





ARTICLE



<https://doi.org/10.1057/s41599-024-03806-8>

OPEN

Misinformation and higher-order evidence

Brian Ball ¹✉, Alexandros Koliouisis¹, Amil Mohanan ¹ & Mike Peacey²

This paper uses computational methods to simultaneously investigate the epistemological effects of misinformation on communities of rational agents, while also contributing to the philosophical debate on ‘higher-order’ evidence (i.e. evidence that bears on the quality and/or import of one’s evidence). Modelling communities as networks of individuals, each with a degree of belief in a given target proposition, it simulates the introduction of unreliable mis- and disinformants, and records the epistemological consequences for these communities. First, using small, artificial networks, it compares the effects, when agents who are aware of the prevalence of mis- or disinformation in their communities, either deny the import of this higher-order evidence, or attempt to accommodate it by distrusting the information in their environment. Second, deploying simulations on a large(r) real-world network, it observes the impact of increasing levels of misinformation on trusting agents, as well as of more minimal, but structurally targeted, unreliability. Comparing the two information processing strategies in an artificial context, it finds that there is a (familiar) trade-off between accuracy (in arriving at a correct consensus) and efficiency (in doing so in a timely manner). And in a more realistic setting, community confidence in the truth is seen to be depressed in the presence of even minimal levels of misinformation.

¹Northeastern University - London, London, UK. ²University of Bristol, Bristol, UK. ✉email: brian.ball@nulondon.ac.uk

Introduction

The informational environment has changed dramatically during the first quarter of the 21st century. Most notably, following the (historically recent) advent of the Internet, we now get our information - including about 'current affairs' or 'news' - online, often through various social media platforms (cf. Cairncross (2019)). Concerns have been raised in this context about the presence of misinformation (including 'fake news') - in some cases due to concerted disinformation campaigns - and its potential ill effects. Such false/inaccurate information may have an impact on people's attitudes and behaviours - for example, in relation to the climate emergency, democracy, health, or security issues - and various strategies have therefore been proposed for addressing the problem. In particular, it has been suggested that we need to improve communication strategies (for science and news media), policies and practices (for tech platforms), oversight and/or regulation (by charities and government), and information literacy (for the general public).

At the same time, there has been discussion in the philosophical literature of so-called 'higher-order evidence' (HOE) and how best to respond to it (see e.g. Christensen (2010); Feldman (2005); Fitelson (2012); Kelly (2005); Skipper & Steglich-Petersen (2019)). In particular, otherwise rational agents are sometimes presented with information that is not concerned directly with the issue they are considering, or investigating (it is not 'first-order evidence' in this sense), but which instead bears on the 'evidential relations' ((Christensen, 2010), p.186) at play in the circumstances they find themselves in - e.g. by bearing on the quality of the information at their disposal, or the appropriateness of certain ways of responding to it (it is, or appears to be, evidence of a 'higher-order').¹ For example, an agent who needs to rely on information from a sensor to reason towards a decision may be given (defeasible) evidence that the sensor is broken or that their reasoning is impaired. Two broad approaches to thinking about such circumstances have been articulated (see Horowitz (2022)): some theorists seek to deny the import of the purported evidence, and have argued that the agent should simply ignore it; while others have suggested that it must be accommodated, raising the question of how this can best be done.

In the present paper, we look to shed light on both the problem of misinformation and its effects, and the debate surrounding higher-order evidence, by bringing them into closer contact with one another.² The result, we think, is mutually illuminating: for, on the one hand, bringing our method for studying misinformation to bear on the debate on higher order evidence, we reveal how thorny the issues there are - resolving them will require a non-trivial decision on what makes for a better epistemic outcome; while, on the other hand, in considering the effects of adopting different strategies from that debate in coping with the problem of misinformation, we can appreciate just how pernicious a problem it is - as we shall see, both strategies yield worse outcomes (by at least one metric) in the presence of misinformation. But not all of the news of which we are the bearers is bad: we also observe that there are features of the informational environment in a real world setting that might be exploited to begin to address epistemological concerns and societal challenges - provided we adopt a community (i.e. network, or graph) level perspective.

In what follows, we begin with some cases, before abstracting the general character of the problem they present, and sketching its relevance. We then explain our computational methodology, including the ways in which we operationalize both the key notion of misinformation (and the related concept of disinformation) and a pair of strategies for processing higher-order evidence. In the third section, we analyse and interpret the findings from our investigations of some small, artificial

communities exposed to misinformation, revealing a trade-off between the information-processing strategies explored. Then in the fourth section we present some observations regarding the effects of misinformation within a real world setting, before finally concluding.

Case 1: One restaurant town. Imagine: you live in a town with just one restaurant. Not only that, the restaurant is not that good: only (exactly) half of the dishes it prepares are tasty. But there is good news: a second restaurant is opening! This new restaurant promises to be (slightly) better... though it could be (slightly) worse. Which restaurant should you go to?

You and the other townsfolk go to the new restaurant if, from your perspective, it is likely to be better than the old one - in which case you get further evidence about the proportion of the meals it serves that are tasty. And you share reports about the dishes, good and bad, with one another.

But now imagine: an oracle reveals that only a certain fraction of the inhabitants of the town are reliable when issuing their reports - though, sadly, they do not say which. How should you respond to the restaurant's various reviews?

Case 2: Drug trial (and error). Imagine: you are a doctor, and a new drug has recently become available for the treatment of a certain disease (from which patients either recover fully, or die). This seems like good news, since the previous drug has a recovery rate of just 50%. The new drug may be better - though there is a risk it is worse. Which drug should you administer?

You and your colleagues in the medical community decide to administer the new drug, if you suspect it is better than the old one - though not otherwise, as it is important to give patients the best available treatment. You all report on the outcomes regularly (e.g. at medical conferences).

