Believing Conspiracy Theories: A Bayesian Approach to Belief Protection

Nina Poth & Krzystof Dolega
Ruhr-Universität Bochum

Abstract

Despite the harmful impact of conspiracy theories on the public discourse, there is little agreement about their exact nature. Rather than define conspiracy theories as such, we focus on the notion of conspiracy belief. We analyse three recent proposals that identify belief in conspiracy theories as an effect of irrational reasoning. Although these views are sometimes presented as competing alternatives, they share the main commitment that conspiracy beliefs are epistemically flawed because they resist revision given disconfirming evidence. However, the three views currently lack the formal detail necessary for an adequate comparison. In this paper, we bring these views closer together by exploring the rationality of conspiracy belief under a probabilistic framework. By utilising Michael Strevens' Bayesian treatment of auxiliary hypotheses, we question the claim that the irrationality associated with conspiracy belief is due to a failure of belief revision given disconfirming evidence. We argue that maintaining a core conspiracy belief can be perfectly Bayes-rational when such beliefs are embedded in networks of auxiliary beliefs, which can be sacrificed to protect the more central ones. We propose that the irrationality associated with conspiracy belief lies not in a flawed updating method according to subjective standards but in a failure to converge towards well-confirmed stable belief networks in the long run. We discuss a set of initial reasoning biases as a possible reason for such a failure. Our approach reconciles previously disjointed views, while at the same time offering a formal platform for their further development.

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

1. Introduction

Over the course of the past decade, conspiracy theorising has become part of the public discourse; the advent of social media and the erosion of trust in epistemic authorities has opened the door to the rise of political populism and widespread scepticism about mainstream narratives, the effects of which are well felt among the ongoing COVID-19 pandemic. The prevalence of conspiracy theories has led to an explosion of research on the topic, with understanding and identifying the nature of such theories becoming an issue of primary concern. Given the duplicitous nature of the phenomenon, it is unsurprising that scholarly attempts at defining conspiracy theories have become a source of contention. There is little agreement about the explanatory status of such theories (see e.g. Butter [2021]; Dentith [2014]; Fenster [2008]), and whether they belong to a general category (Stokes [2016]) or should be evaluated on a case by case basis (Basham [2016]; Dentith [2016]).

Some researchers proposed to sidestep the search for a unifying definition, by focusing their attention on understanding the psychological factors that may contribute to the endorsement of beliefs in conspiracy theories (see e.g. Brotheron and French [2014]; Cichocka, Marchlewska, and de Zavala [2016]; Douglas, Sutton, and Cichocka [2018]; Van Prooijen and van Vugt [2018]). One claim which has gained traction in the literature is that agents' beliefs about conspiracy theories result from irrational reasoning. The received view is that such belief is principally epistemically wrong. Two important components of this negative characterization are, on the one hand, the widely shared assumption that proponents of conspiracy theories resist belief revision in light of disconfirming evidence[1] and, on the other hand, that systems of conspiracy beliefs tend to highly correlate (i.e. people who believe in one conspiracy tend to believe in others), even though these beliefs are often considered to be semantically and logically unrelated (Goertzel [1994a]), or even mutually inconsistent(Douglas et al. [2012]).

However, despite the relevance of the topic for wider research on cognition, the literature attempting to explain these features of conspiracy beliefs is as fragmented as that exploring the possible definitions of conspiracy theories themselves. Some researchers explain the resistance to counter evidence as a symptom of the *monological* nature of a belief system and its resulting tendency to ignore “context in all but the shallowest respects" (Goertzel [1994a], p. 740) or to self-insulate the system from the information flow of the surrounding context (Napolitano [2021]), for instance, by creating echo chambers (Lackey [2021]). Other researchers explain believers’ simultaneous endorsement of a set of contradictory beliefs by appealing to their acceptance of more abstract, higher-order, beliefs that support the consistency among more specific beliefs that appear to be inconsistent at a lower level (Douglas et al. [2012], [2019]). For example, Douglas et al. ([2012]) consider it possible to simultaneously believe that Diana is dead and alive by appealing to the higher-order belief that the government is engaging in a cover-up story, which justifies either of these alternatives.

