Cognition 245 (2024) 105693

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

Confirmation bias emerges from an approximation to Bayesian reasoning

Charlie Pilgrim [a,b,*], Adam Sanborn [c], Eugene Malthouse [c], Thomas T. Hills [c,d]

[a] The Mathematics of Real-World Systems CDT, The University of Warwick, Coventry, CV4 7AL, UK
[b] Experimental Psychology, University College London, London, WC1H 0DS, UK
[c] Psychology, The University of Warwick, Coventry, CV4 7AL, UK
[d] The Alan Turing Institute, The British Library, 96 Euston Road, London, NW1 2DB, UK

ARTICLE INFO

Dataset link: https://github.com/chasmani/PUBLIC_confirmation_bias_and_the_BIASR_model

Keywords:
Confirmation bias
Source reliability
Information processing
Bayesian
Cognitive model
Bounded rationality

ABSTRACT

Confirmation bias is defined as searching for and assimilating information in a way that favours existing beliefs. We show that confirmation bias emerges as a natural consequence of boundedly rational belief updating by presenting the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). In this model, an individual's beliefs about a hypothesis and the source reliability form a Bayesian network. Upon receiving information, an individual simultaneously updates beliefs about the hypothesis in question and the reliability of the information source. If the individual updates rationally then this introduces numerous dependencies between beliefs, the tracking of which represents an unrealistic demand on memory. We propose that human cognition overcomes this memory limitation by assuming independence between beliefs, evidence for which is provided in prior research. We show how a Bayesian belief updating model incorporating this independence approximation generates many types of confirmation bias, including biased evaluation, biased assimilation, attitude polarisation, belief perseverance and confirmation bias in the selection of sources.

1. Introduction

Confirmation bias is the search for and assimilation of information in a way that favours the preservation of prior beliefs (Nickerson, 1998). It has been described as one of the most pernicious (Nickerson, 1998) of the cognitive biases, with impacts felt in many social domains including religion (Batson, 1975; Nickerson, 1998), politics (Lord, Ross, & Lepper, 1979; Nickerson, 1998; Nyhan & Reifler, 2010; Taber & Lodge, 2006), climate change (Cook & Lewandowsky, 2016; Hart & Nisbet, 2012), health and medicine (Liberman & Chaiken, 1992; Malthouse, 2022; Nickerson, 1998), justice (Nickerson, 1998), stereotyping (Darley & Gross, 1983), conspiracy theories (McHoskey, 1995), and science (Mahoney, 1977; Nickerson, 1998). Understanding the underlying cognitive mechanisms that drive confirmation bias is therefore of fundamental theoretical and practical interest.

Confirmation bias encompasses numerous distinct but closely related behaviours (Fischhoff & Beyth-Marom, 1983; Friedrich, 1993; Klayman & Ha, 1987; Nickerson, 1998). Though many such behaviours have been identified, we focus on five here which have received wide empirical support (see Klayman (1995) for a review): (i) *biased evaluation*: judging information that opposes one's views more critically than that which supports them (Koehler, 1993; Lord et al., 1979; Russo, Medvec, & Meloy, 1996; Taber & Lodge, 2006); (ii) *biased assimilation*: whereby people are less influenced by opposing than confirmatory sources (Lord et al., 1979; Taber & Lodge, 2006); (iii) *attitude polarisation*: extreme views both for and against a hypothesis can become more extreme upon seeing the same evidence (Lord et al., 1979; Taber & Lodge, 2006); (iv) *belief perseverance*: the reluctance to change beliefs in the face of disconfirmatory evidence (Anderson, Lepper, & Ross, 1980; Batson, 1975); and (v) *confirmation bias in the selection of sources*: preferring sources of information that confirm existing beliefs (Redlawsk, 2002; Taber & Lodge, 2006).

Given confirmation bias' wide prevalence and potential negative impact, a natural question is why confirmation bias exists at all? Sufficiently costly tendencies should be expected to disappear under evolutionary pressures (Nickerson, 1998), unless they are themselves an adaptive solution to a more costly alternative. While confirmation bias may be an impediment to finding the truth, the adaptive force on cognition is primarily towards pragmatic survival and only secondarily concerned with truth seeking (Friedrich, 1993). In light of this, we may ask is confirmation bias truly a dysfunction, or does it serve some adaptive purpose?

Explanations for confirmation bias have been put forward at the social (Mercier & Sperber, 2011; Norman, 2016; Peters, 2020), individual (Festinger, 1962; Friedrich, 1993; Kunda, 1990; Nickerson, 1998) and information processing levels (Cook & Lewandowsky, 2016; Gerber & Green, 1999; Henderson & Gebharter, 2021; Jern, Chang, &

* Correspondence to: Experimental Psychology, 26 Bedford Way, University College London, WC1H 0DS, UK.
E-mail address: c.pilgrim@ucl.ac.uk (C. Pilgrim).

https://doi.org/10.1016/j.cognition.2023.105693
Received 25 April 2023; Received in revised form 30 October 2023; Accepted 11 December 2023
Available online 19 January 2024

Kemp, 2014; Koehler, 1993). These levels of analysis are qualitatively different but are nonetheless connected. Social behaviour emerges from individual behaviour, and individual behaviour emerges, in part, from information processing. In this paper we present a normative explanation at the information processing level, although our description complements many existing social and individual explanations.

Before we go further, it is helpful to discuss the definition of "bias". In the psychological literature, the word bias is used to mean a variety of things (Hahn & Harris, 2014). This ranges from the everyday usage of the term as a leaning or tendency in one direction, to the precise use in statistics of a systematic departure from accuracy. Within the context of research on beliefs, bias is usually accepted to mean a departure from a normative model (Hahn & Harris, 2014), which is often Bayesian rationality (Hahn & Harris, 2014; Klayman, 1995). This definition introduces difficulties because behaviour that is irrational given one belief-updating model may be rational given a different belief-updating model. This has been the case for biased evaluation (Koehler, 1993) and attitude polarisation (Cook & Lewandowsky, 2016; Henderson & Gebharter, 2021; Jern et al., 2014). We aim to sidestep the issue by not claiming that the behaviours are fundamentally biased under all possible belief-updating models. Rather, we will define behaviours as departures from specific rational belief updating models described in previous literature.

Our contribution is multifold. We present a model of information processing that can generate a large range of empirically confirmed confirmation bias type behaviours, more so than other explanations. In particular, we explore existing Bayesian models of inference in a world with uncertain beliefs and unreliable sources of information (Bovens, Hartmann, et al., 2003; Hahn, Merdes, & von Sydow, 2018; Koehler, 1993; Merdes, Von Sydow, & Hahn, 2020; Olsson, 2011). We argue that maintaining full rationality is impossible for realistic agents due to the high memory demands of remembering dependencies between beliefs. As a consequence, humans are forced to make approximations in order to maintain complex world models. We demonstrate how these approximations to rationality can introduce small biases that magnify as data is processed sequentially over time. In different task domains, these biases encompass the five confirmation bias behaviours we list above.

We will begin with a discussion of information processing models of belief updating in the literature. This will lead to a description of the BIASR model and an interrogation of each of its assumptions. We will then evaluate each of the 5 confirmation bias type behaviours we list above, defining each and showing how the BIASR model can generate the behaviour. We will end with a general discussion of the model and its position in the literature.

An example may help build intuition. Alice has a neutral belief about vaccine safety. She talks to her new neighbour Bob, who tells her that vaccines have not been thoroughly tested and that they are dangerous. Alice is at first not entirely convinced, but she does become slightly more wary about vaccines. The next week Bob again tells Alice about the dangers of vaccines. Alice is more receptive now as she already has a slight belief that vaccines are dangerous, and she starts to see Bob as reliable because his information matches her slight belief. This continues for several weeks until Alice is convinced that vaccines are dangerous and that Bob is a very reliable source of information. When Alice now hears on the news that vaccines are safe she is not convinced — after all, both she and Bob cannot both be wrong, especially considering how knowledgeable Bob is. Alice's beliefs about the reliability of Bob and the dangers of vaccines are correlated. If she forgets this correlation then she does not give enough consideration to the counterfactual world where Bob is wrong and vaccines are safe. In this example Alice exhibits biased evaluation, biased assimilation and belief perseverance.

2. Models of information processing

Bayes' theorem provides the objectively optimal way to update beliefs given new evidence, where beliefs are described in terms of degrees of uncertainty. For human cognition, inference affects behaviour, which in turn affects adaptive success. One could therefore expect that adaptive pressures over our evolutionary history would drive our inference mechanisms towards Bayesian rationality.

The **simple version of Bayes' theorem** is

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}. \tag{1}$$

If one applies this simple rule to a single hypothesis, H, then data from all sources is treated equally. There is no judgement of evidence quality. Under this model, any unequal treatment of evidence is considered biased evaluation. Indeed, bias as the unequal consideration of evidence is a definition of confirmation bias (often implicitly) used in the literature (Lord, Lepper, & Preston, 1984; Lord et al., 1979; Miller, McHoskey, Bane, & Dowd, 1993; Plous, 1991).

The simple version of Bayes' theorem is not adequate in terms of describing either observed or desirable behaviour (see Table 1 and below). Indeed, it is reasonable to judge the quality of evidence based on assessments of the reliability of the source (Fischhoff & Beyth-Marom, 1983). Lord et al. (1979) state,

> "Our subjects' main inferential shortcoming ... did not lie in their inclination to process evidence in a biased manner. Willingness to interpret new evidence in the light of past knowledge and experience is essential for any organism to make sense of, and respond adaptively to, its environment. Rather, their sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially "justified" the processing bias".

Here Lord et al. (1979) are suggesting **biased evaluation prior to assimilation**. Data is first evaluated based on prior beliefs, with unlikely data considered as less reliable evidence. And then the result of this evaluation determines the weight of the evidence in updating those same prior beliefs. This idea was formalised mathematically by Gerber and Green (1999), who present a Bayesian model of belief updating combined with biased learning. The biased learning is represented as a weakening of the strength of evidence that disconfirms prior beliefs, before updating those same beliefs within the Bayesian machinery. They provide an example of a politician's supporters considering whether the politician is corrupt or not. In this example, evidence in support of corruption is discounted by a factor, $\alpha < 1$.

Russo et al. proposed a similar model involving "predecisional distortion of information" in relation to choice among alternatives (Russo, 2014, 2018; Russo et al., 1996). Prior preferences influence the evaluation of data, and this evaluation influences how the data is used to update beliefs, generating a bias towards initial preferences (Russo, 2014, 2018). These ideas can describe biased evaluation and biased assimilation, and can go some way to describing belief perseverance (Carlson, Meloy, & Russo, 2006) (Table 1). Though useful, these models are not Bayesian and do not have a clear normative basis.

The Bayesian framework does, however, allow us to incorporate evidence evaluation in the form of **Bayesian updating including source reliability**. Koehler (1993) argues that a normative account of belief updating should consider an individual's prior beliefs about source reliability as well as evidence evaluation. Koehler (1993) proposes a rational Bayesian model that includes source reliability and which is able to generate biased evaluation (Table 1). This model supports proposals in the literature that judging evidence based on prior beliefs is not necessarily irrational, such that it can be rational to consider unlikely evidence more critically (Fischhoff & Beyth-Marom, 1983; Klayman, 1995; Koehler, 1993; Lord et al., 1979).