But now imagine: an oracle tells you (and others in the community) that not all of your colleagues report their findings accurately. They even tell you what percentage of the doctors in your community are reliable reporters - but they do not name names of those who can and cannot be trusted. What should you do when confronted with the reports you receive?

Case 3: New coins for old. Imagine: you are a gambler. In particular, you - and everyone else in your community - likes to toss coins; and you (all) have a predilection for heads. Hitherto, all the coins available for tossing have been scrupulously fair, with a 50-50 chance of landing heads or tails. But now, some new coins have been minted, and an employee at the mint has told you (correctly) that they are biased, and by how much... though not in which direction. Which coins should you toss, old or new?

You and your fellow gamblers all toss the new coins if and only if you find it more likely than not that they will produce more heads overall than the old ones. You all report regularly on the outcomes of your coin tosses.

But now imagine: an oracle informs you that only some of your fellow gamblers report the results of their coin tosses accurately; and they even tell you what fraction of the gamblers do so - though they do not tell you which are trustworthy and which are not. How should you respond to the information provided in the reports you receive?

The problem revisited. All of these concrete problems share a structure. In the abstract, the basic problem is (or can be modeled) as follows: rational agents have two actions available to them, A and B. A is known to yield a positive outcome half of the time (i.e. with probability 0.5), while B yields a positive outcome

with probability $0.5 + \epsilon$ - though for all the agents know, it may do so with probability $0.5 - \epsilon$ instead. Each agent has a credence in the hypothesis that B is better than A (i.e. has a probability greater than 0.5 of yielding a positive outcome), and if that credence is above 0.5 they take action B for all of their trial actions; otherwise they take action A. Agents issue reports to one another about the outcomes of their trials. As rational agents, they update their beliefs in light of the evidence at their disposal. But it is known that only a certain fraction of the agents in the community are reliable in their reports. What should the agents in such a community do? Should they ignore the higher-order evidence at their disposal? Or should they seek to accommodate it somehow?

We address this issue by way of computer simulations of such communities of agents - as described in the next section. But it is worth briefly noting why, despite its somewhat artificial character, we think investigating this problem can be illuminating. First, it renders certain aspects of the philosophical debate surrounding higher-order evidence computationally tractable. In particular, we are able, in what follows, to operationalize two strategies for coping with higher-order evidence, and thereby bring new methods to bear on the discussion of that topic in illuminating ways.³ Second, when it comes to the problem of misinformation, we think it is helpful to abstract from some of the complexities of human psychology. The idealizations made in our models allow us to reveal how just how problematic the degradation of our informational environment can prove, even in communities of agents that are fully rational (on at least one widely accepted way of thinking about what this involves⁴). Our results may therefore constrain the space of plausible solutions to the pressing practical problem this otherwise theoretical issue poses.

Models and methods

We begin our simulations by artificially generating, or else importing, a graph representation of our (hypothetical, or respectively real) community. Each agent in the community is represented by a node in this graph, with edges, or connections between nodes, representing channels of communication through which reports are issued. Simulations are initialized by assigning (uniformly at random) to each node a credence (between 0 and 1) in the hypothesis that (coin, restaurant, or drug) B is better than A (so that the probability of its yielding a positive outcome - such as the coin landing heads, the dish being tasty, or the patient recovering - is $0.5 + \epsilon$).⁵ At each successive step in the simulation, agents whose credence that B is better is greater than 0.5 perform action B, conducting a fixed number of trials.

In the basic model, based on the mathematical work of Bala and Goyal (1998) - and building on the simulation technique of Zollman (2007) - agents observe the results of their trials and report the number of successes to their network neighbours (those to whom they are connected by an edge). They then update their beliefs using Bayes' rule, conditionalizing on the evidence they generate as well as that which they receive from their neighbours.⁶ This process then repeats, until either all nodes have a credence greater than 0.99, or all have a credence below 0.5 (so that none takes action B, and no new evidence is generated) - or an upper limit on the number of simulation steps is reached.

O'Connor & Weatherall (2018) generalized this approach to allow updating to proceed on the basis of Jeffrey's rule, rather than Bayes'.⁷ On their approach, agents display a tendency to homophily: in particular, they are more trusting of others whose beliefs are similar to their own. This idea is operationalized through a formula which sets the final probability assigned by an

agent to the evidence received from each neighbour as a function of the distance between the beliefs of that agent and that neighbour (i.e. the absolute value of the difference between the credences of the two agents involved in the interaction).⁸ Elsewhere (Ball et al. (forthcoming)) we have explored the performance of simulations based on this model - but it is important to note that the models we are interested in here are different in character than those that have been investigated previously. While our models allow evidence to be discounted using Jeffrey's rule, the basis on which this occurs is different: rather than being due to homophily, as it is for O'Connor and Weatherall, any discounting of evidence in our models is due to an attempt to accommodate higher-order evidence about the reliability of the messages, or first-order evidence, available.

Two kinds of unreliable agents. In our ('testimonials') models, some agents are unreliable. Our simulations accordingly have a 'reliability' parameter, which determines the probability that any given node will reliably report their findings to their neighbours. When reliability is set to 1.0, the result is equivalent to the original Bala-Goyal models. But when reliability falls below 1.0, we need to decide how our unreliable agents will behave. Here we explore two different reporting behaviours which unreliable agents might engage in.

On one model, the unreliable agents can be thought of as misinformants, who do not intend to deceive. Perhaps they are incompetent: although they toss the new (biased) coin B, they report their observations of the results of simultaneous tosses of (fair) coin A by mistake; or they have terrible memories, and what they report for each trial is unrelated to what actually occurred on that occasion, instead being 'remembered' at random. Or maybe their reports are 'bullshit' (Frankfurt, 2005): having no regard for the truth, their issuers simply make up the result of each trial (using a process resembling that of a fair coin toss). In any such case, the (first-order) 'evidence' they provide will be neutral overall,⁹ with the successes reported being drawn from a binomial distribution with a probability of 0.5 for each trial; it will therefore neither favour, nor tend to disconfirm, the hypothesis that B is better.