In this paper, we aim to explore the relationship between beliefs about conspiracy theories (and conspiracies more generally) and their rationality. While we agree that the characterization of conspiracy theoretic belief as irrational has intuitive appeal, we think that defining such beliefs as resulting from irrational reasoning might be premature. For our analysis, we focus on how agents incorporate new information to update their beliefs by borrowing the tools of Bayesianism and reinterpreting sets of propositions as probabilistically related beliefs. Our treatment of conspiracy belief originates in Strevens’ ([2001]) Bayesian analysis of auxiliary hypotheses and its recent application in cognitive science (Gershman [2019]). These views explain the robustness of high-probability beliefs to revision in the light of counterevidence based on the availability of low-probability beliefs that can easily be thrown overboard to protect the beliefs central to the system. 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 possible point of departure to explain the correlations between apparently incoherent belief systems and the resulting divergences in people’s reasoning about the world. This treatment not only reconciles the monological belief and higher-order views within a single formal framework, bringing previously disjointed views closer together but also opens new avenues for the study of conspiracy belief as a unified phenomenon.

Before we proceed, we wish to add a clarificatory remark. Typically, a conspiracy theory is understood as a set of propositions about conspiracies (Dentith [2019]; Harris [2018]; Keeley [1999]). For example, take the conjunction of the proposition that there is a conspiracy around 9/11 and the proposition that the towers were a controlled demolition and the proposition that jet fuel can't melt steel beams. Together, this set forms a theory that is endorsed to explain the underlying causes of 9/11. The belief that there is a conspiracy around 9/11 is an example of what we mean when we talk about conspiracy belief. To avoid pejorative use of the label, we assume that the negative valence associated with claims about conspiracy theories is a separate issue of characterising additional aspects of a subset of conspiracy beliefs. Thus, we do not want to engineer a special concept of conspiracy theoretic belief,

[1] Not everyone believes that resistance to counter-evidence is a problem of conspiracy belief (e.g. see Dentith [2019]; Harris [2018]; Keeley [1999]).

nor do we assume that there is a principled difference between beliefs in conspiracies and conspiracy theories. If any other difference exists, it should emerge as a conclusion rather than constitute a starting point of our inquiry.

Introductory remarks in place, we will now bring out the three views about the nature of conspiracy theoretic beliefs in section 2. After that, we will move on to outline the Bayesian treatment of the relationship between networks of associated beliefs and disconfirmatory evidence in section 3. We then apply this analysis to the received negative characterization of conspiracy belief in section 4. Section 5 reconciles the monological, higher-order and self-insulated views, while section 6 discusses the possible sources of conspiracy belief formation. We then clarify in section 7 whether our account renders conspiracy beliefs as irrational, followed by the discussion of two objections that our view might be facing in the penultimate section. We end with a brief conclusion about the possible implications of this paper for future research.

2. Irrationality of conspiracy belief and its sources

As already mentioned, the three versions of the negative characterisation stress the doxastic structure of conspiracy beliefs and their epistemic support or lack thereof. They are:

(a) the monological thought view on which beliefs in conspiracy theories directly support one another to form a self-sustaining network (Goertzel [1994a]; [1994b]);

(b) the higher-order view on which inconsistent conspiracy beliefs are only related in so far as they are independently supported by a broader higher-order belief that makes them consistent (Douglas, Sutton, and Cichocka [2017]; Douglas et al. [2019]); and

(c) the self-insulation view on which such beliefs are isolated from disconfirming evidence and other doxastic states (Napolitano [2021]).

The first view, (a), assumes that belief in conspiracy theories is emblematic of a reasoning style in which a set of beliefs comprise a self-sustaining network of contents that mutually support each other to afford a coherent explanation of contingent phenomena which 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 (Goertzel [1994a]). 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" (Goertzel [1994b], p. 166). Ted Goertzel ([1994a], p. 740) adds that "[i]n a monological belief system, each of the beliefs serves as evidence for each of the other beliefs." 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.

Contrary to the monological thinking hypothesis, the higher-order view (b) assumes that conspiracy beliefs are only related to each other to the extent that they cohere with a higher-order belief that indirectly provides their mutual support; so, there is no direct evidential relationship between particular conspiracy beliefs. Analysis conducted by Wood et al. ([2012]) has shown that beliefs in mutually contradictory conspiracies correlate positively only when more general beliefs, for example, that there is a conspiracy or cover-up, are not taken into consideration. However, when such beliefs are considered with the more general one, the correlation between the higher-order belief and the more detailed contradictory beliefs is higher than between the more particular ones themselves. Wood et al. present their explanation as an alternative to the monological view: "The monological nature of conspiracy belief appears to be driven not by conspiracy theories directly supporting one another but by broader beliefs supporting conspiracy beliefs in general" (Wood et al. [2012], p. 767). Importantly, the higher-order view renders such beliefs as irrational not only because they allow subjects to endorse mutually incompatible explanations of the same event, but also because their adoption is driven by social and existential factors, such as in-group dynamics or the need for security (Douglas, Sutton, and Cichocka, [2017]).

