Bayesian Belief Protection: A Study of Belief in Conspiracy Theories

Abstract

Several philosophers and psychologists have characterized belief in conspiracy theories as a product of irrational reasoning. Proponents of conspiracy theories apparently resist revising their beliefs given disconfirming evidence and tend to believe in more than one conspiracy, even when the relevant beliefs are mutually inconsistent. In this paper, we bring leading views on conspiracy theoretic beliefs closer together by exploring their rationality under a probabilistic framework. We question the claim that the irrationality of conspiracy theoretic beliefs stems from an inadequate response to disconfirming evidence and internal incoherence. Drawing analogies to Lakatosian research programs, we argue that maintaining a core conspiracy belief can be Bayes-rational when it is embedded in a network of auxiliary beliefs, which can be revised to protect the more central belief from disconfirmation. We propose that the irrationality associated with conspiracy belief lies not in a flawed updating method, but in a failure to converge towards well-confirmed, stable belief networks in the long run. This approach not only reconciles previously disjointed views, but also points towards more specific descriptions of why agents may be prone to adopting beliefs in conspiracy theories.

Keywords: conspiracy belief, conspiracy theory, Bayesianism, prior probabilities, rationality

1. Introduction

Over the course of the past decade, there has been an explosion of research on belief in conspiracy theories (henceforth CTs; see Goreis & Voracek, 2019 for an overview), reflecting an urgency to understand the phenomenon. This mounting pressure is motivated by the presence of conspiracy theorising in public discourse, the potential of social media for spreading such beliefs, the associated erosion of trust in epistemic authorities, and the role that these factors play in spreading scepticism regarding the official story about the ongoing COVID-19 pandemic.

Despite the heightened interest, there remains little consensus about the nature and irrationality of CT beliefs, e.g., what makes such beliefs intuitively `bad’ or `good’, with existing attempts to explain key features of such belief being highly fragmented. Common points of contention concern the epistemic justification of such beliefs given their apparent resistance to counterevidence (Napolitano, 2021; but also Keeley, 1999; Harris, 2018), lack of falsifiability (Feldman, 2011) and truth-aptness (Cassam, 2019; but see Hagen, 2022), and the extent to which such theories, insofar as they belong to a general category (Stokes, 2016), are explanatory at all (Butter, 2021; Fenster, 2008). Attempts to understand the psychological factors that may contribute to people’s endorsement of belief in CTs are likewise inconclusive as to whether it might result from irrational reasoning (e.g., Cichocka et al., 2016; Douglas et al., 2019; Van Prooijen and van Vugt, 2018). For example, CT beliefs tend to highly correlate (i.e. people who believe in one conspiracy tend to believe in others), even when they are semantically and logically unrelated (Goertzel, 1994a), or even mutually inconsistent (Wood et al., 2012). From these perspectives, the problem with belief in CTs is not their disregard for the evidence *per se*, but their monological nature (Hagen, 2018), an aspect that has also been visible in recent analyses of COVID-19 CTs (Miller, 2020).

Contrary to these negative attitudes, philosophers and psychologists also argue on diverse grounds in favour of certain epistemic and psychological benefits that might equally make such beliefs a source of rational reasoning. From these latter perspectives, belief in CTs is indeed often responsive to the available evidence (Levy, 2021; Suthaharan et al., 2021), sometimes poses the best explanation of the events (Dentith, 2016), and may even lead to the truth (Dentith, 2019). These views instead highlight the subjective importance of background beliefs as well as the relevance of sociocultural structures to evaluate the explanatory status of belief in CTs on a case-by-case basis (see also Basham, 2016). Most importantly, they point to the protective status that belief in CTs can bear on one's social belonging or personal identity.

Our aim in this paper is to explore the rationality of belief in conspiracy theories in more depth from the perspective of Bayesian cognitive science. Instead of preemptively accepting a view on *whether* such belief is rational or irrational, we begin by asking under what formal conditions such belief *would* become irrational in the first place.[1] Our main aim is to clarify the debate by making these conditions more precise through the use of tools from Bayesian analysis in cognitive psychology and philosophy of science. Our hope is that this will offer a shared platform on which previously disjointed views can be brought closer together. Specifically, we focus on how agents incorporate new information to update their beliefs, and we argue that the implicit background structures can be seen as playing the role of auxiliary hypotheses which, under certain conditions, can be rightly discarded *to protect core beliefs*. Building on works from Strevens (2001) and Gershman (2019), we analyse the structure and rationality of belief in CTs in analogy to Lakatosian research programs and explain the robustness of high-probability beliefs to disconfirmation by counterevidence in reference to low-probability beliefs that can easily be discarded to protect core beliefs. We suggest that, if belief in conspiracy theories should be deemed irrational at all, it is not because of a failure to revise beliefs given disconfirming evidence. Rather, we consider the initial biases and assumptions that guide agents' inferences as a crucial point of departure to make sense of the correlations between apparently incoherent belief systems on the one hand and people's apparent unresponsiveness to evidence contrary to what these systems entail.

[1] For the same reasons, we do not assume that there is a principled difference between beliefs in conspiracies and conspiracy theories. If any such difference exists, it should emerge as a conclusion rather than constitute a starting point of our inquiry.

Our treatment is in line with previous attempts to elucidate the rationality of CT belief in terms of their protective psychological nature. However, while several of these views have been presented as being `Bayes-like' or Bayesian-compatible' (e.g., Napolitano, 2021; Dentith, 2016; Levy, 2019), none offers a formalization of the protective nature associated with belief in CTs.[2] A further advantage of our treatment is that it is highly unifying; as we show, it also reconciles the traditional and the higher-order framing of the monological belief view, bringing previously disjointed views on belief in CT within a single formal framework. Finally, while Bayesian models in cognitive science have previously been appreciated for their high unificatory credentials (Colombo & Hartmann, 2017), this has not been shown for the domain of belief in CTs. We therefore think that this is also an interesting case study for proponents of Bayesian cognitive science.