On a second model, by contrast, the unreliable agents can be thought of as disinformants, who seek to deceive and mislead. For each trial, they report heads if the outcome was tails, delicious if the dish was awful, or recovery if the outcome was death - and vice versa. They lie about that trial - knowing that their reports will, in aggregate, tend to disconfirm the correct hypothesis. Their reports are drawn from the distribution that results from a chancy process with probability $0.5 - \epsilon$.

Other behaviours are of course possible on the part of unreliable agents:¹⁰ but here we focus (for convenience) on these two simple approaches.

Two information processing strategies. How should epistemically rational agents respond to the higher-order evidence that a certain fraction of the members of their community are unreliable? As we have seen, the philosophical debate suggests two broad strategies.

A first is to simply ignore the higher-order evidence:¹¹ that is, agents might behave exactly as they would if they were unaware of the unreliability, fully trusting the evidence supplied by all agents; we might regard such trusting agents as 'gullible' - though, of course, as we have described the situation, they are knowingly so. Why might this seem a plausible approach (in the present context)? A number of authors (e.g. Burge (1993); Coady (1992); Reid (1983) have held that hearers are *prima facie* justified in believing what speakers tell them. This suggests that the

Table 1 Descriptive statistics for all simulations.

Model	Misinfo	Reliability	Count	Mean	Std	Min	25%	50%	75%	90%	Max
Bala-Goyal	None	1	500	425	376	93	234	323	489	657	3476
Gullible	Binomial	0.75	500	637	872	96	266	405	642	1073	9509
Gullible	Binomial	0.5	500	1181	1824	88	402	639	1198	2385	17823
Gullible	Binomial	0.25	500	1950	2531	135	547	1151	2304	4462	20000
Aligned	Binomial	0.75	500	631	1090	94	303	454	672	1041	17054
Aligned	Binomial	0.5	500	1313	3114	150	437	630	1009	1512	20000
Aligned	Binomial	0.25	500	3385	5450	235	905	1433	2253	9739	20000
Gullible	NegEps	0.75	500	1095	1560	97	387	662	1142	2193	17765
Gullible	NegEps	0.5	500	3184	4107	123	826	1605	3752	7011	20000
Gullible	NegEps	0.25	500	3195	3775	97	970	1830	3919	7402	20000
Aligned	NegEps	0.75	500	542	627	119	292	420	613	930	10484
Aligned	NegEps	0.5	500	816	8010	129	441	639	866	1382	8319
Aligned	NegEps	0.25	500	1954	2321	381	863	1284	2086	3396	20000

information conveyed to them by testimony, and not the mere fact of the testimony itself, is evidence for them.¹² Moreover, the appropriate response to evidence, in a probabilistic setting, is to conditionalize on it using Bayes' rule. And the hearer has no specific reason to doubt that any given speaker in particular is reliable - so, on the face of it, that seems the appropriate response in the case of each piece of testimony received.¹³

An alternative strategy - one which attempts to accommodate the available higher-order evidence - is to discount the available (first-order) 'evidence' on the grounds that it cannot be fully trusted. In particular, agents do not know which evidence to accept as (fully) trustworthy, and which to reject as inaccurate and misleading; but they do know the level of reliability in the network. One heuristic, then, is to simply align the level of confidence they place in any given piece of evidence with the known level of reliability. This provides a method of operationalizing Jeffrey's rule in a manner that reflects the higher-order evidence possessed by the agents.¹⁴

On this approach, the information received is processed in a manner that dampens its effect on our other attitudes - and in particular, on our credence concerning the hypothesis that is the target of our inquiry, because we have (a general) reason to doubt that it bears on that hypothesis in the manner that we would otherwise be inclined to think it does. In this respect, the current strategy is not unlike that of being cautious in drawing logical consequences from given premises in circumstances in which it is known that there is a non-trivial chance that doing so will lead one astray (cf. Christensen (2010)).

Clearly, each of these strategies coincides with that pursued in Bala Goyal models when reliability in the network is 100% (or 1.0). But as the reliability retreats from this ideal of perfect truthfulness, the two strategies diverge: and their merits can be assessed, at least in part, by how they perform (in terms of truth-conduciveness) under various circumstances. This is what we attempt below.

Our simulations. We carried out simulations on complete networks of size 64, setting epsilon to 0.001, and the number of trials conducted at each simulation step to 64.¹⁵ We ran 500 simulations of this kind for each collection of further parameters, as follows. First, we ran Bala Goyal model simulations, with no mis- or disinformation present, and agents updating their credences by conditionalizing on the available evidence using Bayes' rule. (As noted above, these simulations are equivalent to employment of either of our other models with a reliability setting of 1.0.) We then also ran simulations in which agents remained 'gullible', with each type of unreliable nodes: binomial misinformants; and negative epsilon disinformants. And similarly, we ran simulations

in which agents 'aligned' the posterior probability of the (first-order) evidence they encountered to the reliability level in the network, in the presence of unreliable nodes of each of the two types described. We did this in each case setting reliability to three quarters (0.75), half (0.5), and one quarter (0.25). We let simulations run until a consensus emerged, whether true (B is better) or false (A is better), or a maximum of 20,000 steps was reached.¹⁶

We also ran simulations on a larger, real-world network: but we begin by discussing the above simulations on artificially generated complete networks.

Analysis and interpretation

To begin to understand the effectiveness of the two strategies for coping with mis- and disinformation in light of the available higher-order evidence about informer reliability, and the bearing of this on the higher-order evidence debate, we begin with some descriptive statistics concerning our simulations on small, artificial networks. Table 1 shows, for each of the batches of simulations described above (with the simulation parameters given there), the model (including misinformation type and strategy, where applicable), the level of reliability in the network, the number of simulations run with those parameters (count), the mean number of steps taken in those sims, the standard deviation (std) in the number of steps taken, the minimum and maximum numbers of steps, and the number of steps at various percentiles (25, 50, 75, and 90).

