Computational modelling of individual differences in belief updating

David Daubner

4th Year Project Report
Cognitive Science
School of Informatics
University of Edinburgh

2023

Abstract

A growing body of research suggests that people with ASD may have ”enhanced rationality”, in the sense that their behaviour may be less influenced by various cognitive biases than for neurotypical people. It has been proposed that this is the result of reduced influence of prior knowledge on the reasoning of people with ASD, possibly because likelihood information is weighted relatively more strongly than priors during integration in the brain. Research conducted so far provides no conclusive evidence in the favour of this proposed imbalance. Likewise, there has been no research so far exploring possible differences in confirmation bias strength between people with ASD and neurotypical population.

In this project, we design two new behavioural task exploring people's belief updating upon observing a sequence of evidence. In order to validate these tasks, we conduct two pilot experiments, and analyse the collected data using a range of Bayesian models. These models and tasks are proposed as a means to measure individual differences in a) relative weighting of prior and likelihood, and b) the strength of confirmation bias. One of the tested tasks resulted in a significant ”inertia” in the participants' behaviour that our models do not account for. This tendency was also present in the other task, but to a much lesser extent, leading us to believe that minor modifications to the instructions could eliminate it. Our analysis showed that participants' performance in both tasks deviated from optimal, but these deviations were better accounted for by an imbalance in the prior and likelihood weighting than by confirmation bias as defined by our proposed models. In future work, our task without an explicit prior can therefore provide a useful platform to explore individual differences in Bayesian inference as a function of psychological traits. In the current pilot experiment, we found no significant relationship between the participants' weights for prior and likelihood information and their autistic or schizotypical traits. However, this was expected due to the pilot experiment's small sample size, and possible differences therefore deserve further exploration in a larger study. The pipeline we designed and validated is ready to be used for such an experiment.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 29394
Date when approval was obtained: 2021-09-03
The participants' information sheet and a consent form are included in the appendices A and B.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*David Daubner*)

Acknowledgements

First and foremost, I would like to thank my supervisor Peggy Seriès for excellent guidance and support throughout this project. I would also like to express my gratitude to Nikitas Angeletos Chrysaitis for generously sharing his experience and expertise. Last but not least, I would like to thank my friends and family for their support and encouragement.

Table of Contents

Chapter 1

Background and Motivation

1.1 Confirmation bias

Confirmation bias is a term used to describe the tendency to search for and interpret evidence in a way that supports our prior beliefs or expectations [45]. This tendency can be likened to a sort of unconscious 'case-building', when people seek and preferentially treat evidence that supports the hypothesis they currently believe, as opposed to objectively considering all available evidence and then drawing a conclusion.

Confirmation bias constitutes a significant departure from rational decision making, and its effects have been studied in a number of fields, including economics [11], medicine [17], criminal investigation [46], science [43], education [42], and politics [36].

Confirmation bias can manifest itself in a number of different behavioural patterns, which can be broadly categorised into 2 groups: evidence acquisition and evidence interpretation. Biased evidence acquisition has been observed in phenomena such as selective testing focused on cases consistent with the current hypothesis, (e.g. [62]), but also selective exposure to available evidence based on its alignment with one's position (investigated e.g. in online news setting by [63]). On the evidence interpretation side, patterns of biased behaviour that have been observed include restriction of attention to the favoured hypothesis (e.g. [5]), or selective interpretation of evidence based on whether it fits with one's current belief (e.g. [33]).

In our work we are going to focus on another example of biased evidence interpretation, overweighting of confirmatory evidence and underweighting of disconfirmatory evidence. This pattern of behaviour has been observed for example by Lord et al. [38]. The participants in this study were asked if they are in favour of the death penalty, and then shown 2 research reports, one of which provided evidence in favour of the death penalty and the other one against it. When asked to rate the quality of the research, people consistently rated the research that supported their prior belief better than the disconfirmatory one, and the net effect of being exposed to the balanced evidence was actually an increase in attitude polarisation between the 2 groups. There is also research suggesting that this aspect of confirmation bias may be responsible for people's persistence in gambling despite failure [22]. Furthermore, it has been suggested that selective

weighting of evidence may explain peoples' tendency to believe various tricks such as astrology or mind-reading [45]. When a person hears a prediction about themselves, their desire to believe in the special powers of its author (e.g. a horoscope writer) will cause them to pay more attention to those parts of the prediction that are correct, and to ignore those that do not fit the reality.

1.1.1 Measurement

Despite its popularity in psychological research, relatively little work on the measurement of confirmation bias has been done to date. Most of the existing research focuses more on detection of the bias, employing binary scenarios where participants either display confirmatory behaviour or not [51]. These paradigms do not measure the strength of the confirmation bias, and thus they do not allow for exploration of potential individual differences. Knowledge of individual differences in the extent of the confirmation bias is useful for a number of reasons. Recently, in a number of fields there has been a recognition of the negative influence of confirmation bias on decision making, and consequently an effort to decrease this influence (see [37] for a criminal investigation example, or [61] for an education example). A reliable measure of individual's bias strength is necessary for an accurate evaluation of the effects of such efforts. Such a measure would likewise benefit psychiatrical research, where the relationship between proneness to cognitive biases and various psychiatric conditions has been investigated (e.g. [53] in relation to ASD, [65] in relation to schizophrenia, [16] in relation to depression).

Measurement of individual differences in confirmation bias has been attempted by Rassin [51], who used a self-report 10-item questionnaire to measure proneness to confirmatory behaviour. This test achieved acceptable internal and temporal stability, but the resulting scores had only limited correspondence with the participants' behaviour in behavioural decision-making paradigms probing confirmatory tendencies. An objective measure (one based on performance in experimental paradigms) of confirmation bias was proposed by Berthet [4]. He used a task where the participants are supposed to evaluate a hypothesis about an interviewee's personality (e.g. whether they are extroverted) by selecting questions to ask the interviewee. The set of available questions contained 8 that assumed the hypothesis was true (e.g. "What events make you feel popular with people?"), 8 that assumed it was false (e.g. "What things do you dislike about loud parties?"), and 4 neutral ones (e.g. "What are some of your favourite books?"). The confirmation bias score was calculated as the percentage of questions chosen that assumed the hypothesis to be true. When tested experimentally, this measure was shown to have an almost acceptable level of internal consistency, and the range of observed scores indicated a good discriminative ability [4].

1.2 Computational modelling

Although traditional descriptive measures of individual differences in various traits are a useful tool, they do not offer any insight into the underlying mechanisms which produce these differences [47]. This shortcoming can be addressed by computational modelling,

where individual differences are captured by what has been called computational phenotype, i.e. a set of model parameters derived from behavioural or neural data that describes the differences in the mechanisms which produce different behaviours in different individuals. There is a wide range of models that have been used to explore individual differences in the past [56], but in this project we will focus on those based on Bayesian inference.

1.2.1 Bayesian models

The brain has no direct access to the outside world - all the information it has is obtained through noisy sensory channels. That means that it has to overcome the inherent uncertainty of the available evidence and construct models of the external world that will allow it to carry out tasks necessary for survival [8]. In order to do that, it seems necessary that the brain has to somehow represent information about uncertainty, both regarding its beliefs and the incoming evidence, and the Bayesian approach to do that is through probability theory. Through this lens, the brain is carrying out inference under uncertainty, where it probabilistically combines its prior beliefs (hypotheses) and expectations about the world with the new evidence to form new beliefs about the world. Probability theory shows that the best way to do that is by using the Bayes' rule:

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)} \tag{1.1}$$

This rule defines how we can use new evidence (captured by the *likelihood* $P(data|hypothesis)$) to update our current belief (the *prior* $P(hypothesis)$) to form a new belief (the *posterior probability* $P(hypothesis|data)$).

Bayesian inference offers a normative model of human mind, in that it defines the task the brain must solve and the optimal solution how that can be achieved. It is far from obvious that the brain actually solves these tasks in a perfectly rational Bayesian manner. However, given how well it deals with uncertainty, it is unlikely that the brain deviates too much from the norms of good inference, and thus it is reasonable to model it using a normative model like Bayesian inference [8]. Bayesian models have been used in the research of perception ([18], [20]), categorisation ([24], [23]), learning [12], causality [25], language processing [44], and other areas.

When Bayesian models are used to measure individual differences, the parameters of interest are usually the ones describing the shape of the prior or likelihood distribution, e.g. their means and standard deviations (e.g. Stankevicius et al. [60] showed that higher trait optimism score is correlated with higher mean of prior expectations about reward probability in a Pavlovian conditioning task, Karvelis et al. [30] showed that higher ASD traits are correlated with lower standard deviation of the sensory likelihood).

Bayesian techniques have also been used to model cognitive biases, for example by Sharp et al. [58] who used hierarchical Bayesian inference to explain a number of biases, such as the halo effect, anchoring bias, outcome bias, and affective evaluation bias. Most models proposed to account for the confirmation bias focus on the tendency to overweight confirmatory evidence and underweight disconfirmatory evidence. They

usually use the classical Bayesian formula, but augment it with parameters that can capture the bias, and that is also the technique we plan to use in our proposed models. Baccini and Hartmann [3] proposed one such a model, designed to explain how agents modify their belief in a proposition B in light of new argument A in support of B. A perfectly Bayesian observer would adapt his beliefs according to the Bayes' rule:

$$P(B|A) = \frac{P(B)}{P(B) + \chi * P(\neg B)} \tag{1.2}$$

where χ is the likelihood ratio

$$\chi = \frac{P(A|\neg B)}{P(A|B)} \tag{1.3}$$

To model how confirmation bias arises, Baccini and Hartmann suggest that agents' prior belief in the proposition B influences how they perceive the new arguments. If the argument is confirmatory, they will weigh it more highly than the true likelihood ratio dictates, and if it is disconfirmatory they will weigh it less than the likelihood ratio dictates. Additionally, they assume the extent of this over- and underweighting is a function of the strength of the agents' prior belief. To achieve this desired behaviour, Baccini and Hartmann suggest that instead of the true likelihood ratio, people use a perceived likelihood ratio χ' that is given by the following formula:

$$\chi' = 2\chi * \frac{[1 - P(B)]^\gamma}{[P(B)]^\gamma + [1 - P(B)]^\gamma} \tag{1.4}$$

where the parameter γ modulates strength of the bias.

This model has the property characteristic of confirmation bias that for the posterior $P^*(B|A)$ computed using classical Bayesian inference and the posterior $P^{**}(B|A)$ computed using the perceived likelihood the following holds:

- If $P(B) < 0.5$ then $P^*(B|A) > P^{**}(B|A)$
- If $P(B) = 0.5$ then $P^*(B|A) = P^{**}(B|A)$
- If $P(B) = 0.5$ then $P^*(B|A) < P^{**}(B|A)$

Aside from the fact that the model is consistent with main features of confirmation bias, Baccini and Hartmann also provide a principled justification for it. The hypothesis underlying the model is that when an agent modifies their belief in a proposition B, they consider not only an argument A, but also a new variable E. If E is true, it means that the proposition B coheres with the agent's background beliefs. This hypothesis is based on work suggesting that having a coherent system of beliefs is desirable, as it helps the agent make sense of the world [19]. Confirmation bias, which treats arguments supporting our current beliefs preferentially, is thus taken to be an evolutionarily desirable feature.

Under these assumptions, the Bayes rule dictates that the agent should adapt their beliefs in the following way:

$$P(B|A,E) = \frac{P(A|B)P(E|B)P(B)}{P(A|B)P(E|B)P(B) + P(A|\neg B)P(E|\neg B)P(\neg B)} \tag{1.5}$$

, which can be rewritten as

$$P(B|A,E) = \frac{P(B)}{P(B) + \chi'' * P(\neg B)} \tag{1.6}$$

, where

$$\chi'' = \frac{P(A|\neg B)P(E|\neg B)}{P(A|B)P(E|B)} = \chi * \frac{P(E|\neg B)}{P(E|B)} \tag{1.7}$$

This is equivalent to the Baccini and Hartmann model of confirmation bias if

$$\frac{P(E|\neg B)}{P(E|B)} = \frac{2 * [1 - P(B)]^{\gamma}}{[P(B)]^{\gamma} + [1 - P(B)]^{\gamma}} \tag{1.8}$$

, which is the case if

$$P(E|B) = \frac{1}{2}([P(B)]^{\gamma} + [1 - P(B)]^{\gamma}) \tag{1.9}$$

and

$$P(E|\neg B) = (1 - P(B))^{\gamma} \tag{1.10}$$

This choice of $P(E|B)$ and $P(E|\neg B)$ is supported by two main desirable properties. Firstly, if we assume that the proposition B is false, than the probability we assign to the proposition B being coherent with our background beliefs $P(E|\neg B)$ decreases with the prior probability we assign to B. That is, if we a priori believe B to be likely, than the probability that it coheres with our background beliefs even if it is false is lower than if we a priori believe B to be unlikely. Second desirable property is the dependence of $P(E|B)$ on $P(B)$ depicted in Figure 1.1. As we can see, if $\gamma < 1$, $P(E|B)$ reaches its maximum when $P(B) = 0.5$. That is, when our prior beliefs about B are not confident, than it is more likely that B coheres with our background beliefs than if we have a strong opinions about B. Notice that if $\gamma > 1$, $P(E|B)$ reaches its minimum when $P(B) = 0.5$. As this sort of behaviour is not psychologically plausible, we will follow Baccini and Hartmann in restricting the range of γ to (0, 1).

We will use the ideas from Baccini and Hartmann's model to adapt two Bayesian inference models to include confirmation bias, and apply them to a modified version of the beads task, the details can be found in Chapter 3.

