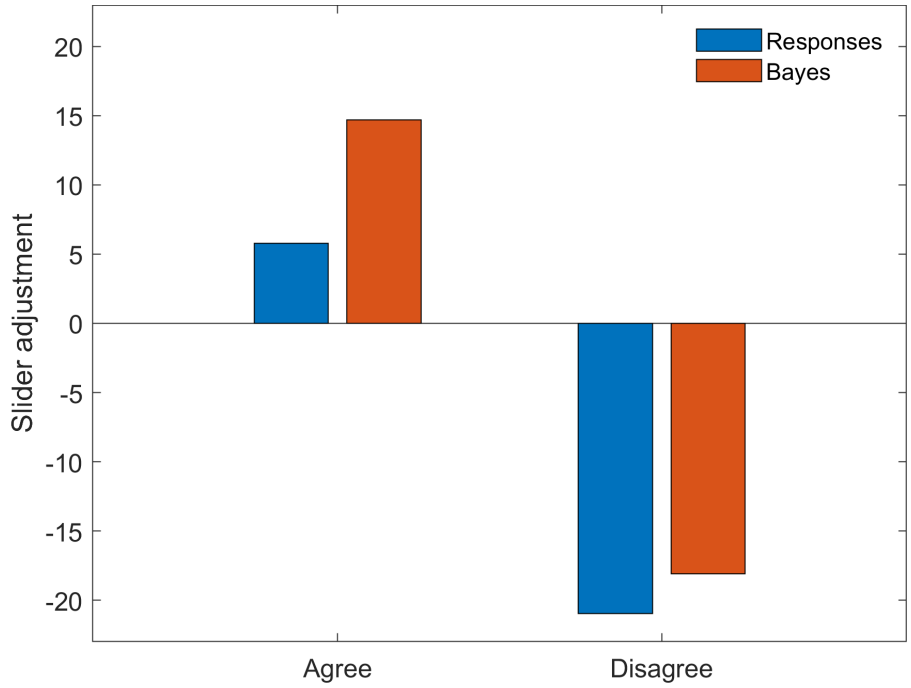

Figure 5.3: Figure 3: Average participant slider adjustment compared to an ideal Bayesian observer. Participants adjusted the slider much less than expected when sensory evidence agreed with priors, but not when it disagreed, consistent with an averaging strategy. The alternative possibility of underweighting the likelihood would affect all trials.

These findings show that despite our task design being based on existing work and our efforts to influence participant behaviour, it did not evoke the expected Bayesian processes in participants, contradicting our first hypothesis. This is a surprising result given the current prevalence of Bayesian models of decision making (Beck et al., 2008; Ma, 2019; Trimmer et al., 2011). The idea that the brain performs Bayesian inference has been highly influential, with theories like the Bayesian brain and predictive coding being proposed as overarching frameworks for understanding information processing in the brain (Friston, 2010).

Prior research has identified various limitations in the brain's Bayesian computations. However, most of these have been deviations in a fundamentally Bayesian process. For instance, researchers have argued that executing Bayesian inference computations is too resource-intensive, leading the brain to employ sampling approximations for integrating priors and likelihoods or for representing priors (Lange and Dukas, 2009; Sanborn and Chater, 2016). Other research has demonstrated that priors can be resistant to change,

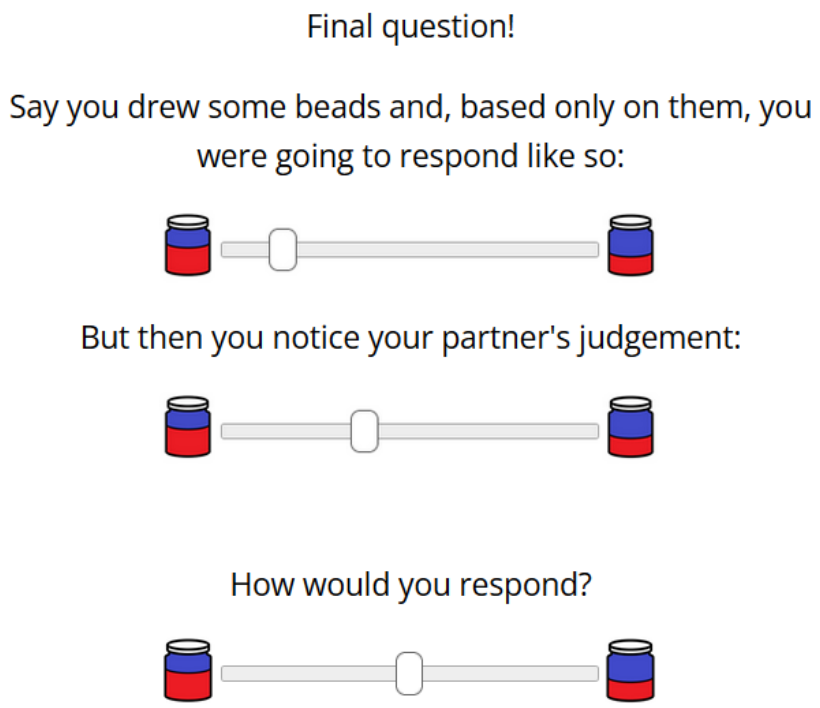

Figure 5.4: Figure 4: The final question designed to discriminate between the averaging and Bayesian strategies. The structure of the question removes the probabilistic interpretations of the bead draw as a factor so that responses are based exclusively on the strategy used by the participant. Bayesian strategies would result in more extreme responses to the left. Averaging strategies would result in responses between the first and the second sliders.

failing to update adequately with new evidence (Yon et al., 2023), that integration of novel combinations of priors and likelihoods is Bayes-suboptimal (Lin et al., 2024), or that cognitive limitations impose costs on information processing, cause individuals to optimise both for accuracy and for the conservation of their cognitive resources (Caplin and Dean, 2015; Lieder and Griffiths, 2020). However, very few studies have shown such a fundamental departure from Bayesian principles in a simple and distinctly probabilistic setting, as observed in our task.

One such study, conducted by Prat-Carrabin and Woodford (2024), bears some similarities to our task. The participants had to sequentially estimate the proportion of green vs red rings in a box. Each trial required participants to draw a single ring and adjust their

proportion estimate based on the ring's color. The procedure repeated with a new box after every 5 trials, continuing for 40 boxes total. The authors observed significant deviations from Bayes-optimal responses, with participants instead updating their estimates by the same amount after each stimulus. This result aligns with our findings, with participants in both tasks responses appearing to be determined largely by the user interface, rather by the probabilities that it represented. However, Prat-Carrabin and Woodford interpreted that behaviour is a product of resource-rational cognition, with participants economising cognitive resources, while still responding relatively accurately to the task.

An alternative explanation of our results draws upon the theory of ecological rationality, which posits that individuals use 'fast-and-frugal' heuristics, instead of analytical, rational approaches in decision-making (Luan et al., 2019). This is similar to the work of Kahneman and Tversky (Kahneman, 2012), which has shown various ways in which individuals' reliance on heuristics leads them to exhibit biases and act contrary to the laws of probabilistically. Examples include base rate neglect (Welsh and Navarro, 2012), wherein individuals disregard base rates and instead make decisions exclusively based on the immediate evidence provided, and anchoring (Furnham and Boo, 2011), where initial information exerts disproportionate influence on participant judgements. However, a growing amount of research is reinterpreting such findings as adaptive heuristics that are misapplied in novel contexts (Haselton et al., 2009; Leimar and McNamara, 2019; Lieder et al., 2018). In a similar vain, it is possible that our participants misinterpreted the task, leading them to rely on inappropriate heuristics. Research has shown that different framings of the same problem can elicit different approaches by participants (Gigerenzer and Hoffrage, 1999), and it is possible that the representation of certainty judgements or the framing of the question in our task could indicate contexts where weighted averaging constitutes an optimal strategy. For example, averaging is appropriate when individuals disagree in the presence of the same information. The difference in our task was that the prior and the likelihood were based on completely independent evidence. Participants were repeatedly informed that this was the case, but the importance of this information might had been unclear to them. Alternatively, participants might have intuitively processed the task as estimating the ratio of blue to red balls in the jar, in which case the Bayesian strategy would indeed consist of the averaging of different opinions.

Investigating these possibilities would require additional research with a larger sample and potentially a task explicitly designed for that purpose. While the characteristics of Bayesian reasoning in the general population is an important research topic in its own right, it is not the primary aim of this thesis. The focus of the current work is on better understanding Bayesian inference in autism, as it is believed to lie at the core of the condition. Given that our task did not reliably evoke Bayesian strategies in the participants, it would not be an effective tool for researching impairments in these processes. We therefore decided not to utilise this task further as part of the current work and instead explore alternative paradigms. Nonetheless, these findings raise interesting questions outside the scope of this thesis. Which apparently Bayesian tasks result in non-Bayesian behaviour by the participants? What is the mechanism through which underlying Bayesian processes give rise to such behaviours? Or, alternatively, which fundamental cognitive processes are Bayesian and which are not?

* * *

Chapter 4 investigated circular inference across autistic traits using a probabilistic decision-making task. Despite the theoretical promise of this framework and its success in explaining aspects of schizophrenia, we found no evidence for increased signal reverberation in autism or across autistic traits. In **Chapter 5**, we attempted to investigate circular inference in a social task, but participants behaved in a deeply non-Bayesian way. One potential explanation for both of these results is the highly explicit nature of these tasks. It is possible that Bayesian processes are at the core of low-level information processing, but high-level explicit strategies can vary, leading Chapter 5 participants to utilise different heuristics to respond to our task. Similarly, although participants in Chapter 4 behaved in a Bayesian way, it is possible that additional explicit mechanisms masked the differences across autistic traits or between diagnostic groups.

In the systematic review of Chapter 2 we had shown that tasks where participants had to implicitly learn environmental statistics were more likely to find differences in autism or across autistic traits compared to studies where the relevant information was explicitly provided. We had also shown that tasks that did not measure the imbalance between priors and likelihoods, but still investigated the learning of environmental regularities, often found differences between autistic and non-autistic individuals or across autistic traits.

These observations motivated the work presented in **Chapter 6**, where we directly compare implicit and explicit learning of environmental regularities across autistic traits. Inspired by the work of Lawson et al. (2017) and Sapey-Triomphe et al. (2021b), we create an audiovisual task where participants have to learn the relationship between an auditory cue and a visual stimulus. We decided to remove the volatility component of the original tasks, as it would make it more difficult to measure the main effect of explicit instructions on learning and would be challenging to implement an explicit condition. We also changed the difficulty of the task, to ensure that participants had to utilise their knowledge of the regularities to perform optimally, as opposed to simply relying on the visual stimulus.

Chapter 6

Implicit and Explicit Learning in Individuals with Autistic Traits

This chapter includes a manuscript to be submitted for publication: Angeletos Chrysaitis, N. & Seriès P. (in prep.) Influence of truthful and misleading instructions on statistical learning across the autism spectrum

Chapter 6 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 the effects of both their presence and their veracity on statistical learning, 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 drew attention to them and resulted in strong priors, correctly reflecting the task's statistics, regardless of the information's veracity. Autistic traits did not significantly affect the influence of priors.

Keywords: Bayesian inference, Learning, Instructions, Explicit, Implicit, Autism

6.1 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 (Knill & Pouget, 2004). 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 (Forkstam & Petersson, 2005; VanPatten & Smith, 2022). The results show evidence for two distinct processes (Foerde et al., 2006; Maddox & Ashby, 2004), affecting different brain areas (Critchley et al., 2000; Fletcher et al., 2005; Hazeltine, 1997). These processes depend on different factors, with explicit learning being correlated with IQ (Gebauer & Mackintosh, 2007; Reber et al., 1991) and the resulting knowledge decaying over time (Liu et al., 2023; Musen & Treisman, 1990), 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 lead to inhibition of implicit learning and worse participant performance (Caljouw et al., 2016; Fletcher et al., 2005; Reber, 1976; Song et al., 2007). However, to our knowledge, no studies have directly compared

implicit and explicit learning when environmental regularities are based on the frequency of the stimuli, instead of specific rules that they follow. Moreover, no studies have separated the influence of instruction presence and instruction content 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 (Lawson et al., 2014; Pellicano & Burr, 2012; Van de Cruys et al., 2014). 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 (Lawson et al., 2017; Palmer et al., 2017). 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 (Angeletos Chrysaitis & Seriès, 2022). 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 (Angeletos Chrysaitis & Seriès, 2022). This seems to align with non-Bayesian studies showing intact explicit but occasionally impaired implicit learning in ASD (see Kourkoulou, 2010 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 both the presence and the veracity of the instructions. 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.