Table 1
Comparison of information processing models of confirmation bias. Checkmarks denote which behaviours have been explained using the different models.

	Simple version of Bayes' theorem	Bayesian updating including source reliability (Koehler, 1993)	Bayesian networks (Cook & Lewandowsky, 2016; Henderson & Gebharter, 2021; Jern et al., 2014)	Biased evaluation prior to assimilation (Gerber & Green, 1999; Lord et al., 1979)	Belief-based sequential updating with source reliability (Bovens et al., 2003; Hahn et al., 2018; Merdes et al., 2020; Olsson, 2011)	BIASR. Bayesian updating with an independence approximation and source reliability
Biased evaluation		✓	✓	✓	✓	✓
Biased assimilation				✓	✓	✓
Attitude polarisation			✓			✓
Belief perseverance				✓	✓	✓
Selection of sources						✓

This idea is extended by Bovens et al. (2003) and Olsson (2011) to account for sequential belief updates when receiving data over time. In both these cases, individuals maintain separate beliefs about a central hypothesis and source reliability, and upon receiving information those beliefs are updated simultaneously (Merdes et al., 2020). This type of Bayesian updating has been described as a **"belief-based" strategy** for inference in a world with unknown source reliability (Hahn et al., 2018; Merdes et al., 2020). As data is received, beliefs about the central hypothesis and source reliability both change over time, which influence how subsequent data is interpreted. It has been shown that these models demonstrate order effects such that the order in which information is received changes the final belief (Hahn et al., 2018). This has been described as a form of confirmation bias (Hahn et al., 2018; Merdes et al., 2020) (connected to belief perseverance), and the updating process is a departure from Bayesian rationality in the same way as the model that we present. In this research, the normative argument is that the rational approach would be too much of a cognitive burden, as it would require remembering all data received so far and updating from initial priors whenever data is received (Hahn et al., 2018). However, we contend that this normative argument is not complete as it does not consider the alternative rational approach of maintaining a joint belief distribution over the central hypothesis and source reliability.

More generally, **Bayesian networks** are graphical representations in which nodes represent variables and edges represent dependencies between these variables, allowing us to reason about events influenced by multiple factors (Pearl, 2009). Directed edges are commonly used to represent causal relationships between variables. In some representations, undirected or dashed edges can capture non-causal conditional dependencies, such that gaining information about one variable can inform us about another variable. In Fig. 1, we use solid directed edges to denote causal dependencies and dashed undirected edges to indicate non-causal information dependencies. Cook and Lewandowsky (2016) used Bayesian networks to explain attitude polarisation in participants who were given evidence about climate change. They demonstrated that polarisation can be rational given a Bayesian network, because individuals' beliefs about the evidence they observed were influenced not only by whether they believed climate change was true, but also by their worldview and trust in scientists (Cook & Lewandowsky, 2016).

Jern et al. (2014) generalise this idea by describing the set of Bayesian networks that can lead rational agents to attitude polarisation — crucially this set of networks share the property that upon receiving some data, beliefs about more than one hypothesis are updated simultaneously. In order to generate rational attitude polarisation, individuals require differences in prior beliefs about the "central" hypothesis in question, and importantly also some difference in other "auxiliary" prior beliefs (Gerber & Green, 1999; Henderson & Gebharter, 2021). For example, those with strong views about the dangers of climate change may also believe that scientific evidence is more reliable than those who are less worried about climate change (Cook & Lewandowsky, 2016). Bayesian networks can also be used to describe biased evaluation (Jern et al., 2014) (Table 1).

So far our discussion of information processing has centred on Bayesian rationality. However, this is not necessarily the appropriate normative standard when modelling human probabilistic reasoning. We must also take into account realistic cognitive constraints (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020; Daw, Courville, & Dayan, 2008; Klayman, 1995). People update many hypotheses simultaneously (Gershman, 2019), which can be computationally demanding (Dasgupta et al., 2020). Dependencies between hypotheses mean the computational scale of inference can quickly overwhelm any realistic agent, who will be forced to make approximations to optimal Bayesian inference. To understand how human cognition may overcome this limitation we can take inspiration from computer science, a field with much experience in the approximation of computationally expensive Bayesian reasoning (Sanborn, 2017). This path from computer science to human cognition is a well worn road, and algorithms such as Markov Chain Monte Carlo have shed light on human cognitive processes including behavioural biases (Sanborn, 2017).

3. The BIASR model assumptions

We present the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability; see Fig. 1), which rests on the following assumptions:

1. Source reliability. Upon receiving information, we update our beliefs about the reliability of the source.
2. Simultaneous updating. We update our beliefs about source reliability and the central hypotheses at the same time.
3. Independence approximation. Simultaneous updating introduces dependencies between our beliefs about (a) central hypotheses and (b) source reliabilities. Our model approximates these dependencies away by taking marginal beliefs and assuming independence.
4. Sequential Updating. Data is received and processed sequentially over time. The independence approximation is applied between sequential updates.

Each of these assumptions will be explored in the following sections. The belief updating process under BIASR is visualised in Fig. 1.

3.1. Source reliability

It is rational to hold and update beliefs about the reliability of sources of information (Hahn et al., 2018; Merdes et al., 2020). Doing so allows us to weigh the quality of evidence based on the source and protects against the influence of unreliable sources that may foster misinformation. The inclusion of source reliability in belief updating models has been suggested as a possible rational basis for the conjunction fallacy (Bovens et al., 2003; Fischhoff & Beyth-Marom, 1983; Jarvstad & Hahn, 2011; Tversky & Kahneman, 1983). There is also an abundance of empirical evidence that people track source reliability (Liberman & Chaiken, 1992; Lord et al., 1979; Mahoney, 1977; Taber & Lodge, 2006).

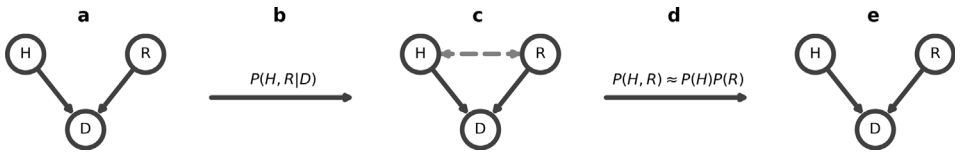
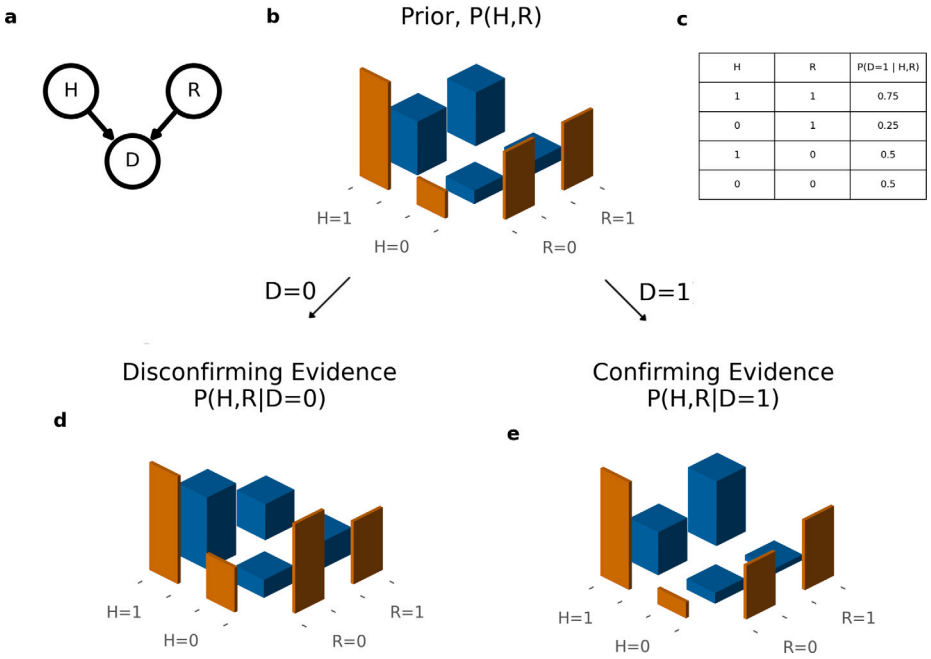


Fig. 1. The BIASR belief updating model. (a) Data received is believed to be causally influenced by both the true value of the hypothesis at hand and the source reliability in a collider Bayesian network. (b) When receiving data, beliefs about the hypothesis and source reliability are updated simultaneously. (c) This updating introduces information dependencies (grey dotted line) between beliefs. (d) These belief correlations can be approximated away by assuming independence. (e) This approximation simplifies the Bayesian network structure. The process repeats when more data is received.



H	R	P(D=1 | H,R)
1	1	0.75
0	1	0.25
1	0	0.5
0	0	0.5

Fig. 2. Simultaneous updating of source reliability, R, and the central hypothesis, H. Blue (thick) bars show the joint belief distribution. Orange (thin) bars show the marginal belief distributions. (a) The Bayesian network structure describes how the individual believes the data is influenced by the true values of the central hypothesis and source reliability, given by (c) the conditional probability distribution. (b) Prior beliefs favour the central hypothesis, $P(H = 1) = 0.8$, and are neutral about the source reliability, $P(R = 1) = 0.5$. (d) The posterior following disconfirming evidence shows how disconfirmatory data can be explained away as coming from an unreliable source, with only a small impact on belief in the central hypothesis. (e) The posterior following confirming evidence is updated towards stronger belief in both the central hypothesis and the reliability of the source. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Simultaneous updating of source reliability and central hypotheses

In the absence of an objective standard of truth, we can judge a source's reliability based on our assessment of the plausibility of the information received (Hahn et al., 2018). If someone tells us that Elvis Presley is outside, it is a fair guess that we will not believe them. Instead, we are likely to downgrade our belief in them as a reliable source. We update our belief about their source reliability and Elvis simultaneously. Our strong prior belief in the hypothesis that Elvis is not outside is protected by an auxiliary hypothesis in the reliability of the source. In this way our belief in source reliability can absorb disconfirmatory evidence about strongly held central beliefs (Fig. 2). Empirical evidence shows that people do consider information about source reliability when updating beliefs about a hypothesis, and vice versa (Collins, Hahn, Von Gerber, & Olsson, 2018).

The idea that our beliefs are not updated in isolation is known, in the context of scientific epistemology, as the Duhem-Quine thesis (Gershman, 2019). No hypothesis can be tested in isolation and upon receiving evidence we update a set of beliefs together, sometimes partitioned into central and auxiliary hypotheses, while maintaining overall coherence. The auxiliary hypotheses (e.g., source reliability) can act to absorb disconfirmatory evidence, allowing us to maintain central beliefs (Fig. 2). This can be rational; if a scientist detects faster than light travel it is sensible to question the accuracy of the measurements (Gershman, 2019; Lord et al., 1979). Empirically, scientists question whether disconfirmatory evidence is the result of an error before abandoning a central hypothesis (Dunbar, 1995). And people test hypotheses more extensively when told that discomfirmatory evidence may be in error (Gorman, 1989).

