# Coherence as a Constraint on Scientific Inquiry

Forthcoming in *Synthese*

Borut Trpin
LMU Munich
Munich, Germany;
University of Maribor
Maribor, Slovenia;
University of Ljubljana
Ljubljana, Slovenia
borut.trpin@lrz.uni-muenchen.de

Martin Justin
University of Maribor
Maribor, Slovenia
martin.justin1@um.si

**Abstract**

We investigate the epistemic role of coherence in scientific reasoning, focusing on its use as a heuristic for filtering evidence. Using a novel computational model based on Bayesian networks, we simulate agents who update their beliefs under varying levels of noise and bias. Some agents treat reductions in coherence as higher-order evidence and interpret such drops as signals that something has gone epistemically awry, even when the source of error is unclear. Our results show that this strategy can improve belief accuracy in noisy environments but tends to mislead when evidence is systematically biased. We explore the implications for the rationality of coherence-based reasoning in science.

**Keywords:** coherence, higher-order evidence, belief updating, Bayesian networks, scientific reasoning

## Declarations

- **Availability of data and material:** The simulations were implemented in Python using the `bnlearn` and `Mesa` libraries. The source code is available at `https://github.com/Martin-Justin/CohABM/`.
- **Competing interests:** None.

- **Funding:** We acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0144 and project No. J6-60107).

- **Authors' contributions:** Both authors contributed equally.

- **Acknowledgements:** Thanks to the audiences at "East European Network for Philosophy of Science Conference" (Kraków, Poland, 2024), "Methodological Trends and Challenges in Contemporary Philosophy" (University of Ljubljana, 2024) and "Formal Models of Scientific Inquiry" (Ruhr University Bochum, 2025), for their helpful comments on earlier versions of this paper.

# 1 Introduction

Probabilistic methods are central to contemporary science and its philosophical understanding. Scientists use probabilistic tools to model uncertain phenomena, manage noisy data, and assess evidential support. Philosophers, in turn, have drawn on Bayesian frameworks to reconstruct learning processes, formalize scientific reasoning, and articulate normative constraints on belief change. Yet, despite their successes, such approaches continue to face significant challenges, particularly when it comes to reasoning under uncertainty, bias, or incomplete information.

A growing body of work explores how agents, human or artificial, might cope with such challenges using cognitive heuristics. One such strategy involves coherence-based filtering, the tendency to discount evidence that would disrupt the internal coherence of one's beliefs. This phenomenon is well-documented in cognitive science and psychology, where it is associated with confirmation bias (Festinger et al., 1956), myside bias (Stanovich et al., 2013; Baccini et al., 2023), and motivated reasoning (Mandelbaum, 2019). But coherence also has a long-standing philosophical pedigree as a criterion of epistemic justification (BonJour, 1985) and as a feature of explanatory reasoning (Thagard, 1989, 2002).

While earlier work challenged the justificatory role of coherence by highlighting its limited connection to truth (Bovens and Hartmann, 2003; Olsson, 2005), later contributions have argued that coherence may still have epistemic value, either as a heuristic in conditions of partial knowledge (Angere, 2008) or as a defeater in social epistemic contexts (Goldberg and Khalifa, 2022). Our aim in this paper is to advance this line of inquiry by examining whether and when reductions in coherence provide higher-order reasons to reject some evidence. In doing so, our approach also connects to Thagard's pioneering ECHO model of explanatory coherence (Thagard, 1989), though, as we discuss later, it differs in important ways.

Although discussions of higher-order evidence in epistemology, such as those arising from disagreement with epistemic peers, typically focus on indicators of an agent's unreliability, including memory, perception, or reasoning errors (for recent reviews, see Dorst 2024; Horowitz 2022; Ye 2022), coherence considerations may also play a higher-order role. In scientific contexts, researchers sometimes treat sharp drops in coherence as signals that something has gone wrong, even if they cannot immediately identify a specific error, and the issue may not stem from their unreliability. In this way, coherence can function as a form of higher-order evidence not necessarily tied to the agent but may concern the evidence itself. Yet, as we will show, this strategy can both help and hinder inquiry.

To illustrate this point, we begin in Section 2 with two contrasting historical motivating examples. In the first, the physics community's rejection of superluminal neutrino measurements (despite their statistical strength) was arguably warranted due to the incoherence of the result with established theory. In the second, decades-long overreliance on Millikan's original measurements of the electron's charge—despite better data emerging later—illustrates the downside of excessive coherence-based conservatism. These cases raise the broader question: when can coherence-based evidence filtering lead to more accurate beliefs? To address this, we develop a computational model of belief updating under uncertainty. Drawing on Bayesian networks, we simulate agents who repeatedly revise their beliefs in response to new evidence. Some agents update on all evidence, even if the evidence goes against their expectations; others apply a coherence-based filter, rejecting updates that would reduce the coherence of their current belief set.

Our results show that coherence-based filtering can be epistemically beneficial, but only under specific circumstances. In highly noisy environments, where a substantial amount of evidence is erroneous, it helps agents resist misleading evidence and maintain accurate beliefs. But in low-noise settings, or when evidence is systematically biased, coherence filtering consistently leads agents away from the truth. These findings help explain why coherence-based reasoning may function as a productive heuristic in some domains (e.g., mature sciences) while becoming an epistemic liability in others (e.g., biased or politicized research contexts).

By examining how coherence interacts with probabilistic belief updating, our study contributes to ongoing work on probabilistic reasoning in the sciences. It aligns with research on the normative role of heuristics in scientific inference, the epistemology of bias and misinformation, and the limits of Bayesian rationality under real-world constraints. More broadly, it shows how formal tools can be used to model cognitive strategies and to evaluate their reliability across epistemic contexts.

We proceed as follows: Section 2 presents our motivating examples. Section 3 introduces the formal coherence measures and our modeling framework. Section 4 describes the results of our simulations. Section 5 explores our findings' normative and philosophical implications. Section 6 concludes.

# 2 Motivating Examples

To motivate our analysis, we begin with two historical episodes illustrating how coherent considerations influence scientific reasoning. In both cases, the episodes may plausibly be reconstructed as instances in which the decreased coherence of a new result with existing theory shaped how the scientific community responded. However, while coherence-based skepticism proved beneficial in the first case, it arguably hindered progress in the second. These examples suggest that coherence can serve as a heuristic for managing uncertainty but that its epistemic value is highly context-dependent.

The first example involves the OPERA experiment's 2011 claim that neutrinos had been observed traveling faster than the speed of light. Based on measurements between CERN and the Gran Sasso Laboratory in Italy, the result reported a statistical significance well above the standard discovery threshold (Adam et al., 2011; Brumfiel, 2011). If correct, it would have overturned one of the central tenets of modern physics, namely, the invariance of the speed of light, as codified in Einstein's theory of relativity. Despite the apparent strength of the statistical evidence, the physics community responded with widespread skepticism. Many researchers suspected a methodological error precisely because the result was deeply at odds with well-established theoretical commitments. This skepticism proved well founded: subsequent investigations revealed that the anomaly was due to a faulty fiber-optic cable connection. Once corrected, the measurements aligned with relativistic expectations (Cartlidge, 2012).

In this case, coherence-based reasoning played a clearly beneficial epistemic role. Faced with a surprising and disruptive result, scientists did not accept the evidence at face value. Instead, they treated its incoherence with accepted theory as a defeasible reason to question its reliability. Their resistance to belief revision was not irrational conservatism but a reasonable response to the possibility of experimental error. Here, coherence considerations functioned as a form of higher-order evidence, which prompted deeper scrutiny that ultimately revealed the true source of the anomaly.

The second example illustrates a more problematic side of coherence-based evidence filtering. In the early twentieth century, measurements of the electron's elementary charge took several decades to converge on the correct value, partly due to the authoritative influence of Robert Millikan's 1913 oil drop experiment (Millikan, 1913). While Millikan's original measurements were based on careful experimentation, they contained an incorrect value for the viscosity of air (Feynman, 1985, Cargo Cult Science). As a result, his calculated value for the elementary charge was slightly off. Later experiments often produced more accurate measurements that diverged from Millikan's findings, but these deviations were downplayed or dismissed as outliers. The scientific community's strong preference for coherence with Millikan's authoritative result fostered a kind of implicit deference, significantly delaying the measurement's correction.

Contrary to the superluminal neutrinos episode, this case exemplifies the negative side of coherence-based evidence filtering: non-misleading evidence that conflicted with established consensus and prior expectations was unintentionally discounted. In contrast to the OPERA case, coherence considerations here led to epistemic inertia rather than productive scrutiny. Rather than helping to detect misleading initial results, the desire for coherence insulated an inaccurate belief from revision.

These two episodes highlight the double-edged nature of coherence-based reasoning in science. In the neutrino case, a drop in coherence correctly signaled an underlying flaw in the evidence and protected inquiry from being misled. In the electron charge case, a decreased coherence with an authoritative but flawed result led to misplaced confidence and slower progress. These contrasts raise a broader question: under what conditions does coherence-based filtering of evidence support more accurate belief formation, and when does it become a source of bias or error? This question motivates the formal model we develop in the next section.

# 3 A Simulational Study

## 3.1 Formal Measures of Coherence

To explore the epistemic role of coherence, we first require a clear conceptualization of coherence itself. Intuitively, epistemic coherence captures how well the propositions within an information set "hang together" (BonJour, 1985). Consider, for instance, the difference between these two information sets (from BonJour 1985, p. 96):

$\mathbf{S_1} = \{$[All ravens are black], [This bird is a raven], [This bird is black]$\}$, and

$\mathbf{S_2} = \{$[This chair is brown], [Electrons are negatively charged], [Today is Thursday]$\}$.

Set $\mathbf{S_1}$ is intuitively much more coherent than $\mathbf{S_2}$ because the propositions in $\mathbf{S_1}$ support each other, while those in $\mathbf{S_2}$ lack meaningful connections. Clarifying and formalizing this intuitive notion has generated substantial philosophical debate (e.g., Lewis 1946; Rescher 1973; BonJour 1985; Thagard 1989; Lehrer 2000). Within Bayesian epistemology, coherence is typically formalized through quantitative measures that map probability distributions over propositional variables onto numerical coherence scores (see Olsson 2022 for a comprehensive recent review). Because Bayesian measures explicitly accommodate uncertainty and are readily operationalizable in computational models, we adopt this Bayesian approach in our analysis. That is, Bayesian coherence measures formally quantify how strongly propositions within an information set "hang together," enabling rigorous exploration of coherence's role in belief updating.