Finally, unlike the other two positions, the self-insulation view (c) postulates that what is crucial for the irrationality of conspiracy theories is not the relationship among conspiracy beliefs, but how their associated credence is (or rather is not) updated. Napolitano ([2021]) follows Keely ([1999]) in claiming that conspiracy beliefs display the so-called *probabilistic irrelevance* condition according to which an agent's degree of belief in a conspiracy theory does not seem to change regardless of whether disconfirming observations are taken to bear on that belief (i.e. whether the probability that the conspiracy is true is taken to be conditional on the evidence or not). Napolitano ties this to the idea that "a conspiratorial explanation can only be immune to being disconfirmed by any new evidence if it remains so general that it makes no specific predictions" ([2021], p.10), while also pointing out that it is difficult to explain how agents could form such generalbeliefs without succumbing into forming more specific beliefs that could be easily disconfirmed. For such 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 far from being exhaustive, it is sufficient to highlight that the three views share some crucial features despite their many differences. Firstly, they all analyse conspiracy beliefs through the lens of flawed reasoning processes which are taken to be crucial for understanding the target phenomenon. Secondly, they share the important assumption that the cognitive processes which give rise to conspiracy theories are irrational and should be demarcated from rational reasoning, which can be observed 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 and b) or the beliefs and evidence (a and c). Finally, despite the shared focus on the operations which produce and sustain beliefs in conspiracy theories, none of the three views offers a detailed analysis or model of the process it describes.[2] This is an important deficiency of the three 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 conceptually, and in formal terms? Under what conditions is rejecting counterevidence acceptable, and when is it not? Generally, we adopt the Bayesian framework to provide answers to these questions. We think that conspiracy belief's self-insulation is best understood as a form of "explaining away" (Pearl [1988]), rather than ignoring the evidence. Furthermore, we think that 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 explain this by drawing an analogy with Lakatos' ([1976]) idea of degenerating research programs and Strevens' ([2001]) Bayesian treatment of the Quine-Duhem problem, which we outline in the next section. We propose that our analysis reconciles (a) and (b), and though we question the claim that the belief-updating process is irrational, we agree with (c) that an adequate assessment of the rationality of conspiracy belief should take into

[2] A notable exception here is Ben Goertzel's 1994 book which is an attempt at applying complex systems theory to the problem of distinguishing rational from intuitive thought. Unfortunately, this work has produced little impact in terms of formal modelling specific to the topic of conspiracy beliefs (though it has inspired some recent work aimed at formally distinguishing conspiracy narratives from conspiracy theories on the internet, see Tangherlin et al. [2020]). Similarly, 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.

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 views discussed in the previous section 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 mutual dependence between theories and observations, the degree to which an observation can confirm or disconfirm a theory, and so on.

One of the more famous among such issues is the Quine-Duhem thesis according to which a scientific hypothesis cannot be empirically tested in isolation but instead relies on additional background assumptions (e.g. about the accuracy of some other hypotheses or the experimental methodology) which facilitate empirical inquiry (Harding [1976]). Central scientific beliefs are entrenched in a network of associated beliefs, hence the evidence for or against them cannot be evaluated without regarding the whole network of auxiliary beliefs. 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 (Duhem [1953]; Lakatos [1976]).

In his 2001 article, Strevens offers such a method by virtue of a Bayesian formalisation of the Quine-Duhem problem. For this, he needs 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 somewhat 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 conspiracy belief.[3]

We understand the 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. The interesting consequence is that auxiliary hypotheses can be discarded to protect central beliefs in a way that 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*).

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

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

(II) $Pr(h|e) = Pr(ha|e) + Pr(h\neg a|e)$

and

(III) $Pr(a|e) = Pr(ah|e) + Pr(a\neg h|e), since\ Pr(ah + a\neg h) = 1.$

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).

What is the 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

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

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

(V) $Pr(\neg ha) = 1 - Pr(ha) = 1 - Pr(a|h)Pr(h),$

and we can derive

(VI) $Pr(h|e) = \left[\frac{1-Pr(a|h)}{1-Pr(a|h)Pr(h)}\right] Pr(h).$[4]

Thus, under the given assumptions and so long as Pr(*h*) ≠ 1 and Pr(*a*) ≠ 0, belief in *h* should always decrease upon disconfirmation of *ha* by *e*. Furthermore, the blame should be apportioned 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. These two features of the model are illustrated in Figure 1, which is adapted from Gershman ([2019], p. 15) (see also Strevens' [2001] original, p. 526). The graph models the robustness of a central hypothesis to disconfirmation as the ratio Pr(*h*)/Pr(*a*|*h*).