Following existing models in the literature (Koehler, 1993; Merdes et al., 2020), our model assumes that individuals believe evidence, $D \in \{0, 1\}$, is influenced by both the truth value of the central hypothesis, $H \in \{0, 1\}$, and the reliability of the source, $R \in \{0, 1\}$ (Fig. 1). In our model all sources are less than perfectly reliable, but a "reliable" source, $R = 1$, has less noise and is more likely to report the true value of the central hypothesis than an "unreliable" source, $R = 0$, which has more noise. These probabilities can be quantified as $P(D = 1|R = 1, H = 1)$ for the reliable source and $P(D = 1|R = 0, H = 1)$ for the unreliable source (and symmetrical in the case that $H = 0$). A quantitative example is given in Fig. 2. The true values of the central hypothesis and the source reliability are causally independent. Our individual has prior beliefs in the central hypothesis, $P(H)$, and the source reliability, $P(R)$. Assuming initial independence between these beliefs, this is a collider type Bayesian network, $H- > D < -R$. Notably, this is one of the set of Bayesian networks that Jern et al. (2014) proved can lead to attitude polarisation. Beliefs are simultaneously updated by Bayes' rule,

$$P(H, R|D) = \frac{P(D|H, R)P(H, R)}{P(D)}. \quad (2)$$

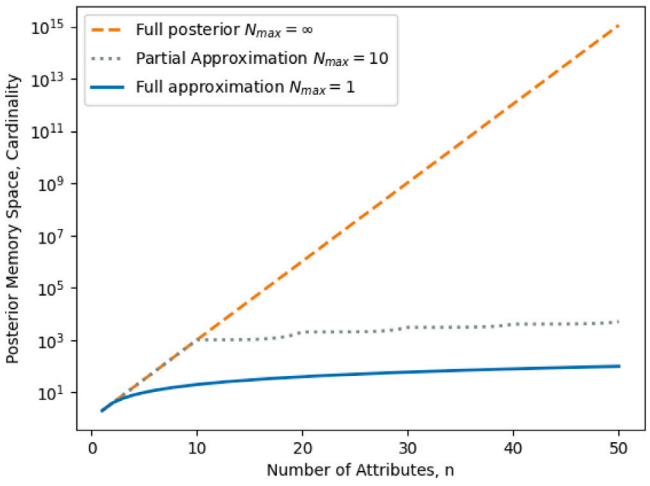


Fig. 3. Memory space scaling as a function of number of binary attributes. With no approximation (red dashed line) memory requirements scale exponentially (a straight line on this log-linear figure). A full mean field approximation (blue solid line) scales linearly. A partial approximation (grey dotted line) scales exponentially up to d and then linearly after, in this case with $d = 10$.

Initially, the individual's beliefs in H and R are independent so that $P(H, R) = P(H)P(R)$ and we can simplify the update rule to

$$P(H, R|D) = \frac{P(D|H, R)P(H)P(R)}{P(D)}. \tag{3}$$

A general property of causal graphs of this structure, including collider type Bayesian networks, is that upon receiving data, beliefs in H and R are no longer independent (Pearl, 2009) (Fig. 1). Mathematically, the individual's beliefs no longer fulfil the independence relationship and $P(H)P(R) \neq P(H, R)$. If we later learn that Elvis is alive, we should update our belief in our friend's reliability. Moreover, upon receiving subsequent evidence from the same source about the same hypothesis, in order to remain Bayesian rational we can no longer use the simpler update rule (Eq. (3)) and must instead consider the full joint belief distribution (Eq. (2)).

Notably, this formalism is different to previous work on belief-based updating (Hahn et al., 2018; Merdes et al., 2020). In that work, beliefs are assumed to be stored marginally, i.e. individuals have some belief about the hypothesis in question and a separate belief about the reliability of a source. Given this belief structure, the rational benchmark is to remember the entire history of data received and to carry out full inference on all the data at each timestep (Hahn et al., 2018), which Merdes et al. (2020) argue is unrealistic and not normative. We argue that the more complete rational benchmark also includes the possibility of maintaining a joint belief distribution as described here. However, as we will see in the following section, normative arguments based on cognitive limitations recover the marginal belief structure.

3.3. Independence approximation

In our minimal example (Fig. 2) we are considering only one central hypothesis, H, and one source reliability, R, each of which can take values of either 1 or 0. In this case, the joint belief distribution is relatively small, with 4 possible combinations, $\{(H = 1, R = 1), (H = 1, R = 0), (H = 0, R = 1), (H = 0, R = 0)\}$, shown as the blue (thick) bars in Fig. 2. In our actual day to day reasoning we track many more hypotheses, each with many possible values, and evidence from many different sources. If we combine all these then there are many possible combinations to track in the joint belief distribution.

We can ask how the size of the hypothesis space scales as we add more attributes to a world model. In computer science, the amount of computational resources that an algorithm uses is known as the computational complexity. This can be measured in terms of processing time or memory space. Big O notation is a way of comparing the computational complexity of algorithms as the size of inputs to that algorithm grows. As we add new attributes to our world model the memory space requirements scale exponentially, as $\sim O(k^n)$, where n is the number of attributes and k is the number of hypotheses per attribute. This is the curse of dimensionality (Bellman, 2015 (1957)). If we were tracking 300 binary attributes about the world then the joint belief distribution would have size $2^{300} \approx 10^{90}$ — more than the number of atoms in the observable Universe.

This combinatorial explosion in memory space will quickly exhaust any reasonable level of cognitive resources (or for a fixed cognitive resource, will limit the richness and resolution of the agent's world model). A realistic agent would therefore require an approximation to optimal Bayesian reasoning, and this should be included in a normative account. Alternatively, one could remember the entire history of data received from all sources and carry out the full inference with the initial priors each time new data is received, but this also imposes an unrealistic computational burden (Merdes et al., 2020).

Variational approximations are an approach to approximating Bayesian reasoning that can reduce computational requirements associated with large posterior distributions (Ormerod & Wand, 2010). One option is to take a mean field approximation of the joint posterior distribution by partitioning variables and assuming that the partitions are independent (Ormerod & Wand, 2010; Sanborn & Silva, 2013). This kind of approximation has been applied before to understanding how human behaviour can emerge as a consequence of cognition overcoming computationally intractable problems, in the realm of associative learning (Sanborn, 2017; Sanborn & Silva, 2013). At one extreme, if we assume all variables are independent then our memory requirements now scale linearly as $\sim O(kn)$, a vast improvement. We can now track 300 binary attributes with a belief distribution of size 600. If we limit belief partitions to d variables, then computation scales linearly as $\sim O(n(d^k))$. This type of partial or structured mean field approximation (Sanborn, 2017) will preserve dependencies between some variables while avoiding the curse of dimensionality. Fig. 3 shows an example of this partial approximation with $d = 10$ variables, as compared to no approximation and a full mean field approximation. It should be noted that we are still updating beliefs simultaneously — the mean field approximation disentangles these beliefs following simultaneously updating.

The quality of this compression in terms of loss of information will depend upon the degree to which the attributes being inferred are actually independent, and the type of approximation we make. A common choice of measure to guide the approximation is to minimise the Kullback–Leibler divergence between the full posterior and the approximated posterior (Ormerod & Wand, 2010; Sanborn, 2017). This is achieved by taking marginal belief distributions for each attribute. By assuming independence the full joint belief distribution can be approximated from the marginal belief distributions (Fig. 1),

$$P(H, R) \approx P(H)P(R). \tag{4}$$

Human cognition is unlikely to always use a full approximation, such that people would be unable to remember any dependencies between beliefs — the key point is that people are unlikely to remember all the dependencies and will have to make some approximations. Do these approximations include forgetting dependencies between source reliability and more central beliefs? There is evidence that people do not always correctly associate sources of information with their beliefs and instead people can experience source confusion where the belief remains but the source is mis-attributed (Johnson, Hashtroudi, & Lindsay, 1993).

3.4. Sequential updating

A common assumption underlying Bayesian inference is exchangeability, i.e. that the order that data is received is irrelevant (Gelman, Carlin, Stern, & Rubin, 1995). Data can be processed in any order, or all at once, and the final beliefs will be the same. This assumption holds in the rational case if the data generating process is static, as in our model. However, exchangeability does not hold under the BIASR model because the independence approximation introduces path dependency, such that biases accumulate over successive steps. Therefore, the order that data is received and sequentially processed influences the final beliefs. A general consequence of sequential updating with approximations to Bayesian inference is the potential loss of exchangeability and the introduction of effects that are dependent on the order of processing (Daw et al., 2008).

There is evidence that people do not process data all at once, but update sequentially. Empirical evidence for this includes the primacy effect (Bruner & Potter, 1964), where the order that data is seen has an influence on final beliefs. In the realm of decision making, Russo (2014) describe a stepwise evolution of preference paradigm. This stepwise updating has been shown empirically in many contexts, with experiments showing that people sequentially update their preferences and their opinion on the diagnosticity of the data (Russo, 2014).

4. Evaluation of five forms of confirmation bias

In this section we evaluate the BIASR model in relation to the five forms of confirmation bias outlined in Table 1. For each form of confirmation bias, we first discuss the literature and empirical evidence. We then define a mathematical requirement for this behaviour in the context of Bayesian rationality. Following this definition, we simulate the behaviour under different models of information processing.

For each form, we simulate how an individual could update their beliefs about whether a central hypothesis is true or false, $H \in \{1, 0\}$, and whether a source is reliable or not, $R \in \{1, 0\}$. If a source is reliable, they transmit the true state of the hypothesis 75% of the time, $P(D = 1|H = 1, R = 1) = 0.75$. If the source is unreliable, they transmit the true value only 50% of the time, $P(D = 1|H = 1, R = 0) = 0.5$. In most cases we use a neutral prior on the source reliability, $P(R) = 0.5$, and a strong prior belief in the central hypothesis, $P(H) = 0.8$. In the case of attitude polarisation, we also include a strong prior belief against the central hypothesis, $P(H) = 0.2$. And in the case of belief perseverance we start with a neutral prior in the central hypothesis, $P(H = 1) = 0.5$. In all the examples, the simulated individual receives multiple datums sequentially from either a single source or two sources. The values used, and the problem setup itself, are intended to minimally demonstrate the behaviours as clearly as possible. The behaviours are robust and emerge under a wide range of parameters (see the Appendix for a sensitivity analysis).

We consider 3 information processing models:

1. Simple version of Bayes' theorem. Beliefs in source reliability are not updated at all, i.e. the prior belief, in this case $P(R) = 0.5$, remains the same. Beliefs in the central hypothesis are updated according to Bayes' rule,

$$P(H|D) = \frac{P(H)\sum_R P(D|H, R)P(R)}{P(D)}. \tag{5}$$

2. Rational updating including source reliability. Beliefs about the central hypothesis and source reliability are updated simultaneously. This introduces a dependency in the joint belief distribution, $P(H, R|D)$. This dependency is remembered between successive datums by updating using Bayes' rule over the full joint belief distribution (Eq. (2)). Given the data generating process, this is the rational way to update beliefs. As such, exchangeability holds and this is equivalent to updating on all data received using initial priors.

3. BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). Beliefs are updated as in the rational case, but dependencies between the central hypothesis and source reliability are forgotten between successive datums. We take marginal beliefs

$$P(H) = \sum_R P(H, R) \tag{6}$$

and

$$P(R) = \sum_H P(H, R). \tag{7}$$

We ignore dependencies by using these marginal beliefs in the independent version of Bayes' rule (Eq. (3)).

All code and data can be downloaded from https://github.com/chasmani/PUBLIC_confirmation_bias_and_the_BIASR_model

4.1. Biased evaluation (Biased assimilation)

In the confirmation bias literature, the terms biased evaluation and biased assimilation are often used interchangeably. We can strictly define evaluation as the judgement of the quality of the evidence and assimilation as concerning the degree of belief change in the central hypothesis at hand. These are separate, but connected, beliefs. In many studies what could strictly be thought of as biased evaluation is sometimes called biased assimilation (Lord et al., 1984; Miller et al., 1993). This is understandable as the meanings of the two overlap: if a piece of evidence is rated as "more convincing" (Lord et al., 1984) or "more persuasive" (Miller et al., 1993), is that evaluation or assimilation? From a cognitive dissonance perspective, assimilation and evaluation are connected through coherence in beliefs — a disconfirmatory piece of evidence creates a cognitive dissonance that can be resolved through biased evaluation (Kunda, 1990). Or more simply, contrary evidence is explained away as coming from an unreliable source. An early mention of confirmation bias is found in the writings of Bacon (1878 [1620]), who also links evaluation and assimilation,

> "Once a human intellect has adopted an opinion (either as something it likes or as something generally accepted), it draws everything else in to confirm and support it. Even if there are more and stronger instances against it than there are in its favour, the intellect either overlooks these or treats them as negligible or does some line-drawing that lets it shift them out of the way and reject them. This involves a great and pernicious prejudgment by means of which the intellect's former conclusions remain inviolate". *Francis Bacon* (Bacon 1878 [1620])

There is strong empirical evidence for biased evaluation, some of which also supports biased assimilation. Mahoney (1977) found that scientists judged studies more harshly when the findings disagreed with their own theoretical positions. This was followed by Lord et al. (1979), who ran an experiment with two sets of students — those with strong prior opinions either for or against capital punishment. Both groups were shown the same set of evidence that consisted of studies for and against capital punishment. When the students were asked to rate the quality of the evidence, the studies that agreed with their position were rated higher than those that disagreed. Students were also asked to self-report on their degree of attitude change following reading the studies, finding that the students rated confirmatory studies as having a greater influence. Lord, Ross and Lepper went on to replicate those findings and explore confirmation bias in different contexts in a range of papers, Lord et al. (1984) and Vallone, Ross, and Lepper (1985).

Gilovich (1983) recruited volunteer students to gamble on American football games, and found evidence of biased evaluation in the post-match description of losses and wins, with losses more likely to be explained away. They were even able to influence participants' future

likelihood of gambling on a match by mentioning that a previous match was decided by a "fluke" play that could have gone either way, and so bringing into question the reliability of the previous match result as a predictor of future results. Liberman and Chaiken (1992) found that caffeine drinkers were more critical of messaging that linked caffeine to health problems (Liberman & Chaiken, 1992). Koehler (1993) found a bias in scientists evaluating studies that either agreed or disagreed with their prior positions. McHoskey (1995) found that prior beliefs had a strong effect on people's ratings of the persuasiveness of evidence for and against a conspiracy. Malthouse (2022) found biased evaluation in the assessment of evidence for the efficacy of vaccines. These studies represent just some of the empirical evidence for biased evaluation.

It has often been pointed out that judging evidence based on prior beliefs is not irrational, as it can be rational to consider unlikely evidence more critically (Fischhoff & Beyth-Marom, 1983; Klayman, 1995; Koehler, 1993; Lord et al., 1979). Nevertheless this effect is still often called *biased evaluation* — judging confirmatory sources more favourably, and disconfirmatory sources less favourably. We will follow this naming convention and define a sufficient condition given our minimal model:

$$P(R|\boldsymbol{D}_{for}) > P(R|\boldsymbol{D}_{against}), \tag{8}$$

where $\boldsymbol{D}_{for}$ is a set of data that agrees with a prior hypothesis, and $\boldsymbol{D}_{against}$ disagrees to the same extent.

Fig. 4 shows biased evaluation effects with a strong initial prior belief in the central hypothesis, $P(H) = 0.8$, and a neutral prior belief in source reliability, $P(R) = 0.5$. With the BIASR model, we see biased evaluation as the confirmatory sources (Fig. 4**a**) are judged to be more reliable than the disconfirmatory sources (Fig. 4**b**). Given our model setup, the message receiver will eventually be persuaded and come to trust the source. This is because at worst an unreliable source is sending only noise. If we instead allowed anti-reliable sources, who consistently lie, then the overall effect would be stronger and it is possible for trust in a source to consistently move towards 0.

Notably, when given confirmatory information the belief in the reliability of the source is lower in the BIASR model than the rational Bayesian network model (Fig. 4**a**). This was unexpected and is related to an underestimation of probability mass for the correlated beliefs $P(H = 1, R = 1)$ in the BIASR model. This is explored in the Appendix.

Given our model, a sufficient condition for biased assimilation is if an individual updates their beliefs in the central hypothesis more so than they would do under the rational version of Bayes' theorem with the Bayesian network. In the case of confirmatory evidence

$$P(H|\boldsymbol{D}_{for}) > P(H|\boldsymbol{D}_{for})_{rational}. \tag{9}$$

And in the case of disconfirmatory evidence,

$$P(H|\boldsymbol{D}_{against}) > P(H|\boldsymbol{D}_{against})_{rational}. \tag{10}$$

Biased assimilation is simulated in Fig. 4. The BIASR model shows a stronger posterior belief in the central hypothesis than rational updating under both the Bayesian network and simple models, for both confirmatory and disconfirmatory evidence.

What causes these dynamics? When receiving confirmatory data a positive correlation is induced between H and R — it is more likely that the source is either correct and reliable, $P(H = 1, R = 1)$, or incorrect and unreliable, $P(H = 0, R = 0)$, than the alternatives. With the BIASR model, the agent forgets about this correlation, i.e. the agent forgets that their belief in the central hypothesis is partly due to their belief in the source reliability, and vice versa. One consequence is that the agent does not give enough consideration to the counterfactual world where the central hypothesis is wrong and the source is unreliable. A similar pattern happens with disconfirmatory evidence. During the independence approximation, probability mass is effectively moved away from correlated beliefs where those beliefs go against an indiviudal's priors. This is explored further in the Appendix, where we explore belief updating across the entire joint belief distribution, $P(H, R)$.

We gave an intuitive example of confirmatory evidence in the Introduction with Alice and Bob. Here we give an example in the case of disconfirmatory evidence. Alice has a strong belief that vaccines are dangerous. She meets a new acquaintance, Chris, who tells her that vaccines are actually safe. This goes against Alice's strongly held views and so she naturally questions how reliable Chris is, and only updates her beliefs about vaccines slightly. The next time they meet, Chris again raises points about vaccine safety. This information again goes against Alice's views, and this time she already has question marks over Chris's reliability and is able to dismiss the evidence more easily. Over time, Alice is able to hold onto her belief that vaccines are dangerous and dismiss Chris as an unreliable source. Under the BIASR model, she does not remember the relationship between her belief in Chris' reliability and her beliefs in vaccine safety. As a consequence, she gives little consideration to the possibility that Chris is reliable and vaccines are safe.

4.2. Attitude polarisation

In the study that we described above by Lord et al. (1979), people's evaluation of the evidence for and against capital punishment depended on their prior beliefs: Those who were pro-capital punishment self-reported that the evidence swayed them to be more fervent in their beliefs, and those who were against capital punishment stated that they also became more fervent, in the opposite direction — the two groups diverged in their beliefs after seeing the same data. This attitude polarisation has been replicated in the context of climate change (Cook & Lewandowsky, 2016), gun control (Taber & Lodge, 2006), affirmative action (Taber & Lodge, 2006), the Iraq war (Nyhan & Reifler, 2010), the JFK assassination (McHoskey, 1995), homosexual stereotypes (Munro & Ditto, 1997), drug use (Taber, Cann, & Kucsova, 2009), freedom of speech (Taber et al., 2009) and nuclear energy (Plous, 1991).

A sufficient condition for *attitude polarisation*, given our model, is that individuals with different prior beliefs in the central hypothesis update in opposite directions,

$$\begin{cases} P(H|\boldsymbol{D}) > P(H), & \text{if} \quad P(H) > 0.5 \\ P(H|\boldsymbol{D}) < P(H), & \text{if} \quad P(H) < 0.5 \end{cases}. \tag{11}$$

$\boldsymbol{D}$ is the same set of evidence shown to both individuals, which can include more than one source and multiple datums from each source, both for and against hypotheses.

When considering a single hypothesis in isolation, it is a property of Bayesian updating that different prior beliefs will converge given the same data (or more precisely, not diverge). However, if we have a more complicated belief structure then it can be rational for individuals to update in opposite directions. This was confirmed by Jern et al. (2014), who prove a family of Bayesian network motifs that can lead to attitude polarisation. They go on to analyse the results of Lord et al. (1979) and offer two potential Bayesian network structures that could create attitude polarisation in this experiment. For instance, if an individual who has a strong pro-capital punishment prior also has a belief that the consensus is biased against capital punishment, then studies that are anti-capital punishment can be explained away as resulting from the biased consensus, while studies that are pro-capital punishment are strong evidence in support of capital punishment (Jern et al., 2014). If individuals who are anti-capital punishment also believe that there is a bias in consensus, in this case a bias in favour of capital punishment, then it is rational for these individuals to also strengthen their beliefs when seeing the same data (Jern et al., 2014). The pro- and anti- groups can rationally update their beliefs in opposite directions. Other recent work explaining polarisation has proposed additional mechanisms such as relying on participants optionally stopping evidence accumulation when they are polarised (Kvam, Alaukik, Mims, Martemyanova, & Baldwin, 2022) or having an additional private source of affective information (Melnikoff & Strohminger, 2023).