Two central intuitions frequently guide coherence measures (see, e.g., Schippers 2014). The first is the deviation from independence: propositions in a coherent set are probabilistically dependent, and the degree of dependence (in whichever specific way this is formalized) may be used as a proxy for the degree of an information set's coherence. For example, the propositions in $\mathbf{S_1}$ above clearly depend on each other. Following Shogenji (1999) who formalizes this intuition, we can then measure the exact degree of coherence in the following way:

$$\mathit{coh}_S(\mathbf{S}) := \frac{P(A_1, \ldots, A_n)}{P(A_1) \times \cdots \times P(A_n)} \tag{1}$$

where $\mathbf{S} = \{A_1, \ldots, A_n\}$ represents a set of propositions. This measure captures the former intuition of how much propositions deviate from probabilistic independence (the threshold value at which the set is neither coherent nor incoherent is 1).

The second intuition is that coherence measures the relative probabilistic overlap among propositions. Propositions are coherent if they tend either to be jointly true or jointly false, meaning their joint probability makes up a large proportion of their union's probability (Olsson, 2002; Glass, 2002). For instance, consider the following highly coherent set:

$\mathbf{S_3} = \{$[The restaurant is crowded], [The restaurant is noisy]$\}$.

This set is intuitively coherent since these propositions tend strongly either to both hold true or false: more or less, all crowded restaurants are usually noisy and vice-versa. Consequently, their joint probability is about as high as their union's probability.

By contrast, consider a much less coherent set:

$\mathbf{S_4}$ = [The restaurant is crowded], [The restaurant has many free tables].

These propositions tend systematically to disagree: a crowded restaurant rarely has many free tables, and a restaurant with many free tables is rarely crowded. Thus, the propositions do not overlap much, making their joint probability very low relative to their union's probability, and hence the set is intuitively much less coherent. Olsson (2002) and Glass (2002) formalize this in the following measure:

$$coh_{OG}(\mathbf{S}) := \frac{P(A_1, \ldots, A_n)}{P(A_1 \vee \cdots \vee A_n)} = \frac{P(A_1, \ldots, A_n)}{1 - P(\neg A_1, \ldots, \neg A_n)}. \qquad (2)$$

A recent measure proposed by Hartmann and Trpin (forthcoming) combines both intuitions:

$$coh_{HT}(\mathbf{S}) := \frac{P(A_1, \ldots, A_n)}{1 - P(\neg A_1, \ldots, \neg A_n)} \Big/ \frac{P(A_1) \times \cdots \times P(A_n)}{1 - P(\neg A_1) \times \cdots \times P(\neg A_n)} \qquad (3)$$

This hybrid measure compares the actual relative overlap to the relative overlap there would be if the propositions were probabilistically independent and their marginal probabilities fixed. We include this measure in our simulations because it has been shown to be a reliable truth-tracker (Hartmann and Trpin, forthcoming).

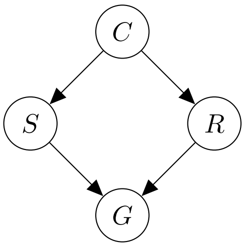
There are, of course, also many other coherence measures available in the literature, several of which rely on averaging coherence scores across subsets of propositions (e.g., Fitelson 2003; Meijs 2006; Douven and Meijs 2007; Schupbach 2011; Koscholke et al. 2019). These subset-based measures, however, are computationally demanding for larger information sets. All these other measures are also excluded from our analysis because our simulations indicate that even the three relatively simple measures presented here, despite their differences, yield highly consistent results. The main difference in the compared updating strategies is, therefore, plausibly not in the specific measures of coherence but rather in the use of coherence-based evidence filtering vs. not using such a filter.

## 3.2 The Model

Based on the insights from the two motivating examples presented above and previous work on the formal measures of coherence, we developed a computational simulation to further investigate the role of coherence in scientific reasoning. In the proposed simulation, agents try to form an accurate picture of the ground truth (the world) by gathering information about it. The simulation is roughly inspired by two existing frameworks, the Bayesian Normative Argument Exchange across Networks (NormAN) modeling framework (Assaad et al., 2023) and the bandit modeling framework (Zollman, 2007, 2010). From NormAN, we take the idea of representing the world using a Bayesian network. On the other hand, the evidence-gathering process in the simulation resembles playing a single-armed bandit, where information takes the form of a set of stochastically generated states of the world. In this section, we will first present the fundamental entities and dynamics of the simulation and then explain how we extended it to explore the role of coherence considerations.

The world in the simulation consists of a set of probabilistically related events. It is represented via a Bayesian network (BN), consisting of a directed acyclic graph (DAG) and a corresponding conditional probability distribution over a set of binary propositional variables from the BN nodes (see Pearl 1988 for a general introduction to Bayesian networks theory, and Hartmann 2021 for an introduction to their philosophical applications). Nodes in the DAG represent the events in the world, which may be true or false, while edges represent probabilistic dependencies between them. The conditional probability distribution (CPD) then contains information about the likelihood of individual events given different values of related events.

Agents in the simulation already have an accurate representation of the events in the world and their probabilistic relations—in other words, they are aware of the structure of the Bayesian network (BN) in



| Probability | Value |
|---|---|
| $P(\mathrm{C})$ | 0.5 |
| $P(\mathrm{S}\vert\mathrm{C})$ | 0.1 |
| $P(\mathrm{S}\vert\neg\mathrm{C})$ | 0.5 |
| $P(\mathrm{R}\vert\mathrm{C})$ | 0.8 |
| $P(\mathrm{R}\vert\neg\mathrm{C})$ | 0.2 |
| $P(\mathrm{G}\vert\mathrm{S},\mathrm{R})$ | 0.99 |
| $P(\mathrm{G}\vert\mathrm{S},\neg\mathrm{R})$ | 0.9 |
| $P(\mathrm{G}\vert\neg\mathrm{S},\mathrm{R})$ | 0.9 |
| $P(\mathrm{G}\vert\neg\mathrm{S},\neg\mathrm{R})$ | 0 |

Figure 1: The "sprinkler" network, where $C, S, R, G$ are propositional variables with corresponding values C: "It is cloudy", ¬C: "It is not cloudy", S: "The sprinkler is turned on", ¬S: "The sprinkler is not turned on", R: "It rains", ¬R: "It does not rain", G: "The grass is wet" and ¬G: "The grass is not wet," and the corresponding probabilities of its CPD.

question, though not its exact probability distribution. To illustrate how such a structure might look and to clarify the elements of a Bayesian network, we employ a simple and widely used textbook example, the so-called "sprinkler" network.[1]

The sprinkler network (see Figure 1) describes a simplified scenario involving weather conditions and wet grass, where each node represents a binary event: whether it is cloudy ($C$), whether it rains ($R$), whether the sprinkler is turned on ($S$), and whether the grass is wet ($G$).[2] The directed edges between nodes indicate probabilistic dependencies: the likelihood of rain or sprinkler use depends on whether it is cloudy, and the likelihood of wet grass depends on whether the sprinkler is activated or it is raining.

We chose this example because it is intuitive, clearly illustrates the basic properties of Bayesian networks (e.g., conditional independence and dependency structures), and is extensively used in the literature on Bayesian modeling. Throughout our simulation, agents know precisely these dependency structures but are initially uncertain about the exact conditional probabilities linking these events.

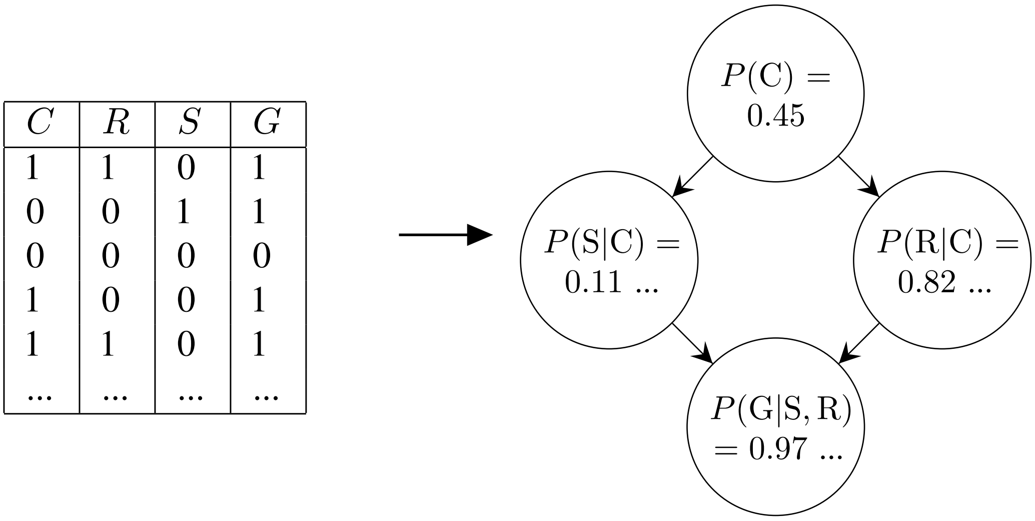
In the course of the simulation, agents gradually learn about the probabilities of the events in the world by observing it many times. More specifically, we model the learning of the agents by having the agents sample from a CPD that is associated with the world and then fitting these observations to the model via maximum likelihood estimation (MLE).[3] For example, one sample the agents might gather is $S_1$ = [Cloudy=True, Sprinkler=False, Rain=True, Wet Grass=True], another is $S_2$ = [Cloudy=False, Sprinkler=True, Rain=False, Wet Grass=True], and so on. In every round of the simulation, agents gather a number of such observations. Using this information, they then form a new belief about the distribution by fitting the observations to the BN. Effectively, this means that the agents develop an updated subjective CPD. Figure 2 illustrates how this may work in practice. Note that the agent's CPD will tend to deviate from the CPD of the true BN.

As we were interested in the effect of the agents' prior beliefs on their accuracy, we set their prior CPDs at the start of the simulation before they gathered any evidence from the world. The prior CPDs are generated by randomly changing the parameters of the world's CPD within some interval. For example, if in the true distribution $P(\mathrm{R}|\mathrm{C}) = 0.8$, agents might start with a distribution where $P(\mathrm{R}|\mathrm{C})$ is sampled from a uniform distribution between $[0.8 - i,\ 0.8 + i]$, where $i \geq 0,\ 0.8 - i \geq 0,\ 0.8 + i \leq 1$, and $i$ is a parameter of the model, which can be determined by the modelers. This kind of change is applied to

[1] This BN was, to the best of our knowledge, first introduced in the canonical form in Russell and Norvig (1995), although the general set-up is also discussed in Pearl (1988).

[2] We follow the convention of listing propositional variables in italics and the instantiations in roman script.