6.2 Methods

6.2.1 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. 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, $p = 0.47$, $BF_{10} = 0.09$) or across conditions ($p = 0.83$, $BF_{10} = 0.006$). Individuals self-reported an autism diagnosis or that they were in the process of getting one did not differ across conditions (mean 17.3%, $p = 0.78$, $BF_{10} = 0.01$).

6.2.2 Design

The experiment consisted of a cue-stimulus association task (Figure 6.1). In all conditions, the experiment begun 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 begun at a very easy level, with the gradings being presented at a very high contrast and at 70° or 120°. During the training, the

Figure 6.1: Example stimulus (A) and structure of individual trial (B). *Stimuli in the experiment were presented in significantly lower contrast than this example.*

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.

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.

6.2.3 Computational Modelling

Besides simple statistical methods, we modelled the evolution of the participants' prior beliefs using the Hierarchical Gaussian Filter (HGF, Mathys et al., 2014). 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 corresponding 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 = $0.5 + 0.5 / (1 + \exp(\log(\text{contrast}) - c_0))$, with c_0 functioning essentially as a baseline contrast level that was fit to the data.

The usual implementation of the HGF models the noise in the responses as decision noise. It assumes that the participants follow a probability matching strategy, where they sample their

responses using the posterior probability. A different possibility is that participants always choose the option highest probability. In that case, the noise in the responses would be a direct result of sensory noise. We compared both implementations of the model. Moreover, given that our 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 (Stephan et al., 2009) in each of the four conditions.

6.3 Results

6.3.1 Behavioural

The main expected influence of the prior was biasing the participants towards the tilts that were associated with the tones. This would result in participants having higher accuracy when tilts and tones were congruent and lower when they were not (Figure 6.2). By 'congruent' here we mean the tone-tilt pairings that appeared in 75% of the trials (in the I^0R^+, I^+R^+, and I^-R^+ conditions) or that participants were told they 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.

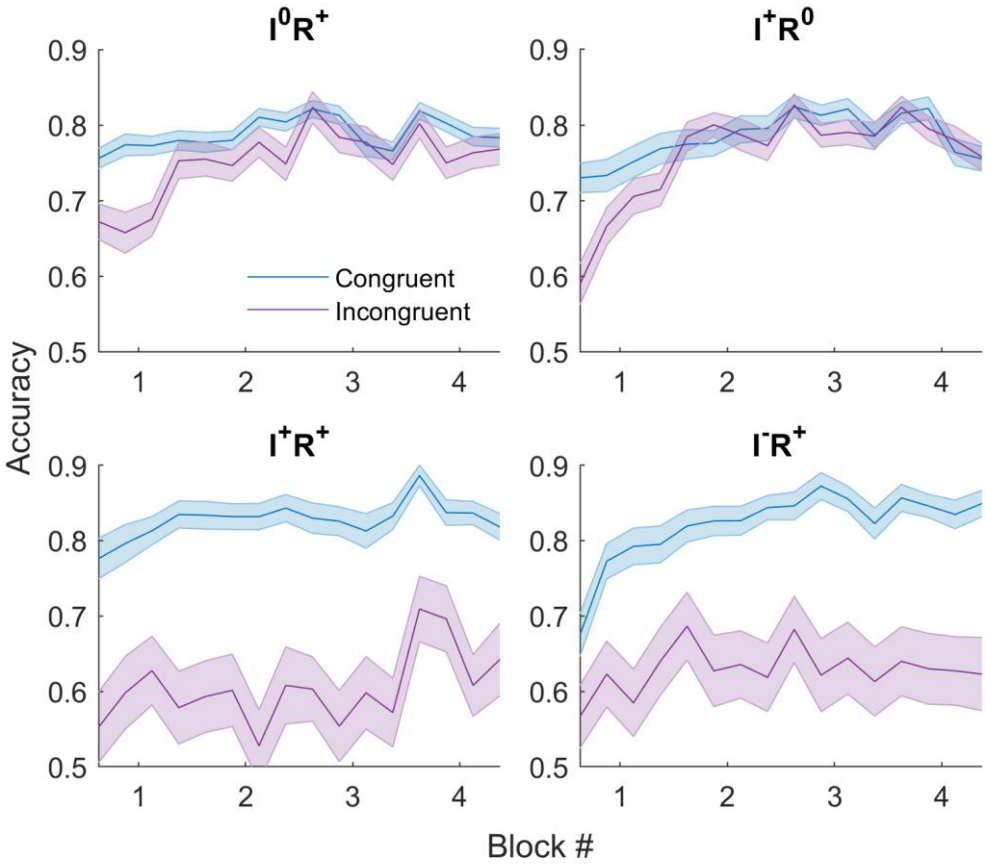

Figure 6.2: Accuracy comparison between congruent and incongruent trials in the four conditions. *Participants showed small accuracy differences in the implicit condition, large accuracy differences in the two conditions where explicit information was combined with regularities, and no differences in the $I^{+}R^{0}$ condition after the first block.*

Looking at the accuracy differences (Figure 6.2), we see that after the first block participants showed biases in the direction of the actual environmental regularities (Figure 6.3). They had significantly higher accuracy in the congruent trials than in the incongruent ones in the $I^{+}R^{0}$ (mean diff = 0.024, $t(118) = 2.73$, $p = 0.007$, $BF_{10} = 3.6$), $I^{+}R^{+}$ (mean diff = 0.226, $t(50) = 6.46$, $p = 10^{-8}$, $BF_{10} = 10^{5}$), and $I^{-}R^{+}$ conditions (mean diff = 0.204, $t(58) = 5.04$, $p = 10^{-6}$, $BF_{10} = 6971$), but not in the $I^{+}R^{0}$ condition (mean diff = 0.004, $t(105) = 0.46$, $p = 0.65$, $BF_{10} = 0.11$). Biases in the $I^{+}R^{+}$ and $I^{-}R^{+}$ conditions were larger than those in the $I^{0}R^{+}$ condition ($t = 7.59$, $p = 10^{-12}$, $BF_{10} = 10^{10}$ and $t = 5.78$, $= 10^{-8}$, $BF_{10} = 10^{5}$, respectively), but they did not differ among

each other ($t = 0.39$, $p = 0.69$, $BF_{10} = 0.1$). 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^+: $t(118) = –0.82$, $p = 0.42$, $BF_{10} = 0.13$; I^+R^0: $t(105) = –1.39$, $p = 0.17$, $BF_{10} = 0.26$; I^+R^+: $t(50) = –1.60$, $p = 0.12$, $BF_{10} = 0.51$; I^-R^+: $t(58) = –0.21$, $p = 0.84$, $BF_{10} = 0.13$).

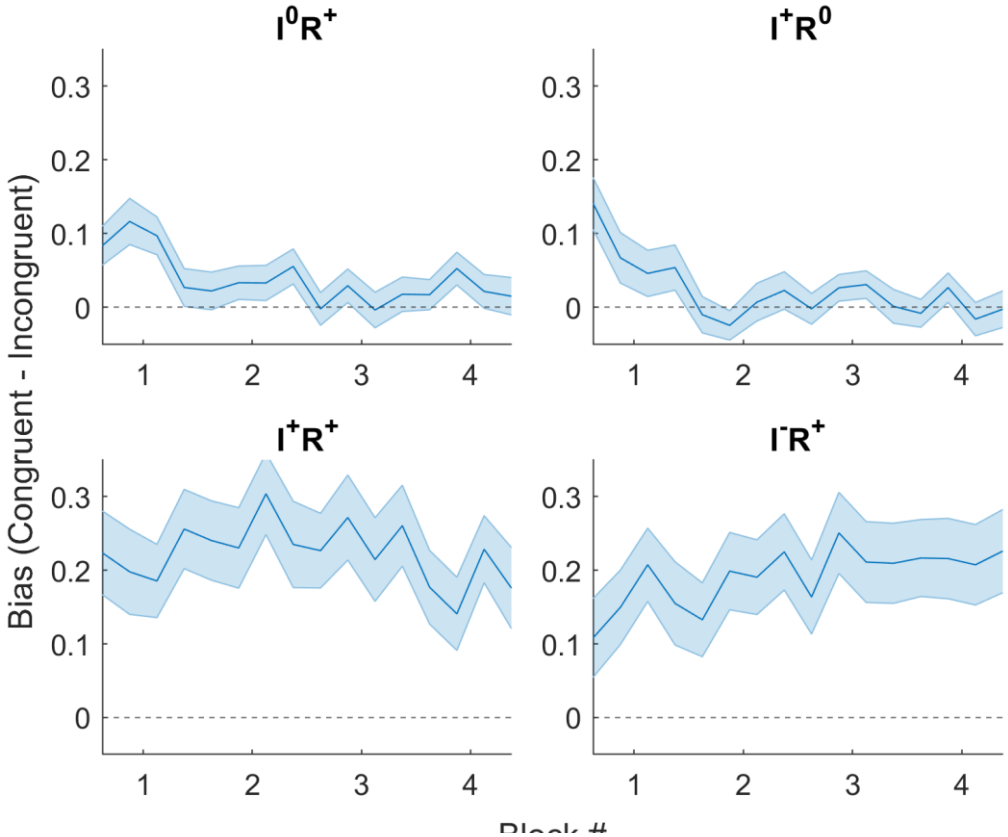

Figure 6.3: The evolution of bias over time in the four conditions.

From Figures 6.2 and 6.3, 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^+ ($t = –0.47$, $p = 0.65$, $BF_{10} = 0.1$) and I^+R^0 conditions ($t = 0.34$, $p = 0.74$, $BF_{10} = 0.1$), while a weak negative and a weak positive relationship was found in the I^+R^+ ($t = –2.35$, $p = 0.041$, $BF_{10} = 2.1$) and I^-R^+ conditions ($t = 2.07$, $p = 0.065$, $BF_{10} = 1.1$), 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 6.A1 in 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.

6.3.2 Computational

Comparing different models (Section 6.C in 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^0R^+ 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 6.4). 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, $t = 4.3$, $p = 10^{-5}$, $BF_{10} = 568$; I^+R^+: mean prior = 0.57, $t = 2.7$, $p = 0.01$; I^-R^+, $BF_{10} = 4.8$; mean prior = 0.44, $t = -2.0$, $p = 0.052$, $BF_{10} = 1.21$). But when that belief disagreed with the actual regularities, it was quickly amended (I^+R^+ vs I^-R^+, blocks 2-4: $t = 1.1$, $p = 0.25$, $BF_{10} = 0.18$). Again, we see that priors in both the I^+R^+ and the I^-R^+ conditions were much stronger than priors in the implicit condition ($t = 7.9$, $p = 10^{-13}$, $BF_{10} = 10^{10}$ and $t = 5.2$, $= 10^{-6}$, $BF_{10} = 10^4$).