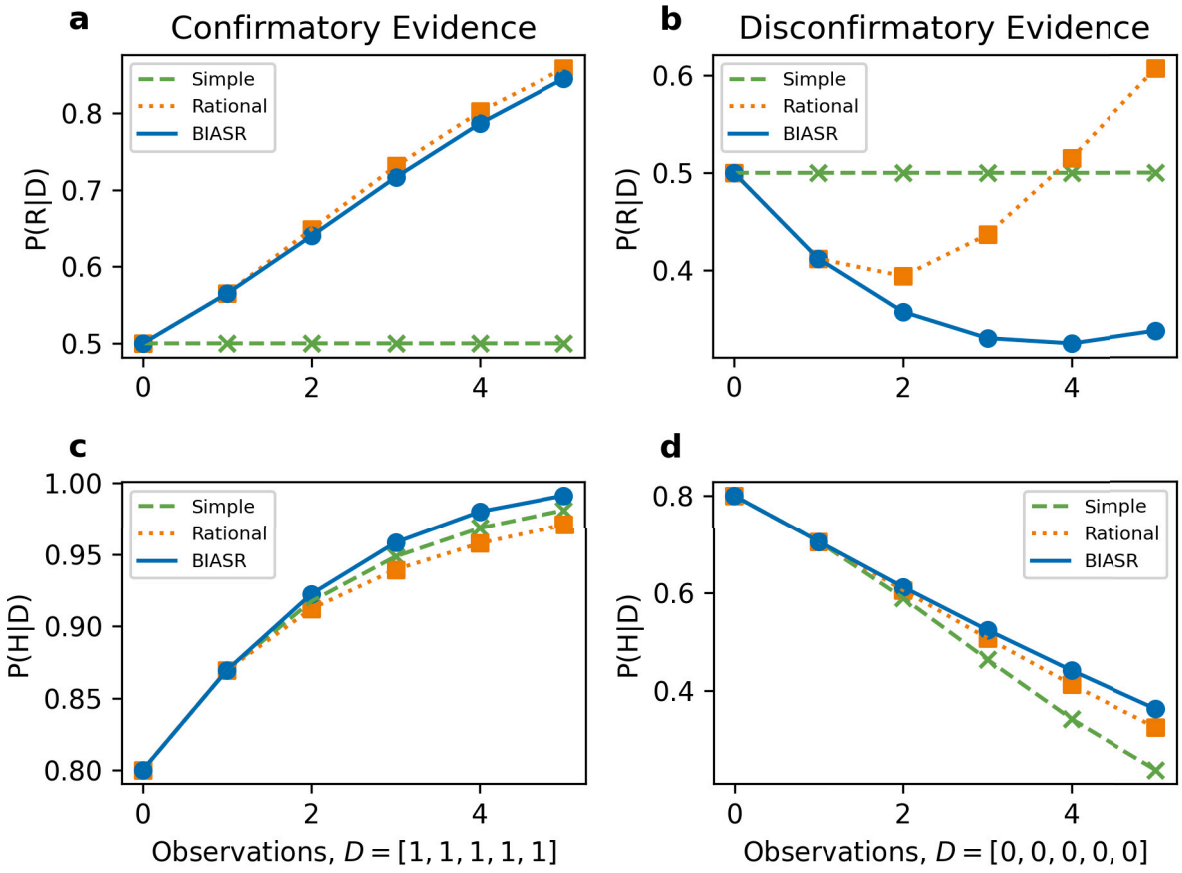


Fig. 4. Assimilation and evaluation for confirmatory evidence and disconfirmatory evidence for sequential information from the same source. The BIASR model shows biased evaluation with (**a**) confirmatory sources judged to be more reliable than (**b**) disconfirmatory sources (the solid blue line is higher in **a** than **b**). The BIASR model shows biased assimilation, with a stronger posterior belief in the central hypothesis than the simple and rational models following both (**c**) confirmatory and (**d**) disconfirmatory evidence (i.e the solid blue line is the highest line in both **c** and **d**).

The rational basis for attitude polarisation was explored further by Henderson and Gebharter (2021) using a Bayesian network where evidence is influenced by the true values of the central hypothesis and source reliability, as in the BIASR model. They conclude that attitude polarisation can arise only if the individuals have different prior beliefs in both the central hypothesis and source reliability (Henderson & Gebharter, 2021). This is a property of the Bayesian networks that generate rational attitude polarisation (Cook & Lewandowsky, 2016; Henderson & Gebharter, 2021; Jern et al., 2014) — they require different priors not only in the central hypothesis but also auxiliary beliefs.

The Bayesian network structures described by Jern et al. (2014) give a good explanation for attitude polarisation when central and auxiliary priors are different between polarising groups. However, the BIASR model generates attitude polarisation under the stricter condition that the pro- and anti- individuals differ only in their prior beliefs in the central hypothesis, and have the same auxiliary prior beliefs. In our model,

$$\begin{cases} P(H|\boldsymbol{D}) > P(H), & \text{if} \quad P(H) > 0.5, P(R) = r \\ P(H|\boldsymbol{D}) < P(H), & \text{if} \quad P(H) < 0.5, P(R) = r \end{cases}, \tag{12}$$

where $P(R) = r$ is the same prior belief in source reliability for both individuals.

As shown in Fig. 5, the BIASR model leads to attitude polarisation when data is presented from two different sources, even when individuals only differ in their prior belief in the central hypothesis, $P(H = 0.8)$ and $P(H = 0.2)$. The simple and rational Bayesian models do not. As such the BIASR model meets both the general and stricter conditions we have defined for attitude polarisation. Strong prior beliefs either for or against the central hypothesis become more extreme overall.

Intuitively, let us consider the case where Alice starts with a strong prior belief in the dangers of vaccines. She is given two studies to read, one for and one against vaccine safety. The first study begins by stating that vaccines are safe — Alice starts to think that the study is not reliable, as she is confident that vaccines are dangerous. After reading on, the study makes another point about vaccine safety, Alice is now more easily able to dismiss this as she already has doubts over the study's reliability. As Alice reads on, she becomes convinced that the study is not credible and the later information has very little impact on her beliefs. The second study raises questions about vaccine safety. As Alice reads this study, her confidence in its credibility grows as it provides information that aligns with her existing beliefs in the dangers of vaccines, and she uses this evidence to bolster those same beliefs. Alice's beliefs about the dangers of vaccines and the reliability of the studies become correlated, and if she forgets about these correlations then attitude polarisation emerges.

4.3. Belief perseverance

People can persevere in their beliefs with greater tenacity than the evidence would warrant (Klayman, 1995). Belief perseverance is typically defined with a temporal aspect in the sense that once a belief is formed it will persist even once the evidence that formed its basis is discredited (Anderson et al., 1980; Ross, Lepper, & Hubbard, 1975). In an early experimental study (Ross et al., 1975) gave participants false

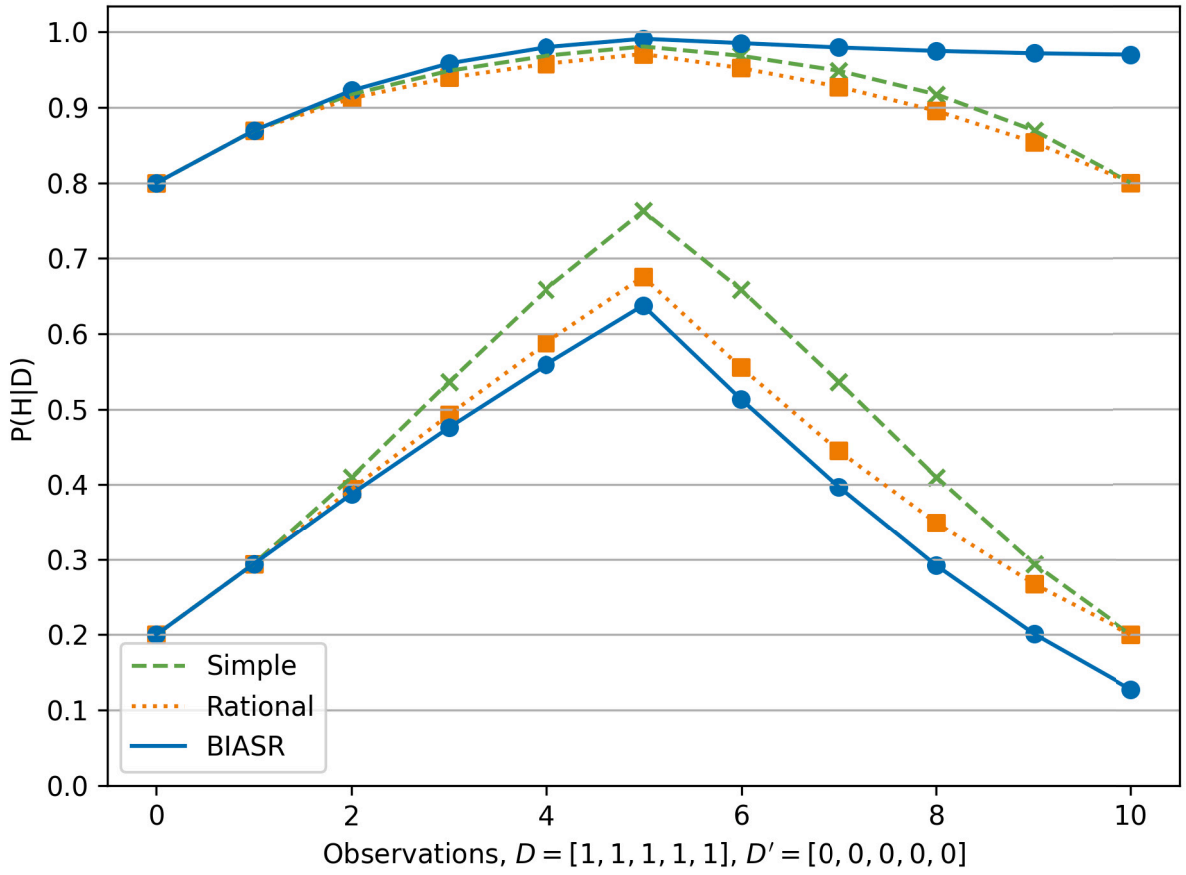


Fig. 5. Attitude polarisation. Data from a source is followed by data from a new second source with the opposite view. With a strong initial prior (top), the BIASR model shows positive biased assimilation from the first data source followed by negative biased assimilation, overall increasing belief in the central hypothesis. With a low initial prior belief (bottom) we also see biased assimilation of both sources of data, overall decreasing belief in the hypothesis. Both positions become more extreme from seeing the same set of data under the BIASR model, showing attitude polarisation.

feedback on a task (either good, average or bad). This (reasonably) influenced the participants' opinion of their task performance, but the participants held onto these opinions even after they were told that the feedback was fictitious. This effect was explored further by Anderson et al. (1980), who gave participants fictitious data suggesting that firefighters who were courageous were more likely to be successful in their jobs. This induced participant beliefs that persisted even once the data was revealed to be fictitious.

It has been noted that belief perseverance is connected to the primacy effect (Nickerson, 1998), where data observed earlier has a larger impact on belief than data seen more recently. Bruner and Potter (1964) showed participants images, and found that they were slower to recognise those images when they came into focus slowly, as compared with participants who saw the same image without first seeing it out of focus. They attributed this effect to the perseverance of hypotheses generated while the image was out of focus. This was followed in the late 60s with a series of studies that tested participants' ability to form opinions through sampling, finding that early data could induce beliefs that were then held onto more strongly than would be Bayesian rational in light of later evidence against the belief (Geller & Pitz, 1968; Jones, Rock, Shaver, Goethals, & Ward, 1968; Peterson & DuCharme, 1967).