1.3 Autism Spectrum Disorder

Autism spectrum disorder is neurodevelopmental condition affecting approximately 1% of the population [35]. Its symptoms are heterogenous and vary across individuals, but the most essential features are a) persistent impairment in social interactions and b)

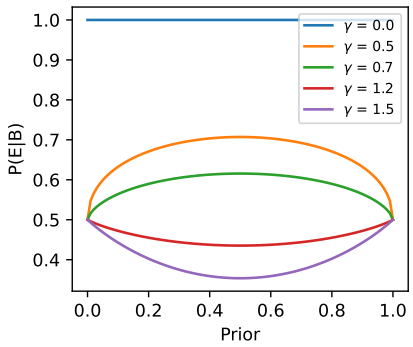


Figure 1.1: The dependence of the probability that the proposition B coheres with the agent's background beliefs given that B is true $P(E|B)$ on the agent's prior belief in B $P(B)$

restricted repetitive patterns of behaviour, activities or interests [1]. Although the social symptoms (a) are most salient to an outside observer, the non-social features (b) are also an important feature, helpful in understanding the phenomenology of people with ASD and diagnosis. One possible explanation for the presence of non-social symptoms is that people with ASD perceive the world in a fundamentally different way from others, prompting research attempting to identify these possible differences [39]. The results include reports that individuals with ASD are faster at locating a target concealed in a more complex figure [57], are not impaired by the oddity of "impossible" figures like the Penrose triangle when asked to reproduce them to the same extent as control participants [41], judge the shape of a slanted circle more accurately than control participants when there are no ambient visual cues [52], and have an enhanced sensitivity to changes in pitch [6].

Pellicano and Burr [48] proposed a mechanistic account of these findings based on Bayesian theories of perception (see 2.2.1). They suggest that the reason why individuals with ASD have a perception that could in many experimental scenarios be described as more accurate is that they have broader priors in the Bayesian sense. This would mean that their prior expectations and beliefs about the world have a smaller impact on how they interpret sensory evidence, making their perception more precise. Broader priors could also explain the sense of being overwhelmed by their surroundings that is often reported by people with ASD. In general, prior expectations make the world more predictable, removing small sensory noise present in the sensory signals. Attenuated priors would result in a more noisy perception of the world, potentially leading to the feeling of being overwhelmed by all the new sensory information.

Many of the phenomena that Pellicano and Burr attribute to attenuated priors are also amenable to explanation that assumes normal priors but reduced sensory noise [7]. The common property of these explanations is that the sensory likelihood is weighted more strongly than the priors during integration in the brain, but in many situations it is difficult to distinguish whether this is due to broader priors or sharper likelihood. Subsequent research does not offer any conclusive evidence in favour of either of these theories, or even of the proposed prior/likelihood imbalance that lies at the centre of

Bayesian theories of autism [10].

1.3.1 Cognitive biases

There has been a number of studies that reported enhanced rationality in people with ASD, in the sense that their decision-making was less prone to various cognitive biases. Weaker influence of biases was observed in experiments testing for example the conjunction fallacy [40], sunk-cost fallacy [21], framing bias [15], and optimism bias [34]. However, to the best of our knowledge, there has been no investigation so far into the relationship between ASD and the strength of the confirmation bias. One of the aims of this project is to design and validate an experiment capable of such an investigation.

Rozenkrantz et al. [53] describe two theoretical frameworks that could account for the findings of enhanced rationality in people with ASD. The first theory postulates that the increased objectivity in decision-making is caused by reduced influence of emotional and reward systems on reasoning. Brains of people with ASD have been reported to have reduced activations in the amygdala [31] and in response to rewards [32], suggesting that this condition does indeed affect how the brain processes and integrates emotions and reward signals. Reduced influence of emotions on reasoning could lead to less loss aversion, and thus explain improved performance in tasks involving the framing bias, or sunk-cost fallacy, although this theory has not been explicitly tested yet.

However, enhanced rationality of people with ASD has been observed also in contexts where emotionality does not seem to play any role, such as the conjunction fallacy or attraction effect. Many of these effects could be explained by a second theory described by Rozenkrantz et al. [53], which postulates that enhanced rationality is the result of reduced influence of prior and contextual knowledge on reasoning, and thus a more bottom-up, detail-oriented style of reasoning. This theory is consistent with the various findings of perceptual differences of people with ASD described at the beginning of this section, and also with the Bayesian account of autism, but extends their claims from perceptual to cognitive domain. A second aim of this project is to assess the suitability of the experiment we design to investigate individual differences in prior and likelihood weighting.

1.4 Contribution

Confirmation bias is an ubiquitous feature of human cognition that has been observed in a wide array of both experimental and real-world contexts. However, little research has been done into measurement of possible individual differences of the bias. We plan to design a new behavioural task to develop such a measure. This work differs from previous attempts to measure confirmation bias strength (e.g. by Berthet [4]) in that we will use a computational model, rather than a purely descriptive measurement of behavioural differences. This approach has the potential to aid our understanding of the underlying neurobiological mechanisms causing the bias. It will also allow us to assess the hypothesis that enhanced rationality in people with ASD is the result of reduced influence of prior knowledge on their reasoning.

The objective of this study is to design, implement, and validate a pilot experiment capable of:

- Measuring individual differences in the strength of the confirmation bias.
- Measuring individual differences in the relative weighting of prior and likelihood.

The plan is to design new behavioural tasks capable of exploring how people update their beliefs, and then validate them by conducting two pilot experiments. We will assess whether the participants understood the experimental tasks, and then use the collected data to investigate two Bayesian models of confirmation bias and two alternative Bayesian models that do not include the bias. We will perform multiple model comparison techniques to assess which of the models captures the participant data the best. We will also perform model recovery to assess whether this model can be reliably identified as the best using the model comparison techniques we used. Then we will estimate this model's parameters using maximum likelihood estimation, and conduct parameter recovery to assess whether this model is capable of reliable parameter estimation. Once we obtain reliable estimated parameters, we will explore the relationship between these parameters and participants' psychiatric traits. Since we plan to conduct only a small sample pilot experiment with limited statistical power, we do not expect any statistically significant findings. However, our work will provide basis for potential larger data collection in the future.

Chapter 2

Methods

2.1 Participants and Questionnaires

In the first part of the project, we run 2 pilot experiments, one for each version of the experimental task described below. In these pilot experiments we asked people to participate voluntarily. For Task 1 (Belief updating without prior), we had a total of 13 participants, 7 males and 6 females, in the age interval of 20 to 52. For Task 2 (Belief updating with prior), we had a total of 10 participants, 6 males and 4 females, in the age interval of 20 to 50.

The study was approved by the Informatics Research Ethics (RT number 29394). In the experiment, the participants first signed a written consent, and then we asked them to fill in two questionnaires. The first questionnaire was the Autism Spectrum Quotient (AQ) used to measure participants' autistic traits [54]. The second questionnaire was the 21-item Peter et al. Delusions Inventory (PDI) [50], used to measure schizotypal traits, specifically proneness to delusions. Written consent and both questionnaires were hosted on Qualtrics, and after their completion the participants were redirected to the task hosted on Pavlovia.

The AQ is a self-report measure of autistic traits designed for adults with IQ on average or above average level. It consists of 50 items, each of which has four possible responses: "definitely agree", "slightly agree", "slightly disagree", and "definitely disagree". The scoring is binary - both mild and strong endorsement of an autistic trait is scores as +1, while both mild and strong rejection is scored as 0. The maximum score is thus 50.

The 21-item PDI questionnaire is used as a measure of schizotypal traits present in the general population. It is a shortened version of the original 40-item questionnaire [49] designed to measure delusional ideation in the normal population. The selection of the 21 items was based on an experiment on 272 healthy individuals [50]. PDI measures three dimensions of delusions: distress, preoccupation, and conviction. Each of the 21 items can be answered as Yes or No, and in the case of a positive answer, respondents are supposed to respond 1 to 5 for propmts about each of the three dimensions. There are 5 scores that can be calculated based on the responses. The PDI Yes/No scores are calculated by adding 1 for each Yes answer, and 0 for each No answer, leading to a range

between 0 and 21. Each of the three dimensions is scored in the same way. For each item, No answer adds 0 to the score of each dimension, whereas in case of a Yes answer a number between 1 and 5 is added to the score, depending on the respondent's response on the corresponding prompt. The score of each dimension thus ranges between 0 and 105. Finally, by adding the PDI Yes/No score with the scores for each of the three dimensions, a grand total PDI score can be calculated, ranging from 0 to 336.

2.2 The sequential beads task

To measure the individual differences in confirmation bias strength and prior and likelihood weighting, we decided to design two new tasks exploring how people update their beliefs upon observing a sequence of new evidence. We based our tasks on the beads task, where participants sample coloured beads and use this evidence to guess which jar they have been drawn from. Variations of this task have been used in the past to investigate individual differences in behaviour. For example, Simonsen et al. [59] used a social version of the task to investigate differences in how people integrate different kinds of information, specifically the direct information from the sample the participant was shown with the indirect information from the guesses of other people. Croft et al. [13] used a variation more similar to what we plan to use in this project, which they called Probability Estimation Task, to measure 'disconfirmatory updating' in schizophrenia. They presented the participants with an illustration of two jars with 100 beads each, with inverse proportions of coloured beads at a 80:20 ratio. The participants were then informed the computer will randomly draw a bead from one of the jars, show it to them, and then return it to the jar. The participants were then shown a sequence of 30 beads, and after each draw they were asked to rate how certain they were about which jar the beads were being drawn from. They were also told that the computer may change the jar it draws the beads from at any point during the experiment.

For the purposes of this project, we modified the Probability Estimation Task of [13]. We decided to test 2 versions of the task in the pilot experiments. The simpler Task 1 (Belief updating without prior), used to make sure that participants approach the task generally in the same way our models do, was similar to the Probability Estimation Task. There were two jars with inverse proportions of coloured beads at a 60:40 ration, as seen in Figure 2.1. This ratio was chosen because it maintains uncertainty for longer than a more unbalanced one would - if the ratio was e.g. 90:10, few draws would be sufficient to make a reasonably accurate judgement, limiting the amount of data about participants' decision making we can gather from each trial. The participants were told that the computer will choose one of these jars at random, and then draw beads from it one by one. After each draw, participants were asked to rate how certain they were about which jar the beads are being drawn from, as seen in Figure 2.3. Unlike in the task used by Croft et al. [13], in our task the participants are told that the jar from which beads are drawn will stay fixed for the entire duration of each trial, because investigation of how people deal with uncertainty related to changes in the environment is outside the scope of this project.

In Task 2 (Belief updating with prior), the participants will still have to make the choice between two jars with different proportions of coloured beads, jar type A and jar type

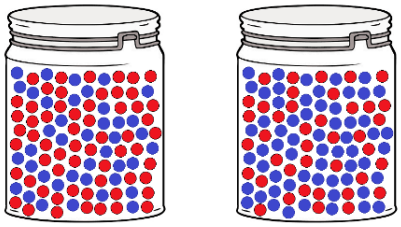


Figure 2.1: Both tasks were based on two possible types of jars, one with red:blue ratio 60:40 (left), one with the inverse ratio (right)

B. However, in order to induce prior expectations in the participants, we included an extra step at the beginning of each trial, before the sequence of draws starts. We tell the participants that there are 10 jars in total we can choose from, some of them of type A, some of type B (Figure 2.2). In each trial we will choose one of the 10 jars, and then draw beads from it for the entire duration of the trial, one by one, asking the participants how certain they are about which type of jar the beads are being drawn from after each draw. At the beginning of each trial, we will tell the participants the proportion of the two types among the 10 jars, and this knowledge should induce a prior belief in the participants about which jar is more likely to be used in a trial, even before any evidence is presented. For example, if there are 8 jars of type A and 2 of type B, the prior probability of jar of type A being drawn is 0.8. We ask the participants to state their confidence before any beads are drawn, to confirm that they take the prior into account. Using this task setup should allow us to examine how the participants integrate their prior expectations with new evidence to update their beliefs about a binary variable (i.e. which of the two jars is used to draw the beads in this trial) at each step of the evidence sequence. In this version also, the same jar is used for the entire duration of a trial.

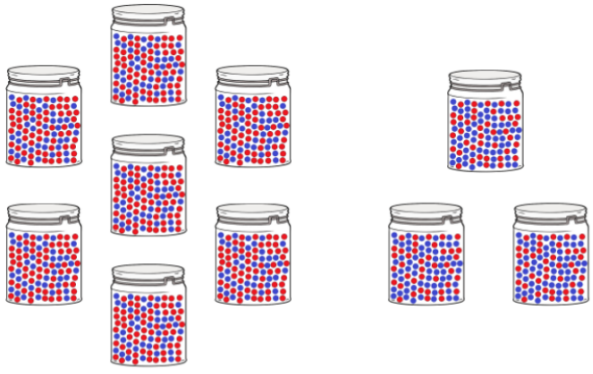


Figure 2.2: In Task 2, at the beginning of each trial the participants received information about the red-majority:blue-majority ratio among the ten jars from which the jar used in that trial will be drawn. This is an example of a trial where there were 7 red-majority jars

To ensure that participants understand the task, we start with detailed instructions and also two practice trials before the real experiment. All information given to the

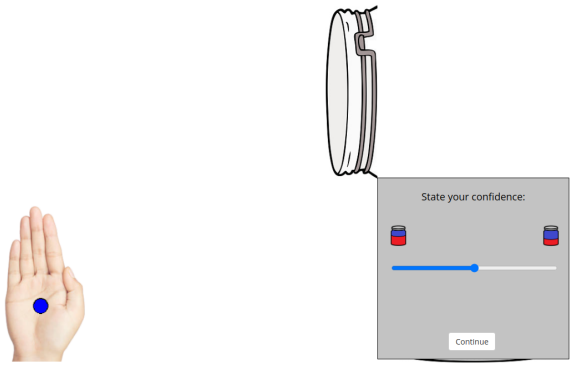


Figure 2.3: Each time a bead was drawn, participants were prompted to state their confidence about which jar is being used in the current trial

participants, prior and evidence, is accompanied by animations (an image of the 10 jars for prior, a hand drawing and holding a bead for evidence). The confidence rating is done using a slider bar ranging from ”I am sure it is the majority red jar” to ”I am sure it is the majority green jar”. A response in the middle of the bar means ”I don’t know”.