[3] Alternative updating rules could be used here. We ran the simulations using a Bayesian Parameter Estimator for a subset of parameters explored in Section 4 with results remaining consistent.



| $C$ | $R$ | $S$ | $G$ |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 |
| ... | ... | ... | ... |

$P(\mathrm{C}) =$ 0.45
$P(\mathrm{S}|\mathrm{C}) =$ 0.11 ...
$P(\mathrm{R}|\mathrm{C}) =$ 0.82 ...
$P(\mathrm{G}|\mathrm{S},\mathrm{R})$ $=$ 0.97 ...

Figure 2: An agent estimates the CPD (right) from sampled observations (1 stands for the variable being true, 0 for it being false; left) via MLE.

the whole distribution. To make sure that agents' priors affect their later belief updates, we then generate some number of samples from this modified prior CPD at the start of the simulation and add them to the evidence agents receive from the world. Controlling agents' priors in this way allows us to model different scientific contexts, from mature science with a broad consensus about the domain of study to nascent lines of research, where research supports a wide variety of hypotheses and theories. In the former case, the priors may be close to the truth (the value of $i$ is close to 0), while in the latter, they could be well off (higher values of $i$).

The process of gathering evidence and updating beliefs about the probabilities of the events in the world is then repeated for multiple steps. At the end of the simulation (the number of steps is determined at the onset as a parameter of the model), we evaluate how closely agents' beliefs approximate the actual conditional probability distribution of the ground truth using the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence measures the discrepancy between two probability distributions, quantifying the information loss when approximating the true distribution. It is particularly useful here as it captures how well agents' beliefs reflect the correct probability distribution of the ground world.

To explore how coherence affects scientific inquiry, we also implemented an alternative learning process. Whereas the default process described above involves updating beliefs solely by fitting evidence to the known network structure, this alternative process adds a step that explicitly considers how newly gathered information impacts the coherence of an agent's beliefs. Specifically, in this step, agents first determine the most probable state of the world based on the learned distribution. In the running example of the "sprinkler" BN, this may be that it is cloudy, the sprinkler is off, it is raining, and the grass is wet. Then, they check how coherent this state is using one of the coherence measures presented above. Agents accept the updated belief if it is more coherent than the state that was the most probable according to their prior belief. If not, they reject the evidence and the update, effectively remaining in their prior state.

Formally, the agent calculates the coherence of the most probable joint state under their current probability distribution $P_0$. Suppose this state is $\{C{=}1, R{=}1, S{=}0, G{=}1\}$. The agent computes $coh^{P_0}(\{C{=}1, R{=}1, S{=}0, G{=}1\})$ using one of the presented coherence measures and their current probability distribution $P_0$, prior to incorporating new observations. After fitting the new data to the network, the agent obtains a revised conditional probability distribution, yielding a posterior distribution $P_1$.

They then determine the most probable joint state under $P_1$ and calculate its coherence, e.g., if the same joint state remained most probable, they would calculate $coh^{P_1}\{C{=}1, R{=}1, S{=}0, G{=}1\}$. If this posterior coherence is at least as high as the prior one, i.e., if $coh^{P_1}(\{C{=}1, R{=}1, S{=}0, G{=}1\}) \geq coh^{P_0}(\{C{=}1, R{=}1, S{=}0, G{=}1\})$, then the agent accepts the updated distribution $P_1$. Otherwise, they reject it and retain their prior distribution $P_0$.

Agents using this alternative learning process treat coherence as a higher-order reason to dismiss new evidence when it would reduce the perceived coherence of their beliefs. This mirrors the reasoning observed in the two motivating examples discussed above. However, rejecting relevant evidence simply because it disrupts coherence conflicts with the Principle of Total Evidence, the idea convincingly defended by Carnap (1947) and later by Good (1967) that all available evidence should be taken into account when estimating the probability. This makes coherence-based filtering a questionable strategy in general. To assess whether and when it might nonetheless improve epistemic outcomes, we simulate environments in which the incoming evidence may be misleading, reflecting the kinds of distortions and noise that often arise in real-world scientific inquiry.

We extended the model in one additional way to capture this possibility. Specifically, agents in the model can receive two types of erroneous or misleading evidence. In one case, we randomly changed some percent of the evidence points agents collected. For example, let's say the current world state the agents observe is $S_1$ = [Cloudy=True, Sprinkler=False, Rain=True, Wet Grass=True]. In an extremely noisy environment, in which agents would be misled about 50 $\%$ of their evidence, they wouldn't observe $S_1$ but a modified set where (on average) two out of four values would be changed (for True to False or the other way around). The percentage of evidence that changes—the level of "noise" in the environment—is determined as a model parameter. We call this type of misleading evidence "noisy evidence". It represents a possible deviation of the measured value from the actual one that is not predictably biased in any direction (i.e., random measurement error).

In the other case, agents receive evidence that is systematically misleading. Specifically, they have some probability of drawing samples not from the ground truth but from an alternative Bayesian network that is biased. For example, suppose a sprinkler factory is trying to downplay the role of rain in making the grass wet. Then where in the ground truth of the model $P(\mathrm{G}|\neg\mathrm{S},\ \mathrm{R}) = 0.9$, in the alternative, misleading BN, this probability may be changed to $P(\mathrm{G}|\neg\mathrm{S},\ \mathrm{R}) = 0.4$; that is, if it rains but the sprinkler is off, the grass is much less likely to be wet than in the true case. We call this "systematically misleading evidence".

In contrast to noisy evidence, misleading evidence presents a picture of the world that systematically deviates from the truth. There are different possible sources of such bias in scientific inquiry. A well-known example concerns the so-called publication bias (Easterbrook et al., 1991). Scientific journals strongly prefer to publish positive results. This means that a lot of negative results never get reported, which creates a biased scientific record. An alternative important source of bias is in industry funding, management, or promotion of scientific research (Holman and Elliott, 2018). For example, industry-funded studies in clinical drugs and medical devices research have a higher chance of reporting positive efficacy results than non-industry studies (Lundh et al., 2017). While drawing samples from an alternative Bayesian Net with a different probability distribution is an idealization of such scenarios, it can be understood as representing a scientist conducting, for example, an industry-biased study.

## 3.3 Simulation Setup and Procedure

To investigate the epistemic consequences of coherence-based reasoning, we constructed a simulation that models agents attempting to arrive at accurate beliefs about an underlying world. The simulation formalizes a series of epistemic choices—concerning how agents treat incoming information, how they weigh coherence, and how they respond to uncertainty—and embeds them in a dynamic learning environment.

This allows us to examine whether and under what conditions coherence-based filtering serves as an epistemically productive strategy.

**Epistemic Environment.** The world is represented as a Bayesian network whose structure is known to the agents but whose underlying CPD is not. Agents begin with an initial (prior) CPD, which may deviate more or less from the truth, depending on a model parameter. In each round, they gather evidence by observing sampled world states and use this to update their beliefs. The degree to which the evidence is accurate, noisy, or biased is also under experimental control. We report results using two textbook networks of increasing complexity: the "Sprinkler" (4 binary nodes) and "Asia" (8 binary nodes) networks, although any Bayesian may be used.[4]

**Epistemic Strategies.** We model two types of agents. Normal agents update on all available evidence. Coherence agents, by contrast, treat coherence as a form of higher-order evidence: if an update reduces the coherence of their current beliefs, they reject it. In this way, coherence functions as a defeater, overriding first-order evidence when the belief system as a whole becomes less integrated.

**Simulation Dynamics.** Each round of the simulation unfolds as follows (see also Figure 3 for a flowchart):

1. Agents receive a batch of evidence in the form of sampled world states.
2. They propose an updated belief distribution via maximum likelihood estimation.
   (a) Normal agents adopt the updated distribution.
   (b) Coherence agents compare the coherence of the most probable joint state under the new distribution to the previous one and accept the update only if coherence does not decrease.
3. Belief accuracy is evaluated from the modeller's perspective as the Kullback–Leibler divergence between the agent's current probability distribution and the true distribution, which is not available to the agents.

**Iteration.** This process is repeated for $N$ rounds. Each round represents an inquiry cycle in which agents encounter new evidence, attempt to integrate it, and in which their epistemic position is evaluated. No additional stopping conditions are imposed.

This simulation framework offers a controlled environment for exploring how coherence considerations interact with belief revision. By formalizing coherence-based reasoning in probabilistic terms and embedding it in an iterated learning process, we are able to assess not only whether coherence can function as a useful heuristic but also when its use is epistemically appropriate—or counterproductive.

## 4 Results

Having established the model and its rationale, we now turn to the outcomes of our simulations. We aim to assess whether agents who employ coherence as a defeater perform better or worse than those who do not across a variety of epistemic environments. These environments are defined by the accuracy of the agents' priors, the reliability of the evidence they receive, and the presence or absence of systematic noise or distortion.

[4] Interested readers are invited to experiment with other textbook or custom Bayesian networks, keeping in mind that larger networks are computationally rather demanding. The simulations were implemented in Python using the `bnlearn` and `Mesa` libraries. The source code is available at `https://github.com/Martin-Justin/CohABM/`.

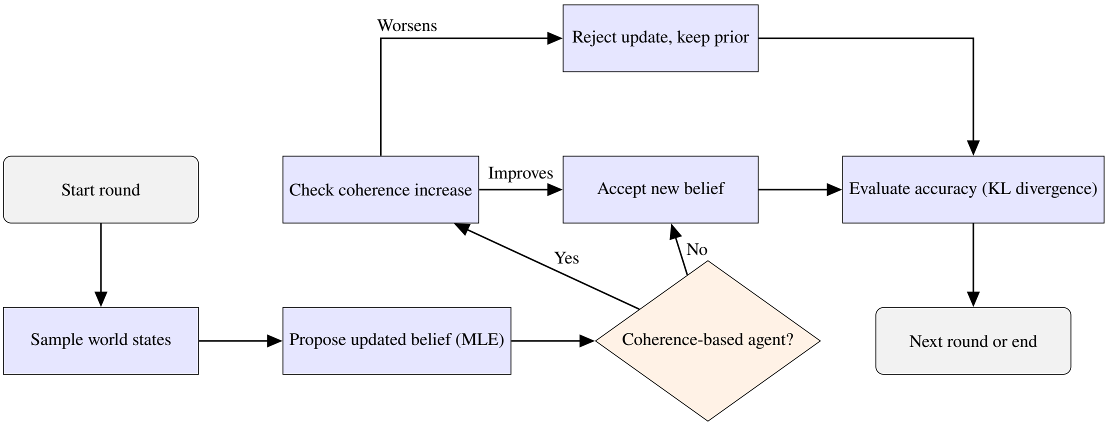


Figure 3: Flowchart of simulation dynamics for coherence and normal agents.