A quick look reveals, for instance, that (i) in general, these distributions have a long right tail, with some simulations requiring far more steps in order to complete than e.g. the median, or even than the 90th percentile simulation, and indeed (ii) some simulations ran for the full 20,000 steps allocated without converging on a consensus opinion. Equally, though, we can see that (iii) Bala-Goyal simulations were on average the fastest, with the lowest mean and median (i.e. 50th percentile) number of steps. Indeed (iv) the presence of mis- and disinformation considerably delays epistemic decision making in our simulated communities: for instance, while 9 out of 10 simulations converged one way or the other (either to A or to B) by 657 steps when agents were fully reliable and fully trusting, much smaller fractions of simulations completed after this many steps in the presence of mis- or disinformation (e.g. less than a quarter in negative epsilon disinformation simulations with reliability 0.25).

Tables 2 and 3 give the same descriptive statistics, but now restricting attention to those simulations that converged to the incorrect (A) and correct (B) opinions respectively. Here we see that, as expected, no simulation ran for the full 20,000 steps

Table 2 Descriptive statistics for simulations converging to A.

Model	Misinfo	Reliability	Count	Mean	Std	Min	25%	50%	75%	90%	Max
Bala-Goyal	None	1	11	808	417	328	506	706	949	1456	1597
Gullible	Binomial	0.75	14	2465	3226	203	627	898	1925	7838	9509
Gullible	Binomial	0.5	69	2103	2179	142	679	1474	2457	4379	11600
Gullible	Binomial	0.25	149	2707	2903	137	1028	1710	3209	5772	19601
Aligned	Binomial	0.75	13	4050	5510	330	931	1951	3985	12991	17054
Aligned	Binomial	0.5	11	39323	4521	397	1111	1430	5166	11094	13243
Aligned	Binomial	0.25	14	7997	6276	605	2881	6673	13762	15790	18279
Gullible	NegEps	0.75	64	2489	3161	204	970	1755	2469	4465	17765
Gullible	NegEps	0.5	293	3140	3113	154	1110	2062	4010	6803	19380
Gullible	NegEps	0.25	483	2959	3146	97	983	1807	3829	6781	19776
Aligned	NegEps	0.75	12	2371	2953	351	885	1271	1922	5612	10484
Aligned	NegEps	0.5	8	1309	925	726	794	892	1389	2154	3472
Aligned	NegEps	0.25	31	6950	4564	753	3804	5394	9447	14011	15880

Table 3 Descriptive statistics for simulations converging to B.

Model	Misinfo	Reliability	Count	Mean	Std	Min	25%	50%	75%	90%	Max
Bala-Goyal	None	1	489	416	371	93	232	321	476	641	3476
Gullible	Binomial	0.75	486	585	636	96	262	399	619	1051	4949
Gullible	Binomial	0.5	431	1033	1718	88	366	597	1070	1945	17823
Gullible	Binomial	0.25	349	1523	1816	135	470	847	1786	3326	16368
Aligned	Binomial	0.75	487	540	387	94	300	450	655	956	3818
Aligned	Binomial	0.5	477	782	700	150	428	610	942	1378	9099
Aligned	Binomial	0.25	443	1627	1320	235	862	1315	1890	2778	13510
Gullible	NegEps	0.75	436	890	1009	97	364	610	987	1718	11796
Gullible	NegEps	0.5	194	2123	2962	123	527	1118	2045	5158	19098
Gullible	NegEps	0.25	9	889	948	205	454	459	522	2350	2832
Aligned	NegEps	0.75	488	497	349	119	289	415	594	865	4798
Aligned	NegEps	0.5	492	808	806	129	437	633	863	1376	8319
Aligned	NegEps	0.25	467	1545	1096	381	833	1238	1915	2768	9241

(since here we are restricting attention to those simulations that stopped because of having converged, either to A or to B). We can also see that Bala-Goyal simulations are again (on average) the fastest (having the lowest mean and median), both amongst simulations converging to A and amongst simulations converging to B.

We can also compare the steps for simulations converging to A vs B within a given batch of simulations (i.e. collection of simulations run with the same parameter settings), checking for significance with a Mann-Whitney U-test. We did this, and found that the B-converged simulations were significantly ($p < 0.05$) faster (on average) than the A-converged simulations in all cases.¹⁷

Accuracy and proportions. When attempting to determine the effects of the presence, type, and level of mis- and disinformation, and of the adoption of different strategies for coping with it in light of higher-order evidence about it, one question we can ask is: how accurate is the simulated community of inquirers? And one way of measuring this is by asking: when the community arrives at an answer, in what proportion of cases is that answer correct?¹⁸

Thus, in the Bala Goyal model simulations that we ran, we found that all 500 ran to completion (i.e. converged one way or the other), with the community converging (incorrectly) to A in 11 cases, and to (the correct verdict) B in 489 cases. In short, these communities with no unreliable agents (and so no misinformation) got the correct answer 97.8% of the time (when they got an answer at all).

Table 4 Proportions of converged simulations converging to B.

Model	Misinfo	Reliability	Count	Bs	As	Proportion
Bala-Goyal	None	1	500	489	11	97.8%
Gullible	Binomial	0.75	500	486	14	97.2%
Gullible	Binomial	0.5	500	431	69	86.2%
Gullible	Binomial	0.25	500	349	149	70.0%
Aligned	Binomial	0.75	500	487	13	97.4%
Aligned	Binomial	0.5	500	477	11	97.7%
Aligned	Binomial	0.25	500	443	14	96.9%
Gullible	NegEps	0.75	500	436	64	87.2%
Gullible	NegEps	0.5	500	194	293	39.8%
Gullible	NegEps	0.25	500	9	483	1.8%
Aligned	NegEps	0.75	500	488	12	97.6%
Aligned	NegEps	0.5	500	492	8	98.4%
Aligned	NegEps	0.25	500	467	31	93.8%

Table 4 shows, for each batch of (500) simulations that we ran, the number that converged to B and the number that converged to A, as well as the proportion of those that converged which converged to B. (Some simulations ran for 20,000 steps without all nodes having credence above 0.99 or all nodes having credence less than 0.5, so that the number A + B is not always equal to 500.)