The experimental task was implemented using the jsPsych version 6 framework [14]. This JavaScript framework is commonly used to implement behavioural and psychophysical experiments that can run in a web browser. We hosted the task on Pavlovia servers. Our implementation of the task was based on the work of Matthew Whelan, with modifications done by Fateme Soltani and Nikitas Chrysaitis. These authors implemented a non-sequential version of the jar task, similar to that of Simonsen et al [59]. However, our task was in many respects different from those already implemented, mainly in that sequential rather than single-trial, and included a different mechanism to provide prior information. We reused some of the already existing elements, but also made significant timeline modifications and implemented new elements to match the specifics of our tasks. We also implemented an extra task at the end of the experiment meant to assess whether participants combine the two sources of information (proportion of red jars among all 10, and the sequence of drawn beads) in a Bayesian manner.

2.3 Computational modelling

We will use four models based on Bayesian inference (described in detail in part 1.2.1). Two of these models were used by Chrysaitis et al. [10] to investigate individual differences in hierarchical Bayesian processing and their relation to higher autistic traits. These two models are a simple Bayesian model (SB), a weighted Bayesian model (WB). The other two models are extensions of these models, which include the possibility of confirmation bias. This extension was based on the model proposed by Baccini and Hartmann [3] discussed in part 1.2.1. These two models are a simple Bayesian model biased (SBB) and a weighted Bayesian model biased (WBB).

The simple Bayesian model is based on classical Bayesian inference. The probability of the bead being drawn from jar A given its colour is red has the form:

$$P(A|red) = \frac{P(red|A)P(A)}{P(red} = \frac{P(red|A)P(A)}{P(red|A)P(A)+P(red|B)P(B)} \tag{2.1}$$

However, following Chrysaitis et al. [10] we will use log odds ratio instead of pure probability:

$$\frac{P(A|red)}{P(B|red)} = \frac{P(red|A)P(A)}{P(red|B)P(B)} = \frac{P(red|A)}{P(red|B)} * \frac{P(A)}{P(B)} \tag{2.2}$$

$$ln(\frac{P(A|red)}{P(B|red)}) = ln(\frac{P(red|A)}{P(red|B)}) + ln(\frac{P(A)}{P(B)}) \tag{2.3}$$

$$L_e = L_s + L_p \tag{2.4}$$

Where L_e is the log odds ratio of the posterior belief about which jar is being used, L_s is the log odds ratio of the sensory evidence, and L_p is the log odds ratio of the prior belief. To include the confirmation bias in this model, we will follow Baccini and Hartmann [3] and use a perceived likelihood odds instead of the real ones. In their original model, the perceived likelihood ratio χ' was calculated using the following expression:

$$\chi' = 2\chi * \frac{P(B)^\gamma}{[P(A)]^\gamma + [P(B)]^\gamma} \tag{2.5}$$

, where γ is the bias strength. However, Baccini and Hartmann define the likelihood ratio as $\chi = \frac{P(red|B)}{P(red|A)}$. In order to obtain the perceived likelihood odds as defined by us, we need to take the inverse of χ':

$$[\frac{P(red|A)}{P(red|B)}]' = [\chi']^{-1} = \frac{1}{2}\chi^{-1} * \frac{[P(A)]^\gamma + [P(B)]^\gamma}{P(B)^\gamma} \tag{2.6}$$

which can be rewritten as the inverse of the original ratio χ times a weight w_b, which is a function of a prior belief $P(A)$:

$$[\chi']^{-1} = w_b\chi^{-1} \tag{2.7}$$

$$w_b = \frac{1}{2}\frac{[P(A)]^\gamma + [P(B)]^\gamma}{P(B)^\gamma} = \frac{1}{2}([\frac{P(A)}{P(B)}]^\gamma + 1) = \frac{1}{2}(e^{\gamma * L_p} + 1) \tag{2.8}$$

The log perceived likelihood odds L'_s thus become:

$$L'_s = ln(w_b * \frac{P(red|A)}{P(red|B)}) \tag{2.9}$$

where

$$w_b = \frac{1}{2}(e^{\gamma * L_p} + 1) \tag{2.10}$$

And the Biased Simple Bayes model has the form:

$$L_e = L'_s + L_p \tag{2.11}$$

The weighted Bayesian model extends the SB model by allowing for the possibility that prior and sensory information are weighted differently when integrated. This model has the form:

$$L_e = F(L_p, w_p) + F(L_s, w_s) \tag{2.12}$$

where

$$F(L, w) = ln(\frac{we^L + 1 - w}{(1-w)e^L + w}) \tag{2.13}$$

The F function returns the log odds ratio L weighted by its "trustworthiness" w.

The biased weighted Bayesian model simply uses the log perceived likelihood odds L'_s in place of L_s:

$$L_e = F(L_p, w_p) + F(L'_s, w_s) \tag{2.14}$$

but the bias weight w_b is calculated using the weighted log prior odds instead of the raw ones:

$$w_b = \frac{1}{2}(e^{\gamma * F(L_p, w_p)} + 1) \tag{2.15}$$

2.4 Model fitting

We fitted the models to the participants' data using Maximum Likelihood Estimation (we minimised negative log likelihood (NLL)). Since our models make point estimates of the participant confidence, we augmented them with a data model consisting of a normal distribution centred on the point estimate.

$$\hat{y} = argmin_y(NLL) = argmin_y[\sum_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{L_x^{(i)} - L_e^{(i)}}{\sigma})^2}] \tag{2.16}$$

where NLL is the negative log likelihood, calculated by summing the probability of the participant response $L_x^{(i)}$ given normal distribution centred on the corresponding model estimate $L_e^{(i)}$ with standard deviation σ over all participant responses.

The standard deviation of the normal distribution σ was calculated using participants' responses in the following way:

$$std = \sqrt{\frac{\sum_i^n (L_x^{(i)} - L_e^{(i)})^2}{n-1}} \tag{2.17}$$

We implemented all the models in Python 3. The optimisation was done using the Scipy `optimize` package's `minimize` function, which uses the Broyden–Fletcher–Goldfarb–Shanno optimisation algorithm [2]. We set the maximum number of function evaluations to 10,000. To prevent finding local minima, for each participant we started the fitting procedure at 10 random initial parameter values sampled from a uniform distribution (with range [0.5, 1] for w_p and w_s and [0, 1] for γ).

2.5 Model comparison

We need a way to determine which of the considered models fits the participant data the best. The simplest approach would be to use the negative log likelihood of the data given each model at its best fitting parameter settings. However, since the data we would be using to evaluate the model are the same we used to estimate its parameters, we would be risking overfitting. A model with more free parameters can usually fit the data it has been trained on better than simpler models. Since our models are nested, the more complex models are actually guaranteed to always have a smaller or equal negative log likelihood to the simpler models. However, it is possible that the difference in the models' performance is not generalisable to data they have not been trained on. Therefore we need to find a way to somehow take model complexity into account along with fit to the data.

One such measure very popular in cognitive modelling is the Bayesian Information Criterion (BIC) [55], which includes a penalty for free parameters.

$$BIC = 2 * NLL + k * log(N) \tag{2.18}$$

where NLL is the negative log likelihood of the data given the model, k is the number of free parameters in the model, and N is the number of data points the model has been trained on.

The probability that BIC will select the correct model approaches one as the sample size approaches ∞. However, for small samples, the heavy penalty on model complexity imposed by BIC often leads it to pick a model that is too simple [26]. An alternative measure is the Akaike Information Criterion.

$$AIC = 2 * NLL + 2 * k \tag{2.19}$$

AIC imposes a lower penalty on complexity than BIC, which leads it to pick models that are too complex when the sample is large [26]. Since our sample is quite small, AIC may be more suitable than BIC, but in case of conflict, it would be useful to have an alternative.

One alternative measure that prevents overfitting is evaluating model performance on a held-out validation dataset that was not used for parameter fitting. Although not widely used in cognitive modelling, this technique efficiently penalises model complexity if it does not improve general performance, since models overfitted to a single dataset will not do well on other data. Major problem with this technique is that it reduces

the amount of data that can be used for parameter estimation, which is especially problematic for psychological research where the data is usually scarce. This problem can be avoided by using k-fold cross-validation [26], where the data is split into k parts, and then the model is trained on all except one of these parts and evaluated on this part. This is repeated for each part, and the final model performance is the average of its performance on each part. We will use this technique (each participant's data will be one fold) to support the results obtained with BIC and AIC.

2.6 Parameter and model recovery

Model and parameter recovery is an important part of any computational modelling pipeline [64]. If we plan to draw any conclusions based on the parameter values we estimate for the participant data, we need to make sure that the models' parameters are identifiable, i.e. there is a unique mapping between the estimated parameters and the data patterns that the estimation is based on. We can assess whether this is the case by parameter recovery. First, we use the model to simulate data with known parameter values, and then we fit the model to these simulated data. In the end, we compare the estimated parameters with the true parameters that produced the data. An ideal outcome of parameter recovery is a tight correlation between the simulated and estimated parameters, and no bias.

To choose a relevant range for the parameters used to simulate data, we based it on the parameters fitted to the participant data. At each iteration of the parameter recovery, we randomly sampled each parameter's value from the estimated participant parameters, and added small Gaussian noise to it. The resulting parameter combination was then used to simulate the data. Since our model outputs a point estimate of the confidence instead of a probability distribution, we augmented it with the same data model we used for the Maximum Likelihood Estimation.

For each of the two task versions, we simulated participant data for a 200 different parameter combinations. Each simulation consisted of 20 trials with 10 bead draws each, just like in the actual experiment. Before fitting the model to the simulated data, we added noise equal to the standard deviation calculated using equation 2.17 for the WB model. For the no prior task, parameter recovery of the best-fitting WB model was good for both parameters ($r_{w_p} = 0.71$, $r_{w_s} = 0.97$, and there was no bias, Figure 2.4). The recovery of w_p was worse in the lower range of the spectrum, but the w_ps estimated from actual participant data are almost all near 1, so this should not be a problem. We observed no significant correlation between the two parameters ($p = 0.11$).

For the task version with prior, parameter recovery of the best-fitting WB model had high correlation for both parameters ($r_{w_p} = 0.98, r_{w_s} = 0.97$). There was no significant correlation between the two parameters ($p = 0.46$).

Model recovery is necessary to assess whether the techniques of model comparison we used to pick the best model are reliable. The procedure is similar to parameter recovery. For each model, we sample 200 parameter combinations and then use them to simulate data. All models are then fitted to the data and their AIC score is used to assess which fits it the best.

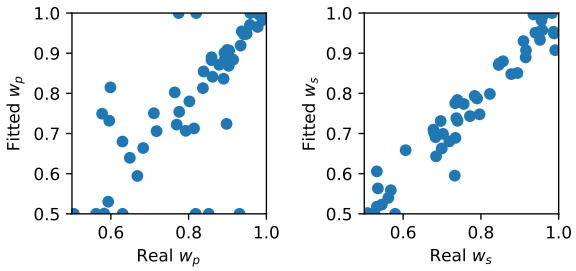


Figure 2.4: Parameter recovery for the WB model in Task 1.

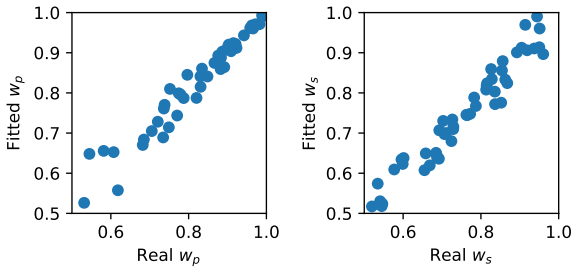
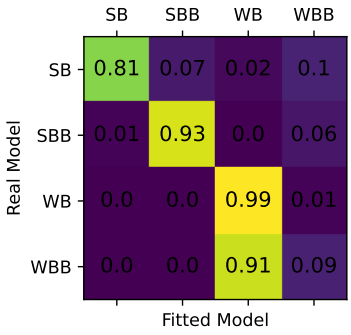
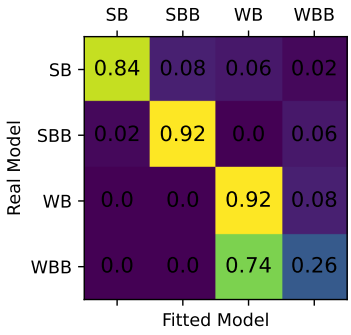


Figure 2.5: Parameter recovery for the WB model in Task 2.

Ideally, the model that produced the data should also always be the one that fits them the best (i.e. the confusion matrices in figures 2.6a and 2.6b should be identity matrices). However, as we can see in the two figures, the model recovery is good for all models except weighted Bayes biased. Data produced by WBB is in most cases fitted the best by the unbiased weighted Bayes model. This is potentially problematic - if our model comparison picks WBB as the best model, we cannot have much confidence in this result, since its recovery is so poor. At the same time, if WB is picked as the best model, we cannot be sure that the data was not actually produced by WBB.