Introductory remarks in place, we will begin our analysis by outlining the two contrastive views about the nature of belief in CTs in section 2. In section 3, we outline the Bayesian treatment of the relationship between networks of associated beliefs and disconfirmatory evidence. In section 4, we apply this analysis to a received negative characterization of conspiracy belief, while focusing on its evidence-responsiveness. Section 5 shows how our treatment unifies the two views, while section 6 discusses a set of inductive biases as the possible sources of the formation of belief in CTs. We subsequently clarify the extent to which the Bayesian treatment provides a benchmark to identify and assess the degrees of epistemic rationality associated with belief in CTs. We end with a brief conclusion

[2] Some approaches have used Bayesian tools explicitly to analyse the protective nature of *delusional* beliefs (e.g., McKay, 2012). Although the two kinds of phenomena are oftentimes related (see Bortolotti et al., 2021, for a comparative analysis), explicit Bayesian analysis of belief in CTs is still widely lacking. To fill this gap, we concentrate our efforts in this paper only on the case of belief in CTs, not precluding that, due to their apparent similarities, our analysis might be beneficial to study other cognitive phenomena such as delusional belief.

concerning implications for future research in epistemology, psychology, and cognitive science.

2. Irrationality of conspiracy belief and its sources

As mentioned, the negative characterisation stresses the doxastic structure of CT beliefs and their epistemic support or lack thereof. They are:

(1) their *monological nature*, where beliefs in conspiracy theories mutually support one another to form a self-sustaining network (T. Goertzel, 1994; B. Goertzel, 1994); under some construals, this network might even contain mutually inconsistent beliefs that are separately supported by a broader higher-order belief that makes them consistent (Wood et al., 2012); and

(2) their insensitivity to disconfirmation, which appeals either to the beliefs' insensitivity to criticism (and even reframing it as supporting evidence, see Keeley, 1999) or to their self-insulation, postulating that such beliefs are isolated from disconfirming evidence and other doxastic states (Napolitano, 2021).

The first feature, (1), assumes that belief in conspiracy theories is emblematic of a reasoning style in which a set of beliefs comprises a self-sustaining network of contents that mutually support each other to afford a coherent explanation of contingent phenomena that could be otherwise difficult to explain or would threaten the cohesiveness of the existing belief system. Conspiracy theorists are said to represent a *monological* reasoning style because those who believe in one conspiracy theory are more likely to endorse beliefs in other conspiracies (T. Goertzel, 1994). As Benjamin Goertzel, who originated this idea, explains, a *monological* belief system is "a belief system which speaks only to itself, ignoring its context in all but the shallowest respects" (1994, p. 166). Ted Goertzel adds that "in a monological belief system, each of the beliefs serves as evidence for each of the other beliefs" (1994, p. 740). What is crucial for this account is that monological beliefs are opposed to *dialogical* ones in which evidence for different beliefs is examined in independent contexts.

As a special case of the monological view, the higher-order hypothesis postulates that conspiracy beliefs relate to each other to the extent that they cohere with a higher-order belief that indirectly provides their mutual support. Here, the focus is not on the direct evidential relationship between particular beliefs, but on the support, they receive from a belief that entails their predictions. Wood et al. (2012) show that even mutually contradictory beliefs correlate positively in this way. For instance, people are more likely to simultaneously agree that Osama Bin Laden was both dead and alive when the US forces arrived at the al-Qaida compound if they also believe that the related statements issued by the US government are suggestive of a cover-up operation.

Unlike (1), (2) postulates that what is crucial for the irrationality of beliefs in CTs are not their internal relationships, but the way in which their associated credences are (or rather are not) updated in light of novel evidence. The earliest version of this postulate can be found in Keeley (1999), who claims that "all potentially falsifying evidence can be construed as supporting, or at worst as neutral evidence" (p. 121) of a CT. Napolitano (2021) restates that belief in CTs renders evidence *probabilistically irrelevant*, meaning that such evidence turns out equally likely when conditioned on the belief as when conditioned on its negation. While this suggests that CT beliefs are unfalsifiable, Napolitano questions whether this condition is sufficient to explain why an agent's degree of belief in a CT would remain constant regardless of whether disconfirming observations bear on that belief.[3] As she points out, under the irrelevance condition "a conspiratorial explanation can only be immune to being

[3] It is important to note that this is not the same as the belief being probabilistically independent from the evidence. Independence corresponds to: Pr(*belief*|*evidence*) = Pr(*belief*), while irrelevance corresponds to: Pr(*evidence*|*belief*) = Pr(*evidence*|¬*belief*).

disconfirmed by any new evidence if it remains so general that it makes no specific predictions" (2021, p.10), while also voicing scepticism about the possibility of agents acquiring such general beliefs without forming more specific beliefs that could be easily disconfirmed. Thus, in contrast to Keeley, Napolitano postulates that for CT beliefs to be maintained they need to be *self-insulated,* and the process of belief-updating cannot admit *any* disconfirming evidence.

While the above summary is not exhaustive, the two views share some crucial features despite their many differences. Firstly, they all analyse conspiracy beliefs through the lens of flawed reasoning processes taken to be crucial for understanding the phenomenon. Secondly, they share the important assumption that the cognitive processes which give rise to belief in CTs are irrational and should be demarcated from rational reasoning in everyday as well as in scientific inquiry. Thirdly, they all place special importance on the notion of consistency and inconsistency, either between beliefs themselves (as in a) or the beliefs and evidence (as in b). Finally, despite the shared focus on the operations which produce and sustain beliefs in conspiracy theories, none of these views offers a detailed analysis or model of the process it describes.[4] Although Napolitano does present her view in terms of conditional in-\dependence between beliefs and evidence, the Bayesian framing of the insulation of conspiracy theoretic beliefs is only used for exposition and does not formalise how insulation happens. This is an important deficiency of the two competing views since it is not entirely clear that they are, in fact, incompatible.

We hope to clarify some of the questions that are left open by previous views. Why do some beliefs appear to be evidentially self-insulated? How can this process be understood

[4] An exception is Benjamin Goertzel's (1994) application of complex systems theory to distinguish open from closed minds. Despitelittle impact for the topic of belief in CTs, it has inspired work formally distinguishing conspiracy narratives from conspiracy theories on the internet (Tangherlin et al., 2020).

conceptually, and in formal terms? Under what conditions is rejecting counterevidence acceptable? From the Bayesian perspective we propose, there is nothing special to self-insulation *per se*, which is apparent in many forms of belief (e.g., scientific belief), however, as a psychological feature, it can become pathological in extreme forms. We propose that our analysis reconciles (1) and (2), and though we question the claim that the belief-updating process is irrational, we agree with (2) that an adequate assessment of the rationality of conspiracy belief should take into account the way its associated credence is updated. We start by showing that resistance to counter evidence is principally compatible with Bayesian norms of rationality.