[4] See also Gershman ([2019], p. 15).

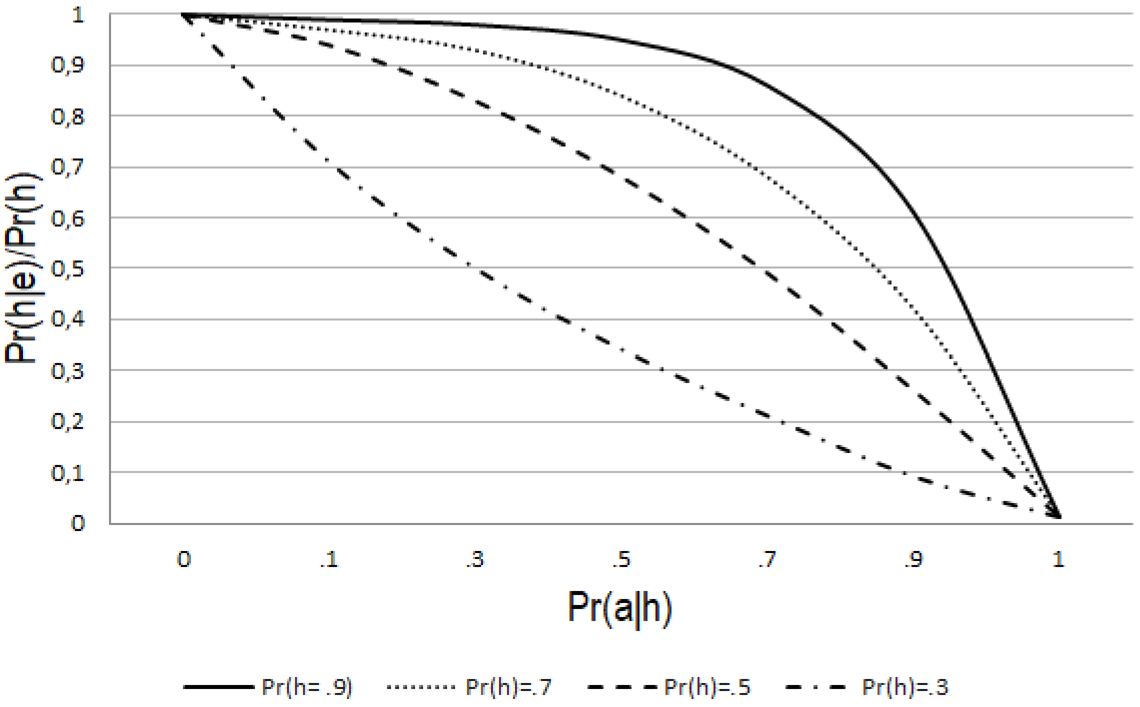

Figure 1. The ratio of the posterior to the prior of *h* as a function of *Pr(a|h)* for different values of the prior. Adapted from Gershman ([2019], p. 15) and Strevens ([2001], p. 526).

Following this Bayesian treatment, 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 the priors might be, a version of a common worry within subjective versions of the Bayesian framework. We respond to this worry in section 4, where we suggest that, with *enough* counterevidence, 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. This extension allows us to set a benchmark for conditions under which belief in a conspiracy can become pathological in extreme cases, even if it seems to be internally coherent at one point in time.

In line with Gershman's ([2019]) recent suggestion, Strevens' treatment can be generalised to many cognitive domains, from conceptual change in childhood to stereotype updating and extinction learning, and conspiracy belief. Generally, the Quine-Duhem problem appears to be well-suited to analyse reasoning about conspiracies, which does not seem to depend on singular, isolated, beliefs, but rather requires a system of interconnected beliefs that support each other. In the following, we ascribe to auxiliaries the function of a protective belt for a conspiracy belief, to account for its robustness to counterevidence. A key observation about the Bayesian treatment is that it explains robustness not as a form of ignorance towards unexpected findings, but rather to explain away, and reinterpret, counterevidence in a way that is consistent with the central belief.

4. Implications for the rationality of conspiracy belief

We now want to highlight three general implications of our view for how conspiracy belief can be understood. 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*. Consider a simple thought experiment. Let *h* be the central hypothesis that Princess Diana is still alive. Let *a* be the auxiliary hypothesis that the public media is trustworthy and transparent, and hence offers a reliable evidential source. Let *e* be the evidential statement that there is photographic evidence showing the damaged car after her severe accident. Furthermore, let *a'* be the auxiliary hypothesis that Diana faked her death. 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 the two kinds of 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]).[5]

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.