(a) Task 1: Belief updating without prior.



(b) Task 2: Belief updating with prior.

Figure 2.6: Model recovery confusion matrices using AIC score. We can see that for both tasks, the model recovery is good for all models except the Weighted Bayes Biased.

Chapter 3

Results

The aim of this project is to design and validate the pipeline for two new behavioural tasks designed to analyse belief updating, one with and one without prior. To this purpose, we conducted two pilot experiments and used the collected data to answer the following questions:

- Do people understand the task? Do they solve it in a way compatible with the models we are considering?
- Are the participants behaving optimally, or are there deviations that can be captured by our models?
- Is there any relationship between these deviations and the participants' psychiatric traits?

First we will evaluate the results from the version of the task without prior, and then the one with prior.

3.1 Task 1: Belief updating without a prior

3.1.1 Questionnaires

First we analyse the AQ and PDI scores to understand our participant samples. In the sample that did the task without prior, the mean AQ score was 23.08, which is significantly higher ($p = 2.5e-4$) than the mean AQ score of general population, 16.94 [54]. The average PDI Yes/No score of our sample was 6.92, which is not significantly different ($p = 0.84$) from the average PDI Yes/No score of general population, 6.7 [50]. The mean PDI scores for distress, preoccupation, and conviction were respectively 18.08, 19.46, and 21.23, and were not significantly different from the general population. The distribution of AQ and PDI Yes/No scores for the no-prior sample can be seen in Figure 3.1.

In the no-prior sample, we observed relatively strong significant ($p = 0.02$) correlation between AQ and PDI Yes/No scores ($r = 0.62$). This relationship could make interpretation of potential results problematic, since it would be difficult to disambiguate

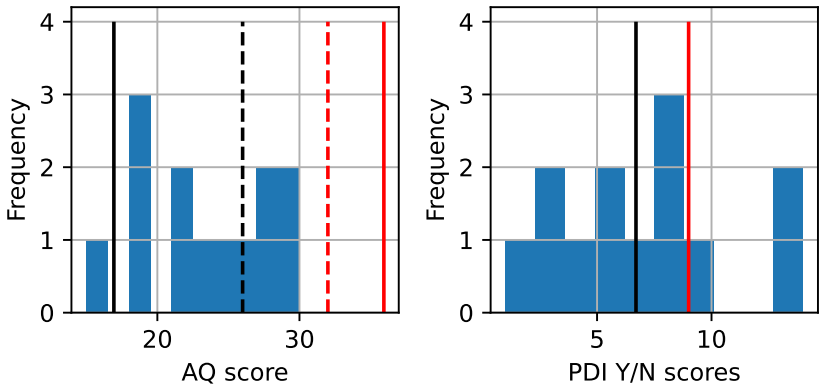


Figure 3.1: The AQ and PDI Yes/No scores obtained from the two questionnaires for the sample participating in Task 1. In both figures, the solid black and red lines indicate the mean score in general and clinical population respectively. In the figure on the left, the dashed black line (AQ scores below 26), delineates the absence of autism spectrum disorder, while the dashed red line (AQ scores above 32) indicates clinically significant level of autistic traits [54] [50]

whether the observed differences are related to differences in autistic or shizotypical traits.

3.1.2 Behavioural analysis

Left/Right bias Since we generate the stimuli in the task randomly, the task is not perfectly symmetrical. The mean of all stimuli (considered as 1 for a red bead and 0 for a blue bead) was 0.5146. To investigate left/right bias, we use the performance of the Simple Bayes model as an unbiased baseline. The mean response of this optimal model was 0.5168, which was not significantly different ($p = 0.25$) from the mean participant response, 0.5237. We thus observe no left/right bias in participant responses.

Final response distribution Next we explore the distribution of participants' responses at the end of each trial, once they have seen all the evidence. This is shown in Figure 3.2. Since the task is symmetrical, we analyse the absolute confidence calculated as the distance of the response from complete uncertainty, i.e. 0.5 ($|r - 0.5|$). This value ranges from 0 for no confidence to 0.5 for complete confidence. For comparison, we also plot the absolute confidence values of the optimal Simple Bayes model.

We can see that compared to the SB model, the participants' responses are more spread out over the whole absolute confidence range. Another difference is that participants' estimates are more concentrated around the edges of the range, i.e. absolute confidence or no confidence, whereas SB estimates tend towards the center of the range. This can be linked to a tendency some participants described in their feedback questionnaire, where they mentioned that if they become confident enough, they moved the slider to the edge for the corresponding jar.

Update size distribution We also looked at the distribution of the participants'

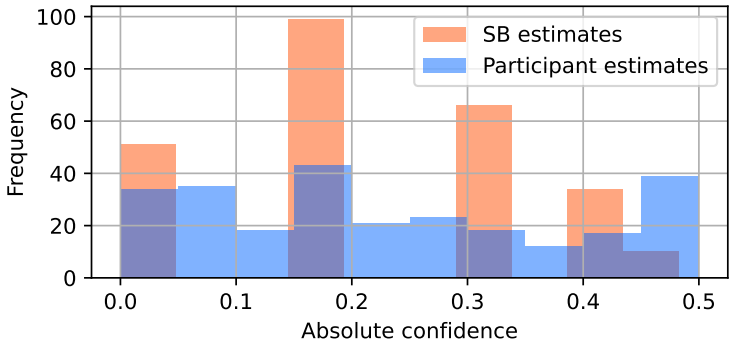


Figure 3.2: The distribution of absolute confidence at the end of each trial of the no prior task version, for both participants (left) and the Simple Bayes model (right)

confidence updates after observing evidence. We find that participants' average update is 0.0029, which is not significantly different from 0 ($p = 0.3$). This indicates that there is no increase/decrease bias, so we can consider just the distribution of update size, i.e. the absolute value of the update. This is depicted in Figure 3.3.

Participants' mean update size is 0.0869, which is not significantly different from the optimal model ($p = 0.12$). We can see that participants' update sizes seem to cluster in the area around this mean. However, unlike in the optimal model updates, there is a second peak near 0. This is caused by the fact that unlike the optimal model, participants sometimes do not update their confidence estimate at all after observing evidence - there seems to be inertia about their estimates until enough new evidence accumulates. Some participants also reported this strategy in the feedback survey.

This finding suggests that the participants do not always report a change in their current confidence until they see more evidence. Our models assume no such inertia, which means that they do not account well for some of the collected data. However, Figure 3.4a shows that for most people this tendency was not too strong, although there is one outlier who did not update their confidence in more than half the trials.

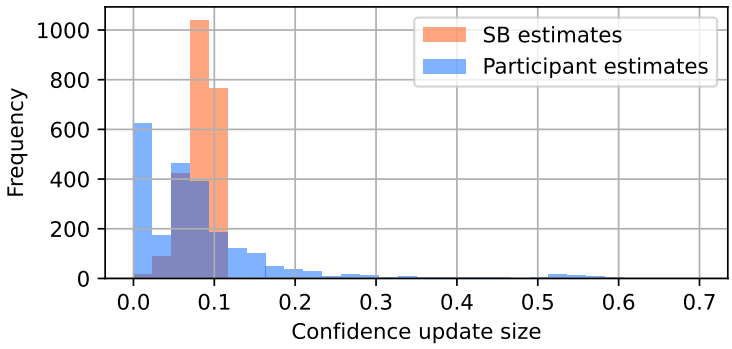
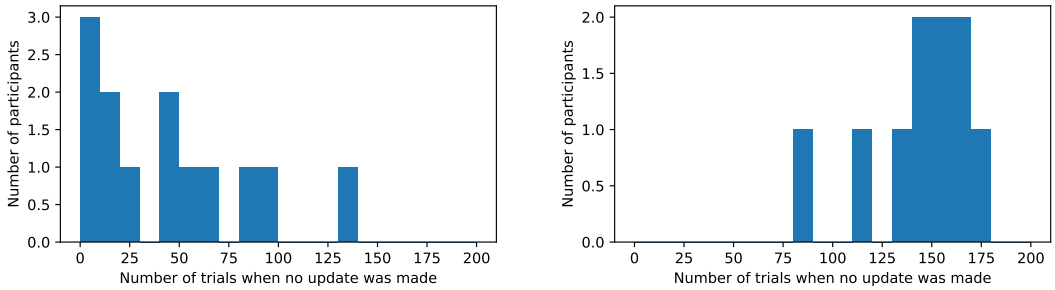


Figure 3.3: The distribution of update sizes after observing evidence in Task 1, for both participants and the Simple Bayes model

Summary Model-free analysis indicates that participants understand the task. In some



(a) Task 1: Belief updating without prior.

(b) Task 2: Belief updating with prior.

Figure 3.4: The distribution of the total number of trials when no update was made among the participants

participants' estimates we detected an inertia - sometimes they did not update their estimates after observing new evidence but waited until more evidence was available.

3.1.3 Computational analysis

3.1.3.1 Model comparison

To assess how the participants decision making differed from the optimal solution, we fitted our four models (Simple Bayes, Simple Bayes biased, Weighted Bayes, and Weighted Bayes biased) to the participants' data. Then we compared the results using BIC, AIC, and cross-validation, as illustrated in Figure 3.5. For each metric, the performance is plotted as the difference between the score of each model and the score of the best model for that metric. For BIC, which penalises model complexity heavily, the best score was achieved by Weighted Bayes. For the more lenient AIC, Weighted Bayes Biased came out as the best model. We also evaluated the models using cross-validation, where WBB also came out as the best model, although the difference between WBB and WB was negligible.

In Figure 3.5 we show that the best fitting models vary across different participants. We plot the difference in AIC between each model and the best-fitting model according to AIC, WBB. We see that WBB was the best fit only for three participants' data. For 7 it was WB, for two it was SBB, and SB for one.

It is surprising that WBB was picked as the best model by AIC despite the fact that model recovery suggests that most of the datasets produced by WBB are actually better fitted by WB. However, we see that WBB was the best fit for only 3 participants, while for 7 it was WB. This suggests that most participants are best fitted by WB, but there are some outliers which are better fitted by WBB, and those skew the overall model comparison. Most participants had low bias strength, and one way to interpret the model recovery results is that WB and WBB are too similar to reliably distinguish unless the bias is quite strong. Since WBB has poor model identifiability, we decided to exclude it from further analysis.

The best-fitting model is thus WB. However, all of the measures mentioned above allow

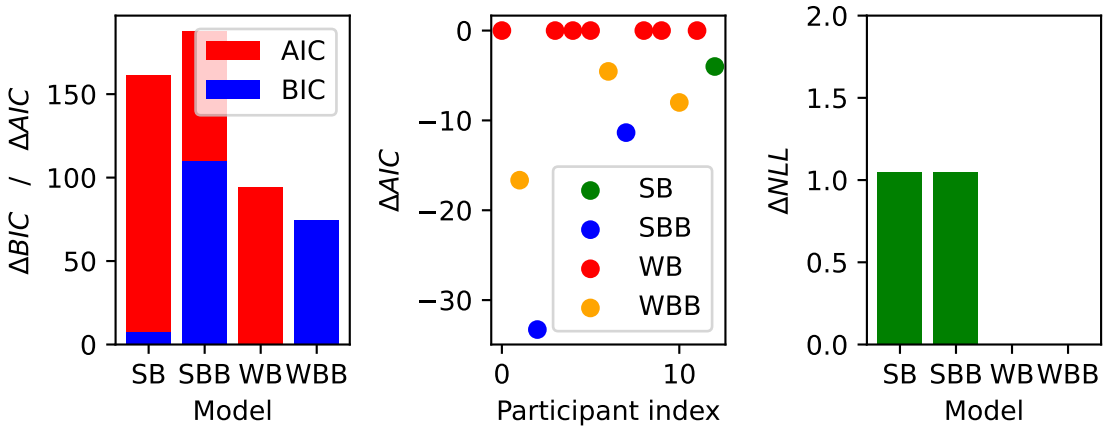


Figure 3.5: Model comparison for no prior task. On the left, we see the difference between the BIC/AIC score of each model and the best-fitting model according to each metric. In the middle, we see the best fitted model for each participant, together with the difference between its AIC and the AIC of best-fitting WB. On the right, we see the difference in the cross-validation average NLL between each model and the best-fitting WBB model.

us to measure only the relative goodness fit of the models, and thus to determine which of the models we are considering is the best. However, since we are considering only a limited selection of models, it is possible that even the best-fitting model still fails to capture important characteristics of the participant behaviour. To prevent this outcome, it is important to perform model validation of the best fitting model [64].

One way to assess how well the model captures the participants' behaviour is to calculate the correlation between the model's predictions and the participant estimates. For the WB model, $r = 0.79$. This is a fairly good correspondence, as can also be seen in Figure 3.6. This figure shows the accuracy of the WB model's predictions of the participant responses.

3.1.3.2 Parameter estimation and psychiatric traits

In this section we use the WB model to evaluate whether there is any relationship between the estimated parameters for each participant's data and their psychiatric traits.

First we look at whether there is any relationship between the participants' psychiatric traits and the best-fitting model. Figure 3.7 depicts this. With the limited data we have, it is difficult to judge whether the data points best-fitted by each model form any clearly discernible clusters. Since the single participant best fitted by SB had very low AQ and PDI Yes/No scores, whereas 3 of the 4 participants best fitted by the biased models (SBB and WBB) had high AQ scores, there is some indication that higher AQ scores may lead to suboptimal decision making. However, the small sample does not give much weight to these speculations.