3. The Bayesian treatment of auxiliary hypotheses

The outlined views place special focus on how conspiracy beliefs are evaluated in relation to other beliefs or the available evidence. This mimics some of the well-known problems in the philosophy of science such as the Quine-Duhem thesis (Duhem, 1953) according to which a scientific hypothesis cannot be empirically tested in isolation from additional background assumptions.[5] One of the results of this interdependence is the underdetermination of scientific prediction by the (confirming or disconfirming) evidence. Suppose we have a central belief *h* and an auxiliary hypothesis *a*, such that their conjunct *ha* entails prediction *p*, which *h* alone does not. If *p* is contradicted by evidence *e*, then *e* disconfirms *ha*. But this says nothing about which of the two conjuncts - *a* or *h* - is refuted. The problem calls for a method of rationally distributing the blame between the central hypothesis and the auxiliary constructs.

[5] It is important to note that our analysis builds on an *analogy*, as opposed to an *identity*, between psychological and scientific reasoning.

Clarke (2002) has noticed that CT belief's resistance to counter evidence mimics Lakatos' (1976) conception of degenerating research programs in which *h* is protected from revision by an ever changing set of auxiliary hypotheses $A = \{a_1, a_2,.. a_n\}$ that can accomodate problematic evidence. However, as Clark and others (Harris, 2018; Napolitano, 2021) have pointed out, Lakatos did not provide a clear set of criteria that could elucidate at what point it becomes irrational to defend a degenerating research program. The search for this kind of criteria has been taken up by Bayesian philosophers of science, most notably Howson and Urbach (1993). Here, we focus on Streven's (2001) addition to this tradition, which also offers an answer to both of the outlined problems.

Strevens starts with a set of assumptions. Firstly, the simplified assumption that *e* entails ¬(*ha*), that is, that *e* affects *h* purely in virtue of falsifying *ha*, and not in some other way. Secondly, that *h* and *a* are not independent of each other, and that they are positively probabilistically dependent so that when $Pr(a)$ increases $Pr(a|h)$ will increase as well. And thirdly, that there is a limited range of alternatives to *a* while each of them, together with *h*, assigns a well-defined probability to *e*. In what follows, we accept these assumptions to allow for an elegant analysis of CT belief.[6]

We understand blame-shifting via an analogy to Lakatosian research programs in which auxiliary hypotheses form a 'protective belt' (Lakatos, 1976) that can absorb the evidential disconfirmation of a central hypothesis. For example, someone might maintain the core belief that Princess Diana is still alive under the auxiliary assumption that the government and its public institutions are involved in a cover-up story, which justifies discarding the alternative auxiliary that the photographic evidence of her funeral is reliable.

[6] A detailed discussion of these assumptions can be found in the exchange between Fitelson and Waterman (2005, 2007) and Strevens (2005).

Instead, the assumption of a cover-up protects the central belief that Princess Diana is alive due to the expectation that the evidence is fake. In this case, the auxiliary hypothesis that the evidence source is trustworthy is discarded to protect the central belief. In the following, we show that this process entirely conforms with Bayesian norms of rationality.

Formally, we model the relationship between the degree of belief in the conjunct *ha* upon receiving evidence *e* with Bayes' theorem:

$$\text{(I)} \qquad Pr(ha|e) = \frac{Pr(e|ha)Pr(ha)}{Pr(e|ha)\,Pr(ha) + Pr(e|\neg(ha))\,Pr(\neg(ha))},$$

where the posterior probability of *ha* given *e* is a function of the prior probability of *ha* regardless of *e*, Pr(*ha*) and the likelihood of observing the evidence if *ha* was true, Pr(*e*|*ha*). This is normalised relative to the sum of the likelihoods and priors associated with *ha* and those associated with its negation, ¬(*ha*).

The first step to solving the problem of apportioning blame between the two hypotheses is to formally separate *h* from *a*. We can do this by marginalising over *a* under the assumption that the probability of *ha* and *h*¬*a* sums up to 1 (following the sum rule). We obtain

$$\text{(II)} \qquad Pr(h|e) = Pr(ha|e) + Pr(h\neg a|e)$$

and

$$\text{(III)} \qquad Pr(a|e) = Pr(ah|e) + Pr(a\neg h|e), since\; Pr(ah + a\neg h) = 1.$$

Thus, by marginalising, we `extract' the influence of the central versus auxiliary hypothesis from the overall belief system. Gershman calls this the 'crux' of the Bayesian answer to underdetermination: "A Bayesian scientist does not wholly credit either the central or auxiliary hypotheses, but rather distributes the credit according to the marginal posterior probabilities'' (2019, p. 16). On this basis, it is possible to identifythe impact of *e* on the posterior probability of *h* when *ha* is disconfirmed (i.e., when *ha* entails ¬*e* but *e* is observed). Since *e* is observed, we can replace it by ¬(*ha*), such that

$$\text{(IV)} \qquad Pr(h|e) \;=\; Pr(h|\neg ha) \;=\; \left[\frac{Pr(\neg(ha)|h)}{Pr(\neg(ha))}\right] Pr(h).$$

Pr(¬(*ha*)|*h*) says that if *h* is true, then ¬(*ha*) can only be obtained if ¬*a* is the case. If we assume *h*, it follows that Pr(¬(*ha*)|*h*) = Pr(¬*a*|*h*) = 1-Pr(*a*|*h*). If we insert this into equation (IV) and under the product rule, we obtain

$$\text{(V)} \qquad Pr(\neg ha) \;=\; 1 - Pr(ha) \;=\; 1 - Pr(a|h)Pr(h),$$

and we can derive

$$\text{(VI)} \qquad Pr(h|e) \;=\; \left[\frac{1-Pr(a|h)}{1-Pr(a|h)Pr(h)}\right] Pr(h).$$[7]

This model apportions the blame in proportion to the relative prior probabilities assigned to *a* and *h*. The higher the prior probability, *Pr*(*h*), the less *h* is blamed to the disfavour of *a* when *ha* is refuted. Conversely, if *a* is already highly probable, the blame is put on *h*, and so the negative impact of *e* on *h* increases relative to the certainty about *a*. In other words, as Pr(*a*|*h*) increases, the probability of the set of alternative auxiliaries multiplied with Pr(*h*) decreases. The robustness of a central hypothesis to disconfirmation can be summarised as the ratio Pr(*h*)/Pr(*a*|*h*), which is illustrated in Figure 1. The interesting consequence of viewing the structure of CT belief in this way is that, it might outwardly seem as if such beliefs are unresponsive to counterevidence, when in fact they follow consistent reasoning in which auxiliaries are rejected to protect the core belief from refutation.