We tested whether the proportions differed significantly from that in the basic Bala-Goyal model using the chi-squared test. We found that when the aligned strategy was used in the presence of binomial misinformation, there was no significant difference in

the proportion of the simulations that converged which arrived at the correct answer than in the basic case of the Bala-Goyal model; and in the presence of negative epsilon disinformation, there was only a significant difference ($p = 0.00$, $\chi = 9.05$) when reliability in the network was at its lowest (0.25). Even then, the community arrived at the correct answer B in 93.8% of cases, compared with 97.8% when all agents were both reliable and fully trusting. It seems the aligned strategy copes well when it comes to accuracy in the long run.¹⁹

By contrast, when we look at the performance of the gullible strategy on this measure of accuracy, we see quite a different picture. In particular, there was a significant difference ($p < 0.05$) in all cases except when the reliability in the network was 0.75 and the unreliable agents produced binomial misinformation; and in those (other) cases, the proportion of the simulations in which the community achieved the correct consensus was less in the presence of mis- or disinformation than in the basic (Bala-Goyal) simulations. Indeed, in the presence of negative epsilon disinformation, the community's ability to discern the truth collapsed completely when reliability was low, with just 1.8% of simulations converging to B when reliability was 0.25.

Thus, we can already conclude that: (A1) in the long run, the aligned strategy of accommodating the higher-order evidence of unreliability did not fare significantly differently, in terms of accuracy, than when the informational environment was pristine (with no unreliability in the network) - except in the most extreme of cases; and (A2) the gullible strategy of denying the import of the available higher-order evidence fared worse in all but the most mild of misinformative environments than in the absence of mis- and disinformation. And indeed, a direct comparison of the two strategies for coping with unreliability in the light of the higher-order evidence available shows that (3) the gullible strategy fared significantly ($p < 0.05$) worse, by this (proportion) metric (of accuracy), in all cases except that of binomial misinformation with reliability 0.75, where the difference (97.2% accurate for gullible vs 97.4% for aligned) was not significant. Figure 1 visualizes the underlying data (from Table 4), allowing us to confirm all of this at a glance.

Efficiency and steps. Accuracy in the long run, however, is not the only measure we can use to assess the effects of mis- and disinformation on opinion formation in a community of rational agents, or the relative effectiveness of the two strategies for responding to higher-order evidence in their presence. One thing that also matters when trying to work out what to think is how long it takes to arrive at the truth (and so, whether we arrive at the truth in a timely manner). Thus we can ask: in those cases in which the community converges to B, how many steps does it

take to do so? This provides a measure of the efficiency of the information processing strategy employed within the community under various informational conditions. And when we look to this metric, we see a somewhat different picture.

More specifically, using the Mann-Whitney U-test for significance, we find the following.²⁰ (E1) When information consumers are gullible, the presence of (binomial) misinformants significantly increases the number of steps required to converge to the truth; and the more misinformants there are, the more steps it takes on average. The presence of (negative epsilon) disinformants likewise increases the number of steps needed to converge to the truth. (E2) When information consumers are sceptical, and align their level of trust with the level of reliability in the networks, again, the presence of binomial misinformants significantly increases the number of steps required to converge to the truth; and the more misinformants there are, the more steps on average it takes. The same holds for the presence of (negative epsilon) disinformants. (E3) We can also report that, in the presence of binomial misinformants, where there was a significant difference in the number of steps to converge to the truth between the gullible and aligned strategies, the (cautious/sceptical) aligned strategy was (on average) slower (i.e. had a larger mean). This was also true in the presence of (negative epsilon) disinformation - except for one anomaly (reliability 0.5). Figures 2 and 3 visualize these efficiency findings, alongside comparable data for simulations converging to A, for binomial misinformation and negative epsilon disinformation respectively.

In sum, then, the presence of mis- and disinformation significantly increases the number of simulation steps it takes for our communities of agents to arrive at the truth (no matter what information processing strategy they adopt); but when we compare strategies, we see that the aligned strategy (which accommodates higher-order evidence) is typically (with one exception) slower, or less efficient, in arriving at the truth than the gullible strategy (which denies the import of higher-order evidence).

Simulations on a large, real-world network

Thus far, we have been discussing simulations run on small (size 64), artificially generated (complete) networks of agents. But our code allows us to import graph representations of real-world networks.²¹ We ran a number of simulations on one such network, the EU Email Core network (Leskovec & Mcauley (2012)). This is a network based on emails sent within an EU research institution, with 1005 nodes in total (see Fig. 4). The network is directed (with 24,929 edges), and information (i.e. first-order evidence) in our simulations flows in the same direction that emails were sent in the original network.