Since we could not discern any clear clusters, we will look at the distribution of

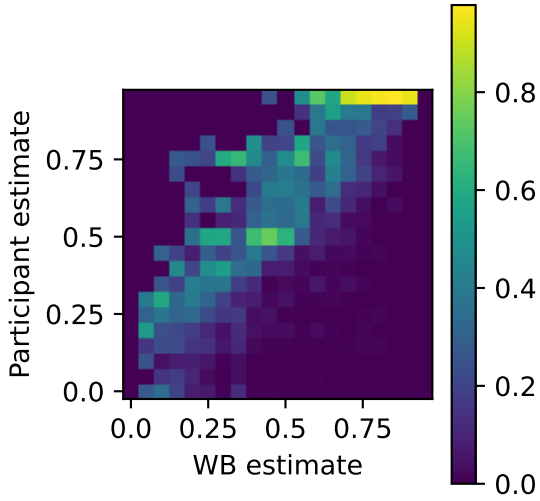


Figure 3.6: Heatmap of the accuracy of the WB model's predictions of participant responses. Both axes are split into 20 bins, and the colour of each bin indicates the proportion of times the participant's estimate fell into the corresponding y-bin when the model's prediction fell into the corresponding x-bin.

	r	p
AQ, w_p	0.15	0.62
AQ, w_s	-0.16	0.61

	r	p
PDI Y/N, w_p	0.15	0.62
PDI Y/N, w_s	0.09	0.78

Table 3.1: Relationship between fitted parameters of the WB model and psychological traits

parameters estimated using the overall best-fitting model, WB, as shown in Figure 3.8. We can see that the estimated prior weights tend to be concentrated around 1 while the likelihood weights are more spread out over the range, though also mostly above 0.8. However, Mann-Whitney U test revealed no significant difference between the two weights ($p = 0.20$).

Next we examine the relationship between estimated parameters and participants' psychological traits. The results are shown in Table 3.1. We can see that there is no significant correlation between either AQ or PDI Yes/No scores and either of the parameters. This is unsurprising considering the small sample size.

Next we divided our participants into a low and high AQ groups based on [54], where AQ score less than 26 is taken as the absence of ASD. 8 of our participants were in the low AQ group, 5 in the high AQ group. We tested if there is any significant difference between the medians of the estimated parameters for each group, depicted in Table 3.2. Using the Mann-Whitney U test, we found no significant difference for either of the 3 parameters. This is not surprising considering the small size of our sample, but larger data collection in the future should be able to assess possible differences more conclusively.

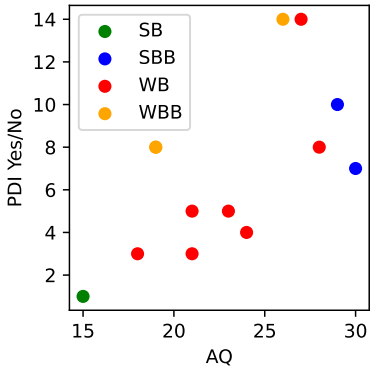


Figure 3.7: The relationship between the best-fitting model for each participant and their psychiatric traits in Task 1

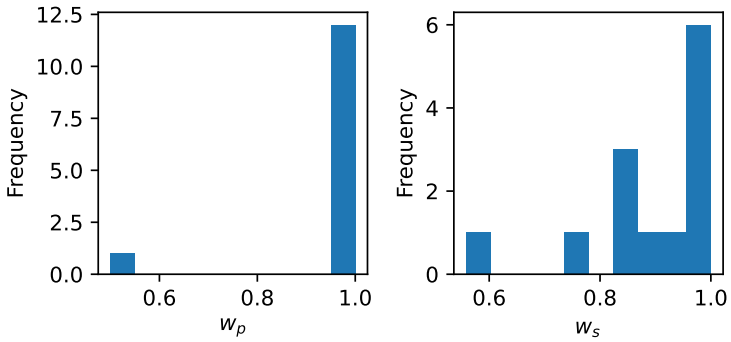


Figure 3.8: The distribution of estimated parameters of the WB model for the no prior task

Next we compared the group AIC score each model achieves on the data from the low and high AQ group separately. Both groups were best fitted by the WB model.

3.1.3.3 Further analysis

In the feedback questionnnaires, several participants described feeling tired towards the end of the experiment, and suggested this might have had some effect on their performance. We tried to assess this by training and evaluating (using AIC) the model separately on the data from first half of the experiment (10 trials) and second half (10 trials). We find that the data from the first half were best fitted by the SB model, whereas for the second half it was WB. This suggests that participants' decision making starts to

	Δ median	p
w_p	0.004	0.21
w_s	0.09	0.76

Table 3.2: Difference of medians of the estimated parameters between the high AQ and low AQ group

	SB AIC	WB AIC
First half	-1106.1	-1092.5
Second half	-1020.3	-1032.8

Table 3.3: Group AIC scores of the SB and WB models when evaluated separately on the first and second half of each participant's trials

deviate from optimal decision making more as the experiment progresses, although the observed differences between the models are relatively small (Table 3.3).

3.1.4 Participant feedback

The purpose of the pilot experiments was to implement and test the pipeline for data analysis, and to assess the suitability of the task before conducting the main experiment. To this purpose, we asked the participants to give us feedback on their experience during the experiment.

In general, the participants found the task engaging, although many of them reported that it was too long and tiring towards the end. In terms of changes that would improve their experience, there were multiple suggestions that we include a progress bar that would let them know how many more trials there are left. Some of the participants also said they would appreciate having ticks on the confidence slider, so that they can pick 0.5 (no confidence) more accurately. In terms of strategies, peoples' reports vary. Some suggested that they adjusted the slider scale just based on how they felt, while others described more elaborate strategies trying to take the precise colour ratios into consideration. We think that this might be the consequence of the fact that our sample for the pilot experiment consisted mainly of university students in technical fields, who are more likely to try to consciously come up with an optimal solution than the general population. Another pattern reported multiple times was that people did not always adjust their confidence after observing new evidence. Rather they described only adjusting their confidence once multiple beads were drawn.

3.1.5 Task 1 Summary

Our analysis showed that in the task without the prior, participants understand the task and mostly solve it in a way that seems compatible with our models. Some people demonstrate inertia in their confidence estimates that our models do not account for. However, in our sample such participants constitute a minority, and we believe that this tendency could be illuminated with minor changes in the experiment that would better incentivise the participants to share their confidence after every bead draw. Our model-based analysis suggests that participants' behaviour deviates from optimal Bayesian inference. However, these deviations seem to usually be better captured by an imbalance in prior and likelihood weighting than by the confirmation bias as defined by our models. As expected due to the sample size, analysis of estimated parameters showed no significant correlations between either prior or likelihood weight and AQ or PDI, and also no significant difference between parameter means of low and high AQ subgroups. Finally, participant feedback suggests that participants find the experiment

too long and computational analysis indicates their behaviour might change towards the end of the experiment as a result. In summary, the pilot experiment suggests that Task 1 is more suitable for exploring differences in prior and likelihood weighting as a function of psychiatric traits than for the measurement of confirmation bias. The pipeline we developed is ready for larger data-collection, which can provide enough statistical power to detect such differences if they exist.

3.2 Task 2: Belief updating with prior

3.2.1 Questionnnaires

In the sample that did the task with prior, the mean AQ score was 24.46, which is significantly higher ($p = 4.68e-6$) than the mean AQ score of general population, 16.94 [54]. The average PDI Yes/No score of this sample was 6.69, which is not significantly different ($p = 0.99$) from the general population average, 6.7 [50]. The mean PDI scores for distress, preoccupation, and conviction were respectively 18.85, 18.08, and 23.15, and neither was significantly different from the general population. The distribution of the AQ and PDI Yes/No scores in this sample can be seen in Figure 3.9.

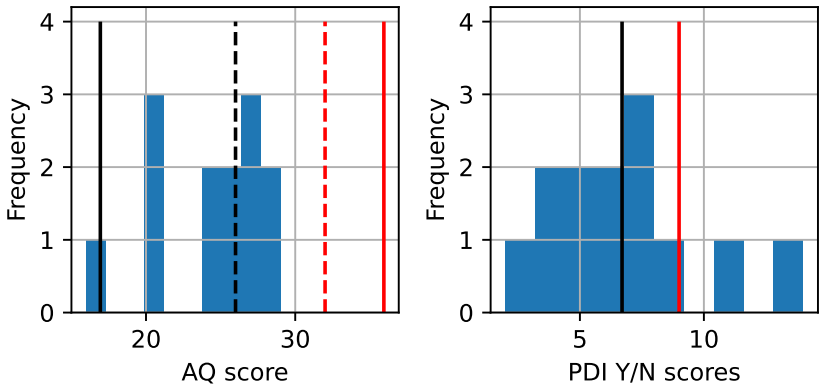


Figure 3.9: The AQ and PDI Yes/No scores measured in the sample participating in Task 2. In both figures, the solid black and red lines indicate the mean score in general and clinical population respectively. In the figure on the left, the dashed black line (AQ scores below 26) delineates the absence of ASD, while the dashed red line (AQ scores above 32) indicates clinically significant level of autistic traits [54][50]

In this sample, we did not observe a significant correlation between the AQ and PDI Yes/No scores, ($p = 0.33$).

3.2.2 Behavioural analysis

Left/Right bias In this version of the task we investigated separately possible left/right bias in the responses after seeing evidence, and also in the prior estimates. The mean of all prior estimates was 0.51, which is not significantly different ($p = 0.71$) than the

mean of optimal model estimates (0.5). For the responses based on seeing evidence, the participant mean is 0.50, which is significantly higher ($p = 0.01$) than the mean of optimal model responses (0.48). However, it is not significantly different from 0.5, and we therefore suggest that the difference from optimal model is caused not by bias, but by other phenomena we observed in participant behaviour, such as inertia.

Final response distribution Next we look at the distribution of the final responses of the participants versus the optimal model, depicted in Figure 3.10. Just as with the no prior version, we analyse absolute confidence ($|r - 0.5|$).

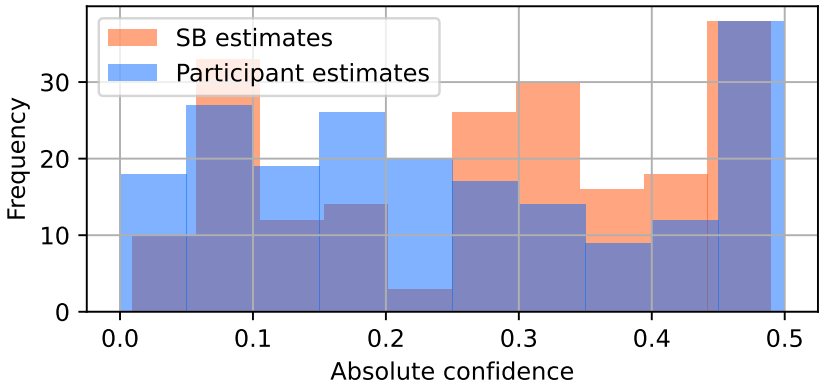


Figure 3.10: The distribution of absolute confidence at the end of each trial of the task version with prior, for the participants (left), and the SB model (right)

We can see that both distributions have multiple peaks. Both peak at absolute confidence. Both distributions also have a second peak in the lower half of the confidence spectrum, but the participant confidences are more spread out than the SB confidences, which cluster around 0.1 and around 0.3. On the whole participant confidence is more concentrated in the lower half of the scale whereas the opposite is true for SB. This can be partly the result of a tendency described in participants' feedback, where some participants at first made a low confidence estimate and then waited until they gathered enough evidence to fully commit to their choice of a jar, expressing no uncertainty.

Prior estimate and integration Since this task introduces new source of information, the prior, we need to investigate whether the participants can and do successfully use this information to make their estimates. We use two metrics to do this, whose distribution can be seen in Figure 3.11. First is the deviation between the participants' confidence before seeing any beads, based on the prior alone, and the real prior. We observe that the mean of this deviation is -0.009, which is not significantly different from 0 ($p = 0.26$), suggesting that participants' estimates are noisy but unbiased. The mean absolute deviation was 0.07, and in Figure 3.11 we can see that vast majority of participants' estimates was within 0.15 of the real prior. Aside from assessing whether the participants can use the prior information on its own, we also want to know if they integrate it with new evidence. We used the absolute update size after observing the first bead to analyse this. We can see in Figure 3.11 that in the vast majority of cases participants do not update their prior estimate at all after observing the first bead. This has been confirmed by qualitative exploration of the participant data, and also in

the participants' feedback. Multiple participants described that the prior information impacted their confidence strongly, and they were not willing to change their estimate until they observed enough evidence, i.e. more than 1 bead. This behaviour is very different than that of the optimal model, which always updates its confidence based on evidence.