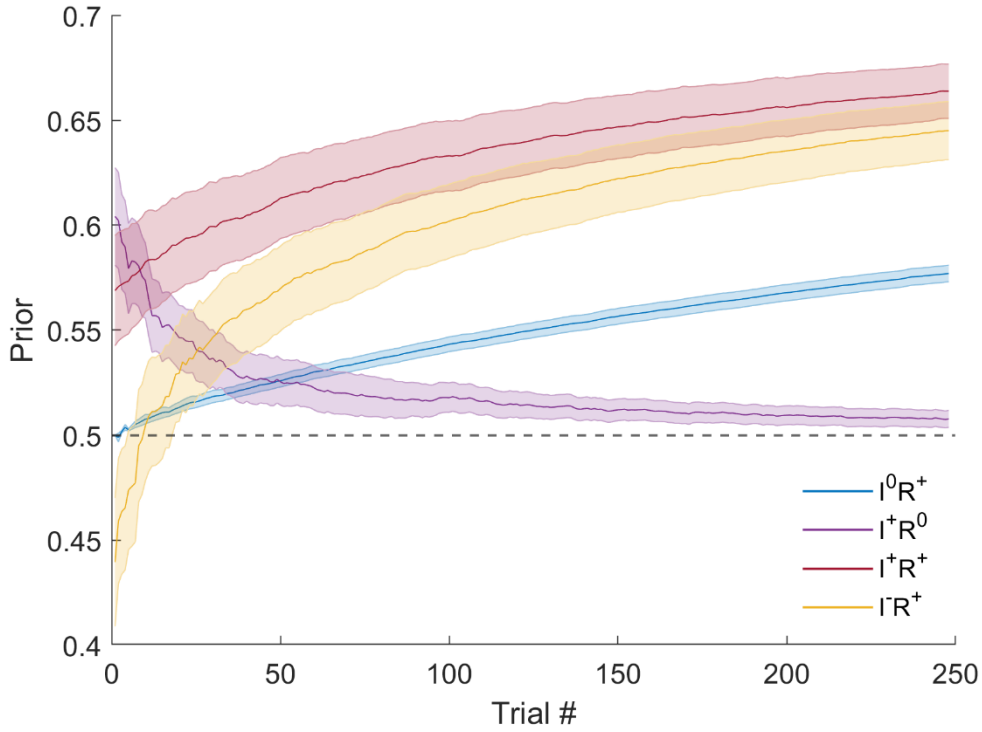

Figure 6.4: The trajectories of model-estimated first-level priors in the four conditions. *Priors were significantly affected by the instructions in the beginning of the task, but that was quickly amended based on the regularities in each condition. Priors in the conditions that combined explicit information and regularities were much stronger than those in the implicit condition.*

One difference with the previous findings is that the computationally estimated first-level prior in the I^+R^0 condition never reaches 0. Moreover, while the biases in Figure 6.3 remained relatively stable after the first block, Figure 6.4 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 6.5), which are equal to variance of second-level beliefs. In all conditions, learning rates decreased during the task (1^{st} block vs 4^{th} block, $t > 12$, $p < 10^{-22}$, $BF_{10} > 10^{14}$), as participants got more information about their environment and formed more precise beliefs. The learning rates of the I^0R^+ condition were the highest ($t > 4.7$, $p < 10^{-5}$, $BF_{10} = 10^3$) and those of the I^+R^0 were the lowest ($t > 3.6$, $p < 0.001$, $BF_{10} > 43$), while the learning rates of the I^+R^+ condition were not significantly different to I^-R^+ ($t = 1.5$, $p = 0.14$, $BF_{10} = 0.3$).

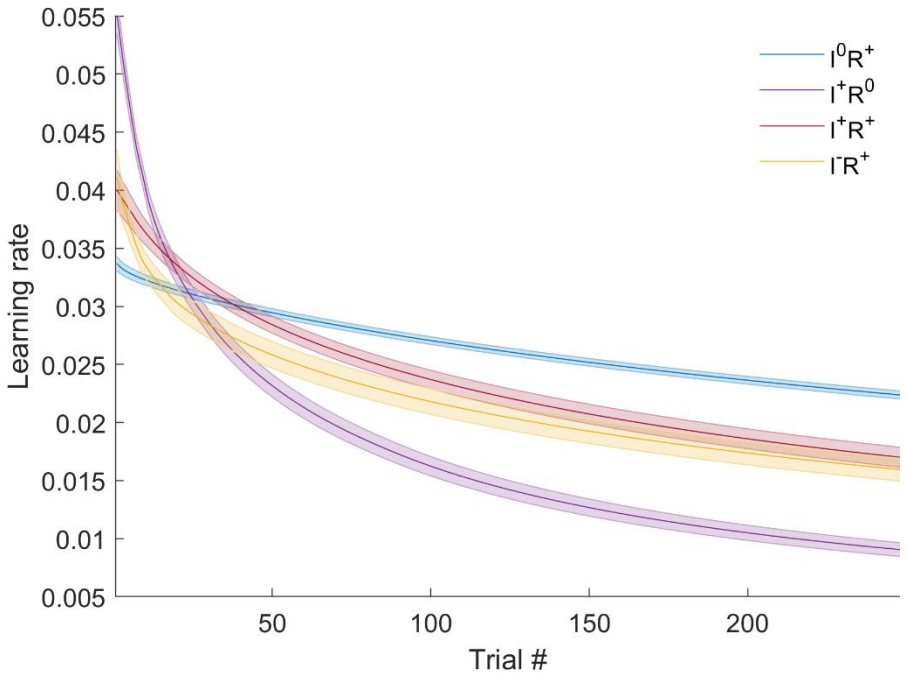

Figure 6.5: The evolution of second-level learning rates (variances) over time in the four conditions. *In all conditions the second-level priors become less uncertain during the task, leading to a decrease in learning rates.*

6.3.3 Autistic traits and diagnoses

To investigate the relationship between autistic traits 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 (Wilcox, 1993). 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 (ps $\geq$ 0.07, BF_{10}s $\leq$ 0.42, Figures 6.6 & 6.7). 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, BF_{10} = 0.35), a relationship that was more weakly present in the model-estimated priors (r = –0.14, p = 0.15, BF_{10} = 0.21). Parameter correlations showed a tendency for lower k_1 with higher AQ in the I^+R^0 condition (r = –0.18, p = 0.07, BF_{10}

= 0.42) and the I^+R^+ condition (r = –0.2, p = 0.16, BF_{10} = 0.29), but the reverse tendency in the I^-R^+ condition (r = 0.18, p = 0.17, BF_{10} = 0.26).

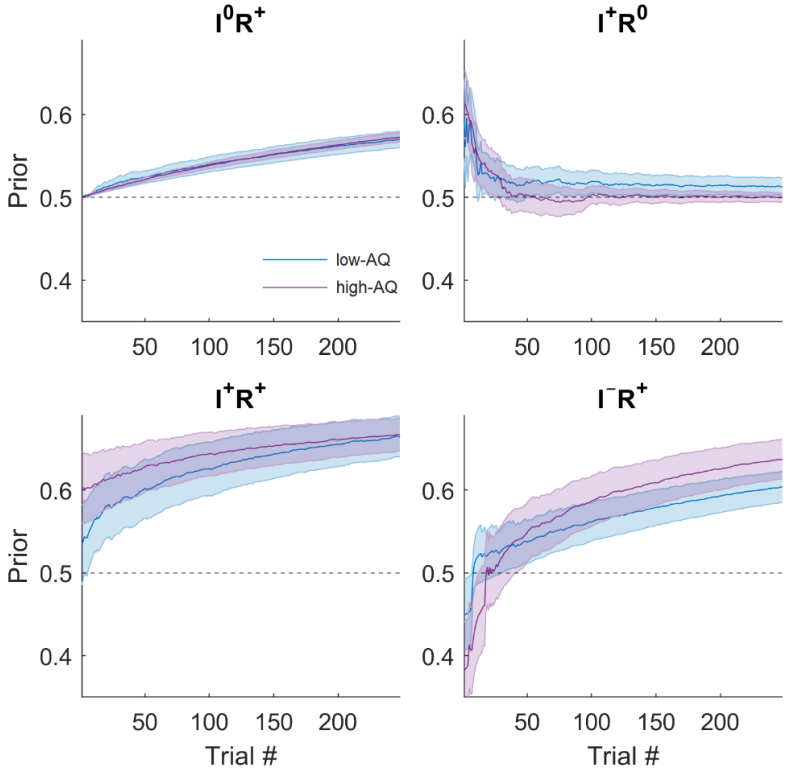

Figure 6.6: Prior trajectories in the low-AQ (≤ 19) and high-AQ (≥29) groups.

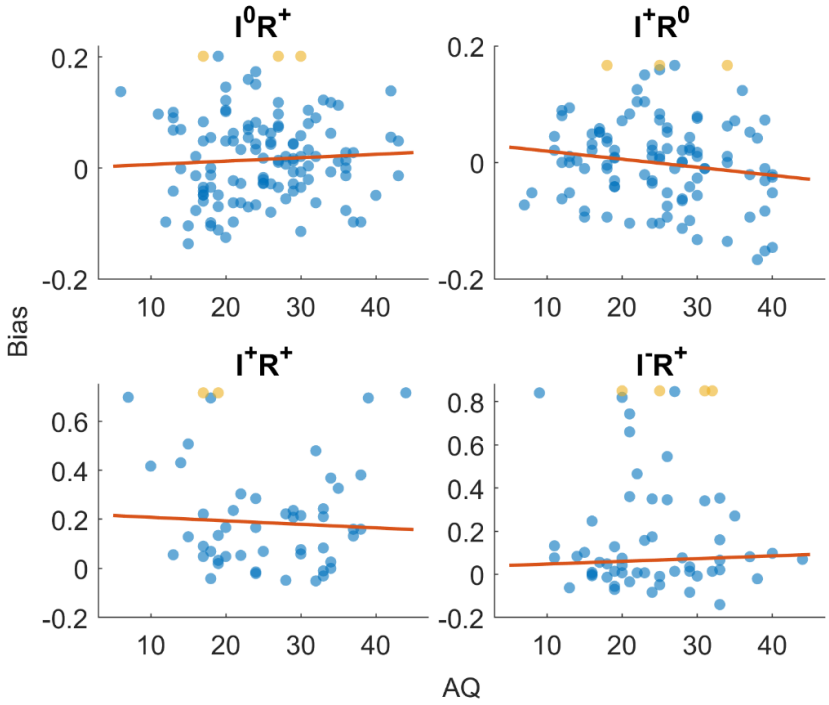

Figure 6.7: Correlations between autistic traits and response biases after the first block. *Orange dots correspond to winsorised outliers, and the red line to a robust linear regression fit. No correlations were significant.*

We also investigated the relationship between AQ scores and learning rates or prior variances (Figure 6.8). The resulting Pearson correlations showed a marginal effect in the $I^{+}R^{0}$ condition ($r = 0.19$, $p = 0.06$, $BF_{10} = 0.51$) and a significant positive correlation in the $I^{0}R^{+}$ condition ($r = 0.19$, $p = 0.03$, $BF_{10} = 0.85$). However, note that the p-value would not survive a correction for multiple comparisons, that the Bayes factor points weakly towards the null hypothesis, and that there was weak evidence against the correlation in the implicit condition being significantly different from the other conditions ($F = 1.4$, $p = 0.22$, $BF_{10} = 0.66$). 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 ($t = 1.7$, $p = 0.09$, $BF_{10} = 0.39$).

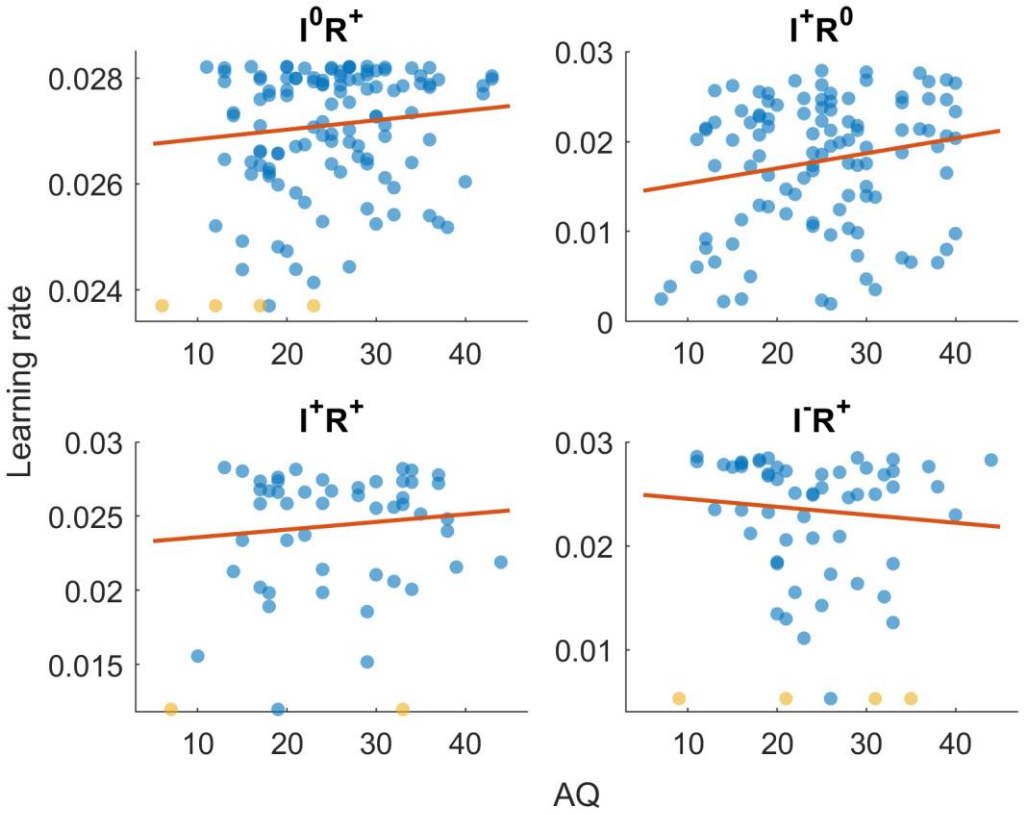

Figure 6.8: Correlations between AQ and average second-level learning rates (variances). *Orange dots correspond to winsorised outliers, and the red line to a robust linear regression fit. In the implicit condition the correlation was r = 0.19, p = 0.03.*

We also compared participants based on their diagnoses. Specifically, we compared participants that self-categorised as not part of the autism spectrum with those that either had or were in the process of getting a diagnosis. Our results once again showed no significant differences between groups for biases, model-estimated first-level priors, likelihoods, or model

parameters ($p > 0.18$, $BF_{10} < 0.35$). However, they did show a weak tendency for higher learning rates in the diagnoses group ($F = 2.55$, $p = 0.12$, $BF_{10} = 0.45$) that did not significantly differ between conditions ($F = 0.2$, $p = 0.89$, $BF_{10} = 0.03$).