[7] Cf. Gershman (2019, p. 15).

[FIGURE 1 HERE]

As an illustration, take the recent rise of belief in Bill Gates conspiracy theories, which might involve as a central belief the claim that Gates has manufactured the COVID-19 pandemic via long-term investments into the creation of vaccines that actually serve to implant microchips to manipulate and infect people with brain tumours. This core belief is surprising, for example, given the existence of photographic evidence of Gates receiving his first Moderna vaccine injection. However, a CT believer could discredit this piece of evidence by adding the auxiliary assumption that Gates elicits control on various governmental agents such as the WHO, Chief Medical Advisor Anthony Fauci, the UK government and Sage Publishing, which, in the extreme, would suggest that such photographic evidence was itself engineered. In terms of our analysis, *h* corresponds to the central belief that Gates manufactured the COVID-19 pandemic, and *a* corresponds to the belief that the photo of his vaccine injection is real. Insofar as *h* is very high a priori for our imaginary CT believer, it is less blamed to the disfavour of *a* when *ha* is undermined by the observation of the photograph.

This Bayesian treatment suggests that reasoning about conspiracies does not depend on singular, isolated, beliefs, but rather requires a system of interconnected beliefs that support each other. It is the internal coherence of the belief system that sets the norms governing changes of degree of belief in a conspiracy theory. That is, it is rational to protect *h* from refutation and maintain the core conspiracy belief, insofar as it does not violate the norms of probability calculus. From this *subjectivist* perspective, whether rejection of an auxiliary to the favour of the core belief should be deemed irrational depends entirely on the assignment of the prior probabilities to *a* and *h*. This raises the question of what the constraints on setting those priors might be, a version of a common worry within subjective versions of the Bayesian framework. We respond to this worry in section 4.1, where we suggest that, with enough counterevidence being available, belief in *h* should be rejected, and rational Bayesian reasoners should, in the long-run, converge to believing the hypothesis that obtains the best track record in terms of its overall evidential support.

4. Implications for the rationality of conspiracy belief

Let us highlight three general implications of our view for understanding belief in CT. Firstly, if a conjunction of auxiliary and central beliefs is falsified by the evidence, then the central belief can be rescued from refutation by replacing the auxiliary conjunct with an alternative that is not inconsistent with *e*. For example, the core belief that Princess Diana is still alive (*h*) seems to be refuted by photographic evidence showing her funeral, *e*, if one were not to discard the auxiliary assumption that the public media is trustworthy and transparent, and hence offers a reliable evidential source (*a*) to the favour of the alternative auxiliary hypothesis that Diana faked her death (a'). It is apparent that ¬(*ha*) entails *e*. But *h* can be rescued from refutation by replacing *ha* with *ha'*, which is compatible with *e*.

Secondly, there is no principled difference between core and auxiliary beliefs in the way that probabilities are assigned to them given the evidence. The evidential impact on *h* increases relative to the certainty about *a* and vice versa; the important difference lies in their initial probabilities. For example, *e* has a great negative impact on Pr(*a*) but only a minor influence on Pr(*h*) when Pr(*a*) < Pr(*h*). Consequently, the probabilities associated with the rivals to *a* will increase. When *ha* is falsified with Pr(*h*) < Pr(*a*), then *h* would instead be blamed more (to the extent its prior is lower). Generally, auxiliary beliefs are more likely to absorb the blame and be readjusted given disconfirming evidence to the extent that they are more questionable to begin with. This contrasts with earlier approaches that see a principal distinction between conspiracy belief and other kinds of belief (such as Napolitano, 2021).

Thirdly, the apparent resistance to belief updating in light of disconfirming evidence complies with Bayesian norms of reasoning. When *ha* is disconfirmed by *e*, *h* can still be rescued by replacing *ha* with *ha'* (equation IV). But it is not principally irrational to seek confirmation for *h* given *e* via *a'* — shifting probability away from *a* is legitimate if the prior for *h* is sufficiently high and there is an alternative to *a* that is consistent with the evidence. In other words, it is *not always* irrational for a conspiracy theorist to shift probability away from auxiliary hypotheses to protect the central belief. The blame is on the protective belt, not on the updating process itself.

For another example, consider the core belief (corresponding to *h* in our model) that certain electromagnetic waves, including those of the recent 5G technology, weaken our immune system and slowly damage our DNA . This belief has generated several novel predictions, for instance, that power lines cause cancer in children, that cell phones and high-speed networks cause brain tumours, autism, or Alzheimer's disease, and that 5G radio waves contribute to the spread of the COVID-19 pandemic. While this kind of reasoning may be diversely motivated (e.g., by a loss of agency and fear of lacking control), the reasoning itself does not have to be incoherent. Generally, it is possible that each of these predictions was initially formed under the auxiliary assumption that they would be empirically tested by trustworthy scientific standards (corresponding to *a* in our model). However, people might not always understand how scientific testing works, and how the results of a study are to be interpreted. Laboring under the influence of certain biases, e.g., confirmation bias and deterministic thinking, some reasoners may tend to ignore much of the established knowledge about electromagnetic waves and their influence on the human body. For instance, they might discount the observation that 5G technologies use weak electromagnetic fields that have not been scientifically associated with a higher chance of developing brain tumours (corresponding to *e* in our model), based on the alternative auxiliary that wavelength weakness does not preclude long-term damaging effects. Attribution biases (suggested by Keeley, 1999) can also lead to adopting alternative auxiliaries postulating a malicious manipulative strategy behind the government's installation of 5G networks and the information communicated in scientific reports. We return to the effects of inductive biases on the formation of auxiliary beliefs in section 5.

From this perspective, the apparent irrationality of conspiracy belief does not necessarily reside in the dismissal of disconfirming evidence. For example, instead of changing the assumption that electromagnetic waves damage our immune systems, an agent could postulate additional hidden causes that could lead to the circulation of manufactured scientific evidence falsely showing a lack of correlation between wireless technology and the spread of the virus. Such additional hidden causes would allow for a consistent reinterpretation of the scientific evidence as irrelevant to the central hypothesis. This, in turn, would support *ha*, namely the belief that there is a conspiracy surrounding 5G technology.