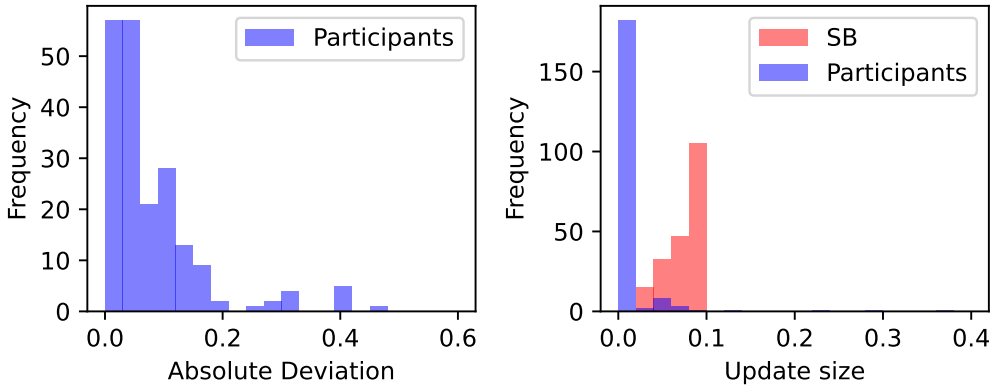


Figure 3.11: The distribution of absolute deviations between the real priors and the participant estimates (left) and the distribution of the absolute update size after observing the first bead in each trial for both the participants and SB model (right)

Update size distribution Next we extend the analysis of participants' confidence updates to all the bead draws. We find out that the mean estimate update -0.0015 is not significantly different from 0 ($p = 41$), which suggests that there is no bias and we can proceed to analyse only the absolute update sizes. The participants' mean absolute update size is 0.036, which is significantly less than the optimal model's mean update size 0.069 ($p = 0.0000$). We can see the full distributions of update sizes for both the participants and the optimal model in Figure 3.12. We can see that most of participants' updates are 0, which confirms the trend observed above. Participants seem to update their confidence estimate only after observing enough evidence, not every time new evidence is available. This tendency is much more pronounced than in the task without prior, as can be seen in Figure 3.4b.

Summary Our model-free exploration of the participant data suggests that people understand the task, but do not make their estimates in the expected manner. The main deviation is strong inertia in their responses confirmed by both qualitative and quantitative exploration - participants mostly do not change their confidence after each new observation, but instead wait until more evidence accumulates before making a new estimate. This tendency is problematic, since it does not make participants' decision making transparent to us and the models we are trying to fit to the data. Furthermore, this tendency seems to be much stronger than in the no prior task.

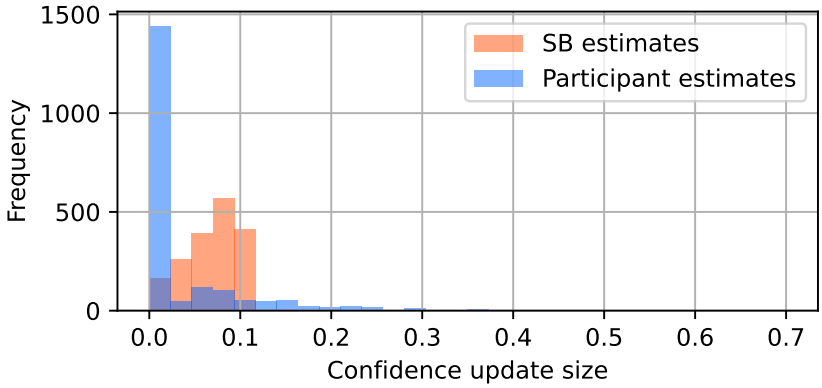


Figure 3.12: The distribution of update sizes after observing evidence in Task 2, for both participants and the Simple Bayes model

3.2.3 Computational analysis

3.2.3.1 Model comparison

Our model comparison methodology is the same as for the no prior version of the task, the results can be seen in Figure 3.13. For this sample, all three methods (BIC, AIC, and cross-validation) picked Weighted Bayes as the best model, with WBB as second. The WBB parameters that fitted participant data the best made the model identical to WB: w_p and w_s were identical to the best-fitting weights of WB, and γ was estimated to be 0. The optimal γ value for the SBB model is also 0, making it equivalent to SB. This suggests that the added bias did not improve the fit of the participant data, although in this version of the task the results of model-based analysis should be taken with skepticism, since our model-free analysis revealed that participants behave differently from either of the analysed models.

When we looked at which model fits each participant the best (using AIC), we found that for 7 participants it is WB, for 2 it is SB, and WBB for just 1. This likewise suggests that participants' responses were not biased in the way defined by our models.

Figure 3.14 depicts the accuracy of WB model's predictions of participant responses. Surprisingly, their correlation is fairly good at $r = 0.86$, considering the observed phenomenon where participants do not update their confidence after every draw, unlike the WB model. It is possible that once participants change their confidence, their confidence update is similar to the sum of all the updates made by the model since the participant's last update, so that broad correlation is preserved despite differences in trajectory.

3.2.3.2 Parameter estimation and psychiatric traits

We will use the best-fitting model WB to estimate and analyse the parameters for each participant's data. First we check if there is any relationship between the participant's psychiatric traits and the model that fits their data the best, Figure 3.15. It seems there are no clearly discernible clusters, so we proceed to analyse all participants using WB

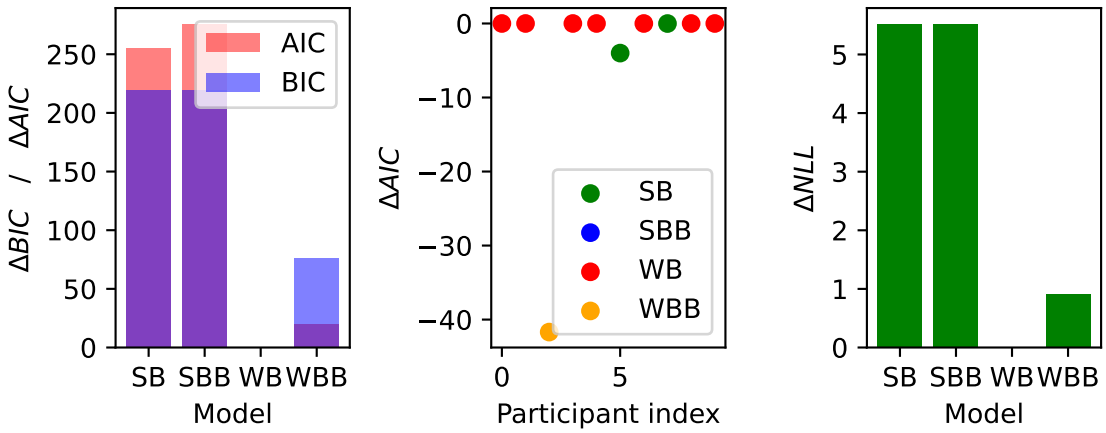


Figure 3.13: Model comparison for Task 2. On the left, we see the difference between the BIC/AIC score of each model and the best-fitting model according to each metric. In the middle, we see the best fitted model for each participant, together with the difference between its AIC and the AIC of best-fitting WB. On the right, we see the difference in the cross-validation average NLL between each model and the best-fitting WB model.

	r	p
AQ, w_p	-0.08	0.8235
AQ, w_s	-0.12	0.7378

	r	p
PDI Y/N, w_p	0.09	0.7900
PDI Y/N, w_s	0.37	0.2981

Table 3.4: Relationship between fitted parameters of the WB model and psychological traits for the task with prior

only.

The distribution of the fitted parameters is shown in Figure 3.16. We can see that the prior weights are all very close to one ($\bar{w}_p = 0.98, std_{w_p} = 0.02$), while the likelihood weights are more spread out ($\bar{w}_s = 0.87, std_{w_s} = 0.14$). However, we recorded a significant correlation between the two parameters ($r = 0.76, p = 0.01$). We observed no such correlation between these two parameters in parameter recovery, which suggests that this correlation is more likely to be due to a real correlation between the two parameters than due to the parameters trading off one another.

Next we examine whether there is any relationship between the estimated parameters and participants' psychological traits. The results are recorded in Table 3.4. We see that, as expected, there is no significant correlation between either AQ or PDI Yes/No score and either of the parameters.

We split the participants into the low and high AQ groups based on the same criteria as in the no prior task. In this sample, 5 participants were in the high AQ group and 5 in the low AQ group. We tested if there is any significant difference between the medians of the estimated parameters for each group, depicted in Table 3.5. Using the Mann-Whitney U test, we found no significant difference for either of the parameters. This is unsurprising considering the sample size.

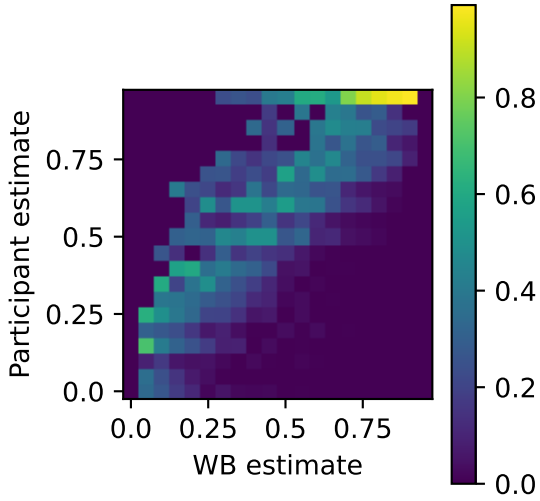


Figure 3.14: Heatmap of the accuracy of the WB model's predictions of participant responses. Both axes are split into 20 bins, and the colour of each bin indicates the proportion of times the participant's estimate fell into the corresponding y-bin when the model's prediction fell into the corresponding x-bin

	Δ median	p
w_p	-0.008	0.84
w_s	0.113	0.67

Table 3.5: Difference of medians of the estimated WB parameters between the high AQ and low AQ group for the task with prior

We also did model comparison for each of the two groups separately. Using AIC, we found that both groups were best fitted by the WB model, which is in line with the evidence above demonstrating no difference between the two groups for our small sample.

3.2.3.3 Further analysis

Similarly to the no prior task, participants' feedback mentioned that they felt like the experiment was too long and tiring, and suggested possible change in performance towards the end as a result. We again assessed this by splitting the participant data into first 10 trials and the second 10 trials, and fitting and evaluating the models separately for each group. We find that both groups are best fitted by the WB model, which does not indicate that there was any change in decision making in the later stages of the experiment. This is contrary to the findings for that task without prior, where the earlier trials were best fitted by SB, whereas the later ones by WB. This conflict is especially surprising given that the task with prior lasted longer than the one without prior. One possible explanation is that the task with prior was more interesting for the participants and thus managed to keep them engaged longer.

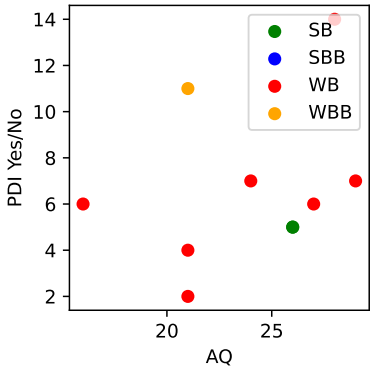


Figure 3.15: The relationship between the best-fitting model for each participant and their psychiatric traits in the task with prior

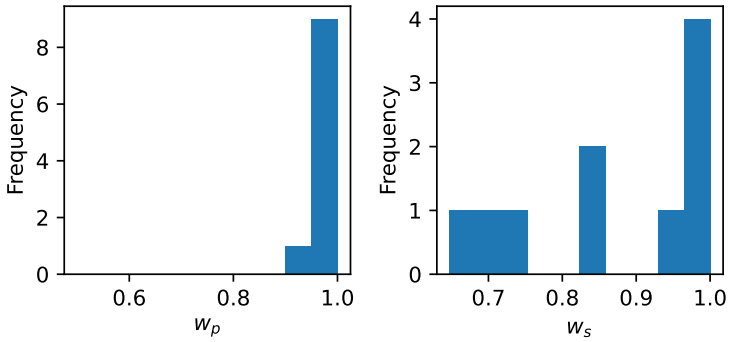


Figure 3.16: The distribution of estimated parameters of the WB model for each participant in the task with prior

3.2.4 Participant feedback

Participants' comments on the general experimental setup were similar to the no prior task. They found the experiment a bit too long, and would have appreciated a progress bar indicating how many trials there are left. When they describe their strategy, 2 comments occur multiple times. First one is that the prior information about how many jars there are of each colour impacted their confidence strongly. The second recurring comment was that the participants did not change their confidence after every draw. Especially at the beginning of each trial, when few beads had been observed, but also later on, participants would describe waiting for about two or three draws before adjusting their confidence estimate.

3.2.5 Task 2 Summary

Our analysis shows that peoples' behaviour in the task with prior is different from that of either of our models. The most significant deviation is the strong inertia we observed both quantitatively and qualitatively. Both participant feedback and their responses suggest that the prior information strongly influences their estimates, and it takes a large amount of evidence accumulated over several bead draws to change their initial estimate

based on the prior alone. The inertia effect was to some extent present also in the no prior task 1, but in task 2 it seems to be reinforced by the explicit prior. Therefore we do not believe that the inertia could be eliminated by any simple modification to the experiment instructions. In future work, task 2 should not be used unless new models are designed that incorporate the inertia observed in participant behaviour. The fact that neither of our current models captures the participants' behaviour also limits the conclusions we can make based on computational analysis. The fact that WB was the best-fitting model suggests that participants' behaviour deviates from optimal Bayesian inference, but these deviations are better captured by a slight imbalance in prior and likelihood weighting than by confirmation bias as defined by our models.

Chapter 4

Conclusions

4.1 Discussion

There is growing research indicating that people with ASD may be less susceptible to various cognitive biases, and thus may have ”enhanced rationality” [53]. We hypothesised that these findings can extend to the confirmation bias, which has, to the best of our knowledge, not yet been analysed with regards to possible differences between people with ASD and general population. In order to conduct such a research, it is necessary to have a reliable measure of individual differences in confirmation bias. One of the objectives of this project has thus been to design and validate models and experimental tasks that would be capable of measuring such differences.

To this purpose, we decided to use Bayesian models of belief updating used previously to explore how people integrate prior and likelihood information in probabilistic reasoning [59] [9]. One of these was the Simple Bayes model, and the other was the Weighted Bayes model. Bayesian models were picked because they provide a principled way to handle integration of prior and likelihood information, which makes them a useful framework for exploring how peoples' prior beliefs bias their interpretation of new evidence. We modified the original models to include confirmation bias based on the model proposed by Baccini and Hartmann [3]. Their model of the confirmation bias proposes that people update their beliefs more based on evidence that is in agreement with their prior beliefs than based on disconfirmatory evidence, and the strength of this effect is a function of the strength of their prior beliefs.