In belief perseverance, the order that beliefs are formed is important. And if beliefs are formed from observed data, then the order that data is received is important. In contrast to the exchangeability principle of Bayesian rationality, i.e. that the order of the data received should not make a difference to the posterior beliefs, we define *belief perseverance* as the observation that data received earlier has a stronger influence on final beliefs than opposing data received later,

$$P(H|\boldsymbol{D}) > P(H|\boldsymbol{D})_{rational} \quad , \quad \boldsymbol{D} = [\boldsymbol{D}_{for}, \boldsymbol{D}_{against}]\,. \tag{13}$$

Fig. 6 shows a simulation of belief perseverance. Starting from a neutral prior, $P(H) = 0.5$, both the simple and rational models end up with the same posterior belief as they began with, after seeing an equal amount of evidence for and against. In the BIASR model, the initial data drives belief in H beyond what is rational. Once the belief is ingrained, negative biased assimilation then slows down disconfirmation of belief. Here, we have simulated data as coming from separate sources. If we instead used a single source then we still observe belief perseverance but the effect is not as strong.

4.4. Confirmation bias in selection of sources

Confirmation bias is usually defined not only in terms of assimilating information, but also in the selection of information in a way that supports existing beliefs. Taber and Lodge (2006) replicated and extended Lord et al.'s (1979) study on attitude polarisation. Participants were chosen who held strong beliefs about either gun control or affirmative action. They were then shown sources for and against those positions, but some participants also had the opportunity to choose the sources they wished to read. Those with strong prior beliefs selected the sources that were likely to agree with their position. Redlawsk (2002) found a similar effect in a behavioural experiment where they simulated a presidential primary election. Once participants had developed a preference for a candidate, they were more likely to

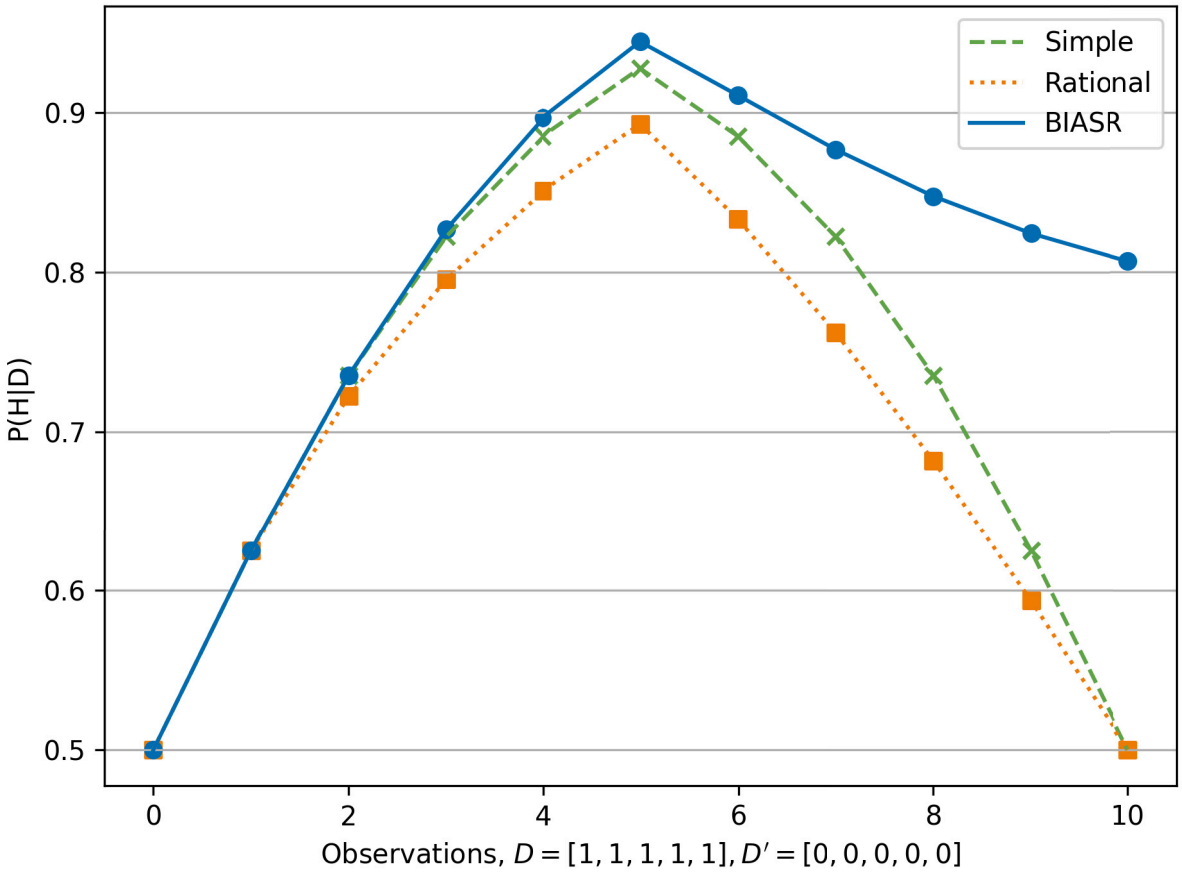


Fig. 6. Belief perseverance. Data for, then data against, the central hypothesis are received from different sources given neutral initial priors in both the central hypothesis and source reliability. Under simple and rational models, the belief in the central hypothesis returns to the prior belief. With the BIASR model, biased assimilation dynamics mean that the data received earlier has a stronger effect on posterior beliefs than data received later.

search for information about that candidate. This form of confirmation bias may go beyond the selection of external sources, and Kunda (1990) also suggested a confirmation bias in the selection of memories and cognitive processes.

In order to extend our model to selection of sources we must add an extra assumption — agents are limited in that they cannot consume data from all sources and must be selective. An optimal selection would presumably be based on some kind of value function on the sources. This is difficult to model as value is subjective and would need to take into account complicated utility functions (Klayman & Ha, 1987).

A Bayesian approach to optimal data selection is to optimise expected information gain (Nelson, 2005; Oaksford & Chater, 1994). The information gain when receiving some data is equivalent to the reduction in uncertainty (defined as the entropy) over hypotheses. When evaluating the potential reduction in uncertainty of a source we do not know what data we will receive, and so we calculate the expected information gain by taking into account all possible results. In the context of our model, the expected information gain is equivalent to the expected Kullback–Leibler divergence between prior and posterior beliefs (full derivation in Nelson (2005)),

$$\mathbb{E}(I_g) = \sum_D P(D) \sum_H P(H|D) \log \frac{P(H|D)}{P(H)}. \tag{14}$$

If agents select sources based on their value in terms of expected information gain, then we can define confirmation bias in the selection of sources as valuing a confirmatory source's relative expected information gain more so than would be rational,

$$\frac{\mathbb{E}(I_g)_{for}}{\mathbb{E}(I_g)_{against}} > \left(\frac{\mathbb{E}(I_g)_{for}}{\mathbb{E}(I_g)_{against}} \right)_{rational}. \tag{15}$$

Fig. 7 shows a simulation of confirmation bias in the selection of sources. A simulated individual with an initial prior belief for a hypothesis, $P(H) = 0.8$, receives data from two sources, with neutral initial prior beliefs in the reliability of each source ($P(R) = 0.5$ for each source). One source provides confirmatory evidence for the hypothesis, while the other provides disconfirmatory evidence (see the Appendix for further simulation details). In the simple model the expected information gain of sources is invariant. In the rational model there is little difference in the expected information gain of sources, and the confirmatory sources are slightly preferred. With the BIASR model, there is a much greater difference in the expected information gain, with confirmatory sources much preferred. An individual that can choose only one source would much prefer the confirmatory source under the BIASR model, if that choice was made based on the expected information gain.

The expected information gain has good theoretical grounding as a measure of optimal data selection, and has been shown to correlate well with empirical results of actual human behaviour (Nelson, 2005). We also investigated an alternative measure, the diagnosticity of a source, with qualitatively similar results (see Appendix).

5. Empirical evidence aligned with the independence approximation

We have shown that the BIASR model can generate a range of confirmation bias type behaviours. If the model is capturing, in some sense, how people actually behave then we would expect to see a difference in behaviour depending on whether information is processed

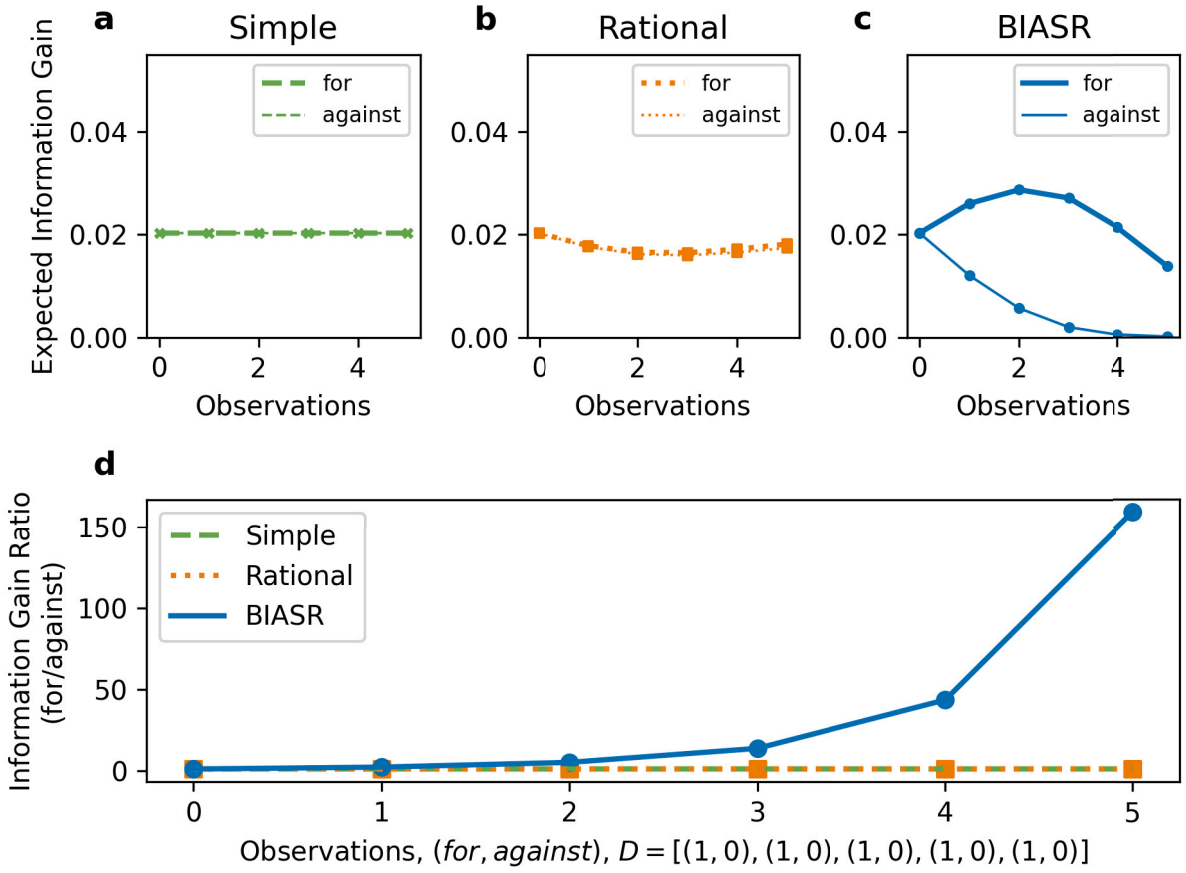


Fig. 7. Confirmation bias in the selection of sources. At each timestep, simulated agents receive one confirmatory datum in support of their prior hypothesis from the "for" source and one disconfirmatory datum from the "against" source. (a) The expected information gain is constant with the simple version of Bayes' theorem. (b) In the rational model, the expected information gain is similar between the sources, with the confirmatory source slightly preferred over the disconfirmatory source. (c) With the BIASR model, the expected information gain of confirmatory sources is much greater than disconfirmatory sources. (d) The ratio of expected information gain (for/against source) is much higher in the BIASR model than in the simple and rational cases, where the ratio stays around 1.

incrementally or all at once. Processing data all at once will give the same result as rational incremental processing, i.e. no bias. However, according to the BIASR model, sequential processing will show path dependence.