We compare the performance of two types of agents across a range of parameter settings (see Table 1). For each simulation run, agents attempt to learn the underlying probability distribution of the world over $N = 50$ rounds. In each round, agents observe 100 samples and update their beliefs accordingly. Each combination of parameters was simulated 30 times to ensure robust averages. Performance is measured by the Kullback-Leibler (KL) divergence between an agent's belief distribution and the true distribution at the final round.

| Parameter | Values |
|---|---|
| Bayesian Net | Sprinkler, Asia |
| Coherence Measures | Shogenji, Olsson–Glass, Hartmann–Trpin |
| Type of Misleading Evidence | noisy evidence, systematically misleading evidence |
| Information Noise | 0.05 to 0.3 (in 0.05 increments) |
| Variation of Prior CPD Values | 0.05 to 0.4 (in 0.05 increments) |
| Sample size | 20, 40, 60, 100 |

Table 1: Parameter values used in the reported simulations. Robustness checks with a wider variety of parameters are presented in the Appendix.

## 4.1 Coherence Under Noise

We begin with the case in which agents receive evidence corrupted by random noise. In each round, a certain proportion of observed values is randomly flipped without systematic bias. This scenario models epistemic environments where error is frequent but directionless—due, for instance, to faulty instrumentation, poor measurement conditions, or human error. We refer to this as "noisy evidence."

Figure 4 reports the results from the "Sprinkler" Bayesian network under varying levels of noise and prior accuracy. Each cell in the heatmap displays the average difference in accuracy between a Coherence agent and a Normal agent after 50 rounds. Positive values indicate cases where the Coherence agent achieved lower KL divergence (i.e., more accurate beliefs).

As the figure shows, coherence-based evidence filtering improves accuracy when the level of noise is high (particularly above 25%) and when prior beliefs are already fairly well-calibrated. In these settings, the coherence filter serves as a protective buffer, allowing agents to insulate their belief system from misleading evidence that would otherwise degrade it. However, this same conservatism becomes a liability

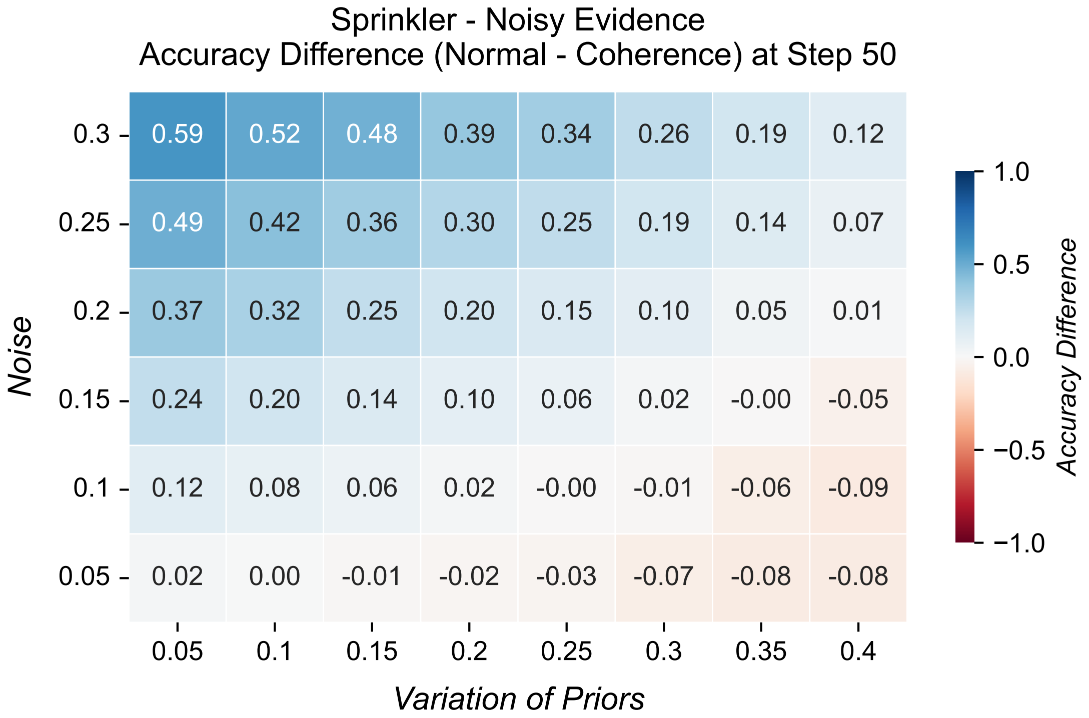


Figure 4: Difference in accuracy (Normal minus Coherence) at round 50 in the "Sprinkler" network with noisy evidence. Positive values favor coherence-based reasoning (KL divergence from truth is, in a sense, a penalty).

in low-noise contexts. When the environment is reliable, rejecting evidence that would lower coherence simply slows the process of convergence toward truth, especially when agents begin with inaccurate priors.

This dynamic is further illustrated in Figure 5, which tracks the evolution of belief accuracy across simulation rounds in two contrasting cases. On the top chart, we see how coherence filtering helps when the environment is noisy: the Coherence agent ultimately avoids being misled. On the bottom, we observe the opposite: in a low-noise setting, coherence filtering acts as an impediment to belief revision, leaving the agent epistemically inert.

To check the robustness of these results, we tested the same combination of evidence noise and variation of priors across a range of different conditions. Specifically, we ran simulations for up to 500 rounds, with a significantly larger "Asia" Bayesian network (Lauritzen and Spiegelhalter, 1988) and with agents receiving less evidence every round. The results proved to be consistent—we invite readers to consult the Appendix for details.

## 4.2 Coherence Under Systematic Misleading Evidence

Next, we compared Coherence and Normal agents in an environment with systematically biased evidence. Before we present the results, we should note how we generated such evidence. As we outlined in Section 3.2, agents receive systematically misleading evidence by having some chance of collecting samples from an alternative Bayesian network. This Bayesian Net has a conditional probability distribution that deviates from the ground truth. We constructed this alternative BN manually by changing certain

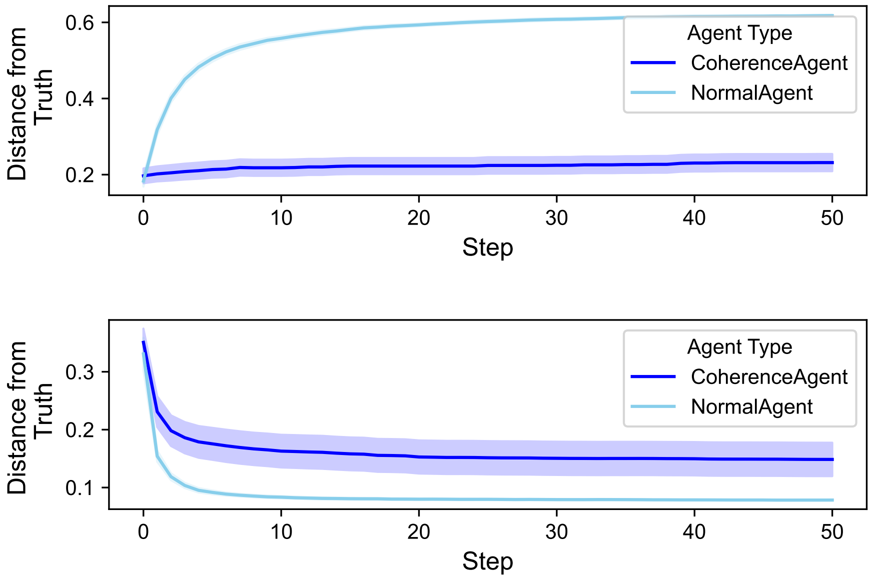


Figure 5: KL divergence from the true distribution over time. Top: high noise (30%), moderately accurate priors (20%). Bottom: low noise (5%), inaccurate priors (30%). The Coherence agent is in blue, and the Normal agent is in orange. The shaded region around the lines represents the 95% confidence interval.

conditional probabilities in the original "Sprinkler" and "Asia" BNs. To control for the variables that might affect the results, we made sure that the modified BNs satisfied two conditions. First, the same joint state should remain the most probable one as in the original network. For example, if $\{C{=}1, R{=}1, S{=}0, G{=}1\}$ is the most probable joint state in "Sprinkler", this should also be the case in the modified network. Second, the most probable state should be less coherent than in the original network. If opposite was the case, one might worry that the effects of misleading evidence on Coherence agents is not due to its systematic nature but due to its being more coherent.[5]

Figure 6 reports the results for systematically misleading evidence in "Sprinkler" Bayesian network under varying levels of prior accuracy and probabilities for agents who received misleading evidence (referred to in the chart as "Noise"). As above, each cell displays the average difference in accuracy between a Coherence agent and a Normal agent after 50 rounds, with positive values indicating cases where the Coherence agent achieved more accurate beliefs.

These results present a stark contrast to those from noisy environments (reported in Figure 4). Here, coherence-based evidence filtering almost indiscriminately hurts agents' inquiry. This suggests that a qualitatively different mechanism is at play. To see why, consider what role an evidence filter might play in an environment with systematically misleading *vis-à-vis* one with randomly noisy evidence. If evidence is randomly noisy, a perfectly reliable evidence filter would, on the one hand, screen off all evidence in an environment with so much noise that following it would cause misleading belief updates. On the other, it would let in all the evidence in situations where following it would improve agents' priors. As we saw, in environments with noisy evidence, coherence-based evidence filtering is somewhat reliable: it resists misleading updates at the cost of resisting some truth-tracking updates as well.

In contrast, in environments where some pieces of evidence are systematically misleading or biased,

[5] The modified BNs we used are available in the same GitHub repository as the code for the model in an easily readable .bif file format. We invite readers to further modify the networks or introduce their own custom ones.