This also shows that the extent to which the evidence impacts, positively or negatively, the central belief depends on the auxiliary beliefs endorsed, such that a change in the field of auxiliary beliefs can produce a change in the interpretation of the data. The extent to which a belief is confirmed depends not only on the difference between the prior probability of that target belief and how probable it is given the available evidence, but also on the probabilities assigned to the protective belt (see figure 1). Data that might be interpreted as supporting a belief, based on a set of auxiliary assumptions A, could be interpreted as defying that belief, based on the auxiliary set B. In the next section, we address some of the limits to probability shifting for belief protection.

4.1 Irrational belief as a desperate rescue

As Napolitano (2021) and others have argued, some CT beliefs appear to be irrational because they resist available evidence specifically when it has a negative bearing. If CT belief is principally compatible with Bayesian norms of rationality, then how can we account for their apparent irrationality?

One idea is that CT belief is irrational because it builds on the endorsement of *ad hoc* assumptions that are motivated by personal desires or wishful thinking (Hahn & Harris, 2014; Kunda, 1990). We can say that an auxiliary belief is ad hoc when it entails unconfirmed claims while being specifically called to rescue a central belief by accommodating the disconfirmatory evidence. When a belief is well confirmed in a stable manner over time, it is well entrenched and not ad hoc. However, if the robustness to disconfirmation is conferred by a strong prior for the central belief, then the endorsement of an ad hoc auxiliary need not be due to motivated reasoning.

Strevens' (2001) example is the discovery of Neptune, whose existence was initially postulated to explain away apparent deviations from the path that Newton's theory of gravitation had predicted for the orbit of Uranus. Strevens characterizes this postulation as a "glorious rescue'' because it correctly shifts most of the blame for a false prediction onto the auxiliary belief that there are seven planets in the solar system, and it generated new predictions that allowed the discovery of Neptune by Herschel's telescopic observations. Analogously, Watergate might be a prime example of a glorious rescue of CT belief. It was correct to replace the hypothesis that the Nixon administration is trustworthy by a cover-up assumption to protect the belief that the 1972 break-in at the Democratic National Committee headquarters at the Watergate Office Building involved a conspiratory act. This hypothesis entailed novel predictions that allowed for its subsequent confirmation by the discovery of the Oval Office tapes which revealed that Nixon conspired. However, not all rescues lead to glorious discoveries. Some attempts to protect a central belief wrongly blame the auxiliaries for a failed prediction. Strevens characterises such cases as "desperate'' because researchers merely “cling to the central hypothesis and discard the evidently superior auxiliary” (ibid.). In this kind of rescue, the blame is *not rationally apportioned* between *a* and *h*. Desperate rescues can be treated as a form of irrational reasoning according to the Bayesian standard.

An example for a desperate rescue in the case of CT belief might be the controversy about Bill Gates’ investments into vaccine production. Upon observing photographic evidence reporting Gate’s receiving the vaccine injection, the central belief, that Gates has funded and planned the COVID-19 pandemic to implant controlling microchips into people (*h*), and the initial auxiliary supplement, that the public media reports are reliable (*a*), are disconfirmed. Following the analysis in section 3, *h* can be protected from refutation by doubting *a*. Then the unexpected event, his reception of the vaccine injection, can be explained by postulating, for example, that he faked it by controlling various public media agents. However, following the analysis in section 3, this rescue is desperate, if the initial faith in the public media was very high to begin with (being equivalent to Pr(*a*) in the model being very high). Then the shift to believing that Gates faked his vaccine reception is unwarranted, since the belief that he has manufactured the pandemic to implant microchips loses most of its credibility (Figure 1), and so trust in the public media would be wrongly discarded.

Of course, labelling this explanation as ‘desperate’ is appropriate only if the belief that the public media reports are reliable is evidentially superior to the belief that Gates manufactured the spread of the virus. The analogy to the history of science suggests that whether a belief is evidentially superior depends on its historical track record. The universal law of gravitation had accumulated a much greater degree of confirmation over the past than the competing alternatives, so its superior track record provided reasons for assigning to it a much stronger prior belief. If agents have set these priors in correspondence with the historical track record, and if this prior turns out to be lower than the priors for available alternatives, then clinging on to that belief (to the disfavour of a better alternative) can be considered desperate, or simply irrational. In the case where the auxiliary that Gates controls various governmental and public agents is specifically called to protect the belief that there is a conspiracy surrounding his investments into vaccine production, this hypothesis might be internally coherent and generate a new prediction — that Gates aims to control the world's population — but its associated central belief is not well entrenched, since little confirmation in favor of this belief has accumulated.

Two important implications of the track-record constraint are that identifying whether a given case of belief protection counts as 'glorious' or 'desperate' in terms of its overall evidential support can often be determined only in the long run. As examples of the conspiracy beliefs surrounding Gates' investments and the Watergate scandal illustrate, the distinction between 'glorious' and 'desperate' rescues might itself be a matter of degree.

5. Reconciling monological and self-insulated systems

Let us now recapitulate the monological and self-insulation approaches outlined in section 3 through the Bayesian lens, where they can be thought of as two ways of describing the same kind of belief system.

On the Bayesian view, both central and auxiliary beliefs mutually constrain one another and follow the same principles for belief revision. For instance, how likely an auxiliary hypothesis is to be rejected directly depends on how well it is entrenched compared to its rivals. This fits with the common characterization of CT belief in terms of a monological system, that is, a system where different beliefs form a self-sustaining network that can absorb varying kinds of evidence. As elucidated in section 3, as $\Pr(a|h)$ increases, the probability of the alternative auxiliaries multiplied with $\Pr(h)$ decreases (Figure 1). For instance, the probability that Gates faked his COVID-19 vaccine injection given that he planned the spread of the virus is considerably higher than the probability that the photograph of him getting the first vaccine is real given that the vaccine contains a dangerous microchip. On one hand, these auxiliaries mutually constrain each other; they cannot be both true in light of the evidence. On the other hand, conflicting hypotheses that share content, e.g., about Gates' evil plans, are part of the same hypotheses space; they are tied to each other by playing the role of arguments in a probability function that is distributed across all of them. Since, following the probability axioms, the probabilities associated with the individual hypotheses must sum up to 1, raising credence in one hypothesis affects credence in the other ones. By the same manner, stipulating *h* (e.g., that Gates has planned the pandemic) directly raises the probability of certain auxiliary hypotheses (e.g., that Gates faked his photograph) to the disfavour of others (e.g., that the public media is transparent). This also holds for the introduction of new auxiliaries to the hypothesis space, insofar as they are compatible with the central hypothesis, that is, if $\Pr(a|h)/\ P(h)$ is sufficiently high. In this sense, the Bayesian view on offer allows us to model how even mutually inconsistent conspiratorial beliefs are interconnected and can support one another as long as they share content with a central hypothesis, thus accommodating the framing of monologicity in terms of higher-order beliefs proposed by (Douglas et al., 2019). We expect that beliefs central to some CT will be more general in scope, not only because they offer hypotheses that better reconcile mutually inconsistent auxiliaries (thus being better entrenched and less prone to disconfirmation), but also because Bayesian methods favor hypotheses that are more general and have a higher chance of generalising to new data (a feature which we discuss in the next section).