An experimental manipulation would be to encourage participants to either (a) process information incrementally or (b) process information all at once. We expect to see more confirmation bias when the information is processed incrementally. We found two previous experimental studies where this distinction was made.

5.1. Redlawsk (2002)

Redlawsk (2002) describes the difference between *on-line processing*, where information is evaluated immediately and sequentially versus *memory processing* where information is remembered and then evaluated all at once when a decision is required. In an experiment, they simulated a presidential election and gave participants information about candidates. In the on-line condition, no further instructions were given as on-line processing is assumed to be the default behaviour. In the memory-based condition participants were encouraged to remember the information that they saw; they were told that they would be tested on it later, as well as being told that they would need to justify their choice to an experimenter. They investigated how participants reacted to incongruent (negative) information once they had developed a preference for a candidate. In the online condition this negative information actually increased the preference for the candidate, while in the memory condition the negative information reduced the candidate rating: In the on-line condition the incongruent information seems to be negatively evaluated to such an extent that it provides evidence in favour of the candidate.

Redlawsk (2002) attribute the difference to an additional accuracy motivation to the memory-based processors, within the framework of motivated reasoning (Kunda, 1990). Within this framework, the memory-based processors are motivated for greater accuracy due to the instruction that they will need to justify their choices, and they achieve this by processing the information all at once (or remembering the dependencies between beliefs). The BIASR model suggests that the bias arises because of path dependence in the on-line condition. The BIASR model also provides an explanation for the greater accuracy of memory processors: remembering data received and/or dependencies between beliefs will avoid independence approximations and therefore also avoid confirmation bias. This highlights that rational processing is possible in the BIASR model if sufficient memory resources can be recruited to manage the number of dependencies (or observations) in a situation. Fig. 8 shows the data as presented by Redlawsk (2002) alongside a simulated replication of the effect with the BIASR model. Here, we used the same model as in the earlier simulations, with the change that unreliable sources are now anti-reliable, so are more likely to give false information, i.e. $P(D = 1|H = 1, R = 0) = 0.35$.

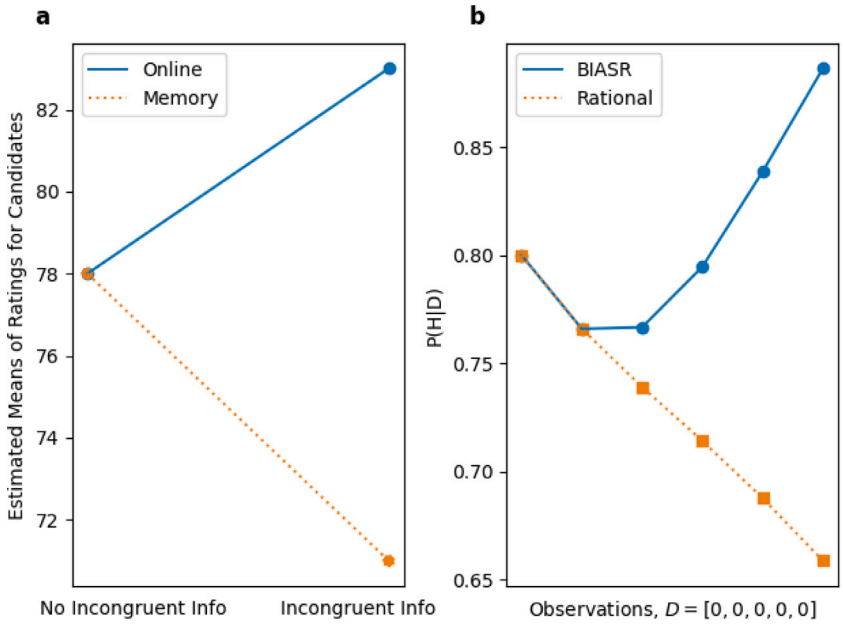


Fig. 8. Replication of the Redlawsk result. (a) The data presented by Redlawsk. Following negative information about a preferred candidate, the online processors increase their rating for the candidate, while the memory processors decrease their rating. (b) A replication with the BIASR model. We start with a prior preference for the candidate and a neutral prior in source reliability (not shown). Unreliable sources are considered anti-reliable, i.e. negative information from an unreliable source actually acts as evidence for a candidate. We simulate receiving a series of negative pieces of information from the source. Similarly to Redlawsk: in the BIASR condition belief in the candidate increases, while it decreases in the rational condition.

5.2. *Carlson et al. (2006)*

In this experiment (Carlson et al., 2006), participants were asked to make a choice between two restaurants after seeing each of the restaurants' attributes. The six attributes were typically neutral but included one for each restaurant that was much in its favour (for example, one restaurant has a professional dessert chef while the other has a small assortment of standard desserts). The order of attributes were manipulated so that the target restaurant had its very positive attribute revealed first, and the opposing restaurant had the attribute in fourth position. As a further treatment, in Study 1 the attributes were shown sequentially, while in Study 3 the attributes were shown together on a single page for each restaurant. They found a significant preference for the target restaurant in Study 1, but not in Study 3. Confirmation bias was not detected when information was presented in one block, but was detected when the same information was presented sequentially.

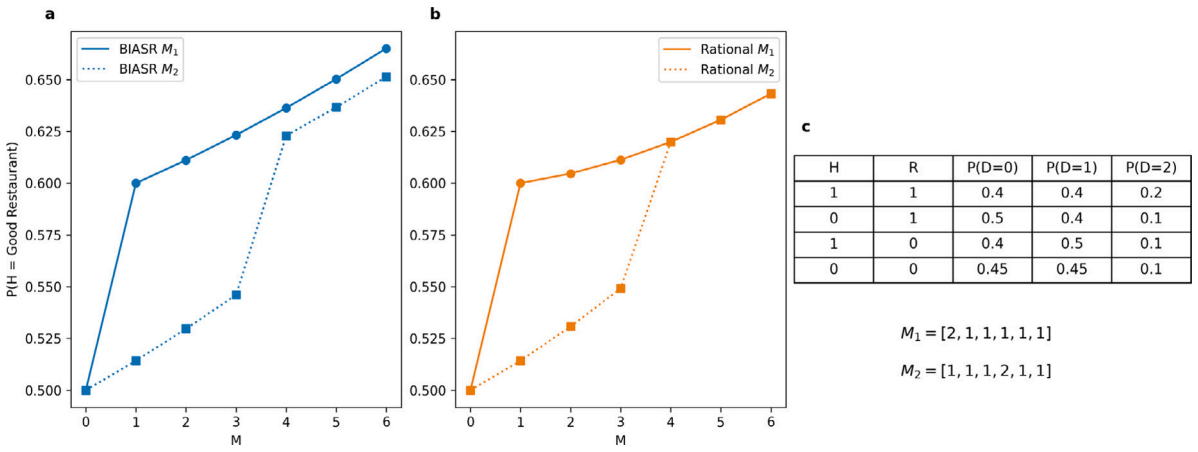
The authors of the study interpret the result within Russo's *predicisional distortion of information* framework. When incorporating information sequentially, a positive first attribute creates an initial preference for the target restaurant that then biases the interpretation of subsequent data so that overall the target restaurant is preferred. This framework is similar to our model and the findings here support both perspectives. Within the BIASR model, the preference for the target restaurant in Study 1 is described by belief perseverance, i.e. the first attribute observed has a greater weight on the final choice than the fourth attribute. Alternatively, when data is shown all at once it is more likely to be processed together, which is equivalent to remembering the history of belief dependencies. We have simulated this result within the BIASR model (Fig. 9). We used a similar setup as in the earlier simulations, but now messages can be negative, slightly positive or very positive, $D \in [0, 1, 2]$ respectively. We chose this setup because it replicates the result with a minimal change to the existing model.

6. General discussion

The traditional normative argument is that rational behaviour should enjoy higher evolutionary fitness (Daw et al., 2008). As argued here and noted before, a normative account should also include cognitive limitations (Dasgupta et al., 2020; Daw et al., 2008; Klayman, 1995), such that when considering computationally intractable problems evolutionary pressures will favour organisms with efficient approximations to rationality (Daw et al., 2008). We have argued that maintaining dependencies within large belief networks is computationally intractable given realistic memory constraints. We showed how human cognition can overcome this limitation through the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). And this approximation leads directly to many confirmation bias behaviours. Our results are general, and similar problems will be encountered by artificial agents with large world models.

Previous information processing models of confirmation bias either introduce irrationality without a complete explanation, or they explain the bias as rational given a certain belief updating structure. Irrationality can be included by, for example, adding a factor to reduce the weight of disconfirmatory evidence (Gerber & Green, 1999). Our contribution offers a principled source of irrationality based on a boundedly rational approximation to Bayesian rationality. This approximation leads to a simplification of the rational model which is equivalent to the "belief-based" updating described in previous research (Bovens et al., 2003; Hahn et al., 2018; Merdes et al., 2020; Olsson, 2011). Additionally, our single model is able to generate many forms of confirmation bias. We do not claim that the BIASR model is the full story, and for example Bayesian networks (Jern et al., 2014) can explain much of the empirical evidence for attitude polarisation. However, the BIASR model demonstrates a variety of other confirmation biases that the Bayesian rational model does not, suggesting that it is capturing an important aspect of boundedly rational cognition.

We have focused on a very simple Bayesian network to demonstrate that confirmation bias can arise from the BIASR model. We do not



H	R	P(D=0)	P(D=1)	P(D=2)
1	1	0.4	0.4	0.2
0	1	0.5	0.4	0.1
1	0	0.4	0.5	0.1
0	0	0.45	0.45	0.1

Fig. 9. Replication of the Carlson, Meloy and Russo result. Information is received about attributes of a restaurant. The attributes that are received are either "1": neutral (or slightly positive); or "2": strongly positive. (a) In the BIASR model the strong initial positive message induces a bias towards restaurant 1 (M_1) which persists. (b) In the rational model, the order that data is received does not influence the final beliefs and both restaurants are judged to be of equal expected quality. The beliefs are simulated using (c) the conditional probability distribution.

claim that this simple model is how people actually update their beliefs. However, the behaviour is robust and emerges under a wide range of conditional probability distributions. The assumptions also hold (and are even strengthened) with more complex belief structures. We present a model based on two *types* of sources (reliable or not). However, inference that includes beliefs about types of sources in general would be susceptible to confirmation bias in the same way (including for example biased or anti-reliable sources).