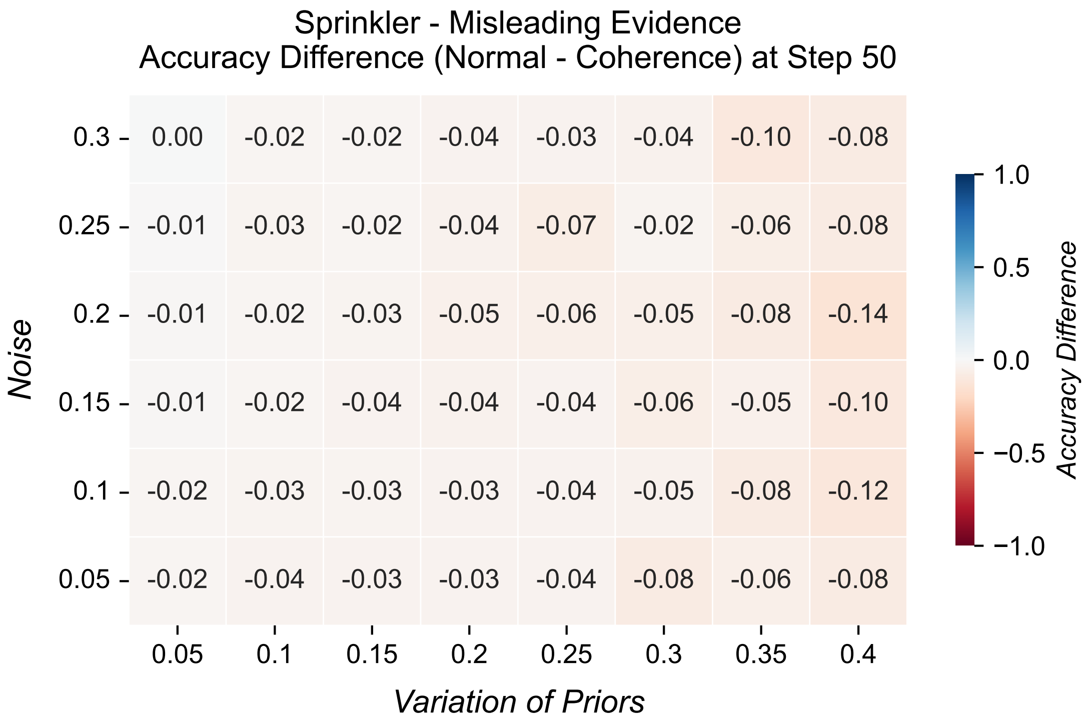


Figure 6: Difference in accuracy (Normal minus Coherence) at round 50 in the "Sprinkler" network with systematically misleading evidence. Positive values favor coherence-based reasoning.

a filter can help discriminate between reliable and biased evidence. A perfect evidence filter in such environments would screen off all pieces of biased evidence and let in all of the reliable evidence. Here, the coherence-based evidence filter fails. It seems that, at least sometimes, it is actually anti-reliable: it lets in misleading evidence and screens off reliable evidence rather than vice versa.

Figure 7 presents some evidence of this anti-reliability. In the Figure, the top chart represents the evolution of one Coherence agent's belief accuracy over time for one combination of parameters. The bottom chart shows the evolution of a Normal agent's belief accuracy over time in the same situation. The figure shows that while the Normal agent reliably advances toward the truth, the Coherence agent actually moves away from it. Note that this dynamic is not present in environments with noisy evidence. There, Coherence agents always make belief updates in the same direction as Normal agents, although much more conservatively.

Filtering out reliable evidence instead of misleading evidence does not happen often. In some situations—especially where agents start with accurate priors—its impact is almost neglectable. However, it becomes much more pernicious when agents start with worse beliefs. These results turn out to be robust across a range of conditions. Interested readers may once again refer to the Appendix.

In sum, coherence-based reasoning can improve belief accuracy—but only under specific epistemic conditions. When it is beneficial, its value lies in its conservatism: it resists misleading updates at the cost of resisting some truth-tracking updates as well. In environments where error is frequent and unsystematic, this tradeoff pays off. In more reliable environments, however, coherence filtering risks entrenching false beliefs. Additionally, in environments where evidence is systematically misleading, coherence-based filtering can act anti-reliably and thus predictably leave agents worse off. Whether coherence should be

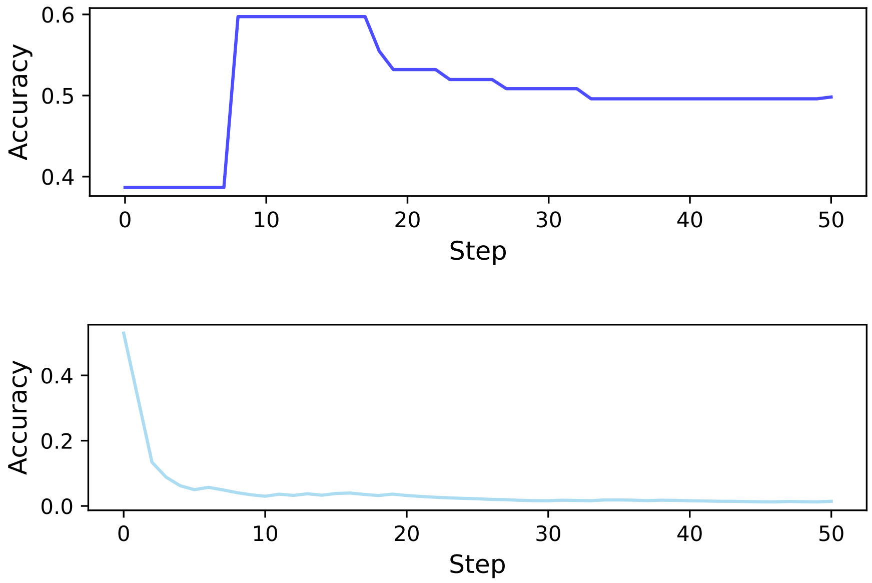


Figure 7: Evolution of one agent's beliefs over time for high chance of receiving misleading evidence (30%), inaccurate priors (35%), and the Hartmann-Trpin coherence measure. Top: Coherence agent. Bottom: Normal agent. Note that, due to random sampling of prior conditional probabilities, agents may start from different levels of prior accuracy.

treated as epistemically virtuous or not thus depends on the structure of the informational world in which an agent finds herself.

# 5 Discussion

Our simulation study set out to evaluate when, if ever, coherence-based filtering of evidence leads to epistemically superior belief states. The results were mixed: coherence can sometimes improve belief accuracy, especially in noisy environments, but its benefits are fragile. In other contexts—particularly when agents have inaccurate priors or encounter systematically misleading evidence—coherence consistently leads them astray. In this section, we explore the broader philosophical significance of these findings, focusing on three questions: (i) whether coherence-based reasoning can be rationally justified as a kind of higher-order evidential practice, (ii) how our results connect to the social epistemology of science, and (iii) whether they problematize the Principle of Total Evidence. Finally, we also look at the relation of our approach to Thagard's (1989) model ECHO, and at the implications of our results for the norms of rationality in science.

## 5.1 The Epistemic Status of Coherence Filtering

In the simulation, we assumed that coherence can be understood as a kind of higher-order evidence and that such evidence acts as a defeater. Both parts of this assumption are controversial; thus, we will say something more about them here.

Higher-order evidence, as usually understood in epistemology, is evidence about one's reasoning or evidential situation (Horowitz, 2022). In contrast to first-order evidence, which tracks the world directly, higher-order evidence tells us something about the agent's epistemic state. Standard cases of higher-order evidence, as discussed in epistemology, usually concern evidence of agents' unreliability. For example, the well-known Hypoxia case (Elga, 2008) concerns a pilot who gets evidence that she suffers from hypoxia—a condition that makes her calculations and predictions unreliable. Similarly, evidence of peer disagreement, which is also commonly discussed in this literature, signals that one of the disagreeing peers has made a reasoning mistake (Christensen, 2016).

In contrast, a drop of coherence, caused by updating on recently received evidence, does not say anything about the agent or the reliability of their reasoning process. Rather, it tells us something about the evidence that caused the drop in coherence—it signals that this evidence might be unreliable or otherwise defective. One might question whether this is a genuine example of higher-order evidence. However, we think this question speaks more to the limited understanding of higher-order evidence in the existing debate than against such an understanding of coherence. First, it is clear that the fact that incoming evidence does not fit with the agent's current beliefs is not first-order evidence—it tells us nothing new about the world. Second, other examples exist where such "evidence of evidence" is used in science. Scientists routinely measure the statistical significance of their results, i.e., how probable such results are if they assume the null hypothesis. As with coherence, these tests tell them nothing about the (un)reliability of their reasoning but act as a kind of higher-order evidence about evidence.

The other part of our controversial assumption concerns the question of whether higher-order evidence has a defeating force. While the thought that higher-order evidence should prompt us to revise our beliefs has strong intuitive support, it has proven remarkably hard to assimilate this insight into a consistent picture of epistemic rationality (see Lasonen-Aarnio 2014 for an early explication of some of the issues). Some philosophers even argue against higher-order defeat altogether (Titelbaum, 2015; Littlejohn, 2018; Lasonen-Aarnio, 2019). Fully addressing this question goes beyond the scope of this paper. Nevertheless, we will outline one possible understanding of coherence-based higher-order evidence where such evidence has a defeating force. Central to this understanding is the question of how we should conceive of the normative force of higher-order evidence.

One natural interpretation of coherence filtering is within a reliability-focused framework. On views of this kind (see, e.g., White 2009; Schoenfield 2018; Ye 2022), higher-order evidence is significant not because it provides propositional support for first-order claims but because it bears on the expected accuracy of the agent's belief-forming methods. In line with this, coherence filtering may be epistemically valuable when the agent's background beliefs are reliable, and the environment is noisy. In these cases, the filter helps screen out noise without obstructing access to truth. However, when the agent's priors are inaccurate or the environment is systematically misleading, the same mechanism becomes counterproductive. Coherence then functions less as a safeguard against noise and more as a gatekeeper against correction. Under some conditions, it can even be actively detrimental to the accuracy of the agent's beliefs. The root of the problem is then clear: coherence may be understood as higher-order evidence in a reliability-focused framework, but it is not a particularly useful form of higher-order evidence.[6]

Our simulations make this point vivid. Depending on the starting point and the nature of the evidence (whether, how, and how much it is misleading), the outcomes vary significantly. Beyond the higher-order evidence debate, this supports the view that coherence has no intrinsic epistemic authority. The effectiveness of coherence-based evidence filtering depends on factors external to coherence itself, such as the reliability of the agent's priors and the noise characteristics of the environment. It is not a mark of rationality *per se* nor a reliable path to truth in general. Instead, coherence functions best as a context-sensitive heuristic—adaptive in certain settings but potentially dangerous when misapplied.

[6] However, *pace* Horowitz (2019), coherence-based higher-order evidence is not predictably misleading. As we showed, there are situations where taking it into account pays.

## 5.2 Social Coherence and the Epistemology of Science

Our results also echo recent work in social epistemology that emphasizes the role of coherence as a defeater in belief assessment. In particular, Goldberg and Khalifa (2022) argue that coherence in science is best understood as a social norm: a belief may be prima facie unjustified if it negatively coheres with information that members of a scientific community are epistemically entitled to expect one another to consider.

Our model is not social in this sense. Agents do not represent or respond to the epistemic positions of others, nor are they assessed against any community-wide body of knowledge. Still, there is a formal analog to negative coherentism at the level of the individual. Coherentist agents filter new evidence through an internal coherence constraint: they are less likely to update on claims that conflict with their existing beliefs. As a result, their beliefs exhibit a kind of temporal coherence—earlier commitments modulate future ones.