Multiple explanations have been proposed to account for the enhanced rationality observed in people with ASD. One of the main theories postulates that the lower susceptibility to bias is the result of reduced influence of prior and contextual knowledge on the reasoning of people with ASD, resulting in a more detail-oriented style of reasoning. This theory is consistent with a Bayesian theory of ASD proposed by Pellicano and Burr [48], which proposes that the differences observed in people with ASD are caused by the fact they have broader priors in the Bayesian sense, and thus their prior beliefs influence their interpretation of new evidence less than in neurotypical people. The Weighted Bayes model we decided to modify to include confirmation bias is capable of capturing such an underweighting of prior information, as well as

overweighting of new evidence, which has been proposed as an alternative to Pellicano and Burr's theory [7]. A second purpose of the experiments we design is thus to explore whether the Weighted Bayes model, or its modification including confirmation bias, show any differences in the relative weighting of prior and likelihood information between individuals with low and high autistic trait scores.

Since Pellicano and Burr proposed their Bayesian theory of ASD, there has been a large amount of research trying to assess whether people with ASD truly overweight likelihood information relative to the prior information. However, the results of this research are mixed. Our experiment concerns the specific case of learned priors, i.e. priors that are acquired during the experiment. A review by Chrysaitis et al. [10] shows that when the priors are learned implicitly, studies are split between those demonstrating an imbalance in prior and likelihood weighting and those failing to demonstrate such an imbalance. However, when the priors were made explicit, as in our experiment, most studies failed to demonstrate any imbalance. One possible explanation for these null results is that Bayesian priors are generally considered to be implicit, and it is unclear whether an explicitly stated prior information, such as the proportion of jars of each type in our case, can be learned as priors influencing Bayesian reasoning. However, another possible explanation of the generally conflicting results in this field of research is the heterogeneity of the research methodology used in studies trying to assess presence of prior/likelihood weighting imbalance. Many of the studies do not use any Bayesian modelling, which can provide more mechanistic detail than simple behavioural analysis. Additionally, even the studies that do use modelling often fail to follow all the recommended modelling guidelines that aim to increase the confidence we can have in model-based analysis of psychiatric data [64]. To avoid such pitfalls, we paid special attention to following these guidelines. Specifically, even though we were mainly interested in modelling confirmation bias, we also analysed alternative models, which did not include bias. Additionally, we performed parameter and model recovery to assess how much trust we can have in the results from our computational analysis, and we validated the best-fitting model by comparing its behaviour to the real participants' data.

These additional measures helped us uncover several shortcomings of our approach that would otherwise go unnoticed. First of all, we discovered that the data is better fitted by the unbiased Weighted Bayes model than by the biased one, and the Simple Bayes model also performed better than its biased version. This suggests that participants' judgements are not influenced by confirmation bias in the manner assumed by our models. Additionally, our model recovery analysis revealed that the Weighted Bayes Biased model has poor recovery, and we therefore had to exclude it from our analysis.

We conducted two pilot experiments to validate our proposed models and experimental tasks. We observed an unexpected "inertia" tendency, where participants did not update their confidence estimate after each new observation, but rather only once more evidence was accumulated over the course of multiple bead draws. This tendency seems to be particularly strong in the task version with prior. Based on participant feedback, we believe this is the result of the fact that the prior information about the relative proportions of the two jar types strongly influences participants' confidence, to the extent that only multiple conflicting bead observations can change it. We therefore do

not think that the task version without prior can be used in further research unless new models capable of incorporating the inertia effect are designed.

The inertia effect was present also in the task version without prior, but to a lesser extent. This leads us to believe that relatively minor modifications to the experiment could eliminate it entirely. One such modification would be changes to the experiment instructions that would encourage the participants to always state their current confidence even if they are unsure. A less subtle modification would be to force the participants to update their confidence estimate after every draw. In the current implementation of the experiment, there is the possibility to proceed to the next bead draw without changing the confidence slider position. This possibility could be removed in future implementations. However, there is a risk that this would force the participants to behave in a manner that does not truly reflect their confidence, and thus we believe that less aggressive modifications should be tried first.

The results from both prior experiments suggest that participants' estimates deviated from optimal Bayesian inference. However, these deviations were better captured by simple imbalance in prior and likelihood weighting of the WB model than by the confirmation bias implemented in the WBB or SBB models. We thus have to conclude that if confirmation bias affects peoples' decision making in the tested tasks, it does not do so in the way assumed by our models.

4.2 Future work

The purpose of the pilot experiments was to validate the pipeline for the two versions of the experiment in preparation for a large-scale experiment. Our results show that the version of the experiment with prior is not ready for large scale deployment, since the participants exhibit inertia that our models do not account for. The no prior version of the task, with some modifications, therefore seems more promising. However, the poor model recovery of the Weighted Bayes Biased model suggests that even if model comparison picked it as the best model for this task, we still could not have full confidence in it. Furthermore, the pilot results suggest that the participants' estimates are not influenced by confirmation bias in the way it is defined in our proposed models.

However, we believe that the no-prior version of the sequential beads task could be used to explore probabilistic belief updating in future research, particularly to explore possible differences in prior and likelihood weighting between people with ASD and neurotypical population. This task is similar to tasks used previously to explore Bayesian theories of ASD and schizophrenia, e.g. the fisher task used by Jardri et al. [29] and Chrysaitis et al. [9], and particularly the social beads task used by Simonsen et al. [59]. However, these tasks consider only a single-trial static integration of prior and likelihood, whereas our task focuses on sequential belief updating. In the fisher task, participants were supposed to state their confidence that a red fish was caught from one of two available lakes. Two sources of information were made available to them: two fish baskets whose size expressed the fisher's preference for each of the two lakes, and the proportion of black and red fish in each of the two lakes. Chrysaitis et al. found that participants did not take the precise prior information expressed by the fish basket

sizes into account, instead just focusing on which of them was larger. Since our task 1 is sequential, the role of the prior after each bead observation is played by the posterior calculated after the previous observation. Since the participants state their posterior after each observation, the prior for the next draw is made explicit and the vagueness present in the fisher task is thus eliminated.

The social beads task by Simonsen et al. [59] has been previously adapted to a version more similar to our task by Soltani et al., as an MSc student at Peggy Series' lab. In their version of the task, participants were asked to state their confidence about which of two jars a sample of multiple beads has been drawn from. Aside from the beads, they also had an additional source of information: the confidence estimates of two other agents. Chrysaitis et al. piloted a similar experiment and discovered that when participants are presented with prior information about which jar the beads are from and a likelihood information in the form of multiple beads drawn from the jar, they tend to average the two sources of information rather than integrate them in a Bayesian way. E.g. if the prior suggests 0.8 confidence in the majority red jar, and 60% of the drawn beads are red, participants tend to update their confidence to approximately 0.7. We observed no such behaviour - when the participants observed a red bead, they either did not change their estimate, or, much more often, updated it in the direction of the majority red jar, regardless of their previous estimate.

We therefore believe that the no-prior version of the sequential beads task could be used in future research to explore Bayesian theories of ASD and potentially also schizophrenia. In our pilot experiment, the Weighted Bayes model had good model recovery, was best according to BIC, and even its AIC was the best for most individual participants. It therefore seems that the most promising continuation of this project is to focus on differences in the weighting of prior and likelihood information between people with ASD and neurotypical population. An important part in planning an experiment that would address this question is determining the necessary sample size.

4.2.1 Power analysis

Power analysis allows us to calculate the sample size that achieves a desired effect size with desired statistical power at a desired significance level α. Therefore we need to determine all the other variables before we can proceed to calculate the sample size.

The power of an experiment is the probability that the null hypothesis will be correctly rejected based on the sample if the actual effect in the population is equal or greater than the specified effect size [27]. The significance level α is the maximum probability that the null hypothesis will be rejected even if it is actually true. The choice of power and significance level is essentially arbitrary, and since our study does not require extraordinary caution to avoid either false positive or false negative results, we will use the most commonly used values: $\alpha = 0.05$ and power of 0.80.

Another variable that determines the necessary sample size is the effect size. Effect size is essentially the magnitude of the association between a predictor and an outcome variable. In our case, there are four hypotheses we could consider. First two of them are whether there is a correlation between participants' AQ scores (the predictor) and

the estimated parameters w_p or w_s (the outcome variable). In this case, the effect size would be the absolute value of the Pearson's correlation coefficient. The other two hypotheses were whether there is a difference between the estimated parameter w_p or w_s (the outcome variable) between the individuals with high AQ and those with low AQ (the predictor). In this case the effect size would be the difference between the outcome variable's means for each of the two groups.

In related previous research, Karvelis et al. [30] explored differences in how prior and sensory information is combined by participants with high and low AQ scores. They conducted a visual statistical learning experiment where participants were asked to estimate the direction of motion of a cloud of coherently moving dots, and analysed the results using a Bayesian model. This model was characterised by four parameters: the mean and uncertainty of the prior; and mean and uncertainty of the sensory likelihood. The study found a significant correlation between the participants AQ scores and the estimated uncertainty of the sensory likelihood ($r = -0.185, p = 0.011$). The uncertainty of the sensory likelihood has the same function as the likelihood weight in our Weighted Bayes model - the higher the uncertainty of the sensory likelihood, the less the likelihood information influences the posterior. We can therefore use the findings of Karvelis et al. to hypothesise that in our task, we should observe a positive correlation between the participants AQ scores and their estimated likelihood weight w_s.

Using the correlation observed by Karvelis et al. as the expected effect size, we calculated that for an experiment to achieve power of 0.8 when we expect to find an effect of size $r = 0.185$ at a significance level $\alpha = 0.05$, we would need results from 86 participants.

An interesting avenue for future research would be to analyse the data from our experiment using a broader variety of Bayesian models. Of particular interest could be the circular inference models that have been proposed as a mechanistic account of schizophrenia[28], and successfully captured differences in Bayesian inference between people with schizophrenia and neurotypical population [29]. The same models have also been applied to the study of ASD [9], but no significant differences have been found. Our task could provide valuable new data on the possibility that these models can explain the mechanisms behind ASD.

4.3 Conclusion

In this project, we conducted two pilot experiments to validate that our proposed models and the two new behavioural tasks we designed are capable of

- Measuring individual differences in prior and likelihood weighting
- Measuring individual differences in confirmation bias

Both experiments were based on a beads task exploring how participants update their beliefs upon repeatedly observing new evidence. We found that the version of the task with explicit prior information resulted in a strong inertia effect that made our models incompatible with the participant behaviour. This effect was also observed in the no

prior version of the task, but to a considerably lesser extent. This leads us to believe that the inertia could be eliminated by minor modifications to the existing no prior task, which could subsequently be used for further research. The results of both pilot experiments suggest that participants behaviour is not affected by confirmation bias as it is defined by our proposed models. This, together with poor model recovery of the WBB model, leads us to conclude that the current tasks and models are not capable of capturing individual differences in confirmation bias strength. Nevertheless, the no prior task we designed still provides a useful platform to explore individual differences in Bayesian inference, in particular the weighting of prior and likelihood information, as it does not suffer from some of the shortcomings present in previously used experiments. In the pilot experiments, we found no significant relation between the estimated prior and likelihood weights and either the autistic or schizotypical traits of the participants. However, this was expected since our power analysis showed that the sample size in the pilot experiments did not result in sufficient statistical power to detect an effect even if it was present. Possible differences in prior/likelihood imbalance between people with ASD and neurotypical population therefore deserve further exploration in a larger study with sufficient power.

Bibliography

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. 2013.

[2] Saman Babaie–Kafaki and Zohre Aminifard. Two–parameter scaled memoryless BFGS methods with a nonmonotone choice for the initial step length. *Numerical Algorithms*, 82(4):1345–1357, 2019.

[3] Edoardo Baccini and Stephan Hartmann. The myside bias in argument evaluation: A bayesian model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44), 2022.

[4] Vincent Berthet. The measurement of individual differences in cognitive biases: A review and improvement. *Frontiers in Psychology*, 12, 2021.

[5] Ruth Beyth-Marom and Baruch Fischhoff. Diagnosticity and pseudiagnosticity. *Journal of Personality and Social Psychology*, 45(6):1185–1195, 1983.

[6] A. Bonnel, L. Mottron, I. Peretz, M. Trudel, E. Gallun, and A. M. Bonnel. Enhanced pitch sensitivity in individuals with autism: A signal detection analysis | journal of cognitive neuroscience | MIT press. *Journal of Cognitive Neuroscience*, 15(2):226–235, 2003.

[7] Jon Brock. Alternative bayesian accounts of autistic perception: comment on pellicano and burr. *Trends in Cognitive Sciences*, 16(12):573–574, 2012.

[8] Nick Chater, Mike Oaksford, Ulrike Hahn, and Evan Heit. Bayesian models of cognition. *Wiley Interdisciplinary Reviews Cognitive Science*, 1(6):811–823, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.79.

[9] Nikitas Angeletos Chrysaitis, Renaud Jardri, Sophie Denève, and Peggy Seriès. No increased circular inference in adults with high levels of autistic traits or autism. *PLOS Computational Biology*, 17(9):e1009006, 2021. Publisher: Public Library of Science.

[10] Nikitas Angeletos Chrysaitis and Peggy Seriès. 10 years of bayesian theories of autism: a systematic review, 2022.

[11] Michael Cipriano and Thomas S. Gruca. The power of priors: How confirmation bias impacts market prices. *The Journal of Prediction Markets*, 8(3):34–56, 2014. Number: 3.