6.4 Discussion

6.4.1 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 instructions 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 (Fletcher et al., 2005; Reber, 1976; Song et al., 2007).

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. This was reflected by their learning rates being significantly higher in the first block compared to the implicit condition, but then significantly lower. The veracity of the instructions did not influence participant learning rates, with misled participants learning at a similar rate to those that were correctly informed about the regularities.

Our results demonstrate that instructions substantially influence participants' statistical learning via drawing attention to the environmental regularities, while their informational content plays only a minor role to the formation of participant beliefs. Researchers should exercise caution when formulating task instructions, as even those that do not provide crucial information to the participants can significantly 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.

6.4.2 Autism

We investigated the relationship between autistic traits and task behaviour, as well as self-reported diagnoses. Our hypothesis was that stronger autistic traits 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, neither did they show a significant difference among participants that self-reported that they had or were in the process of getting a diagnosis and those that did not identify as part of the autism spectrum. The only condition that showed a tendency for weaker biases was the one where no actual regularities were present. These findings challenge

common theories that suggest reduced biases in autism, attributed to either flatter priors or more precise likelihoods.

Taking a closer look to our findings potentially reveals the outline of a more nuanced picture. Our findings hinted at a weak positive correlation between autistic traits and learning rates in the implicit condition, although we found weak Bayesian evidence against this. In our version of the HGF the learning rate of a level is equivalent to that belief's variance. If the correlation was to be confirmed in later studies, it could mean that prior beliefs about the regularities were more uncertain with stronger autistic traits, agreeing with the main hypothesis of flatter priors in autism (Pellicano & Burr, 2012). It could also offer additional support for the more recent theories of impaired volatility processing in autism (Palmer et al., 2017) and specifically overestimating that volatility (Lawson et al., 2017). 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 offer a theoretical explanation of 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 (Angeletos Chrysaitis & Seriès, 2022). However, we have to emphasise that the evidence in our study only hints at this possibility and no conclusions can be drawn unless further studies with larger sample sizes verify this relationship.

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 6.9. 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.

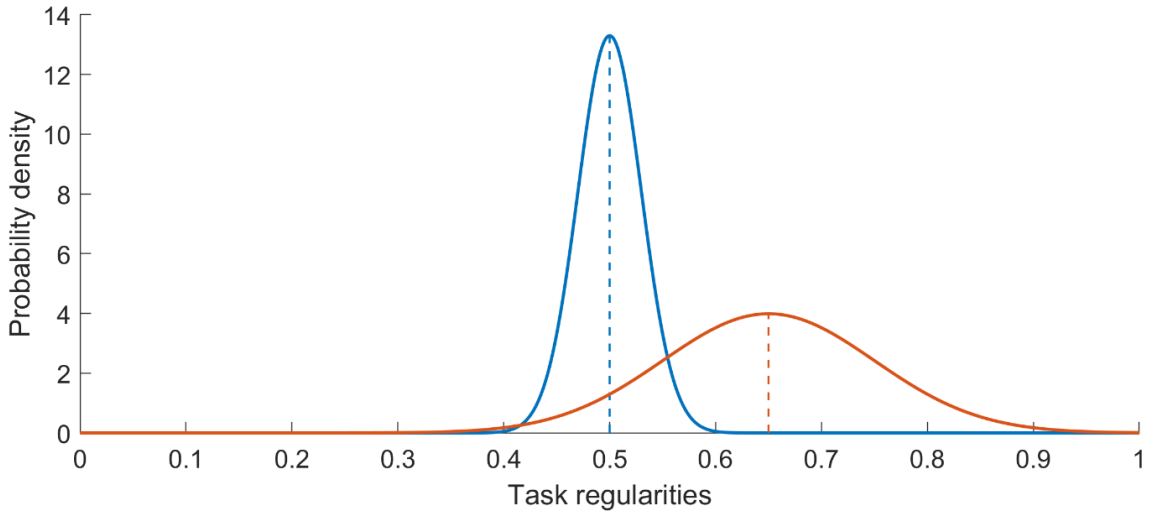

Figure 6.9: Examples of beliefs about task regularities. *The blue line corresponds a strong belief around 0.5, the red line corresponds to a weak belief around 0.65. Dashed lines show the mean of the distributions. In the Hierarchical Gaussian Filter these priors are encoded in the second-level of the 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 potentially hinted at slightly weaker, flatter second-level priors. If this relationship is verified in other studies, it could support the original hypothesis of Pellicano & Burr (2012), but also support 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. This theory might explain how simple measures of response biases commonly employed in behavioural experiments are unable to detect such differences (Angeletos Chrysaitis & Seriès, 2022).

Building on that, our results hinted that this relationship might go away when participants are alerted about the presence of regularities. Therefore, it is possible that individuals with strong autistic traits use explicit strategies to compensate for implicit learning differences. In the seminal paper by Lawson et al. (2017), 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. Further studies could investigate the potential of weaker second-level priors and how these are affected by the presence of explicit instructions in a task better designed to uncover the presence or absence differences in second-level priors.

Chapter 6 References

Angeletos Chrysaitis, N., & Seriès, P. (2022). 10 years of Bayesian theories of autism:a comprehensive review. *Neuroscience & Biobehavioral Reviews*, 105022. https://doi.org/10.1016/j.neubiorev.2022.105022

Caljouw, S. R., Veldkamp, R., & Lamoth, C. J. C. (2016). Implicit and Explicit Learning of a Sequential Postural Weight-Shifting Task in Young and Older Adults. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00733

Critchley, H., Daly, E., Phillips, M., Brammer, M., Bullmore, E., Williams, S., Van Amelsvoort, T., Robertson, D., David, A., & Murphy, D. (2000). Explicit and implicit neural mechanisms for processing of social information from facial expressions: A functional magnetic resonance imaging study. *Human Brain Mapping*, *9*(2), 93–105. https://doi.org/10.1002/(SICI)1097-0193(200002)9:2<93::AID-HBM4>3.0.CO;2-Z

Fletcher, P. C., Zafiris, O., Frith, C. D., Honey, R. A. E., Corlett, P. R., Zilles, K., & Fink, G. R. (2005). On the Benefits of not Trying: Brain Activity and Connectivity Reflecting the Interactions of Explicit and Implicit Sequence Learning. *Cerebral Cortex*, *15*(7), 1002–1015. https://doi.org/10.1093/cercor/bhh201

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, *103*(31), 11778–11783. https://doi.org/10.1073/pnas.0602659103

Forkstam, C., & Petersson, K. M. (2005). Towards an explicit account of implicit learning: *Current Opinion in Neurology*, *18*(4), 435–441. https://doi.org/10.1097/01.wco.0000171951.82995.c4

Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 34–54. https://doi.org/10.1037/0278-7393.33.1.34

Hazeltine, E. (1997). Attention and stimulus characteristics determine the locus of motor-sequence encoding. A PET study. *Brain*, *120*(1), 123–140. https://doi.org/10.1093/brain/120.1.123

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kourkoulou, A. (2010). *Implicit learning of spatial context in adolescents and adults with autism spectrum disorder*. Durham University.

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293–1299. https://doi.org/10.1038/nn.4615

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00302

Liu, H., Forest, T. A., Duncan, K., & Finn, A. S. (2023). What sticks after statistical learning: The persistence of implicit versus explicit memory traces. *Cognition*, *236*, 105439. https://doi.org/10.1016/j.cognition.2023.105439

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*(3), 309–332. https://doi.org/10.1016/j.beproc.2004.03.011

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00825

Musen, G., & Treisman, A. (1990). Implicit and explicit memory for visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 127–137. https://doi.org/10.1037/0278-7393.16.1.127

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, *143*(5), 521–542. https://doi.org/10.1037/bul0000097

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 88–94. https://doi.org/10.1037/0278-7393.2.1.88

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 888–896. https://doi.org/10.1037/0278-7393.17.5.888

Song, S., Howard, J. H., & Howard, D. V. (2007). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning & Memory*, *14*(3), 167–176. https://doi.org/10.1101/lm.437407

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). *Bayesian model selection for group studies*. 14.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675. https://doi.org/10.1037/a0037665

VanPatten, B., & Smith, M. (2022). *Explicit and implicit learning in second language acquisition*. Cambridge University Press.

Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. *British Journal of Mathematical and Statistical Psychology*, *46*(2), 339–349. https://doi.org/10.1111/j.2044-8317.1993.tb01020.x

Chapter 6 Supplementary Information

6.A Stimulus Contrast

Running the experiment online means that participants take part under very different lighting conditions. To ameliorate that, we included a contrast staircase in our experiment. The staircase was set to change rapidly during the training trials, so that it would adjust quickly to the participants' conditions, and then to become more stable during the main experiment. We tried to have the staircase reach a low enough value in the training so that biases would be present during the first trials of the experiment. If the contrast was set too high, the participants would have plenty of information from the gratings and would ignore the auditory cue. This had the unintended side-effect that the contrast was set too low and slightly adjusted during the experiment at a rate of 7%-16% (Figure 6.A1). Interestingly, this happened predominantly in the $I^{+}R^{0}$ and $I^{-}R^{+}$ conditions, in which misleading information was given to the participants. Contrast did not differ significantly across conditions ($F = 1.12$, $p = 0.33$, $BF_{10} = 0.05$).

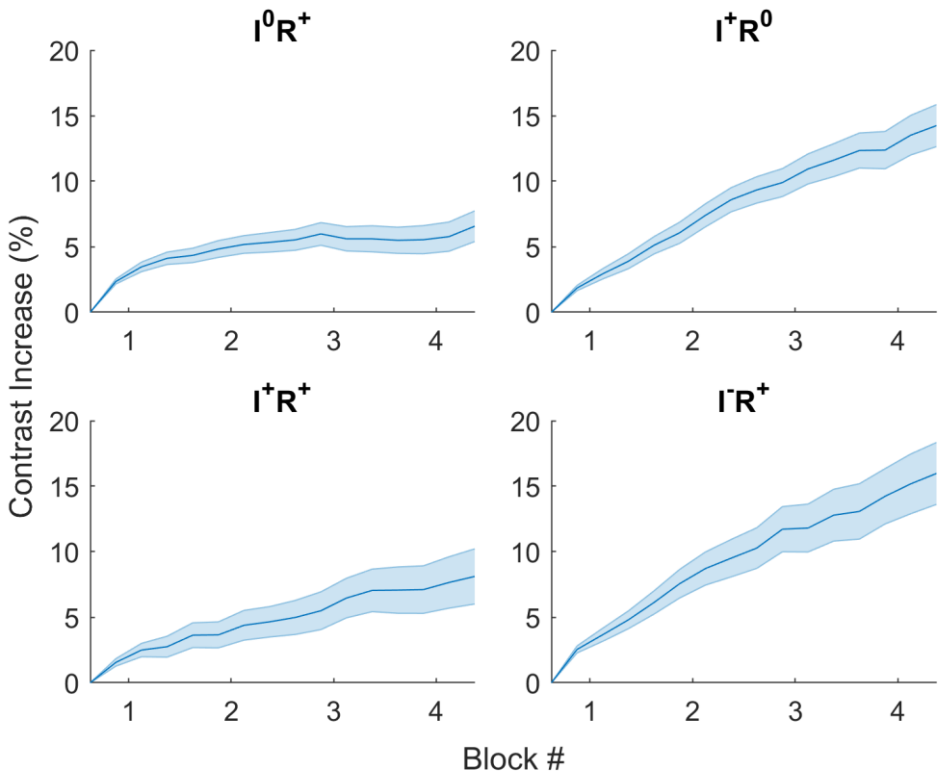

Figure 6.A1: Contrast change during the main experiment.