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 Diana is still alive, an agent could postulate additional hidden causes that could lead to the circulation of fake photographic evidence of her car accident. Such additional hidden causes would allow for a consistent reinterpretation of the photographic evidence as irrelevant to the central hypothesis. This, in turn, would support *ha*, namely the belief that there is a conspiracy theory surrounding Diana's death and that she faked her own death.

[5] Napolitano defends her position by contrast to Clarke's ([2002]) argument that conspiracy theories are degenerating research programs, which is similar in spirit to our view. She writes: "I agree with Clarke that conspiracy theories are often rendered immune to falsification in this problematic way. It has been objected to Clarke that the exact point at which a conspiracy theory becomes a degenerating research program is unclear (Harris [2018]). However, a similar concern does not apply to my account, *since I take conspiracy theories to be the extreme case of conspiracy-beliefs held in such a way as to be completely immune to disconfirmation in nearby possible worlds*. If there is such a point at which a research program becomes a degenerating one, conspiratorial explanations whose believers will retain in light of any disconfirming evidence one could encounter are an example of that." (Napolitano [2021], p. 12, emphasis added). In section 4.1 we suggest an approximation of the point at which a belief system might become relatively more or less degenerate (we assume that an approximation is all one can ask for, given the context-sensitive nature of beliefs). Although we advocate a more nuanced perspective on conspiracy belief, we generally agree with Napolitano that a (conspiracy) belief system which is "completely'' immune to counter evidence must be somehow deviant. Our goal is to account for weaker cases of self-insulation as well, permitting that self-insulation is a matter of degree.

This also illustrates 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. Consequently, 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 conspiracy beliefs appear to be irrational because conspiracy theorists commonly disregard the evidence when it has negative bearing on their beliefs. If conspiracy belief is principally compatible with Bayesian norms of rationality, then how can we account for their apparent irrationality?

One appealing thought is that conspiracy belief is irrational because it builds on the endorsement of unconfirmed *ad hoc* assumptions that are motivated by personal desires or wishful thinking (Hahn and 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]) provides a helpful example from the history of science. The existence of Neptune was postulated by astronomers in the early nineteenth century to explain away apparent deviations from the path that Newton's theory of gravitation had predicted for the orbit of Uranus. The assumption that there was an additional object that irritated the motion of Uranus offered a way to protect Newton's theory from refutation. After all, Newton's calculations could have failed to hold for large distances, thereby generating the wrong prediction for the motion of celestial bodies.

Nevertheless, astronomers rightly decided to maintain Newton's theory and instead postulated an eighth planet, Neptune, to explain away the unexpected observations. One reason for this is that Newtons' theory had obtained far greater evidential support in the past than any of the available alternatives. The fact that it was well entrenched made its failure unlikely. The other, important reason was that the postulation of a hidden disturbing object was fruitful for the independent discovery of an eight planet. Neptune's existence was itself inferred from the observed fluctuations in Uranus' predicted orbit together with hypotheses about its location and size. This suggests that the postulation of Neptune's existence at the time was *not* ad hoc since it could subsequently be confirmed based on Herschel's independent telescopic observations. In fact, Strevens characterises the case of Neptune's discovery as a "glorious rescue'' (Strevens [2001], p. 536) because it correctly shifts most of the blame for a false prediction onto the previous auxiliary belief that there are only seven planets in the solar system. The glory is ascribed to the rivalling Neptune auxiliary that made the correct prediction. Generally, glorious rescues are such that the auxiliary belief is specifically called to protect a central belief, while this belief has been well confirmed in the past. In the following, we refer to such beliefs as well-entrenched, meaning that they have a good track record of support by multiple sources of evidence and produce predictions that are repeatedly confirmed. We assume that well entrenched hypotheses are, ceteris paribus, more probable than those that are not well entrenched, due to their confirmatory records.

But not all rescues lead to glorious discoveries. Some attempts to protect a central belief might indeed 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*. Such a desperate rescue can be treated as a form of irrational reasoning according to the Bayesian standard.