[12] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

[13] Jazz Croft, Christoph Teufel, Jon Heron, Paul C. Fletcher, Anthony S. David, Glyn Lewis, Michael Moutoussis, Thomas H. B. FitzGerald, David E. J. Linden, Andrew Thompson, Peter B. Jones, Mary Cannon, Peter Holmans, Rick A. Adams, and Stan Zammit. A computational analysis of abnormal belief updating processes and their association with psychotic experiences and childhood trauma in a UK birth cohort. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7(7):725–734, 2022.

[14] Joshua R. de Leeuw and Benjamin A. Motz. Psychophysics in a web browser? comparing response times collected with JavaScript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, 48(1):1–12, 2016.

[15] B. De Martino, N. A. Harrison, S. Knafo, G. Bird, and R. J. Dolan. Explaining enhanced logical consistency during decision making in autism | journal of neuroscience. *The Journal of Neuroscience: : the official journal of the Society for Neuroscience*, 28(42):10746–10750, 2008.

[16] K. B. Dohr, A. J. Rush, and I. H. Bernstein. Cognitive biases and depression. *Journal of Abnormal Psychology*, 98(3):263–267, 1989.

[17] Dirk M. Elston. Confirmation bias in medical decision-making. *Journal of the American Academy of Dermatology*, 82(3):572, 2020. Publisher: Elsevier.

[18] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002. Number: 6870 Publisher: Nature Publishing Group.

[19] Jonathan St B. T. Evans. The influence of prior belief on scientific thinking. In Michael Siegal, Peter Carruthers, and Stephen Stich, editors, *The Cognitive Basis of Science*, pages 193–210. Cambridge University Press, 2002.

[20] Jacob Feldman. Bayesian contour integration. *Perception & Psychophysics*, 63(7):1171–1182, 2001.

[21] Junya Fujino, Shisei Tei, Takashi Itahashi, Yuta Aoki, Haruhisa Ohta, Chieko Kanai, Manabu Kubota, Ryu-ichiro Hashimoto, Motoaki Nakamura, Nobumasa Kato, and Hidehiko Takahashi. Sunk cost effect in individuals with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 49(1):1–10, 2019.

[22] Thomas Gilovich. Biased evaluation and persistence in gambling. *Journal of personality and social psychology*, 44(6):1110–1126, 1983.

[23] N. D. Goodman, J. B. Tenenbaum, T. L. Griffiths, and J. Feldman. *Compositionality in rational analysis: grammar-based induction for concept learning*. Oxford : Oxford University Press, 2008.

[24] T. L. Griffiths, A. N. Sanborn, K. R. Canini, and D. J. Navarro. *Categorization as nonparametric Bayesian density estimation*. Oxford : Oxford University Press, 2008.

[25] T. L. Griffiths and J. B. Tenenbaum. Structure and strength in causal induction. *Cognitive Psychology*, 51(4):334–384, 2005.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2009.

[27] Stephen B. Hulley. *Designing clinical research*. Wolters Kluwer/Lippincott Williams & Wilkins, 4th ed edition, 2013.

[28] Renaud Jardri and Sophie Denève. Circular inferences in schizophrenia. *Brain: A Journal of Neurology*, 136:3227–3241, 2013.

[29] Renaud Jardri, Sandrine Duverne, Alexandra S. Litvinova, and Sophie Denève. Experimental evidence for circular inference in schizophrenia. *Nature Communications*, 8(1):14218, 2017. Number: 1 Publisher: Nature Publishing Group.

[30] Povilas Karvelis, Aaron R. Seitz, Stephen M. Lawrie, and Peggy Seriès. Autistic traits, but not schizotypy, predict increased weighting of sensory information in bayesian visual integration. *eLife*, 7:e34115, 2018. Publisher: eLife Sciences Publications, Ltd.

[31] N. M. Kleinhans, L. C. Johnson, T. Richards, R. Mahurin, J. Greenson, G. Dawson, and E. Aylward. Reduced neural habituation in the amygdala and social impairments in autism spectrum disorders. *The American Journal of Psychiatry*, 166(4):467–475, 2009.

[32] Gregor Kohls, Martin Schulte-Rüther, Barbara Nehrkorn, Kristin Müller, Gereon R. Fink, Inge Kamp-Becker, Beate Herpertz-Dahlmann, Robert T. Schultz, and Kerstin Konrad. Reward system dysfunction in autism spectrum disorders. *Social Cognitive and Affective Neuroscience*, 8(5):565–572, 2013.

[33] D. Kuhn. Children and adults as intuitive scientists. *Psychological Review*, 96(4):674–689, 1989.

[34] Bojana Kuzmanovic, Lionel Rigoux, and Kai Vogeley. Brief report: Reduced optimism bias in self-referential belief updating in high-functioning autism. *Journal of Autism and Developmental Disorders*, 49(7):2990–2998, 2019.

[35] Meng-Chuan Lai, Michael V. Lombardo, and Simon Baron-Cohen. Autism. *The Lancet*, 383(9920):896–910, 2014.

[36] A. E. Lerman and D. Acland. United in states of dissatisfaction: Confirmation bias across the partisan divide. *American Politics Research*, 48(2):227–237, 2020.

[37] Moa Lidén, Minna Gräns, and Peter Juslin. The presumption of guilt in suspect interrogations: Apprehension as a trigger of confirmation bias and debiasing techniques. *Law and Human Behavior*, 42(4):336–354, 2018.

[38] C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

[39] Peter Mitchell and Danielle Ropar. Visuo-spatial abilities in autism: A review. *Infant and Child Development*, 13(3):185–198, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/icd.348.

[40] Kinga Morsanyi, Simon J. Handley, and Jonathan S. B. T. Evans. Decontextualised minds: Adolescents with autism are less susceptible to the conjunction fallacy than typically developing adolescents. *Journal of Autism and Developmental Disorders*, 40(11):1378–1388, 2010.

[41] Laurent Mottron, Sylvie Belleville, and Edith Ménard. Local bias in autistic subjects as evidenced by graphic tasks: Perceptual hierarchization or working memory deficit? *Journal of Child Psychology and Psychiatry*, 40(5):743–755, 1999. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1469-7610.00490.

[42] Carolyn B. Murray. Estimating achievement performance: A confirmation bias. *Journal of Black Psychology*, 22(1):67–85, 1996. Publisher: SAGE Publications Inc.

[43] C. R. Mynatt, M. E. Doherty, and R. D. Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology*, 29(1):85–95, 1977.

[44] S. Narayanan and Daniel Jurafsky. A bayesian model predicts human parse preference and reading times in sentence processing. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[45] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[46] Barbara O'Brien. Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy, and Law*, 15(4):315–334, 2009.

[47] Edward H. Patzelt, Catherine A. Hartley, and Samuel J. Gershman. Computational phenotyping: Using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience*, 1:e18, 2018. Publisher: Cambridge University Press.

[48] Elizabeth Pellicano and David Burr. When the world becomes 'too real': a bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10):504–510, 2012.

[49] E. R. Peters, S. A. Joseph, and P. A. Garety. Measurement of delusional ideation in the normal population: introducing the PDI (peters et al. delusions inventory). *Schizophrenia Bulletin*, 25(3):553–576, 1999.

[50] Emmanuelle Peters, Stephen Joseph, Samantha Day, and Philippa Garety. Measuring delusional ideation: The 21-item peters et al. delusions inventory (PDI). *Schizophrenia Bulletin*, 30(4):1005–1022, 2004.

[51] Eric Rassin. Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology*, 64(2):87–93, 2008.

[52] Danielle Ropar and Peter Mitchell. Shape constancy in autism: the role of prior knowledge and perspective cues. *Journal of Child Psychology and Psychiatry*, 43(5):647–653, 2002. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1469-7610.00053.

[53] Liron Rozenkrantz, Anila M. D'Mello, and John D. E. Gabrieli. Enhanced rationality in autism spectrum disorder. *Trends in Cognitive Sciences*, 25(8):685–696, 2021.

[54] Emily Ruzich, Carrie Allison, Paula Smith, Peter Watson, Bonnie Auyeung, Howard Ring, and Simon Baron-Cohen. Measuring autistic traits in the general population: a systematic review of the autism-spectrum quotient (AQ) in a non-clinical population sample of 6,900 typical adult males and females. *Molecular Autism*, 6:2, 2015.

[55] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 1978.

[56] Peggy Series. *Computational Psychiatry: A Primer*. MIT Press, 2020.

[57] Amitta Shah and Uta Frith. An islet of ability in autistic children: A research note. *Journal of Child Psychology and Psychiatry*, 24(4):613–620, 1983. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.1983.tb00137.x.

[58] Paul B. Sharp, Isaac Fradkin, and Eran Eldar. Hierarchical inference as a source of human biases. *Cognitive, Affective, & Behavioral Neuroscience*, 2022.

[59] A. Simonsen, R. Fusaroli, M. L. Petersen, A. Q. Vermillet, V. Bliksted, O. Mors, A. Roepstorff, and D. Campbell-Meiklejohn. Taking others into account: combining directly experienced and indirect information in schizophrenia. *Brain: A Journal of Neurology*, 144(5):1603–1614, 2021.

[60] Aistis Stankevicius, Quentin J. M. Huys, Aditi Kalra, and Peggy Seriès. Optimism as a prior belief about the probability of future reward. *PLOS Computational Biology*, 10(5):e1003605, 2014. Publisher: Public Library of Science.

[61] Suzan van Brussel, Miranda Timmermans, Peter Verkoeijen, and Fred Paas. Teaching on video as an instructional strategy to reduce confirmation bias—a pre-registered study. *Instructional Science*, 49(4):475–496, 2021.

[62] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140, 1960.

[63] Axel Westerwick, Benjamin K. Johnson, and Silvia Knobloch-Westerwick. Confirmation biases in selective exposure to political online information: Source bias

vs. content bias. *Communication Monographs*, 84(3):343–364, 2017. Publisher: Routledge _eprint: https://doi.org/10.1080/03637751.2016.1272761.

[64] Robert C. Wilson and Anne G. E. Collins. Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547, 2019. Publisher: eLife Sciences Publications, Ltd.

[65] Todd S. Woodward, Steffen Moritz, Carrie Cuttler, and Jennifer C. Whitman. The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, 28(4):605–617, 2006. Publisher: Routledge _eprint: https://doi.org/10.1080/13803390590949511.

Appendix A

Participants' information sheet

A.1 Task 1: Belief updating without prior

STUDY TITLE Investigating the relationship between decision-making and personality traits

RESEARCHERS David Daubner, Peggy Seriès

You are being asked to take part in a research study on decision-making and personality. This study was certified according to the Informatics Research Ethics Process, RT number 29394. Please take time to read the following information carefully.

WHAT WILL HAPPEN In this study, you will be asked to answer some questionnaires about psychological traits and afterwards you will be automatically redirected to take part in a decision-making task. In the task you will be shown some coloured glass beads and then you will be asked to make confidence judgements about the jar which they came from. You will be explained the procedure in more detail before starting the experiment.

TIME COMMITMENT The study typically takes about 30 minutes.

PARTICIPANTS' RIGHTS You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be withdrawn/destroyed. You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should send an email to the researchers before participating in the study.

BENEFITS AND RISKS There are no known benefits or risks for you in this study.

STUDY RESULTS The results of this study may be summarised in published articles, reports and presentations. Key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

If you want to find out about the final results of this study, you should send David an email (email address provided below) and he will email you the final report whenever it is available.

DATA PROTECTION AND CONFIDENTIALITY Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number. Your data will only be viewed by the researcher and the supervisors. All data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint).

DATA PROTECTION RIGHTS The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

FOR FURTHER INFORMATION David Daubner or Peggy Seriès will be glad to answer your questions about this study at any time. You may contact David via email at s1965226@ed.ac.uk and Peggy at pseries@inf.ed.ac.uk.

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

A.2 Task 2: Belief updating with prior

STUDY TITLE Investigating the relationship between decision-making and personality traits

RESEARCHERS David Daubner, Peggy Seriès

You are being asked to take part in a research study on decision-making and personality. This study was certified according to the Informatics Research Ethics Process, RT number 29394. Please take time to read the following information carefully.

WHAT WILL HAPPEN In this study, you will be asked to answer some questionnaires about psychological traits and afterwards you will be automatically redirected to take part in a decision-making task. In the task you will be shown some coloured glass beads and then you will be asked to make confidence judgements about the jar which they came from. You will be explained the procedure in more detail before starting the experiment.

TIME COMMITMENT The study typically takes about 30 minutes.

PARTICIPANTS' RIGHTS You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied

to that point be withdrawn/destroyed. You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should send an email to the researchers before participating in the study.

BENEFITS AND RISKS There are no known benefits or risks for you in this study.

STUDY RESULTS The results of this study may be summarised in published articles, reports and presentations. Key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

If you want to find out about the final results of this study, you should send David an email (email address provided below) and he will email you the final report whenever it is available.

DATA PROTECTION AND CONFIDENTIALITY Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number. Your data will only be viewed by the researcher and the supervisors. All data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint).

DATA PROTECTION RIGHTS The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

FOR FURTHER INFORMATION David Daubner or Peggy Seriès will be glad to answer your questions about this study at any time. You may contact David via email at s1965226@ed.ac.uk and Peggy at pseries@inf.ed.ac.uk.

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Appendix B

Participants' consent form

CONSENT By proceeding with the study, I agree to all of the following statements: I have read and understood the above information. I understand that my participation is voluntary, and I can withdraw at any time. I consent to my anonymised data being used in academic publications and presentations. I allow my data to be used in future ethically approved research.