Our treatment likewise captures aspects of the apparent insensitivity to disconfirmation highlighted by Napolitano (2021) and Keeley (1999), which we understand as a matter of belief protection (as opposed to ignorance). In agreement with these views, our view implies that to evaluate the (ir)rationality associated with CT belief from a normative perspective, we should not *only* study the internal relationship among such beliefs, but also to their (dis)confirmatory relationships with the evidence, if only in the long run. Corresponding to our analysis, belief networks might be both monological and self-insulated, to the extent that they are self-sustaining (at least over short periods of time) and exemplify desperate rescues (i.e., they are not well entrenched in the long-run, and hence likely to fail in terms of their novel predictions).

While reconciling these views, our approach does not analyse CT beliefs through the lens of flawed reasoning processes, and instead renders some cases of CT belief – especially the "glorious'' rescues – akin to rational reasoning in everyday as well as in scientific inquiry

6. A benchmark of rationality

So far, we have shown how the Bayesian treatment can unify the monological and higher-order views as well as account for the seeming insensitivity of conspiratorial beliefs to disconfirmatory evidence. However, one of the reasons why epistemologists such as Napolitano are keen on postulating the insulation hypothesis is because they are engaging in conceptual engineering of the notion of (as well as belief in) a conspiracy theory in order to cast it into a concept that is by definition epistemically suspect and derogatory (Napolitano & Reuter, forthcoming). This clashes with our claims that beliefs about conspiracy theories might update in light of disconfirming evidence and belong to fundamentally the same class as other kinds of doxastic states. This, together with the fact that Bayesian probability theory is often taken as a normative standard for rationality, might suggest a view in which believing in conspiracy theories is *always* rational after all. However, we do not endorse such a view.

Our claim is not that all conspiracy beliefs are rational, but rather that beliefs about conspiracy theories are not fundamentally different from any other kind of belief. Furthermore, given a multitude of definitions and a lively debate over the notion of conspiracy theories, we do not wish to engage in any project of engineering the concept. In our view, a belief in a conspiracy is not epistemically different from a belief in a conspiracy theory. After all, there are well-recognized cases of conspiratorial beliefs being true, such as the belief in the Watergate scandal. However, this does not mean that the difference between a true belief about an actual conspiracy and a false belief about an outlandish conspiracy lies only in their truth value. In fact, our account provides a benchmark for tracking the credibility of a central hypothesis and whether it should be abandoned. We can make this explicit by returning to Streven's distinction between glorious and desperate rescues.

Recall from section 4.1 that the two kinds of revisions to auxiliary beliefs differ in how well the central belief is supported by the available evidence. The discovery of Neptune is an example of glorious rescue because, at the time of the adoption of the auxiliary hypothesis about the existence of the planet, Newton's theory had been much better confirmed than any competing theory, which in turn justified adopting surprising auxiliary hypotheses to account for the anomalous disconfirmatory evidence. It is this condition that is violated in what our account clearly stigmatises as a desperate rescue, where the central hypothesis is held onto despite its bad track record and given few sources of evidence and predictions that repeatedly fail to be confirmed.

What is crucial to the view is that the differences between the two kinds of belief revision, glorious and desperate, can be compared in terms of the relative probabilities of the beliefs in question. Thus, as Streven's points out, one is only justified in revising the auxiliary beliefs when the probability of the central hypothesis, Pr(h), is higher than that of the auxiliary, Pr(a), while the degree of justification (or 'glory' of the rescue in Streven's terms) is inversely proportional to the prior probability of the auxiliary hypothesis. While this does not offer, as some philosophers may wish, an a priori distinction between legitimate and illegitimate beliefs in conspiracies, it puts us on track for a comparison of different conspiratorial explanations. While the view advocated here does not clearly state whether the belief in CIA's involvement in the US crack epidemic is a case of glorious or desperate rescue, it clearly states that beliefs in conspiracy theories which are poorly entrenched and can only be maintained through regular adoption of new auxiliary hypotheses, such as the belief that the Earth is flat or the QAnon conspiracy, are desperate and irrational.

7. The effect of inductive biases on the formation of belief in conspiracy theories

In the penultimate section of this paper, we explore some additional insights that Bayesian cognitive science offers for explaining conspiracy belief formation. Specifically, we focus on the initial parameters of the belief system and their constraining role in the inductive process. Such "inductive biases" decide which auxiliary beliefs will be considered as 'good' explanations for observations in the first place, by weighing the posteriors and priors computed for individual beliefs (Tenenbaum et al., 2006). Inductive biases can take multiple forms, but here we concentrate on two examples that seem helpful for characterising important cognitive constraints on the formation of CT beliefs.

The first example is a bias for sparse beliefs, which, following Gershman (2019), encodes a preference for auxiliaries that generate narrow predictions consistent with the evidence. In the extreme, these are auxiliaries that predict all and only the observed data. Evidence suggesting that people endorse such biases comes from studies of concept learning. For example, most people asked to generate number concepts from the range 1, 1000 mention just prime numbers (Perfors & Navarro 2009), illustrating a tendency to place the most weight on very few, highly specific predictions. From the perspective of the philosophy of science, sparse beliefs are valuable because their low initial probability makes them extra informationally relevant to acquired empirical evidence for or against them (Bar-Hillel, 1955; Popper, 1954). Sparse beliefs are verifiable by sparse evidence.