This internal filtering mechanism has similar effects to the negative coherentist constraints described by Goldberg and Khalifa (2022). When agents begin with accurate priors, coherence helps maintain reliable beliefs. But when those priors are poorly calibrated, or when evidence is systematically biased, the same coherence filter leads agents astray. The results of our simulations, therefore, show that the epistemic status of coherence in our model is defeasible in precisely the way Goldberg and Khalifa emphasize: coherence with a prior belief set may be reasonable in some contexts, but it does not guarantee justification and may, in fact, hinder it. This aligns with the minimal version of (informal) negative coherentism that Goldberg and Khalifa defend: that incoherence with a reasonable epistemic position can be a defeater, even if coherence itself is not a mark of justification.

## 5.3 Coherence and the Principle of Total Evidence

Our findings speak to ongoing debates about the Principle of Total Evidence (PTE), the normative claim that one ought to condition belief on all available evidence (Carnap, 1947; Good, 1967). At first glance, our model appears to challenge this ideal: coherence-based agents routinely discard incoming information if it conflicts with their prior beliefs, and in some contexts, this turns out to be beneficial. In doing so, they violate PTE by design.

But this apparent conflict is not so straightforward. In our model, the so-called "evidence" presented to agents is not guaranteed to be veridical. It is sampled from processes that may include significant noise or bias. From the standpoint of classical PTE, such data arguably should not be treated as evidence in the first place. This suggests that the agents' coherence-based filtering can be reinterpreted: not as violating PTE, but as embodying a kind of internal noise-detection mechanism for deciding what to treat as evidence at all.

This reframing aligns with recent efforts to qualify or reinterpret PTE. For example, Sikorski and Gebharter (forthcoming) argue that including all available information can undermine reliability in contexts where evidence sources are interdependent, such as forensic science. Others, like Schurz (2024), defend PTE in idealized settings but acknowledge that it presupposes a well-calibrated background, in case the agent can identify relevant, non-defective information and treat it as approximately certain. Our agents are not in that position. They do not know which data points are genuinely indicative of the world and which are noise. Coherence filtering provides one way to mitigate this uncertainty.

This ambiguity, therefore, speaks in favor of PTE, but it also illustrates a broader point. In real-world settings, agents are not always in a position to determine what counts as total evidence. One may be confronted with information that *seems* evidential but arises from unreliable processes. In such cases, epistemic norms must grapple not just with how to reason *given* one's evidence but with how to determine what *is* evidence. Our results suggest that coherence constraints can function, for better or worse, as one way of navigating that threshold. This shifts the focus from PTE as an unqualified norm to a more procedural perspective: agents may need heuristics to decide what to treat as evidence in the first place.

Under such uncertainty, coherence-based screening offers a defeasible, fallible, but sometimes useful strategy.

## 5.4 Comparison with ECHO

Our approach also bears comparison to Thagard's influential ECHO model of explanatory coherence (Thagard, 1989). In ECHO, evidence has initial priority but may later be "deactivated" if it coheres only with hypotheses that lose support. Hence, evidence can also be filtered in a coherence-based way. However, an important difference from our approach needs to be pointed out. The described approach of ECHO, where evidence is filtered (i.e., deactivated) if it loses support, reflects what Harman (1986, Ch. 4) calls the *foundations theory* of belief revision: each belief must remain underwritten by sufficiently strong justificatory links, and even once-accepted beliefs may be discarded if their supports collapse.[7]

Our model, by contrast, more closely follows what Harman terms the *coherence theory* of belief revision. Once evidence passes the coherence filter it is retained, and belief change proceeds by minimal adjustments that preserve overall system integrity. In this respect our Bayesian approach operationalizes coherence-based conservatism in a way that better matches Harman's account of actual reasoning practices, while Thagard's ECHO represents a more foundationalist, always-ready-to-revisit stance.

At the same time, it is important to stress that Thagard's theory was a pioneering contribution that shaped decades of work on explanatory and computational models of coherence. Its connectionist implementation made coherence a tractable construct in cognitive science, and subsequent extensions of ECHO have enriched our understanding of explanatory reasoning in both science and everyday cognition (e.g., Thagard, 2002). What distinguishes our approach is not a rejection of this tradition but a shift of emphasis: by embedding coherence-based conservatism within a Bayesian network framework, we situate it directly within the dominant modeling paradigm of contemporary formal epistemology and the philosophy of science. This not only makes our model easier to integrate with probabilistic approaches to confirmation, explanation, and scientific reasoning but also highlights how coherence-driven conservatism can be formally analyzed alongside other Bayesian updating rules. In this way, our account builds on the legacy of ECHO while offering a complementary route for understanding the role of coherence in inquiry.

## 5.5 Reframing the Norms of Rational Inquiry

Finally, our results suggest that coherence-based reasoning occupies an uneasy position within epistemic norms. It is not rationally required nor uniformly reliable, but neither is it irrational or epistemically arbitrary. Instead, coherence filtering exemplifies a form of *context-sensitive epistemic conservatism*: a strategy that trades openness to new information for the preservation of belief, with effects that depend crucially on the surrounding informational environment.

What this suggests is a richer picture of epistemic rationality, in which strategies like coherence filtering cannot be evaluated in isolation from their environments. The rationality of an update rule depends not only on its internal logic but also on its ecological fit: how well it performs given the reliability of prior beliefs, the structure of incoming evidence, and the agent's epistemic goals (e.g., truth vs. internal belief stability). This ecological perspective is familiar in cognitive science (e.g., Gigerenzer and Brighton, 2009) and recently also in epistemology (Pils, 2022; Thorstad, 2024), but it is often underemphasized in debates regarding belief revision in the philosophy of science.

Our model helps us see why this aspect is important. Because coherence filtering is sensitive to *how* beliefs change, not just *what* they represent, it builds in a preference for belief-system integrity over short-term responsiveness. In this sense, coherence filtering resembles other conservative strategies in

[7] Note that Harman here discusses theories of belief revision rather than theories of justification. Hence, it is possible that the foundations theory of belief revision (as embodied in Thagard's ECHO model) may be coherentist with respect to justification (or activation, in Thagard's connectionist terms).

science, such as the preference for established paradigms in Kuhn's (1962) theory of scientific change or the use of robustness checks in statistical modeling. These practices slow down belief revision to protect against overfitting or premature shifts, but they can also perpetuate error when the system they protect is flawed, or evidence is biased.

From a normative standpoint, then, our findings support a pluralistic view of rational inquiry. No single updating rule (coherence-based or otherwise) dominates across all epistemic contexts. What matters is not whether coherence filtering is always rational, but when it is rational *to filter for coherence*. This, in turn, depends on the agent's epistemic situation. Understanding rational inquiry in these terms may also illuminate long-standing debates in epistemology and philosophy of science. The tension between total evidence and selective updating, for instance, is often framed in abstract normative terms. Our results suggest that the real epistemic stakes lie not in the ideal of perfect receptivity to evidence but in the pragmatic challenge of distinguishing signal from noise.

## 6 Conclusion

Our findings complicate the idea that coherence is simply a sign of a well-structured belief set. When coherence is treated not merely as a structural virtue but as a defeater for incoming evidence, the question of its epistemic value becomes sharply context-sensitive. The normative question is then no longer just whether coherent beliefs are justified or more likely to be true but whether coherence should license epistemic conservatism. Once coherence is allowed to influence whether new information is taken on board, the epistemic task shifts: it is no longer simply about updating on the total evidence but about deciding whether some coherence-disrupting evidence ought to be treated as evidence at all.

This reframes coherence from a static justificatory concept in epistemology to a dynamic constraint on inquiry. In this light, the old debate over whether coherence justifies belief gives way to a more pragmatist concern: whether coherence-based reasoning helps inquiry move in the right direction.

Our results suggest that sometimes it does. However, in contexts where information is unreliable or adversarially shaped, coherence is a treacherous guide. Its epistemic value turns not on its relation to truth *per se* but on the structure of the informational environment and the agent's place within it. In this sense, coherence inherits the ambiguity of methodological conservatism in science: it can preserve hard-won understanding or obstruct paradigm shifts. What matters is not coherence itself but knowing when to treat it as a warning sign and when to override it in pursuit of better evidence. While identifying the factors that govern this balance exceeds the scope of this paper, our results indicate that doing so is essential for a deeper understanding of the role of coherence in inquiry.

## References

Adam, T., N. Agafonova, A. Aleksandrov, et al. (2011). Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. *arXiv preprint arXiv:1109.4897*.

Angere, S. (2008). Coherence as a heuristic. *Mind 117*(465), 1–26.

Assaad, L., R. Fuchs, A. Jalalimanesh, K. Phillips, L. Schoeppl, and U. Hahn (2023). A Bayesian agent-based framework for argument exchange across networks. *arXiv*.

Baccini, E., Z. Christoff, S. Hartmann, and R. Verbrugge (2023). The wisdom of the small crowd: Myside bias and group discussion. *Journal of Artificial Societies and Social Simulation 26*(4).

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Bovens, L. and S. Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Brumfiel, G. (2011). Particles break light-speed limit. *Nature 22*, 22.

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research 8*(1), 133–148.

Cartlidge, E. (2012). Loose cable may unravel faster-than-light result. *Science 335*(6072), 1027–1027.

Christensen, D. (2016, 9). Conciliation, uniqueness and rational toxicity. *Noûs 50*, 584–603.

Dorst, K. (2024). Higher-order evidence. In M. Lasonen-Aarnio and C. Littlejohn (Eds.), *The Routledge Handbook of the Philosophy of Evidence*. Routledge.

Douven, I. and W. Meijs (2007). Measuring coherence. *Synthese 156*(3), 405–425.

Easterbrook, P., R. Gopalan, J. Berlin, and D. Matthews (1991, 4). Publication bias in clinical research. *The Lancet 337*, 867–872.

Elga, A. (2008). Lucky to be rational. Unpublished paper, available at `https://www.princeton.edu/ adame/papers/bellingham-lucky.pdf`.

Festinger, L., H. W. Riecken, and S. Schachter (1956). *When Prophecy Fails: A Social and Psychological Study of a Modern Group That Predicted the Destruction of the World.* Harper-Torchbooks.

Feynman, R. P. (1985). *Surely You're Joking, Mr. Feynman! Adventures of a Curious Character*. New York: W. W. Norton & Company.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis 63*(3), 194–199.

Gigerenzer, G. and H. Brighton (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science 1*(1), 107–143.

Glass, D. H. (2002). Coherence, explanation, and Bayesian networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, and N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science, 13th Irish Conference, AICS 2002*, pp. 177–182. Berlin: Springer.

Goldberg, S. C. and K. Khalifa (2022). Coherence in science: A social approach. *Philosophical Studies 179*, 3489–3509.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science 17*(4), 319–321.

Harman, G. (1986). *Change in View: Principles of Reasoning*. The MIT Press.