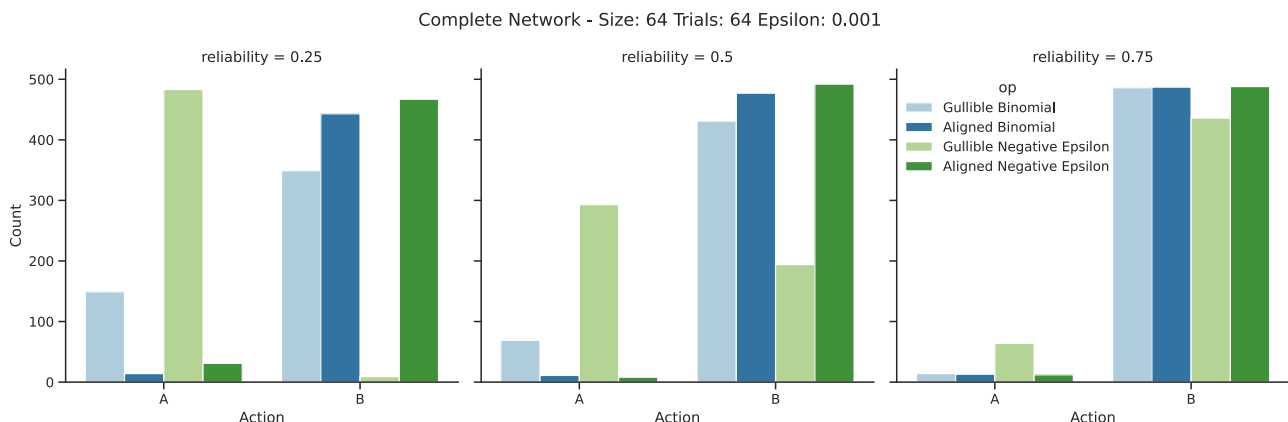


Fig. 1 Numbers of simulations (out of 500) converging to action A and to B for each information processing strategy at each network reliability level.

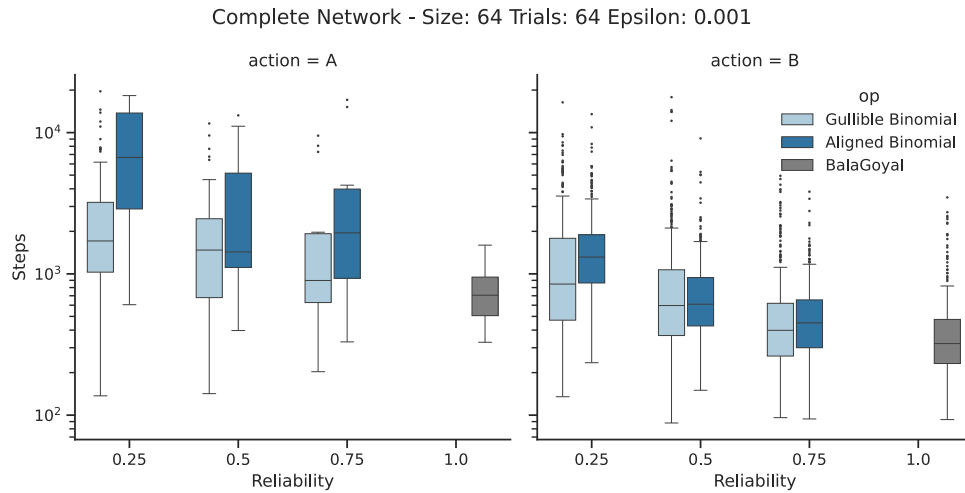


Fig. 2 Steps to converge to A and to B in binomial misinformation simulations, with both gullible and aligned strategies. (Note the log scale on the y-axis).

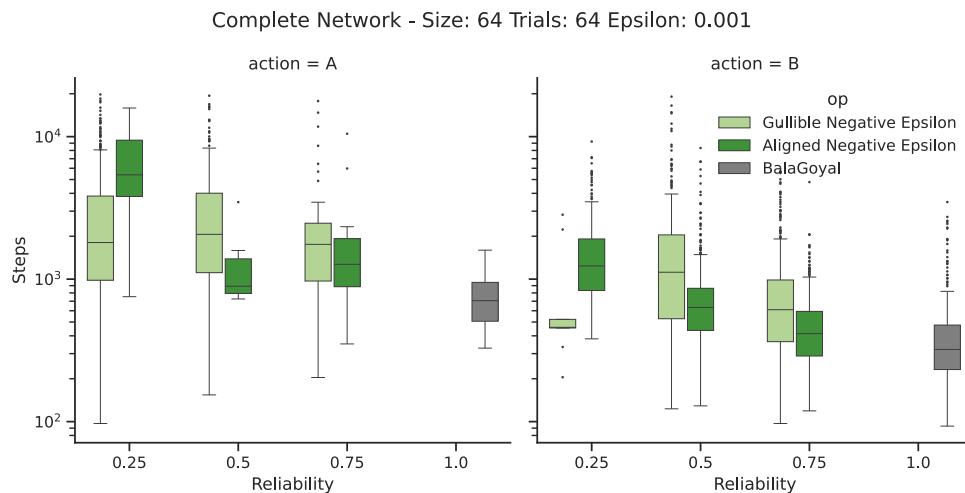


Fig. 3 Steps to converge to A and to B in negative epsilon disinformation simulations, with both gullible and aligned strategies. (Note the log scale on the y-axis).

Since simulations on such (relatively) large networks take considerable time to reach a consensus opinion, we ran just 10 simulations for each parameter configuration of interest, and capped the number of steps at 25,000. As a result, our data cannot be analyzed using the methods above: in particular, we cannot measure accuracy in terms of proportions of simulations converging to the truth; and we cannot measure efficiency in terms of the number of steps required to do so. But we can observe how the average (i.e. mean) credence in the network evolves over time - which we do here. In particular, we explore the effects on this metric of (a) rising general levels of (binomial) misinformation when agents in the network pursue the gullible (or trusting) strategy, and (b) structurally targeted (binomial) mis- and (negative epsilon) disinformation on such communities.