An example for a desperate rescue in the case of conspiracy belief might be the controversy about Diana’s car accident. Upon observing photographic evidence reporting Diana’s accident, the central belief, that Diana is still alive, and its auxiliary supplement, that the public media reports are reliable, are disconfirmed. Following the analysis in section 3, the belief that Diana is still alive (*h*) can be protected from refutation by doubting the assumption that public media reports are reliable. Then the unexpected event, the car accident, can be explained away by postulating, for example, that Diana faked her death. However, following the analysis in section 3, this rescue is desperate, if the initial faith in the public media is very high (being equivalent to *Pr(a)* in the model being very high). Then the shift to believing that Diana faked her death is unwarranted, since the belief that she is still alive loses most of its credibility (Figure 1), and so trust in the public media is 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 Diana is still alive. The analogy to examples from the history of science suggests that whether a belief is evidentially superior will depend on its historical track record. In Newton’s case, the universal law of gravitation had accumulated a much greater degree of confirmation over the past than the competing alternatives, and so its superior track record provided reasons for assigning to it a much stronger prior belief. In other words, the track record of past successes of a belief in predicting phenomena provides an additional constraint on how prior probabilities in the belief network are set. If agents have set these priors in correspondence with the historical track record of the belief, and if this prior turns out to be lower than the 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 Diana was murdered is specifically called to protect the belief that there is a conspiracy surrounding her death, this hypothesis might be internally coherent and generate a new prediction — that there is a murderer of Diana — but its associated central belief is not well entrenched, since little confirmation via empirical evidence has accumulated.

An important implication of the track-record constraint is that the identification of 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. This can be illustrated by cases in which it is still difficult to identify whether the evidence clearly supports a belief in a conspiracy. One example is the controversy surrounding the death of Władysław Sikorski, commander-in-chief of the Polish Army and Prime Minister of the Polish government in exile during the second world war. Assume that the central belief, *h*, states ‘Sikorski’s death was the result of a German/British/Soviet conspiracy’, and the auxiliary, *a*, states ‘Sikorski was murdered in the plane prior to the crash.’ This produces the prediction, *p*: ‘Sikorski was assassinated.’ However, recent exhumation of Sikorski’s body by the Polish government revealed a natural death from the crash (call this *e*), lowering confidence in *p*. Degree of belief in *h* can then be protected by adding the auxiliary *a**, which states: ‘the elevator controls of the plane were sabotaged by special operations executives prior to the flight’, raising confidence in *p*, given *e*. Additionally, the British intelligence keeps important documents on reports of potential sabotage classified. In this case, the evidence in favour or against *h* is mixed, and so there are both reasons to increase and drop confidence in *h*. The evidential support might even cancel out, resulting in no change in the degree of belief in *h*. In this case, we expect a rational believer to postpone deciding whether to accept or reject *h*, while continuing to update their confidence upon novel evidence. Alternatively, insofar as there are no additional reasons to reject or accept *h*, the agent might as well decide at random (for an argument that random decisions can be rational, see Icard [2021]). However, such a choice could not be deemed desperate, since, depending on its long-term track record, holding on to *h* could turn out for the agent to be a case of a glorious rescue. For example, the Watergate scandal seems to be a glorious rescue, even though uncovering it required (at least at first) protection of a conspiracy belief from attempts at disconfirmation.

At what point exactly a core belief should be dismayed or accepted is difficult to determine, but we suggest it depends on its track record of (dis-)confirmation, as opposed to exclusive coherence within the belief system. Our examples of the conspiracy beliefs surrounding Diana's death, Watergate and Sikorski's death illustrate that the distinction between `glorious' and `desperate' rescues is itself a matter of degree, with Watergate being at the far glorious end of the spectrum, and the Diana case being at the far desperate end of the spectrum.

5. Rethinking the divide between monological and higher-order belief systems

Let us now return to the monological and higher-order approaches to conspiracy beliefs outlined in section 3 and look at them through the Bayesian lens. Viewed from this perspective these two proposals can be thought of as two ways of describing the same kind of belief system.

Recall that, on the Bayesian view, both central and auxiliary beliefs mutually constrain one another and follow the same set of principles for belief revision. For instance, the status, and likelihood of rejection, of an auxiliary hypothesis directly depends on how well it is entrenched compared to its rivals. This fits with the characterization that Goertzel ([1994a])and others provide for a monological belief system, that is, a system of thought in which different beliefs are linked into a sustaining network that can absorb varying kinds of evidence. As we explained in section 3, as $\Pr(a|h)$ increases, the probability of the set of alternative auxiliaries multiplied with $\Pr(h)$ decreases (see Figure 1). For instance, the probability that Diana faked her death given that she is still alive is considerably higher than the probability that her funeral is real given that she is still alive. These auxiliaries mutually constrain each other because they cannot be both true in light of the evidence. Conflicting hypotheses that share some content, e.g. about how Diana died, 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 all hypotheses individually 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 Diana is still alive) directly raises the probability of certain auxiliary hypotheses (e.g. that Diana faked her death) to the disfavour of others (e.g. that the 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 conspiratorial beliefs which share content are interconnected and can support one another.