Hartmann, S. (2021, 12). Bayes nets and rationality. In *The Handbook of Rationality*. The MIT Press.

Hartmann, S. and B. Trpin (Forthcoming). Why coherence matters? *The Journal of Philosophy*.

Holman, B. and K. C. Elliott (2018, 11). The promise and perils of industry-funded science. *Philosophy Compass 13*.

Horowitz, S. (2019, 10). Predictably misleading evidence. In M. Skipper and A. Steglich-Petersen (Eds.), *Higher-Order Evidence: New Essays*, pp. 105–123. Oxford University Press.

Horowitz, S. (2022). Higher-order evidence. *The Stanford Encyclopedia of Philosophy*.

Koscholke, J., M. Schippers, and A. Stegmann (2019). New hope for relative overlap measures of coherence. *Mind 128*(512), 1261–1284.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lasonen-Aarnio, M. (2014, 3). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research 88*, 314–345.

Lasonen-Aarnio, M. (2019, 10). Higher-order defeat and evincibility. In *Higher-Order Evidence*, pp. 144–172. Oxford University PressOxford.

Lauritzen, S. L. and D. J. Spiegelhalter (1988, 1). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology 50*, 157–194.

Lehrer, K. (2000). *Theory of Knowledge*. Routledge. Second edition.

Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. La Salle, Illinois: The Open Court Publishing Company.

Littlejohn, C. (2018, 3). Stop making sense? On a puzzle about rationality. *Philosophy and Phenomenological Research 96*, 257–272.

Lundh, A., J. Lexchin, B. Mintzes, J. B. Schroll, and L. Bero (2017, 2). Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews 2017*.

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language 34*(2), 141–157.

Meijs, W. (2006). Coherence as generalized logical equivalence. *Erkenntnis 64*(2), 231–252.

Millikan, R. A. (1913, 8). On the elementary electrical charge and the Avogadro constant. *Phys. Rev. 2*, 109–143.

Olsson, E. J. (2002). What is the problem of coherence and truth? *The Journal of Philosophy 99*(5), 246–72.

Olsson, E. J. (2005). *Against Coherence: Truth, Probability, and Justification*. Oxford University Press.

Olsson, E. J. (2022). *Coherentism*. Cambridge: Cambridge University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers.

Pils, R. (2022). A satisficing theory of epistemic justification. *Canadian Journal of Philosophy 52*(4), 450–467.

Rescher, N. (1973). *The Coherence Theory of Truth*. Oxford: Oxford University Press.

Russell, S. J. and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Schippers, M. (2014). Probabilistic measures of coherence: From adequacy constraints towards pluralism. *Synthese 191*(16), 3821–3845.

Schoenfield, M. (2018, 5). An accuracy based approach to higher order evidence. *Philosophy and Phenomenological Research 96*, 690–715.

Schupbach, J. N. (2011). New hope for Shogenji's coherence measure. *The British Journal for the Philosophy of Science 62*(1), 125–142.

Schurz, G. (2024). The principle of total evidence: Justification and political significance. *Acta Analytica 39*(4), 677–692.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis 59*(4), 338–345.

Sikorski, M. and A. Gebharter (Forthcoming). The criminalist's paradox as a counterexample to the principle of total evidence. *The British Journal for the Philosophy of Science*.

Stanovich, K. E., R. F. West, and M. E. Toplak (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science 22*(4), 259–264.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences 12*(3), 435–467.

Thagard, P. (2002). *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Thorstad, D. (2024). Why bounded rationality (in epistemology)? *Philosophy and Phenomenological Research 108*(2), 396–413.

Titelbaum, M. G. (2015, 2). Rationality's fixed point (or: In defense of right reason). In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology Volume 5*, pp. 253–294. Oxford University Press.

White, R. (2009, 10). On treating oneself and others as thermometers. *Episteme 6*, 233–250.

Ye, R. (2022). *Higher-Order Evidence and Calibrationism*. Elements in Epistemology. Cambridge University Press.

Zollman, K. J. S. (2007, 12). The communication structure of epistemic communities. *Philosophy of Science 74*(5), 574–587.

Zollman, K. J. S. (2010, 1). The epistemic benefit of transient diversity. *Erkenntnis 72*, 17–35.

# Appendix: Robustness Checks

This Appendix presents robustness checks for the results, presented in Section 4. The first part deals with coherence under noisy evidence, and the second with coherence under systematically misleading evidence.

## Noisy Evidence

To check the robustness of results for cases where agents receive evidence corrupted by random noise, we first extend the number of simulation rounds to 500. Figure 8 confirms that the observed advantages (and disadvantages) of coherence filtering are not merely transient: even in the long run, coherence filtering helps when noise is high or priors are good, and hinders when noise is low or priors are poor.

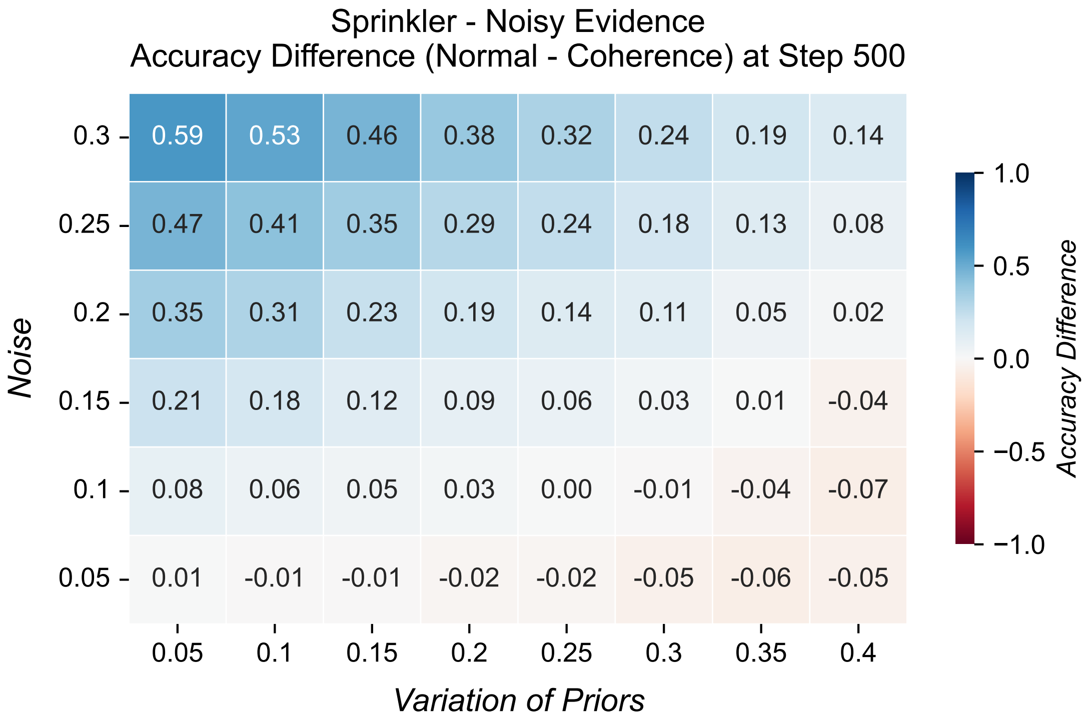


Figure 8: Difference in accuracy (Normal minus Coherence) at round 500 in the “Sprinkler” network with noisy evidence.

We also tested the results under varying sample sizes—20, 40, and 60 samples per round instead of 100. These results are shown in Figure 9. With fewer observations, coherence agents tend to perform slightly worse—reflecting the greater risk of over-rejecting evidence in data-scarce environments—but the overall trends remain stable.

Additionally, we replicated the simulation using the more complex “Asia” network. While the absolute differences are more pronounced—likely due to the increased difficulty of learning in a larger epistemic space—the qualitative patterns are nearly identical (Figure 10).

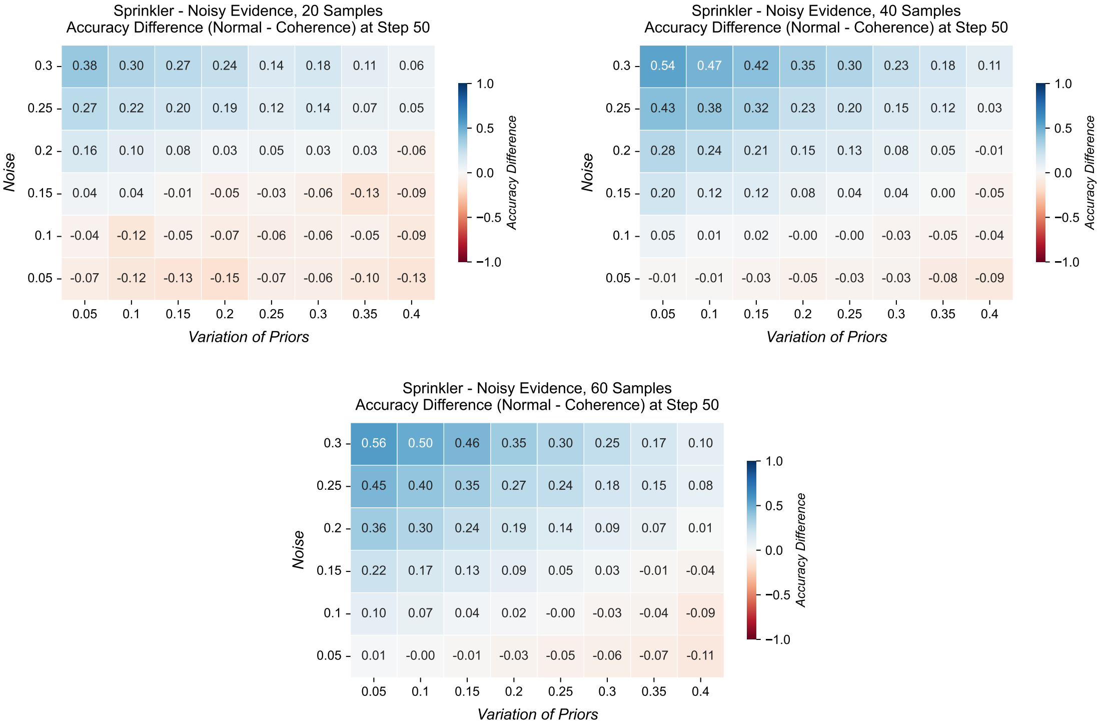


Figure 9: Difference in accuracy (Normal minus Coherence) at round 50 in the “Sprinkler” network with noisy evidence. Top left: agents received 20 evidence samples per round. Top right: 40 evidence samples. Bottom: 60 samples.

## Systematically Misleading Evidence

The same robustness checks were conducted for the cases where agents received systematically misleading evidence. Figure 11, Figure 12, and Figure 13 present the results for 500 simulations rounds, for different amounts of evidence received, and for the “Asia” Bayesian network respectively.