6.B Model Comparison & Recovery

We compared a great variety of HGF variants using Bayesian model comparison in all four conditions of our task. Specifically, we tested the following alternatives:

- fixing or fitting the parameter k_1
- fixing or fitting the response parameter z
- fixing or fitting the likelihood parameter c_0
- fixing or fitting the likelihood parameter s, which corresponds to the slope of the logistic sigmoid
- HGFs with 2 or 3 levels
- fixing or fitting the parameters ω_2, k_2, and μ^0_3, in HGFs with three levels.

These choices led to comparisons of more than 50 variants of the HGF for each of the conditions. However, in most comparisons the variant used in the main paper significantly outperformed the alternatives (posterior probability > 0.9). Here we present the comparisons of the winning HGF variant with two other variants that were especially interesting. One variant is the same as the winning one, used in the main paper, but with the parameter k_1 fixed, as it commonly is in studies that use the HGF (HGF_nokappa). The other is the same as the winning variant but including a volatility level modulated by ω_2, which is once again common in studies using the HGF (HGF_volatility). We also compared these variants with a baseline Rescorla-Wagner model (RW).

Table 6.B1: Posterior model probabilities.

		Models			
		HGF_winning	**HGF_nokappa**	**HGF_volatility**	**RW**
Conditions	$\mathbf{I^0R^+}$	0.95	0.01	0.01	0.01
	$\mathbf{I^+R^0}$	0.57	0.10	0.31	0.01
	$\mathbf{I^+R^+}$	0.75	0.12	0.06	0.02
	$\mathbf{I^-R^+}$	0.74	0.09	0.08	0.05

To ensure that the results of Table 6.A1 were reliable, we performed model recovery using the same models.

Table 6.B2: Percentage of simulated participants that were recovered by each model.

		Recovered Participants			
Simulated Participants	**I^0R^+**	HGF_winning	HGF_nokappa	HGF_volatility	RW
	HGF_winning	**48**	40	11	1
	HGF_nokappa	35	**64**	1	0
	HGF_volatility	**47**	44	9	0
	RW	0	0	0	**100**
	I^+R^0	HGF_winning	HGF_nokappa	HGF_volatility	RW
	HGF_winning	21	**40**	34	4
	HGF_nokappa	20	**58**	21	0
	HGF_volatility	22	**42**	32	3
	RW	0	0	2	98
	I^+R^+	HGF_winning	HGF_nokappa	HGF_volatility	RW
	HGF_winning	27	**44**	24	5
	HGF_nokappa	23	**66**	12	0
	HGF_volatility	24	**46**	25	5
	RW	0	0	0	**100**
	I^-R^+	HGF_winning	HGF_nokappa	HGF_volatility	RW
	HGF_winning	31	**45**	21	3
	HGF_nokappa	23	**62**	14	0
	HGF_volatility	28	**42**	28	2
	RW	0	0	2	**97**

In Table 6.A2, we can see that in most conditions the participants that were simulated by the winning model were not recovered by the same model. This is partly understandable, as the HGF_nokappa model is the same as the HGF_winning with k_1 fixed at 1. Consequently, there are some participants simulated by the HGF_winning that would have their k_1 values close to 1 and would not require that parameter to be free. However, the fact that on average only 32% of participants simulated by the HGF_winning were correctly recovered is concerning for model identifiability. Nonetheless, this does not raise into question our model comparison results. This is because model recovery points to a bias in model comparison *away* from the winning model, which cannot be a factor in HGF_winning being selected over the other models.

Model recovery also showed that HGF_volatility did not recover well. This is an expected result in a task with no volatile regularities. This means that our model comparison results do not constitute evidence against the HGF_volatility and another task would be necessary to understand to what extend the third level of the HGF is necessary to model human behaviour. Finally, model recovery shows that the HGF and the RW models are clearly identifiable among each other, leading credence to our model comparison results.

6.C Parameter recovery and robustness of model behaviour

Parameter recovery showed good identifiability of the starting point of the prior beliefs, but no identifiability of the corresponding uncertainty (Table 6.C1). It also showed medium recovery for k_1 and good recovery for c_0, which determined the effective contrast for each participant.

Table 6.C1: Pearson correlations between simulated and recovered parameter values

		Parameters			
		mu_0	**sa_0**	**kappa_1**	**c_0**
Conditions	**I^0R^+**	-	0.07	0.45	0.92
	I^+R^0	0.84	0	0.43	0.91
	I^+R^+	0.74	0.05	0.44	0.92
	I^-R^+	0.77	0	0.54	0.92

mu_0 is the starting point of the second-level beliefs, sa_0 is the uncertainty at that point.

As our results did not directly depend on parameter correlations, but on correlations with model trajectories, we also looked at how robust these were between simulated and recovered participants. We found that average priors were somewhat robust in the implicit condition (r = 0.47) and quite robust in other conditions ($r \geq 0.72$). Second-level learning rates were weakly correlated between simulated and recovered participants in the I^0R^+ and I^+R^0 conditions (r = 0.26 & r = 0.33), but mediumly correlated in the I^+R^+ and I^-R^+ ones (r = 0.43 & r = 0.59). Further investigations into the robustness of the second-level learning rates yielded an unexpected result (Figure 6.C1). It appears that the recovered second-level learning rates never exceed 0.028, even when the simulated ones are significantly higher than that. This

phenomenon was present in our main results as well (Figure 6.8). As the lowest learning rates were only present in participants with relatively low AQ in our sample, it is possible that if learning rates could be recovered at any value, the correlation between them and AQ would be stronger than it currently is in our data.

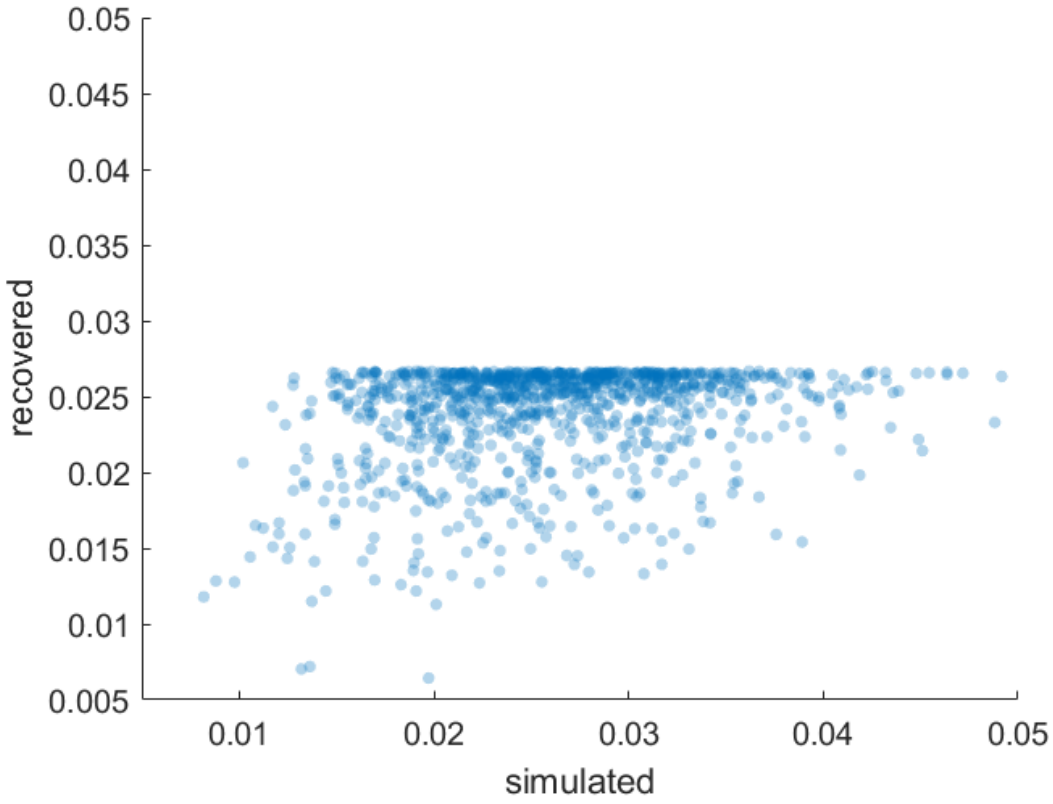

Figure 6.C1: Relationship of second-level learning rates between simulated and recovered participants. *Each dot corresponds to the average second-level learning rate (or variance) of one participant.*

Chapter 7

General Discussion

7.1 Summary of findings

The aims of the present thesis were twofold. The first part, consisting of the first two chapters, was retrospective. In **Chapter 2** we performed an in-depth review of the Bayesian autism literature. We distilled the similarities of the four early theories of autism (Pellicano and Burr, 2012; Brock, 2012; Lawson et al., 2014; Van de Cruys et al., 2014) to the simple 'imbalance hypothesis', which states that the ratio of prior precision to sensory precision is smaller in autism. We found that 51% of the main results showed no significant differences between diagnostic groups or across autistic traits and only 37% exhibited the expected differences, with some showing evidence for the reverse imbalance. Priors learned during the experiments more frequently supported the imbalance hypothesis and even more when no explicit instruction was given to the participants about the nature of regularities. On the other hand, priors that were already developed before the start of the task resulted in reduced biases in autism or autistic traits at a rate of less than 30% when the task involved no social elements. We also found that the studies on average had poor statistical power, with very few of them using computational models to verify their findings, and most of those not following accepted guidelines (Wilson and Collins, 2019). Many studies that did not use computational models were unclear on the nature of the Bayesian processing expected to underlie their results.

In **Chapter 3** we delved into the definition of 'sensory precision'. Researchers have

used the term to refer both to the true reliability of the sensory evidence (what we termed actual sensory precision) and to the weight of this evidence in the agent's internal inferential process (estimated sensory precision). While optimally the two components would be equal to each other, the brain might not be able to perfectly estimate the reliability of the sensory input. Indeed, some theories of autism suggest that estimated sensory precision is divorced from actual sensory precision (Lawson et al., 2014; Van de Cruys et al., 2014). Unfortunately, changes in either of these parameters or in prior precision cannot always be estimated from behavioural data. We highlighted the expected effects of each parameter and reanalysed previously collected data (from Karvelis et al., 2018) in this new light. Our results revealed that the actual sensory precision was clearly identifiable in the Moving Dots task, with the prior precision being somewhat identifiable and estimated sensory prevision having poor identifiability. They also supported the original study conclusions of increased actual sensory precision with stronger autistic traits.

Part two of the thesis included the final three chapters and focused on new approaches in the study of autism. In **Chapter 4** we investigated the presence of circular inference across autistic traits and diagnoses, inspired by the recent success of that theory in schizophrenia (Jardri et al., 2017) and the similarities between the two conditions. We analysed our data using a model based on belief propagation, an alternative to predictive coding (Denève and Jardri, 2016). Our results showed no difference in either circularity or the weighting of priors and likelihoods across the autism spectrum. Then, in **Chapter 5**, we designed a follow-up study using a task with social elements to further investigate circularity in autism, in a setting more relevant to the condition. Despite various ways of emphasising the structure of the task and how participants could optimally approach it, our pilot behavioural results and follow-up interviews with the participants showed that participants did not perform Bayesian inference. Instead, they mostly averaged between prior and likelihood. Because of this we decided to not pursue a full study with measures of autistic traits.

In **Chapter 6** we designed a task to investigate implicit and explicit learning across autistic traits. We included four conditions, with different combinations of regularities and instruction veracity. We saw that participants learned the regularities of the envi-

ronment more quickly and more strongly when their attention was drawn to them by the instructions. Surprisingly, the specific content of the instructions had only temporary and minor effects on the beliefs formed by the participants. Using the Hierarchical Gaussian Filter (HGF) to model the data, we also found that individuals with stronger and weaker autistic traits were similarly influenced by their priors in each individual trial, but our results also hinted at a potential relationship between autistic traits and increased uncertainty in beliefs about the experimental regularities in the implicit condition.

7.2 Study Pitfalls

One common thread throughout this thesis has been the numerous pitfalls that a scientific study can fall into and the difficulties of doing robust scientific work. Some of these pitfalls are general methodological difficulties, such as low statistical power, or systemic issues such as publication bias. These became well-known as potential causes of the replication crisis of the 2010s, when it was realised that many published studies fail to replicate, undermining their trustworthiness (Ioannidis, 2005; Shrout and Rodgers, 2018). This was true in all fields, including psychiatry (e.g., Border et al., 2019), but was most prominent in psychology where replications showed effect sizes that were only half of those in the original studies and an overall replication rate of 37% (Open Science Collaboration, 2015).