Under the assumption of a bias towards sparse belief systems, we can expect such systems to also be highly torn towards determinism, which is a preference to place all weight on only a few beliefs consistent with the data (Gershman, 2019).[8] In the extreme, the system will endorse only auxiliaries that perfectly predict the observed data in its inferences. The rationale for this is that belief networks that predict only a few possible events must assign each prediction a high probability for the distribution to sum up to 1, as required by the laws of Bayesian inference. Together, a bias for sparsity and determinism can lead the system to single out one factor as "the only true cause" for a given set of observations.

The second example is a bias toward simple belief systems. This bias is driven by concern for predictive accuracy and avoidance of overfitting hypotheses to the available data, which can be illustrated by the problem of model selection in Bayesian statistics (see Griffiths & Yuille, 2006). The problem in question is choosing, based on the observations, among a set of hypotheses of varying complexities. Complex hypotheses are more flexible and can be better fitted to the available data. This means that they can make better predictions, provided those future observations follow tendencies present in the existing data. However, complex hypotheses can also lead to worse predictions if the available data is anomalous. Thus, on average, simpler hypotheses will generalise better across a broad range of scenarios and possible observations. This feature has been labelled as *Bayesian Occam's Razor* (BOR): beliefs that are too sparse or fixed are unlikely to generate future observations; beliefs that are too flexible can generate many possible data sets, while also being unlikely to generate a particular data set at random. Interestingly, a recent study by Blanchard, Lombrozo, and Nichols (2017) has shown that when confronted with simple narrative tasks "people's intuitive judgments follow the prescriptions of BOR, whether making estimates of

[8] Gershman does not explain how consistency can be measured, but Strevens (2001, p. 529) assumes that some measure of the degree of confirmation is appropriate.

the probability of a hypothesis or evaluating how well the hypothesis explains the data." (Blanchard et al., 2017).

These examples illustrate that inductive biases can pull the updating process in opposite directions. An optimal inference system should achieve a balance to avoid overfitting (making hypotheses too precise) or underfitting (making them too general) (see Forster & Sober, 1994 in the context of scientific inferences). Depending on which inductive biases are built-in, different inferences can become plausible in light of the same evidential observation. For example, a bias toward sparse hypotheses could explain why "conspiracy theorists use a large set of auxiliary hypotheses that perfectly (i.e., deterministically) predict the observed data and only the observed data (sparsity)'' (Gershman, 2019, p. 23). However, such a system would be unable to generalise towards novel cases, violating a bias toward simplicity. Thus, as Gershman (2019, p. 23) himself suggests, there may be significant individual differences in how strong particular biases are in different individuals. This may be tied to certain personality traits (e.g. "epistemic vices", see Sunstein & Vermeule, 2008; Cassam, 2019) which could predispose some people to have a propensity for forming belief networks more prone to hijacking by CT narratives. Due to space constraints, we leave the exact ways in which such biases might influence inferential processes as a topic for future empirical investigation.

9. Conclusion

In this paper, we have used the Bayesian framework to analyse beliefs about conspiracy theories and present the implications of framing such beliefs in this way for the existing proposals regarding their nature. As we have shown, the Bayesian framing not only offers helpful insight for the existing accounts, but also allows to unify them under a single formal umbrella. Where the Bayesian approach departs from the previous proposals is that it does not principally rule out that conspiracy beliefs can be rational. While previous accounts tend to associate the apparent irrationality of conspiracy belief with a failure to update on novel contrary information, our analysis suggests that the irrationality is not to be found in the updating process itself, but in the desperate attempt to rescue badly confirmed hypotheses by introducing only weakly grounded ad hoc auxiliary beliefs. A tendency to do so may be enhanced by strong individual inductive biases. However, since many of the glorious rescues in science might have at one point in time seemed desperate, an important consequence of our view is that part of the irrationality of conspiracy beliefs might depend on the wider context in which they are formed and independent means of their verification. Consequently, we suggest that more attention should be given to aspects of conspiracy belief other than updating, for example, the role that social factors play in their acquisition.

References

Bar-Hillel, Y. (1955). An examination of information theory. Philosophy of Science, 22(2), 86-105.

Basham, L. (2016). The Need for Accountable Witnesses: A Reply to Dentith. *Social Epistemology Review and Reply Collective*, 5(7): 6-13.

Bortolotti, L., Ichino, A., & Mameli, M. (2021). Conspiracy theories and delusions. *Reti, saperi, linguaggi*, *8*(2), 183-200.

Butter, M. (2021). Conspiracy Theories–Conspiracy Narratives. *DIEGE-SIS. Interdisciplinary E-Journal for Narrative Research/Interdisziplinäres E-Journal für Erzählforschung*, 10.1: 97–100.

Cassam, Q. (2019). *Conspiracy Theories*. Medford, USA: Polity Press.

Cichocka A., Marchlewska M., and de Zavala A.G. (2016). Does Self-Love or Self-Hate Predict Conspiracy Beliefs? Narcissism, Self-Esteem, and the Endorsement of Conspiracy Theories. *Social Psychological and Personality Science*. 7(2):157-166. doi:10.1177/1948550615616170

Clarke, S. (2002). Conspiracy Theories and Conspiracy Theorizing. *Philosophy of the Social Sciences*, 32(2): 131-50.

Dentith, M. R. X. (2016). When Inferring to a Conspiracy might be the Best Explanation. *Social Epistemology,* 30, 5-6: 572-591.Dentith, M. R. (2019). Conspiracy theories on the basis of the evidence. *Synthese, 196*(6), 2243-2261.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., and Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology, 40*, 3-35.

Duhem, P. (1953). Physical Theory and Experiment. in Herbert Feigl and May Brodbeck (ed.), Readings in the Philosophy of Science. New York: Appleton-Century-Crofts, Inc., pp. 235–252.

Feldman, S. (2011). Counterfact conspiracy theories. *International Journal of Applied Philosophy*, 25(1), 15-24.

Fenster, M. (2008). *Conspiracy Theories: Secrecy and Power in American Culture*. Minneapolis: University of Minnesota Press.

Fitelson, B., and Waterman, A. (2005). Bayesian confirmation and auxiliary hypotheses revisited: A reply to Strevens. *The British journal for the philosophy of science*, *56*(2), 293-302.

Fitelson, B., and Waterman, A. (2007). Comparative Bayesian confirmation and the Quine–Duhem problem: A rejoinder to Strevens. *The British journal for the philosophy of science*, *58*(2), 333-338.

Forster, M., and Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1-35.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic bulletin and review*, 26(1), 13-28.