Let us now turn to the higher-order view (Douglas et al. [2019]) on which the consistency of simultaneously held conspiracy beliefs is supported by a higher-order belief. At first glance, it might seem that our view cannot be taken to formalise this position because there is nothing distinctly `higher-order' about central or auxiliary hypotheses. However, the 'higher-orderliness' of beliefs postulated by Douglas and colleagues is itself a kind of misnomer: such beliefs are not, in fact, located on a different level of organisation or stand in a meta-doxastic relation to other beliefs. Instead, such beliefs are characterised as being overall broader in scope than the particular, mutually inconsistent beliefs with which they cohere. As proponents of this view explain, the crux of the position is that for a conspiracy theory believer "…the specifics of a conspiracy theory do not matter as much as the fact that it is a conspiracy theory at all" (Wood et al. [2012], p. 5). Therefore, conflicting beliefs about, say, whether Diana faked her death or was assassinated can cohere with a "higher-order" belief that there is a conspiracy surrounding her car crash. Thus, the higher-order view postulates that for any set of specific and mutually inconsistent beliefs about some conspiracy, we could find a more general belief, the content of which is shared with these particular beliefs. This is remarkably similar to the account developed in the previous sections. On the Bayesian treatment, we expect that beliefs central to some conspiracy theory are 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 favour hypotheses which are more general and have a higher chance of generalising to new data (a feature which we discuss in the next section). In other words, the Bayesian view is compatible with the higher-order view because it postulates that central hypotheses (i.e. those that are better confirmed) will tend to be more general in their contents and therefore tend to be compatible with multiple competing auxiliary beliefs.

However, it is also worth stressing one way in which our view goes beyond the higher-order view.[6] The difference in question is that it also allows for broad beliefs to play the role of possible auxiliary hypotheses. Consider, once again, the conspiracy theory concerning Princess Diana. It would not be surprising to find that many supporters of this conspiracy theory harbour a kind of 'higher-order' auxiliary, call it a'', according to which the government is involved in a cover-up story. This more abstract belief might simultaneously support the belief that the public media is an unreliable evidential source ($\neg a$), and that Diana faked her death (a) and is hence still alive (h). Since there is no *formal* difference between the two kinds of beliefs (section 4), we would expect the Bayesian agents' doxastic dynamics to be such that, over time, a broad belief like a'' would become central in the belief network corresponding to the Diana conspiracy.

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 such a way as to it cast it into a concept which is by definition epistemically suspect and derogatory (Napolitano and 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 on which believing in conspiracy theories is *always* rational after all. However, we do not endorse such a view.

[6] Another departure of our view from the higher-order one is that the Bayesian view has little to say about the possible non-epistemic motivations for adopting certain beliefs. Although this issue is beyond the scope of this paper, we touch on it briefly in the conclusion.

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, as well as accounts of events for which mixed evidence is available, e.g. the possible assassination of Władysław Sikorski, and cannot be clearly ruled out as untrue. However, this does not mean that the difference between a true belief about an actual conspiracy and a false belief about an outlandish conspiracy theory 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 (see footnote 6). 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 role of inductive biases for conspiracy belief

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 conspiracy belief.

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 and Navarro [2009]). They illustrate a tendency to place the most weight in their inferences on only a very few predictions. Furthermore, within the philosophy of science, sparse beliefs are commonly considered valuable because their low initial probability makes them extra informationally relevant on acquired empirical evidence for or against them (Bar-Hillel [1955]; Popper [1954]).

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])[7]. 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 for 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 with the problem of model selection in Bayesian statistics (see Griffiths and 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 that 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 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 quite opposite directions. An optimal inference system should achieve a balance between them to avoid overfitting (i.e. making the hypotheses too precise) or underfitting (making them too general) to the data (see Forster and 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 towards 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). On the contrary, such a system would be unable to generalise towards novel cases and, so, incompatible with a strong bias towards simplicity. However, 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. the so-called "epistemic vices", see Sunstein and Vermeule [2008] and Quassam [2015]) which could predispose some people to have a propensity for forming belief networks which would be more prone to be hijacked by conspiracy theoretic narratives. Unfortunately, due to space constraints, we leave the

[7] Gershman does not explain how consistency can be measured, but Strevens assumes that some measure of the degree of confirmation is appropriate. He suggests that "*if one wants to learn as much as possible about h from the evidence* (in the sense of having the evidence impact one's probability for *h* by as much as possible), one ought to *seek out highly confirmed auxiliary hypotheses*." ([2001], p. 529, emphasis added). The desire to learn as much as possible is the bias towards sparse beliefs (i.e. those that are verifiable by sparse evidence).

exact ways in which such biases might influence inferential processes as a topic for future empirical investigation.

8. Objections

Before we conclude, we would like to briefly highlight and respond to two likely objections to the proposed view.