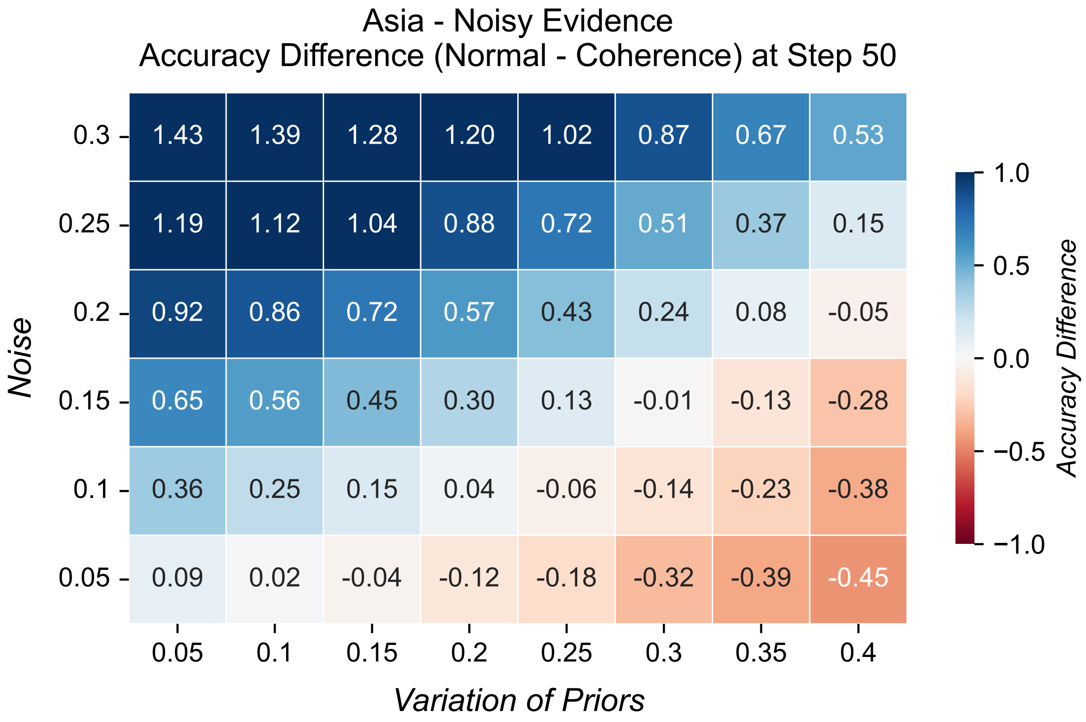


Figure 10: Results for the “Asia” BN with noisy evidence. Coherence is beneficial only under high noise and accurate priors.

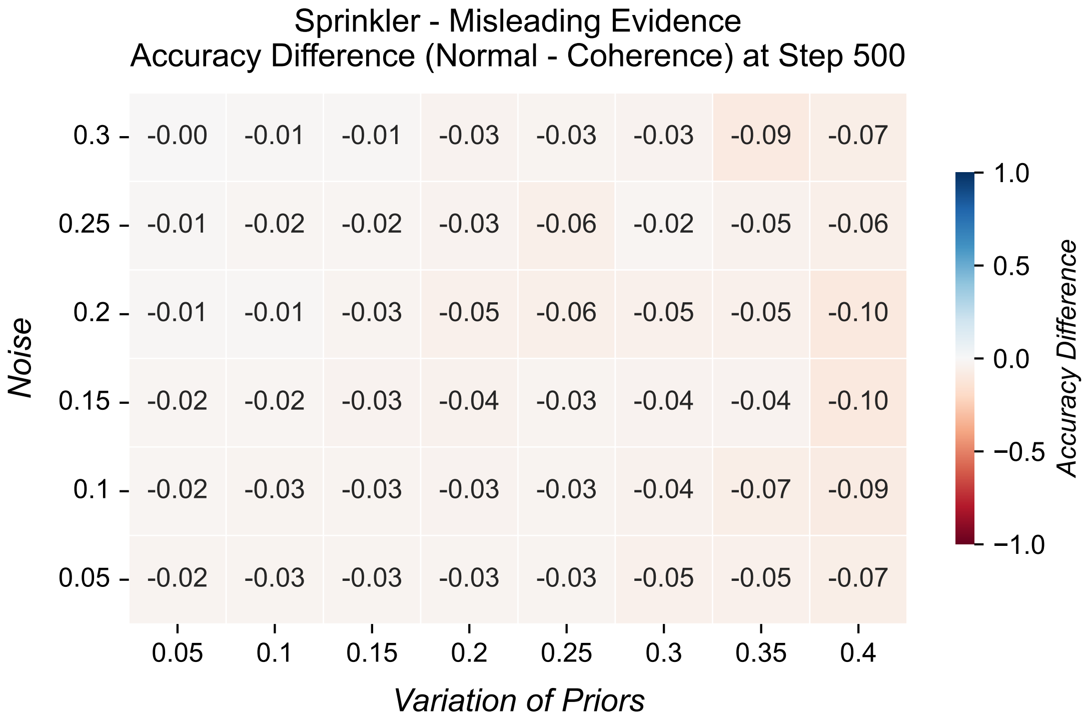


Figure 11: Difference in accuracy (Normal minus Coherence) at round 500 in the “Sprinkler” network with systematically misleading evidence.

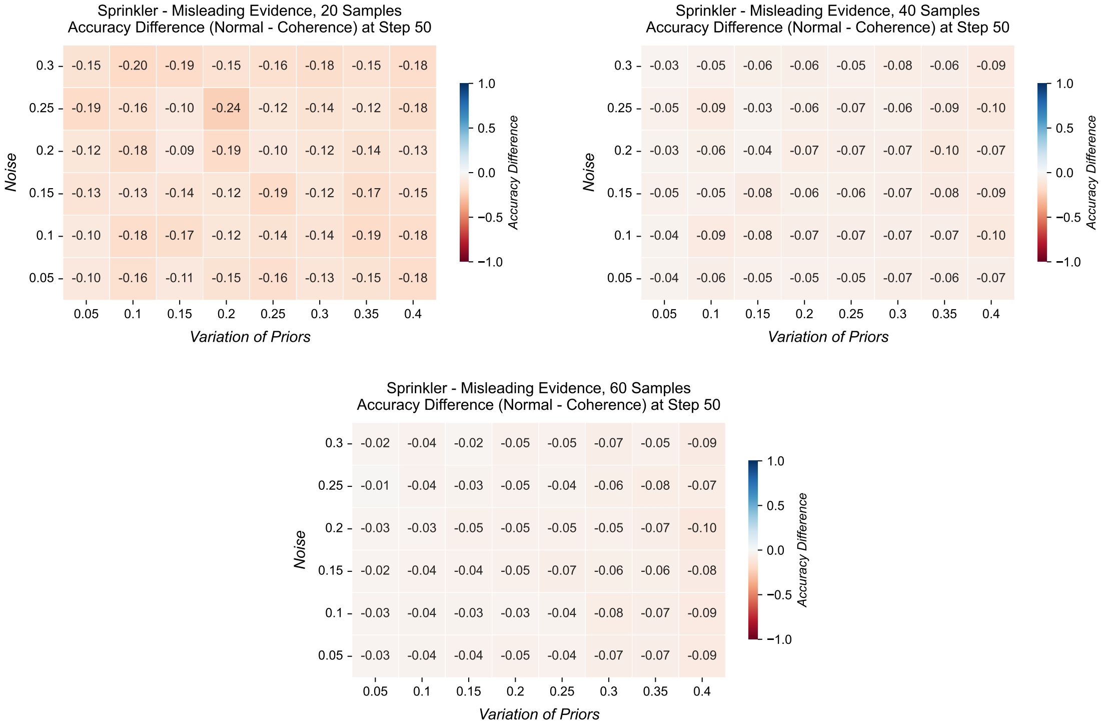


Figure 12: Difference in accuracy (Normal minus Coherence) at round 50 in the “Sprinkler” network with misleading evidence. Top left: agents received 20 evidence samples per round. Top right: 40 evidence samples. Bottom: 60 samples.

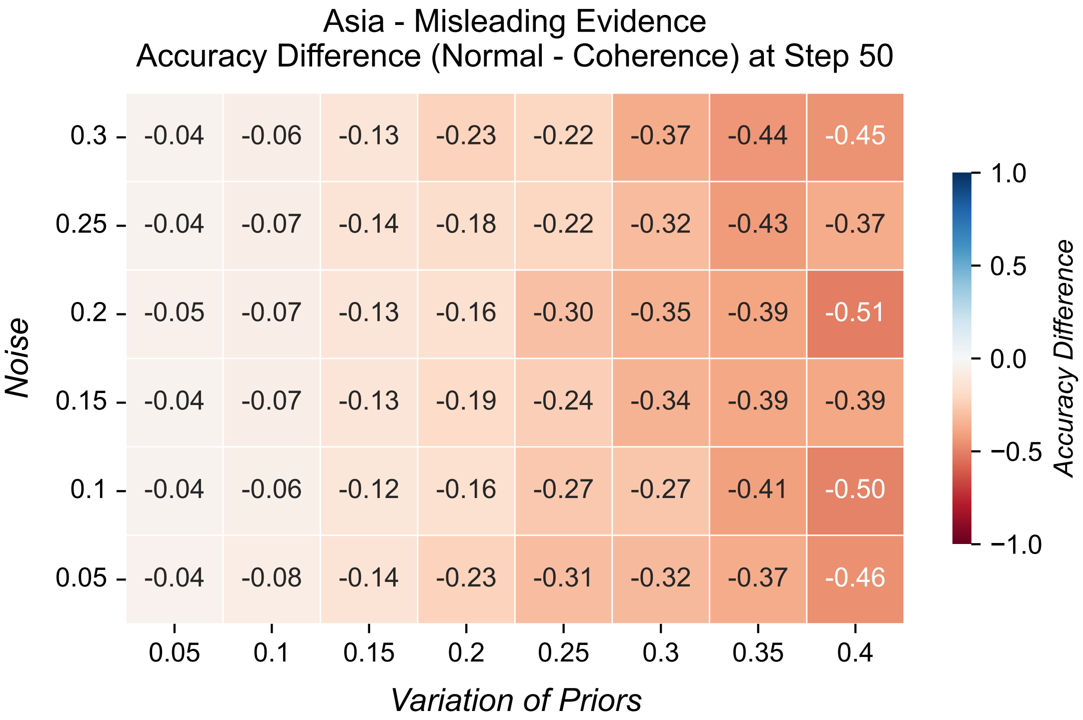


Figure 13: Results for the “Asia” BN with misleading evidence.