In Chapter 2 we showed that such issues were common among the studies we reviewed. The estimated overall statistical power of our pool of studies was only 39%, far from the commonly recommended threshold of 80%. This means that a true effect has only 39% chance of being statistically significant, which in turn lowers the percentage of published results that are true and makes the overall literature of the field less reliable. Fortunately, we found that sample sizes have slowly been improving over the years, from a mean of 37 in 2013-2017 to a mean of 63 in 2018-2021. We also found some weak evidence for selection effects (Hedges, 1992). These can manifest as journal preferences for positive results, self-censorship from researchers who do not attempt to publish null results, or even prioritisation within an article with null results being sidelined post hoc and secondary positive results being presented as the findings of interest. Furthermore, none of the

studies we reviewed were direct replications of previous studies. A few labs run the same task in different populations (ASD vs autistic traits: Amoruso et al., 2019; Bianco et al., 2020; Lawson et al., 2017; Pell et al., 2016) or in the same population but with slight adjustments in the experimental design (Sapey-Triomphe et al., 2021b,a), but none of them had all of their positive findings replicate. Finally, only three of the studies we reviewed were preregistered (Lacroix et al., 2021; Retzler et al., 2021; Tewolde et al., 2018), none finding differences in the expected direction.

The second set of pitfalls is more specific to the field of computational psychiatry and particularly the Bayesian and predictive coding theories of autism. These theories are at their core computational. They view the brain from an information processing standpoint, mental illnesses as alterations in these processes that depart from mathematical optimality, and are expressed using mathematical concepts and terminology. The appeal of such an approach, besides the attempt to bridge the gap between neuroscience and psychology, is the precision that it offers due to its quantitative nature. In principle, formal, theory-driven models make it easier to generate experimental hypotheses from theories, allow for the subtle differentiation between similar hypotheses as they are mathematically connected to the data, and result in clearer communication between researchers (Fried, 2020; Oberauer and Lewandowsky, 2019; Robinaugh et al., 2021; but also see Oude Maatman, 2021). For these reasons, they have been proposed as a potential way to ameliorate the theory and replication crises in psychology. Despite that, in Chapter 2 we saw that most studies that investigate these theories do not use theory-driven computational models of any kind.

Unfortunately, even the studies that employ computational models do so without following best practices (Wilson and Collins, 2019). There are many possible models that can broadly emulate participant behaviour. Therefore, for computational modelling to provide more accurate insights into the workings of the brain, the models have to be compared with each other and the one that is most likely based on the data to be selected. The reliability of this process has to also be verified, most commonly through model recovery. On the opposite side, relying exclusively on model comparison might result in misleading results if none of the chosen models reflects participant behaviour. For this reason, the behaviour of the chosen model has to be inspected and compared

with the behavioural data to assure its validity. Within a given model, parameter estimation has to also be verified via parameter recovery. This is relevant for the studies reviewed, because the most common way that computational models are used to support or reject the Bayesian theories of autism is the estimation of model parameters and their comparison across diagnostic groups or autistic traits. However, our review in Chapter 2 showed that a limited number of computational studies followed these practices, potentially contributing to the low reliability of published results in the field (Karvelis et al., 2023).

An additional issue revolves around the fact that computational theories appear to be underdetermined (Oude Maatman, 2021). This means that many steps lie between a proposed theory and the specific computational hypotheses tested in experiments, presenting researchers with choices between equally valid alternatives. In Chapter 3, we showed how studies make different assumptions when searching for evidence for the theories of autism. For example, some studies equate actual and estimated sensory precision, while others distinguish between them. Some studies assume that percepts are formed based on the mean of the posterior distribution, while others assume they are sampled from it. These degrees of freedom could increase the apparent inconsistency of the results in the field, as such assumptions are often not explicitly mentioned. The inconsistency could be further augmented by liberal interpretations of behavioural results in non-modelling studies. This phenomenon could be partially attributed to the imprecision of language. Due to the interdisciplinarity of the field, Bayesian theories and findings are presented to a broad audience that might not be completely familiar with mathematical terminology.

A final set of pitfalls regards experimental design and expected participant behaviour. In Chapter 5, we saw that not all tasks necessarily evoke Bayesian strategies in the participants. Instead, in that task, participants were influenced by the two sources of information that were intended to take the role of the prior and the likelihood, but combined them in an apparently non-Bayesian way. This, as we observed, was not easily discernible from the data. For most studies, it is unlikely that participants specifically follow the averaging strategy we observed in Chapter 5, as this seems to be mostly applicable to the studies that provide explicit probabilities to participants in each trial. However, other non-Bayesian strategies might be possible. Moreover, even strategies that

are based on Bayesian inference might deviate enough to not be well-approximated by the usual Bayesian models. An example of that is circular inference, which leads to prior and likelihood overcounting and which cannot be accounted for by a simple weighted Bayesian model, as we saw in Chapter 4. In general, greater care should be devoted to discovering if participants are behaving in a Bayesian way and how deviations from that process can be modelled.

In Chapter 6, we showed that instructions can exert a large influence on the acquisition of priors by making their learning explicit. This was the case, even when the specific information about the regularities was false, showing how simply drawing the attention of the participants to the regularities is enough to change the nature of the task. Such effects are not surprising, as various past studies have shown the influence of instructions in perception (e.g., Teigen, 1977; Li and Warren Jr, 2004; Stanley et al., 2010; Kirsch, 2021). Even so, our study was the first distinguish clearly between the presence and the content of the instructions and to clearly document and model these effects in a Bayesian task. Such effects might be of particular importance in the study of autism. The weak evidence we found for faster learning in strong autistic traits was present only when the task was implicit and not when explicit instructions were given about the nature of the regularities, although the difference between conditions was not significant in our sample. Additionally, the difficulties of autistic individuals to understand others and their attention to detail might exacerbate potential differences between the groups. Unfortunately, task instructions are not reported in the majority of studies, nor are they considered an important factor at play. Experiment code is rarely shared along with the data or the analysis code, so that other researchers can understand and replicate the task in detail.

In our experimental studies (Chapters 3, 4, and 6) we tried to avoid these pitfalls. We used large samples (n = 173 & n = 335, with 51 to 119 participants per condition) and highlighted our null results, both when they were the central finding and when they were one part of a nuanced picture. We fit behavioural data with computational models and compared both different models and variants of the same model. We used model and parameter recovery, as well as visual inspections of model behaviour, to ensure the validity of these processes and call attention to their shortcomings. We made our modelling

assumptions explicit and analysed our data under models that represented different sets of assumptions. We were also clear on the instructions given to the participants and made our experiment code available along with our data and analysis code.

7.3 Towards more nuanced Bayesian theories of autism

The early Bayesian or predictive coding autism theories were relatively simple and broad in their formulation. Pellicano and Burr's theory suggested that priors are generally flatter in autism compared to the general population (Pellicano and Burr, 2012). This does not mean that no autistic prior can be stronger than corresponding a neurotypical one, but it implies that no category of priors should systematically diverge from this pattern. Brock proposed an alternative hypothesis that was similarly broad: that likelihoods in autism are uniformly narrower (Brock, 2012). In the same vain, Lawson et al. hypothesised the presence of a general imbalance between the precision of priors and sensory evidence (Lawson et al., 2014).

In Chapter 2, we reviewed the available evidence for these theories from 83 studies, which showed that such broad formulations are likely not accurate. A slight majority of the results showed no differences between diagnostic groups or autistic traits, with the median variance explained by these groups or traits being less than 4%. If a broad precision imbalance lied at the core of autism, we would expect stronger effect sizes. This was also supported by our experimental work in Chapters 4 and 6, which showed no differences in prior influences across autistic traits. On the other hand, our Chapter 2 results also showed that these theories cannot be dismissed altogether. More than a third of the findings were consistent with a precision imbalance in autism, with some additional studies that did not directly investigate the presence of the imbalance showing differences in prior development. Furthermore, Bayesian and predictive coding theories are natural continuations of previous theories, such as Weak Central Coherence and Enhanced Perceptual Functioning (Frith, 1990; Mottron et al., 2006), which themselves relied on a variety of experimental observations. How can we align these apparently contradicting facts?

We consider four possible explanations. One is that not every task necessarily evokes

Bayesian processes in the participants, and the effects of autism appear only on the ones that do. In Chapter 5 we presented a task that required a Bayesian approach to maximise performance and with instructions that actively attempted to steer participants towards this mode of thinking, that however resulted in participants employing non-Bayesian strategies. Despite this, we find this first explanation unlikely based on the wide variety of tasks in the relevant literature. Assuming that non-Bayesian information processing is the rule rather than the exception would go against numerous studies to the contrary (e.g., Ernst and Banks, 2002; Weiss et al., 2002; Alais and Burr, 2004; Hedges et al., 2011; Stocker and Simoncelli, 2006; Girshick et al., 2011). It would also be directly contradicted by some of the included studies that tested that claim and found modelling evidence in favour of Bayesian processes (e.g. Karvelis et al., 2018; Manning et al., 2017). Nonetheless, further research is needed to understand the tasks in which participants employ alternative strategies and what these strategies are.

A second explanation is that differences between autistic and non-autistic individuals may not necessarily extend to differences across autistic traits. As we saw in Chapter 2, most of the Bayesian studies of autism examine trait correlations in the general population rather than comparisons between diagnostic groups, with the minority of clinical population studies yielding more positive findings. Our two experimental studies also largely focused on autistic traits as they allowed us to gather data from large sample sizes, particularly during the COVID-19 pandemic. However, we do not expect that this is a major factor behind our results. First, as discussed in section 1.3, the dimensional view of autism is well-supported, especially for low-level information processing tasks rather than ones that investigate high-level behaviour. Second, the significant differences between diagnostic groups in our review were frequently found in neuroimaging studies, which generally showed more positive findings than behavioural studies. Additionally, our review revealed numerous autistic trait studies with positive findings, showing similar patterns across prior categories. Finally, comparisons in our studies based on self-reported participant diagnoses did not yield positive findings either. These suggest that the differences between autistic trait and diagnostic studies cannot fully explain our results and might stem from the prevalence of neuroimaging designs or potentially the smaller sample sizes in autism studies, which would lead to more type I errors.

The third possibility is that the presence of the precision imbalance depends on the task domain. This could arise because the effects of autism are domain-limited. For example, as demonstrated in Chapter 2, tasks with social elements might show greater differences among diagnostic groups compared to non-social tasks. It could also be an indication of a lack of convergent validity in prior or likelihood precision (Karvelis et al., 2023). That is, findings of reduced or increased precision might not generalise from one task to another. Indeed, studies using multiple behavioural tasks have failed to find a single factor corresponding to reliance on priors in the context of autism (Andermane et al., 2020; Tulver et al., 2019). While this theory has certainly some explanatory power, it cannot fully account for our findings. The fact that autistic individuals had weaker priors in a significant percentage of a very heterogeneous pool of tasks is evidence that these share some underlying factor that is influenced by the condition. We also have ample evidence against strict domain-specificity for the autistic symptomatology. As autism is characterised by a multitude of sensory symptoms, it not unreasonable to expect weaker priors in non-social tasks. Indeed, positive results were only 50% more frequent in social tasks vs their non-social counterparts in Chapter 2. Insisting on this approach, one could suggest that the social and non-social categories are too broad and conclusions can be drawn only on a level that encompasses more narrow categories of tasks. However, any theory that uses this as its starting point should be able to explain in which tasks the precision imbalance appears and why these tasks specifically exhibit similar patterns. An additional complicating factor could be the heterogeneity of the condition itself, which adds noise to any behavioural result. However, large sample sizes should be able to uncover differences between groups, even if those differences affect only a portion of the participants. Our systematic review showed no relationship between sample size and the frequency of positive results, suggesting that this is not a significant cause of the apparent contradiction.