The first major worry is that our view equivocates scientific and cognitive reasoning. One might think that this is a mistake because, for example, it is sometimes argued (e.g. Rini [forthcoming]) that conspiracy theories and scientific theories differ in how they relate to the concept of truth. While truth seems to be an optional disposable constraint on conspiracy reasoning, it is often taken as a regulative ideal in scientific reasoning. Similarly, some authors working on the topic of conspiracy theories (see introduction) may want to question whether the best way to understand conspiracy theories is to liken them to *bona fide* scientific explanations.

Regarding the adequacy of viewing conspiracy theories as explanations, we would like to reiterate that our view is only concerned with beliefs about conspiracies and not the scientific status of conspiracy theories *per se*. Viewing beliefs as a kind of explanation of the available observations is nothing new within the philosophy of mind and cognitive psychology, or the academic discourse on rationality more broadly, hence we do not think that this is a significant problem for our account.

More broadly, we push back on the accusation of equivocation by pointing out that our view relies on *analogy* rather than *identity* between cognitive and scientific reasoning. While Bayesian cognitive science often relies on the assumed continuity between a cognitive view of science and a scientific view on cognition (Gopnik and Meltzoff [1997]; Tenenbaum et al. [2006], [2011]), researchers working within this framework are primarily concerned with exploiting the (dis)similarities between different classes of phenomena, like changes in scientific development and changes in childhood development. Similarly, we do not want to claim that the two kinds of reasoning are the same thing, even if we do believe that the Bayesian normative standard is equally applicable in both cases.

This brings us to the second worry - the normative status of the account on offer. Some readers may point out that, considering our commitment to the Bayesian methodology, it is unclear whether the view developed here should be understood in normative or descriptive terms: are we aiming to provide a normative account about how agents *should* update, revise, or otherwise modify their beliefs about conspiracies in response to the available evidence, or are we aiming to provide an empirically adequate description of these processes?

This is a legitimate worry which grows out of our attempt to avoid an overly strong commitment to either wing of the Bayesian program. On the one hand, we do operationalize rationality using the Bayesian standard, but do not overly commit to the claim that Bayesianism is the only variable normative theory of rationality. On the other hand, our account provides a useful platform for comparing different descriptive accounts of conspiracy theoretic belief, even though it is idealised and might lack details necessary for empirical testing. Perhaps the best way to understand our work is to view it as contributing a *how-possibly* perspective on how conspiracy belief relates to disconfirmatory evidence. In the future, we plan to draw on work in cognitive science on suboptimal reasoning and biases (e.g. by equipping models with overly deterministic priors) to operationalize our account for empirical testing.

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. However, the Bayesian approach departs from the previous proposals in 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.

Brotherton, R., and French, C.C. (2014). Belief in conspiracy theories and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, 28(2), 238–248. https://doi.org/10.1002/acp.2995

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. (2015). Bad Thinkers. *Aeon*, http://aeon.co/magazine/philosophy/intellectual-character-of-conspiracy-theorists/

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.

Denith, M. R. (2016). When Inferring to a Conspiracy might be the Best Explanation. *Social Epistemology*, 30, 5-6: 572-591.

Dentith, M. R. (2014). *The Philosophy of Conspiracy Theories*, Basingstoke: Palgrave Macmillan.

Dentith, M. R. (2019). Conspiracy theories on the basis of the evidence. *Synthese*, *196*(6), 2243-2261.

Douglas, K. M, Sutton, R. M, and Cichocka A. (2017). The Psychology of Conspiracy Theories. *Current Directions in Psychological Science*, 26(6): 538-542. https://doi.org/10.1177%2F0963721417718261

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.

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

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.

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.

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.

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.

Harding, S. (1976). *Can theories be refuted? Essays on the Duhem-Quine thesis.* Springer Science and Business Media.

Icard, T. (2021). Why be random? Mind, 130(517):111–139. https://doi.org/10.1093/mind/fzz065

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.

Lackey, J. (2021). Echo Chambers, Fake News, and Social Epistemology. In *The Epistemology of Fake News* (pp. 206-227). Oxford University Press.

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

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive science*, *14*(1), 57-77.

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.

Pearl, J. (1988). Embracing causality in default reasoning, *Artificial Intelligence*. 35:259-271.

Popper, K. R. (1954). Degree of confirmation. *The British Journal for the Philosophy of Science*, 5(18), 143-149.

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.

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): e0233879. https://doi.org/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 Douglas, K. M. (2017). Conspiracy theories as part of history: The role of societal crisis situations. *Mem Stud*, 10, 323–333.

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