First, then, as indicated, we consider the effect of (binomial) misinformation in the network with agents pursuing the gullible strategy. When reliability in such a network is set to 1.0 - so that 100% of agents are reliable - the resulting model is equivalent to that of Bala and Goyal. This sets a kind of benchmark - and as we can see in Fig. 5, the average credence in each of the 10 simulations proceeds quickly upwards in this case, beginning at (approximately) 0.5, reaching 0.8 well within 5000 steps, and 0.9 at around 10,000 steps. (These and the claims that follow can also be corroborated through an examination of Table 5.²²) As we

introduce more and more unreliable agents, however - so that reliability reduces to 0.75, to 0.5, to 0.25 - we see that average credence in the network increases more slowly.²³ For instance, with reliability at 0.75, the average credence only reaches (roughly) 0.9 after more than 15,000 steps; and it does not typically reach this level within 25,000 steps at lower reliability levels. With reliability at 0.5, the average credence reaches 0.8 only well after 10,000 steps; and with reliability at 0.25 it typically does not do so within the first 25,000 steps. In short, the presence of misinformation reduces the community's confidence in the truth for a considerable period of time.²⁴

Second, we turn to briefly explore one further aspect of the EU Email Core network - namely, its structure. As indicated above, the artificial networks we have used in our small scale simulations have been complete networks: every agent is connected to every other. But real-world networks are not typically complete - and the EU Email Core network is no exception (not everyone who sent an email sent one to everyone else who did). Accordingly, some agents (or nodes) in the network are more connected - and in this sense, more centrally located within the network - than others. There are various metrics of the centrality of a node: here we focus on (out) degree centrality; this is simply the (out) degree of the node (i.e. number of edges originating from it), divided by the (out) degree it would have in the complete network of the

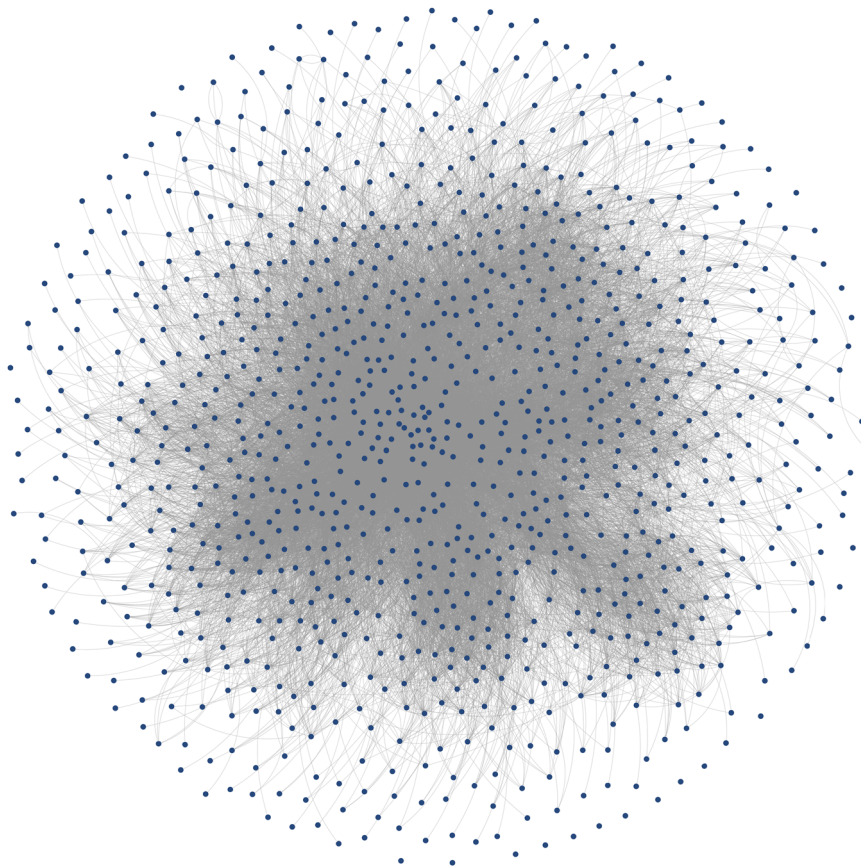


Fig. 4 The EU Email Core network.

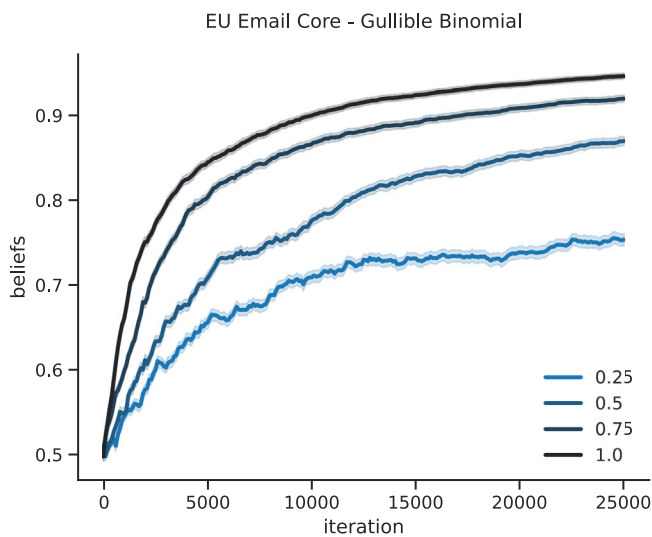


Fig. 5 The average credence in the EU Email Core network with Binomial misinformation under the gullible strategy. (The Bala-Goyal model is represented here as involving the gullible information processing strategy in a network with reliability 1.0).

same size (i.e. the maximum possible out degree for a node in a network of that size). Figure 6 shows the distribution of (out) degree centrality within the network.²⁵

Using this measure, we ran simulations in which we selected the 10 most central nodes in the network and made them unreliable. In 10 of our simulations, these unreliable nodes

Table 5 The minimum, median, and maximum numbers of logged steps required for the EU Email Core community of 'gullible' agents to reach various average credence thresholds with different levels of reliability in the presence of (binomial) misinformation.