This brings us naturally to our third explanation: that more complicated differences in information processing in the brain underlie the autistic symptomatology and that correspondingly nuanced Bayesian theories are needed to understand the condition. Two such theories have been proposed: the High, Inflexible Precision of Prediction Errors in Autism (HIPPEA) account of Van de Cruys et al. Van de Cruys et al. (2014) and

the volatility processing impairments account of Palmer et al. Palmer et al. (2017). According to HIPPEA, autistic individuals' weighting of sensory evidence is not affected by context. Instead, sensory evidence is always treated as important new information by the brain. This comes in contrast to normal processing, in which sensory precision is adjusted based on the environmental noise, with sensory inputs from noisy environments being downweighted as they contain less information. This mirrors our framework in Chapter 3, in which estimated sensory precision should be equal to the actual sensory precision. The authors further argue that the inflexibly high estimated precision leads autistic individuals to overestimate the volatility of the environment, as they do not attribute the variance of sensory inputs to noise, but to the environment changing. In the theory of Palmer et al., on the other hand, autistics are unable to correctly track the volatility of the environment. But sensory information should be taken more into account when the environment is changing, as in that case there is a greater need for the updating of prior beliefs. Because of this, improper volatility beliefs make autistic individuals incapable of properly adjusting the precision of bottom-up information.

As one can see, these theories are largely similar. Both predict that Bayesian impairments in autism are not equally expressed in all experiments, with both sets of authors suggesting that they would be more prominent in volatile environments. Where they differ is that HIPPEA views volatility overestimation as the result of high bottom-up precision, while the theory of Palmer et al. views differences in volatility tracking as the cause of the precision differences. Moreover, Palmer et al. do not specify the nature of the volatility estimation impairments, although a later study building upon this theory suggests that autistics specifically overestimate the volatility of the environment (Lawson et al., 2017).

Few studies have directly tested these theories. Sapey-Triomphe et al. (2021b) showed that the influence of priors differed less in autism compared to neurotypical controls between conditions with different environmental statistics. The authors suggest that his finding supports HIPPEA, but since no theory-driven modelling is employed, it is uncertain if this effect results from less flexible estimated sensory precision, as opposed to less flexible prior precision. In Chapter 2, we found three studies that used the HGF to model participants with autism or high autistic traits (Lawson et al., 2017; Sapey-

Triomphe et al., 2021b; Sevgi et al., 2020). Lawson et al. found increased ω_3 values in autistic participants, a parameter that determines how fast their volatility beliefs change (i.e., meta-volatility). They also found that autistic participants adjusted their learning about the regularities less than controls when regularities became more volatile, but they adjusted their learning about volatility more. Sevgi et al. showed that, in individuals with high AQ scores, the congruency of environmental cues affected less how much the volatility estimates influence their beliefs (κ_2). Sapey-Triomphe et al. (2021) found no relationship between diagnostic groups and either ω_2 (baseline or tonic volatility) or ω_3 (meta-volatility). In addition to these studies, in Chapter 2 we found that differences among groups were most common when priors were learned during the experiment. This is consistent with impairments in volatility processing, since estimating how frequently the environment changes directly affects learning in that environment. In Chapter 3, we showed that autistic traits positively correlate with actual sensory precision, which could speculatively be interpreted as an increased rate of sensory sampling from the environment due to volatility overestimation, in line with the theory of Palmer et al. In Chapter 6, we found a tendency for higher HGF-estimated learning rates about implicit regularities, which could arise from a stronger belief that the environment is changing. However, the regularities in both tasks were stable, making more uncertain any interpretations regarding volatility.

It is difficult to draw a coherent picture from these results. While all but one suggest impairments in volatility processing, the nature of this impairment and its expected effects differ greatly among studies. The theory that autistic individuals overestimate the volatility in their environment has become popular recently (e.g. Kreis et al., 2021; Lawson et al., 2017), but out of the aforementioned results only ours support it, albeit in an indirect way. This, once again, could be a sign of the theories being underdetermined (Oude Maatman, 2021). Presumably, a theory of volatility overestimation would predict higher average volatility beliefs (χ_3, in HGF terms), higher baseline (tonic) volatility (ω_2), larger learning rates, and potentially higher phasic volatility (κ_2). However, most of these effects were not tested in the aforementioned studies and only higher learning rates were shown in their results.

One of the most common predictions associated with the Palmer et al. theory is, once

again, that of reduced biases. However, simply overestimating environmental volatility does not automatically result in weaker biases. For example, in the HGF, biases would be slightly increased when the environment is believed to be more volatile, as participants learn the regularities quicker. This happens despite high volatility leading participants to be less certain about the state of environmental regularities. We explained this phenomenon in Chapter 6 (see Figure 6.9), after we showed that the slightly increased learning rates with high AQ in our sample did not result in differences in the strength of the priors utilised by participants in each trial. In short, in the HGF, a participant's first-level prior (their belief about a singular trial) is determined by the mean of their second-level prior (their belief about the general regularities). Having less certain second-level beliefs has no effect on the first-level prior. For example, one could be extremely certain that they are holding a coin that comes up with tails with exactly 75% probability (a second-level belief) or they could be much more uncertain and think that the coin's probability for tails is uniformly in the range of 55% to 95%. In both cases, when they flip the coin, their expectation should be that will see tails with 75% probability (first-level belief). It is possible that alternative models predict reduced biases as a result of higher volatility estimates. However, the fact that the HGF has been the only model used to investigate the theory of Palmer et al. to this day points again to the problem of underdetermination and implicit assumptions.

Similar problems in the relationship between theory and predictions also appear in HIPPEA. According to the authors, HIPPEA predicts that autistic individuals would cope well in stable environments and show impairments in volatile ones. However, it is not clear why that would be the case. If autistics have permanently high estimated sensory precision, then they would be better suited for volatile environments, where the benefit of new information is higher and the priors less relevant. It is in stable environments in which constant updating would impair everyday function.

These considerations do not necessarily imply that the theories are not in accordance with the behavioural results. The nuance of the proposed accounts provides them with a flexibility that could conceptually fit with most results. These issues are simply another expression of underdetermination, requiring the theories to be further specified, so that their mechanisms are widely understood among researchers and so that they unambigu-

ously generated falsifiable predictions. Besides these two theories, there are additional ways that the Bayesian autism research could be nuanced. In Chapter 4, we investigated circular inference across autistic traits, a different potential impairment in probabilistic inference that goes beyond simple prior and likelihood weightings to account for feedback loops in the inferential process. This theory would predict that the distortion in the representation of priors or likelihoods in the brain is not uniform, but instead depends on the certainty of each signal and the interaction between them. Our results unambiguously showed no differences in circularity across autistic traits, but, interestingly, they also showed no differences in the weights of priors or likelihoods. Importantly though, our task was affected by a few methodological limitations. In particular, the fact that it could be conceived as a cue-combination task rather than one in which priors are combined with likelihoods raises potential issues for autism theories that specifically revolve around prior-likelihood combinations. For our results to be generalisable, they would first need to be verified in additional experimental settings, better aligned with the hypothesised impairments in autism.

An additional way in which complexity could be incorporated in the Bayesian hypotheses for autism could be through the characteristics of the experimental designs. Our review revealed a slightly increased frequency of positive results in implicit learning as opposed to learning where the participants are directly informed on the presence of regularities in the environment. In Chapter 6, we saw that participants behaved very differently depending on the information disclosed in instructions about the nature of the task. This suggests that different mechanisms might drive participant behaviour depending on the task instructions and therefore not all experiments would necessarily exhibit the same differences across autistic traits or diagnostic groups. We also found weak evidence for a marginal correlation between AQ and learning rates in the implicit condition that was not present when the regularities were made explicit (although the difference in correlations across conditions was not significant). While we cannot conclude this from our results, it is possible that individuals with strong autistic traits find alternative strategies to compensate with learning impairments when they are aware of the presence of regularities. This might also explain why the Lawson et al. Lawson et al. (2017) study showed differences in beliefs about meta-volatility and not in the volatility itself. Participants were

explicitly informed of the association between cue and stimulus and also that this might change during the experiment. However, they were not informed of potential changes in volatility, making their learning of them implicit. A final source of nuance could arise from the details of Bayesian modelling itself as it applies to behavioural experiments. Actual and estimated sensory precisions perform different roles in perceptual inference and distinguishing between them would allow for a deeper understanding of behavioural findings and could potentially aid in resolving the apparent contradictions between them.

In summary, early theories of general precision imbalances in autism provide valuable insights into autistic behavior, though they fall short of explaining the full complexity of the condition. More nuanced theories show greater promise, albeit with room for more specific predictions. The field of computational psychiatry and particularly the Bayesian study of autism spectrum disorders is still in its infancy, but is developing rapidly, as evident in the increasing number of computational approaches and the emergence of more detailed, realistic theories. Experimental standards are rising within the field, reflecting a similar movement across psychology, psychiatry, and neuroscience. Moving forward, studies should strive to offer clear explanations of Bayesian impairments and their consequences, potentially using strict mathematical language, so that hypotheses can be deterministically generated from the corresponding theories. These hypotheses should then be tested through behavioural experiments that make thorough use of computational models, being transparent about their assumptions and strong in their justification. Behavioural studies could also benefit from distinguishing between implicit tasks, which likely invoke Bayesian processes in participants, and explicit ones that might employ alternative strategies. This approach could pave the way for a more comprehensive understanding of autism and Bayesian cognition in general.

Bibliography

Abu-Akel, A., Allison, C., Baron-Cohen, S., and Heinke, D. (2019). The distribution of autistic traits across the autism spectrum: evidence for discontinuous dimensional subpopulations underlying the autism continuum. *Molecular autism*, 10:1–13.

Aitchison, L. and Lengyel, M. (2017). With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227.

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262.

American Psychiatric Association, editor (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, fifth edition edition.

Amoruso, L., Narzisi, A., Pinzino, M., Finisguerra, A., Billeci, L., Calderoni, S., Fabbro, F., Muratori, F., Volzone, A., and Urgesi, C. (2019). Contextual priors do not modulate action prediction in children with autism. *Proceedings of the Royal Society B: Biological Sciences*, 286(1908):20191319.

Andermane, N., Bosten, J. M., Seth, A. K., and Ward, J. (2020). Individual differences in the tendency to see the expected. *Consciousness and Cognition*, 85:102989.

Austin, E. J. (2005). Personality correlates of the broader autism phenotype as assessed by the autism spectrum quotient (aq). *Personality and individual differences*, 38(2):451–460.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, 23:169–184.

Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., and Chakrabarti, B. (2009). Talent in autism: hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1522):1377–1383.

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

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen,

M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152.

Bianco, V., Finisguerra, A., Betti, S., D'Argenio, G., and Urgesi, C. (2020). Autistic Traits Differently Account for Context-Based Predictions of Physical and Social Events. *Brain Sciences*, 10(7):418.

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., and Keller, M. C. (2019). No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry*, 176(5):376–387.

Brewster, D. (1826). On the optical illusion of the conversion of cameos into intaglios, and of intaglios into cameos, with an account of other analogous phenomena. *Edinburgh Journal of Science*, 4(7).

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr. *Trends in Cognitive Sciences*, 16(12):573–574.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., and Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, 18(4):590–596.

Caplin, A. and Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203.

Carroll, L. S. and Owen, M. J. (2009). Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome medicine*, 1:1–7.

Chambon, V., Pacherie, E., Barbalat, G., Jacquet, P., Franck, N., and Farrer, C. (2011). Mentalizing under influence: abnormal dependence on prior expectations in patients with schizophrenia. *Brain*, 134(12):3728–3741.

Chater, N., Oaksford, M., Hahn, U., and Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, 1(6):811–823.

Christ, S. E., Kester, L. E., Bodner, K. E., and Miles, J. H. (2011). Evidence for selective inhibitory impairment in individuals with autism spectrum disorder. *Neuropsychology*, 25(6):690.

Cohen, D. J. and Volkmar, F. R. (1997). *Handbook of autism and pervasive developmental disorders*. John Wiley & Sons, Inc.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., and Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry*, 45(4):719–726.

Constantino, J. N., Lajonchere, C., Lutz, M., Gray, T., Abbacchi, A., McKenna, K., Singh, D., and Todd, R. D. (2006). Autistic social impairment in the siblings of children with pervasive developmental disorders. *American Journal of Psychiatry*, 163(2):294–296.

Constantino, J. N. and Todd, R. D. (2003). Autistic traits in the general population: a twin study. *Archives of general psychiatry*, 60(5):524–530.

Couture, S., Penn, D., Losh, M., Adolphs, R., Hurley, R., and Piven, J. (2010). Comparison of social cognitive functioning in schizophrenia and high functioning autism: more convergence than divergence. *Psychological medicine*, 40(4):569–579.

Dawson, G., Meltzoff, A. N., Osterling, J., Rinaldi, J., and Brown, E. (1998). Children with autism fail to orient to naturally occurring social stimuli. *Journal of autism and developmental disorders*, 28:479–485.