Goertzel, T. (1994a). Belief in conspiracy theories. *Political Psychology*, 15(4), 731–742. https://doi.org/10.2307/3791630

Goertzel, B. (1994b). *Chaotic logic*. New York: Plenum.

Goreis, A., & Voracek, M. (2019). A systematic review and meta-analysis of psychological research on conspiracy beliefs: Field characteristics, measurement instruments, and associations with personality traits. *Frontiers in psychology*, *10*, 205.

Griffiths, T. L., and Yuille, A. L. (2006). Technical introduction: A primer on probabilistic inference. *UCLA. Department of Statistics Papers no. 2006010103*. Los Angeles, CA: UCLA.

Hagen, K. (2022). Are “Conspiracy Theories” So Unlikely to Be True? A Critique of Quassim Cassam’s Concept of” Conspiracy Theories”. *Social Epistemology*, 1-15.

Hagen, K. (2018). Conspiracy theorists and monological belief systems. *Argumentation*, *3*, 303-326.

Hahn, U., and Harris, A. J. (2014). *What does it mean to be biased: Motivated reasoning and rationality*. In Psychology of learning and motivation, Brian H. Ross (ed.). Vol. 61, pp. 41-102). Academic Press.

Harris, K. (2018). What's epistemically wrong with conspiracy theorising? *Royal Institute of Philosophy Supplements*, *84*, 235-257.

Urbach, P., Howson, C. (1993). *Scientific Reasoning: The Bayesian Approach*, Chicago and La Salle, IL: Open Court, Second Edition.

Keeley, B. L. (1999). Of Conspiracy Theories. *Journal of Philosophy*, 96: 109–26.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3), 480.

Lakatos, I. (1976). *The Methodology of Scientific Research Programmes*. Cambridge: Cambridge University Press.

Levy, N. (2021). Echoes of covid misinformation. *Philosophical Psychology*, 1-17.

McKay, R. (2012). Delusional inference. *Mind & Language*, *27*(3), 330-355.

Miller, J. M. (2020). Do COVID-19 conspiracy theory beliefs form a monological belief system?. *Canadian Journal of Political Science/Revue canadienne de science politique*, *53*(2), 319-326.

Napolitano, G., and Reuter, K. (*forthcoming*). What is a conspiracy theory? *Erkenntnis*.

Napolitano, M. G. (2021). Conspiracy Theories and Evidential Self-Insulation. In The Epistemology of Fake News (pp. 82-106). Oxford University Press.

Stokes, P. (2016). Between Generalism and Particularism about Conspiracy Theory: A Response to Basham and Dentith. *Social Epistemology Review and Reply Collective*, 5(10): 34-39.

Strevens, M. (2001). The Bayesian Treatment of Auxiliary Hypotheses. *British Journal for the Philosophy of Science*, 52(3).

Strevens, M. (2005). The Bayesian treatment of auxiliary hypotheses: Reply to Fitelson and Waterman. *The British journal for the philosophy of science*, *56*(4), 913-918.

Sunstein, C. and Vermeule, A. (2008). Conspiracy Theories: Causes and Cures. *The Journal of Political Philosophy*. 17(2): 202–227.

Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., ... & Corlett, P. R. (2021). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*, 5(9), 1190-1202.

Tangherlini, T.R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., and Roychowdhury, V. (2020). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLoS ONE,* 15(6). doi:10.1371/journal.pone.0233879

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7), 309-318.

van Prooijen, J.W., and van Vugt, M. (2018) Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13(6): 770-788. doi:10.1177/1745691618774270

Wood, M. J., Douglas, K. M., and Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, *3*(6), 767-773. https://doi.org/10.1177/1948550611434786

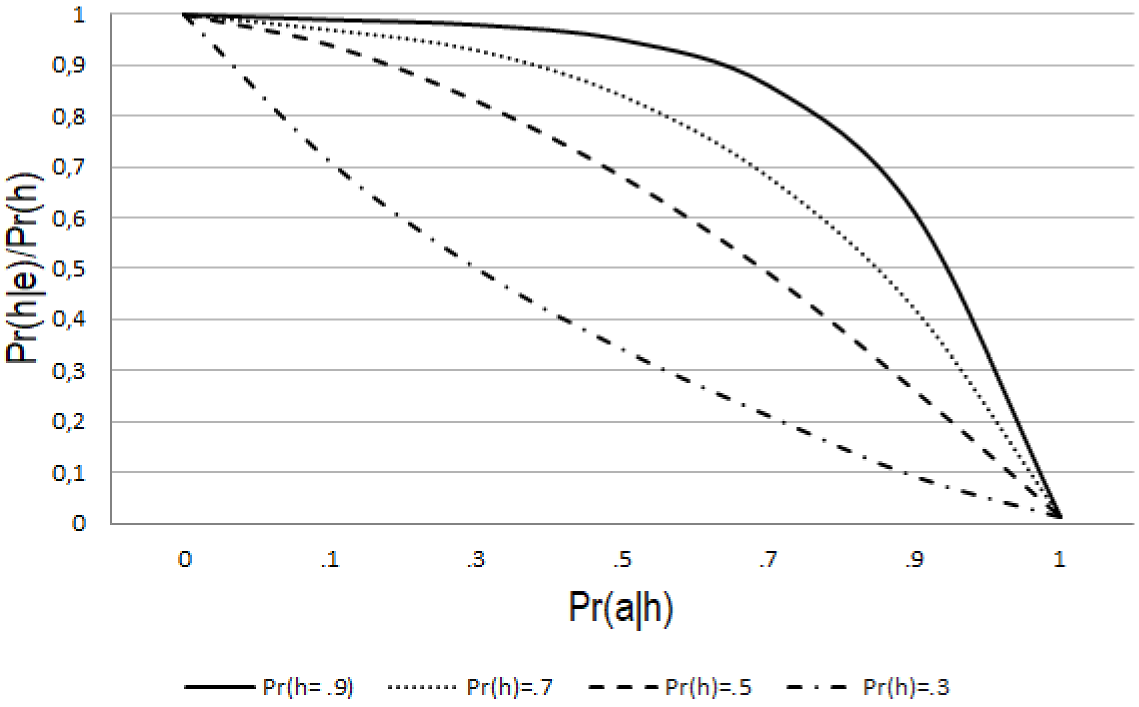

Figure 1. The ratio of the posterior to the prior of *h* as a function of Pr*(a*|*h)* for different values of the prior. Adapted from Gershman (2019, p. 15) and Strevens (2001, p. 526).