We have included an assumption that human information processing is described by the mathematics of Bayesian networks, and that human memory can be analysed in the same way as computer memory. Our feeling is that these principles are fundamental to information processing and so it is reasonable to assume that human cognition is at least partly bound by them.

The BIASR model provides more than a parsimonious account of confirmation bias — it represents a testable hypothesis. Specifically, we connect confirmation bias to memory limitations in boundedly rational belief updating. This suggests future studies to explore this connection, potentially through behavioural experiments or computational simulations. Such research has the potential to deepen our understanding of the root causes of confirmation bias. In turn, a deeper understanding can contribute to the design of more effective interventions that can potentially reduce the negative social impact of confirmation bias and enhance collective cognition.

6.1. Social and individual explanations

The BIASR model is at the information processing level. However, there have also been explanations of confirmation bias at the social and individual level. Our model is not in opposition to these explanations, but instead complements them.

There have been a range of **social explanations** for confirmation bias. Mercier and Sperber (2011) claim that confirmation bias can improve group cognition. If biased individuals argue to support their own belief then the result can be that overall there is a more efficient group search through hypothesis space, which is then reconciled through debate. This idea could describe the scientific process. Indeed, scientists are not immune to confirmation bias (Dunbar, 1995; Koehler, 1993; Mahoney, 1977) and history is littered with individual scientists who steadfastly held onto their beliefs despite disconfirmatory evidence (Nickerson, 1998). Building on this idea, Norman (2016) argues that the purpose of human reasoning in general is to align group intentions and confirmation bias helps in this regard by strongly entrenching group mythology and beliefs that can persevere over time and so maintain group cohesion. Another perspective is that believing something strongly can influence others and help to bring it about (Peters, 2020), a form of self-fulfilling belief (Snyder, 1984).

At the **individual level**, confirmation bias may help to navigate asymmetric error utilities (Friedrich, 1993; Nickerson, 1998) (being wrong about believing there is not a lion is more problematic than being wrong that there is a lion). From an adaptive perspective, a wider utility function is being optimised beyond truth seeking, and confirmation bias helps to drive behaviour towards a beneficial outcome in this wider game. This is almost certainly true if we assume that human behaviour is adaptive. While the threat of being eaten is obvious, the principle applies to other threats such as identity or self-perception. Kunda (1990) made the case that reasoning is motivated only sometimes by accuracy, and other times by a desire to arrive at certain conclusions. In this account, the individual's motivation will determine which cognitive processes are put to use. If accuracy is desired, then deeper processing is carried out. But if an individual has a motivation to e.g. preserve their self-image or identity, then they can introduce biases in their reasoning that lead to the preservation of those beliefs (Kunda, 1990). People are not completely free to believe whatever they want, and are instead constrained by the available cognitive resources and by the need for coherence within beliefs, at least to the extent that they could justify themselves to someone else (Kunda, 1990). The desire for coherence is an older idea that is also a part of the influential cognitive dissonance theory (Festinger, 1962). Biased evaluation, biased assimilation and belief perseverance can be understood as the reconciliation of the dissonant beliefs "I believe that I am someone who holds correct beliefs" and "this evidence disconfirms my beliefs" (Kunda, 1990). However, notably it has been argued that dissonance theory does not easily predict attitude polarisation (Lord, 1989). Motivated reasoning and the avoidance of dissonance are a part of the puzzle, but it still leaves the question open of describing the cognitive processes involved.

As social behaviour emerges from individual behaviour, so individual behaviour emerges from cognitive processing. Kunda's perspective on motivated reasoning (Kunda, 1990) is enriched by our framework. Motivated reasoning relies on the assumption of different cognitive faculties that have differential levels of accuracy and effort. Our model provides a clear account of using extra cognitive resources to improve the accuracy of reasoning. An individual with an accuracy motivation could update their beliefs without applying the independence approximation, and instead use extra memory resources to consider dependencies between beliefs and avoid biases. We can also see the link to emotional states and hot vs cold cognition (Kunda, 1990) — one can imagine a hot-headed individual quickly jumping to false conclusions while a cooler head carefully thinking through the evidence and belief

dependencies. A promising direction for future research is to explore the conditions that either promote or inhibit confirmation bias. The BIASR model predicts that these conditions are tied to memory capacity and motivation.

Given that confirmation bias exists, we can speculate that it would make sense for adaptive pressures to build other behaviours around this bias — nature is parsimonious. A purely rational agent would reason about the world and then decide on their actions based on these beliefs combined with an expected utility distribution. In certain situations it may be more cognitively efficient to shortcut this two-step process by leveraging confirmation bias to drive behaviour based on less than rational beliefs. Given that confirmation bias exists at the individual level, we can speculate that adaptive pressures built useful group dynamics upon it such as argumentation and debate (Mercier & Sperber, 2011), persistent group ideologies and mythologies (Norman, 2016) and even the will to force reality towards our beliefs (Peters, 2020).

7. Conclusion

The BIASR model is based on principled assumptions, generates many confirmation bias type behaviours, and aligns well with both empirical evidence and other explanations in the literature. The main principle of the BIASR model is that put forward by Daw et al. (2008), who contend that rationality is not the appropriate normative standard when studying human and animal behaviour. Instead, where rational computation is expensive we should expect to see efficient approximations to rationality. We demonstrate that an independence approximation is one way in which cognition can overcome intractable computational demands, providing a fuller normative explanation for the "belief-based" updating described in earlier work (Bovens et al., 2003; Hahn et al., 2018; Merdes et al., 2020; Olsson, 2011). Given its general nature, the independence approximation deserves further investigation as a more general cognitive mechanism for boundedly rational reasoning with memory constraints.

CRediT authorship contribution statement

Charlie Pilgrim: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Adam Sanborn:** Conceptualization, Writing – review & editing, Supervision. **Eugene Malthouse:** Conceptualization, Writing – review & editing, Visualization. **Thomas T. Hills:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code and data is avialiable publicly at https://github.com/chasmani/PUBLIC_confirmation_bias_and_the_BIASR_model.

Acknowledgments

T.T.H. was supported on this work by the Royal Society Wolfson Research Merit Award (WM160074) and a Fellowship from the Alan Turing Institute, which is funded by EPSRC (grant number EP/N510129/1). A.N.S was supported by a European Research Council Consolidator Grant (817492-SAMPLING). E.M. was funded by a departmental fellowship awarded by the Department of Psychology at the University of Warwick. C.P. was funded by the EPSRC grant for the Mathematics for Real-World Systems CDT at Warwick (grant number EP/L015374/1).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105693.

References

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of the Personality and Social Psychology*, *39*(6), 1037.
Bacon, F. (1878). *Novum organum*. Clarendon Press.
Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, *32*(1), 176.
Bellman, R. E. (2015). *Adaptive control processes*. Princeton University Press.
Bovens, L., Hartmann, S., et al. (2003). *Bayesian epistemology*. Oxford University Press on Demand.
Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science*, *144*(3617), 424–425.
Carlson, K. A., Meloy, M. G., & Russo, J. E. (2006). Leader-driven primacy: Using attribute order to affect consumer choice. *Journal of Consumer Research*, *32*(4), 513–518.
Collins, P. J., Hahn, U., Von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology*, *9*, 18.
Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179.
Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20.
Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.
Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. *The Probabilistic Mind*, 431–452.
Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. *The Nature of Insight*, *18*, 365–395.
Festinger, L. (1962). *A theory of cognitive dissonance, Vol. 2*. Stanford University Press.
Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239.
Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena.. *Psychological Review*, *100*(2), 298.
Geller, E. S., & Pitz, G. F. (1968). Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance*, *3*(2), 190–201.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
Gerber, A., & Green, D. (1999). Misperceptions about perceptual bias. *Annual Review of Political Science*, *2*(1), 189–210.
Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28.
Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of the Personality and Social Psychology*, *44*(6), 1110.
Gorman, M. E. (1989). Error, falsification and scientific inference: An experimental investigation. *The Quarterly Journal of Experimental Psychology Section A*, *41*(2), 385–412.
Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. In *Psychology of learning and motivation, Vol. 61* (pp. 41–102). Elsevier.
Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678.
Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701–723.
Henderson, L., & Gebharter, A. (2021). The role of source reliability in belief polarisation. *Synthese*, 1–24.
Jarvstad, A., & Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cognitive Science*, *35*(4), 682–711.
Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206.
Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3.
Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect.. *Journal of Personality and Social Psychology*, *10*(4), 317.
Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418.
Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*(1), 28–55.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.
Kvam, P. D., Alaukik, A., Mims, C. E., Martemyanova, A., & Baldwin, M. (2022). Rational inference strategies and the genesis of polarization and extremism. *Scientific Reports*, *12*(1), 7344.
Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin*, *18*(6), 669–679.
Lord, C. G. (1989). The“ disappearance” of dissonance in an age of relativism. *Personality and Social Psychology Bulletin*, *15*(4), 513–518.
Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of the Personality and Social Psychology*, *47*(6), 1231.
Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of the Personality and Social Psychology*, *37*(11), 2098.
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175.
Malthouse, E. (2022). Confirmation bias and vaccine-related beliefs in the time of COVID-19. *Journal of the Public Health*, fdac128. http://dx.doi.org/10.1093/pubmed/fdac128.
McHoskey, J. W. (1995). Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, *17*(3), 395–409.
Melnikoff, D., & Strohminger, N. (2023). Bayesianism and wishful thinking are compatible. http://dx.doi.org/10.31234/osf.io/yhmvw.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.
Merdes, C., Von Sydow, M., & Hahn, U. (2020). Formal models of source reliability. *Synthese*, 1–29.
Miller, A. G., McHoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, *64*(4), 561.
Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636–653.
Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.
Norman, A. (2016). Why we reason: Intention-alignment and the genesis of human rationality. *Biology & Philosophy*, *31*(5), 685–704.
Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.
Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, *8*(2), 127–143.
Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140–153.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Peters, U. (2020). What is the function of confirmation bias? *Erkenntnis*, 1–26.
Peterson, C. R., & DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, *73*(1), 61.
Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of the Applied Social Psychology*, *21*(13), 1058–1082.
Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*(4), 1021–1044.
Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm. *Journal of the Personality and Social Psychology*, *32*(5), 880.
Russo, J. E. (2014). The predecisional distortion of information. In *Neuroeconomics, judgment, and decision making* (pp. 109–128). Psychology Press.
Russo, J. E. (2018). Bayesian revision vs. Information distortion. *Frontiers in Psychology*, *9*, 1550.
Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, *66*(1), 102–110.
Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, *112*, 98–101.
Sanborn, A. N., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, *57*(3–4), 94–106.
Snyder, M. (1984). When belief creates reality. In *Advances in experimental social psychology, Vol. 18* (pp. 247–305). Elsevier.
Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155.
Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.
Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of the Personality and Social Psychology*, *49*(3), 577.