Denève, S. and Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences*, 11:40–48.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Eyring, K. W. and Geschwind, D. H. (2021). Three decades of asd genetics: building a foundation for neurobiological understanding and treatment. *Human Molecular Genetics*, 30(20):R236–R244.

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, 31(4):271–288.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221.

Frith, C. and Dolan, R. J. (1997). Brain mechanisms associated with top-down processes in perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1221–1230.

Frith, U. (1990). *Autism: explaining the enigma*. Cognitive development. Blackwell, Oxford u.a, repr edition. OCLC: 18716337.

Furnham, A. and Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

Gernsbacher, M. A., Stevenson, J. L., and Dern, S. (2017). Specificity, contexts, and reference groups matter when assessing autistic traits. *PloS one*, 12(2):e0171931.

Gigerenzer, G. and Hoffrage, U. (1999). Overcoming difficulties in bayesian reasoning: a reply to lewis and keren (1999) and mellers and mcgraw (1999). *Psychological Review*.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932.

Gogolla, N., LeBlanc, J. J., Quast, K. B., Südhof, T. C., Fagiolini, M., and Hensch, T. K. (2009). Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *Journal of neurodevelopmental disorders*, 1:172–181.

Gregory, R. L. (1970). *The intelligent eye*.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian Models of Cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*, pages 59–100. Cambridge University Press, 1 edition.

Happé, F. G. (1996). Studying weak central coherence at low levels: children with autism do not succumb to visual illusions. a research note. *Journal of child psychology and psychiatry*, 37(7):873–877.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., and Moore, T. (2009). Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Social Cognition*, 27(5):733–763.

Hazen, E. P., Stornelli, J. L., O'Rourke, J. A., Koesterer, K., and McDougle, C. J. (2014). Sensory symptoms in autism spectrum disorders. *Harvard review of psychiatry*, 22(2):112–124.

Hedges, J. H., Stocker, A. A., and Simoncelli, E. P. (2011). Optimal inference explains the perceptual coherence of visual motion stimuli. *Journal of vision*, 11(6):14–14.

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science*, 7(2).

Hoekstra, R. A., Bartels, M., Verweij, C. J., and Boomsma, D. I. (2007). Heritability of autistic traits in the general population. *Archives of pediatrics & adolescent medicine*, 161(4):372–377.

Huys, Q. J., Daw, N. D., and Dayan, P. (2015). Depression: A Decision-Theoretic Analysis. *Annual Review of Neuroscience*, 38(1):1–23.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.

Jardri, R. and Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, 136(11):3227–3241.

Jardri, R., Duverne, S., Litvinova, A. S., and Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, 8(1):14218.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin psychology. Penguin Books, London.

Kana, R. K., Keller, T. A., Minshew, N. J., and Just, M. A. (2007). Inhibitory control in high-functioning autism: decreased activation and underconnectivity in inhibition networks. *Biological psychiatry*, 62(3):198–206.

Kanizsa, G. (1987). Quasi-Perceptual Margins in Homogeneously Stimulated Fields. In Petry, S. and Meyer, G. E., editors, *The Perception of Illusory Contours*, pages 40–49. Springer New York, New York, NY.

Karvelis, P., Paulus, M. P., and Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, 148:105137.

Karvelis, P., Seitz, A. R., Lawrie, S. M., and Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *eLife*, 7:e34115.

Kirsch, W. (2021). On the relevance of task instructions for the influence of action on perception. *Attention, Perception, & Psychophysics*, 83(6):2625–2633.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Kreis, I., Biegler, R., Tjelmeland, H., Mittner, M., Klæbo Reitan, S., and Pfuhl, G. (2021). Overestimation of volatility in schizophrenia and autism? A comparative study using a probabilistic reasoning task. *PLOS ONE*, 16(1):e0244975.

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., Bach, B., et al. (2018). Progress in achieving quantitative classification of psychopathology. *World psychiatry*, 17(3):282–293.

Lacroix, A., Dutheil, F., Logemann, A., Cserjesi, R., Peyrin, C., Biro, B., Gomot, M., and Mermillod, M. (2021). Flexibility in autism during unpredictable shifts of socio-emotional stimuli: Investigation of group and sex differences. *Autism*, page 136236132110627.

Landry, O. and Chouinard, P. A. (2019). Why we should study the broader autism phenotype in typically developing populations. In *Building Bridges: Cognitive Development in Typical and Atypical Development*, pages 36–47. Routledge.

Lange, A. and Dukas, R. (2009). Bayesian approximations and extensions: Optimal decisions for small brains and possibly big ones too. *Journal of Theoretical Biology*, 259(3):503–516.

Lawson, R. P., Mathys, C., and Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, 20(9):1293–1299.

Lawson, R. P., Rees, G., and Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8.

Leimar, O. and McNamara, J. M. (2019). Learning leads to bounded rationality and the evolution of cognitive bias in public goods games. *Scientific Reports*, 9(1):16319.

Li, L. and Warren Jr, W. H. (2004). Path perception during rotation: Influence of instructions, depth range, and dot density. *Vision research*, 44(16):1879–1889.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., and Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25(1):322–349.

Lin, C.-H. S., Do, T. T., Unsworth, L., and Garrido, M. I. (2024). Are we really bayesian?

probabilistic inference shows sub-optimal knowledge transfer. *PLOS Computational Biology*, 20(1):e1011769.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223.

Lord, C., Rutter, M., and LeCouteur, A. (1994). Autism diagnostic interview-revised (adi-r) journal of autism and developmental disorders, 24, 659-685. *Trad fr. Leboyer M.*

Luan, S., Reb, J., and Gigerenzer, G. (2019). Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Academy of Management Journal*, 62(6):1735–1759.

Ma, W. J. (2019). Bayesian Decision Models: A Primer. *Neuron*, 104(1):164–175.

Manning, C., Kilner, J., Neil, L., Karaminis, T., and Pellicano, E. (2017). Children on the autism spectrum update their behaviour in response to a volatile environment. *Developmental Science*, 20(5):e12435.

Markram, K. and Markram, H. (2010). The Intense World Theory – A Unifying Theory of the Neurobiology of Autism. *Frontiers in Human Neuroscience*, 4.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: W.H. Freeman and Company.

Mottron, L. and Bzdok, D. (2020). Autism spectrum heterogeneity: fact or artifact? *Molecular psychiatry*, 25(12):3178–3185.

Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced Perceptual Functioning in Autism: An Update, and Eight Principles of Autistic Perception. *Journal of Autism and Developmental Disorders*, 36(1):27–43.

Oberauer, K. and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5):1596–1618.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

O'riordan, M. A., Plaisted, K. C., Driver, J., and Baron-Cohen, S. (2001). Superior visual search in autism. *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):719.

Oude Maatman, F. (2021). Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It.

Palmer, C. J., Lawson, R. P., and Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, 143(5):521–542.

Pell, P. J., Mareschal, I., Calder, A. J., von dem Hagen, E. A. H., Clifford, C. W., Baron-Cohen, S., and Ewbank, M. P. (2016). Intact priors for gaze direction in adults with high-functioning autism spectrum conditions. *Molecular Autism*, 7(1):25.

Pellicano, E. (2012). The development of executive function in autism. *Autism research and treatment*, 2012(1):146132.

Pellicano, E. and Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10):504–510.

Peters, E., Joseph, S., Day, S., and Qarety, P. (2004). Measuring Delusional Ideation: The 21-Item Peters et aL Delusions Inventory (PDI). *Schizophrenia Bulletin*, 30(4):18.

Powell, G., Bompas, A., and Sumner, P. (2012). Making the incredible credible: After-images are modulated by contextual edges more than real stimuli. *Journal of Vision*, 12(10):17–17.

Powers, A. R., Mathys, C., and Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351):596–600.

Prat-Carrabin, A. and Woodford, M. (2024). Imprecise probabilistic inference from sequential data. *Psychological Review*.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Retzler, C., Boehm, U., Cai, J., Cochrane, A., and Manning, C. (2021). Prior information use and response caution in perceptual decision-making: No evidence for a relationship with autistic-like traits. *Quarterly Journal of Experimental Psychology*, 74(11):1953–1965.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., and Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, 16(4):725–743.

Robinson, E. B., Munir, K., Munafò, M. R., Hughes, M., McCormick, M. C., and Koenen, K. C. (2011). Stability of autistic traits in the general population: further evidence for a continuum of impairment. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(4):376–384.

Sanborn, A. N. and Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, 20(12):883–893.

Sapey-Triomphe, L.-A., Temmerman, J., Puts, N. A. J., and Wagemans, J. (2021a). Prediction learning in adults with autism and its molecular correlates. *Molecular Autism*, 12(1):64.

Sapey-Triomphe, L.-A., Weilnhammer, V. A., and Wagemans, J. (2021b). Associative learning under uncertainty in adults with autism: Intact learning of the cue-outcome contingency, but slower updating of priors. *Autism*, page 136236132110450.

Sasson, N. J. and Bottema-Beutel, K. (2022). Studies of autistic traits in the general population are not studies of autism. *Autism*, 26(4):1007–1008.

Seriès, P., editor (2020). *Computational psychiatry: a primer*. The MIT Press, Cambridge, Massachusetts.

Seth, A. K. (2019). Our inner universes. *Scientific American*, 321(3):40–47.

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., and Schilbach, L. (2020). Social Bayes: Using Bayesian Modeling to Study Autistic Trait–Related Differences in Social Cognition. *Biological Psychiatry*, 87(2):185–193.

Shrout, P. E. and Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69(1):487–510.

Simonsen, A., Fusaroli, R., Petersen, M. L., Vermillet, A.-Q., Bliksted, V., Mors, O., Roepstorff, A., and Campbell-Meiklejohn, D. (2021). Taking others into account: combining directly experienced and indirect information in schizophrenia. *Brain*.

Sotiropoulos, G., Seitz, A. R., and Seriès, P. (2014). Contrast dependency and prior expectations in human speed perception. *Vision Research*, 97:16–23.

Stanley, J., Gowen, E., and Miall, R. C. (2010). How instructions modify perception: an fmri study investigating brain areas involved in attributing human agency. *Neuroimage*, 52(1):389–400.

Stanutz, S., Wapnick, J., and Burack, J. A. (2014). Pitch discrimination and melodic memory in children with autism spectrum disorders. *Autism*, 18(2):137–147.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., and Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, 84(9):634–643.

Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585.

Sucksmith, E., Roth, I., and Hoekstra, R. A. (2011). Autistic traits below the clinical threshold: re-examining the broader autism phenotype in the 21st century. *Neuropsychology review*, 21:360–389.

Sun, J. and Perona, P. (1998). Where is the sun? *Nature Neuroscience*, 1(3):183–184.

Teigen, K. H. (1977). "perception" versus "learning": The effect of different instructions upon the central tendency shift. *Scandinavian Journal of Psychology*, 18(1):301–306.

Tewolde, F. G., Bishop, D. V. M., and Manning, C. (2018). Visual Motion Prediction and Verbal False Memory Performance in Autistic Children: Prediction and false memory in autism. *Autism Research*, 11(3):509–518.

Trimmer, P. C., Houston, A. I., Marshall, J. A. R., Mendl, M. T., Paul, E. S., and McNamara, J. M. (2011). Decision-making under uncertainty: biases and Bayesians. *Animal Cognition*, 14(4):465–476.

Tulver, K., Aru, J., Rutiku, R., and Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition*, 187:167–177.

Turner, M. (1999). Annotation: Repetitive behaviour in autism: A review of psychological research. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 40(6):839–849.

Valton, V., Romaniuk, L., Douglas Steele, J., Lawrie, S., and Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews*, 83:631–646.

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

Vetter, P. and Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27:62–75.

von Helmholtz, H. (1866). *Handbuch der physiologischen Optik*.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604.

Welsh, M. B. and Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, 119(1):1–14.

Wiggins, L. D., Robins, D. L., Adamson, L. B., Bakeman, R., and Henrich, C. C. (2012). Support for a dimensional view of autism spectrum disorders in toddlers. *Journal of autism and developmental disorders*, 42:191–200.

Wilson, R. C. and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547.

Woodbury-Smith, M. R., Robinson, J., Wheelwright, S., and Baron-Cohen, S. (2005). Screening adults for asperger syndrome using the aq: A preliminary study of its diagnostic validity in clinical practice. *Journal of autism and developmental disorders*, 35:331–335.

Yon, D., Thomas, E. R., Gilbert, S. J., De Lange, F. P., Kok, P., and Press, C. (2023). Stubborn Predictions in Primary Visual Cortex. *Journal of Cognitive Neuroscience*, 35(7):1133–1143.