Reliability	Statistic	0.6	0.7	0.8	0.9
1	Min	300	1000	2400	8400
1	Med	550	1250	3200	10200
1	Max	1000	1900	3600	11700
0.75	Min	500	1000	3000	11300
0.75	Med	900	2000	5000	18450
0.75	Max	2400	3600	6100	24300
0.5	Min	500	2700	6500	23100
0.5	Med	1350	3800	11300	<NA>
0.5	Max	3900	10000	19300	<NA>
0.25	Min	1100	4400	17900	<NA>
0.25	Med	2200	6700	<NA>	<NA>
0.25	Max	12000	<NA>	<NA>	<NA>

acted as (binomial) misinformants, while in 10 others, they acted as (negative epsilon) disinformants. In all cases, agents used the gullible information processing strategy (the aligned strategy would hardly discount the evidence at all, with less than 1% of agents unreliable). The results - in terms of average (mean) credences over time - are depicted in Fig. 7, with Bala-Goyal simulation results also shown as a baseline for comparison.

As can be seen, average credence increased more slowly in the simulations in which mis- or disinformants were present than in

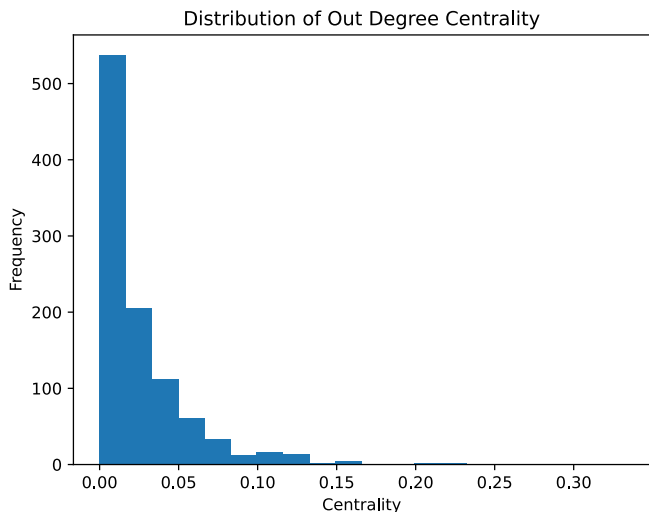


Fig. 6 Out degree centrality distribution in the EU Email Core network.

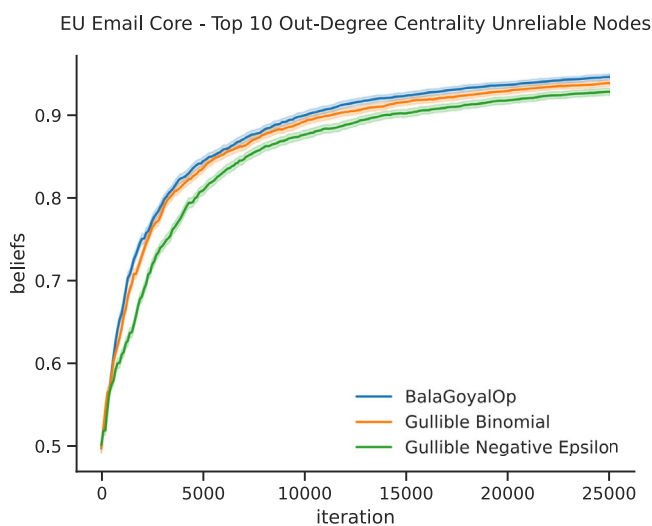


Fig. 7 Average credences in the EU Email Core network with just 10 unreliable central nodes. The cases of binomial misinformation and negative epsilon disinformation are depicted alongside the base case of the Bala-Goyal model.

those in which they were not, with disinformation leading to a greater suppression of confidence in the truth than misinformation.²⁶ Indeed, the summary statistics for these simulations given in Table 6 reveal that the median number of logged steps required to reach a given (mean credence) threshold that we observed was as much as 27% greater (0.6 threshold) than in the Bala-Goyal model with just 10 central (binomial) misinformants in the network, or 60% greater (0.7 threshold) with (negative epsilon) disinformants occupying these central nodes. This compares with up to 81% greater (0.9 threshold) when approximately 25 times as many randomly distributed nodes were unreliable (binomial) misinformants (see Table 5). That these effects are of the same order of magnitude (while the number of unreliable nodes is not) is in itself highly suggestive - both of the dangers associated with structurally sensitive sources of mis- and disinformation, and of the measures that might be taken to improve our informational environment by targeting such sources.²⁷ But further research is required in order to generalize the limited observations we have been able to make here.

Table 6 The minimum, median, and maximum numbers of logged steps required for the EU Email Core community of 'gullible' agents to reach various average credence thresholds with just 10 degree central nodes spreading different types of misinformation.

Misinfo	Statistic	0.6	0.7	0.8	0.9
None	Min	300	1000	2400	8400
None	Med	550	1250	3200	10200
None	Max	1000	1900	3600	11700
Binomial	Min	200	900	2600	9600
Binomial	Med	700	1500	3350	10950
Binomial	Max	1100	2200	4200	15000
NegEps	Min	200	900	2900	9700
NegEps	Med	850	2000	3700	13350
NegEps	Max	2300	4900	7300	18200

Concluding discussion

In this paper, we have explored the effects of introducing mis- and disinformation into a networked community of rational agents. We have assumed the agents are aware of, or have (general) higher-order evidence concerning, the level of unreliability in the information (or first-order evidence) at their disposal, and we have considered two strategies they might employ in this case: they might adopt the (knowingly) 'gullible' strategy of fully trusting the information they receive from their network neighbours (thereby, in effect, denying the import of the higher-order evidence available to them); or they might 'align' their level of trust in the evidence they receive to the level of reliability in the network (thereby accommodating the higher-order evidence at their disposal). Our investigation has involved the analysis and interpretation of data generated in computer simulations run on both small (64 node) artificial networks, and a larger (1005 node) real-world network.

We have found (in our experiments on small, complete networks) that the presence of misinformation significantly increases the amount of time (measured in terms of simulation steps) required for the community to converge to the truth, no matter which information processing strategy they pursue (though simulations in which the aligned strategy was deployed typically took longer than those in which the gullible strategy was used). At the same time, when it comes to the accuracy of the community, measured in terms of the proportion of simulations converging to the truth, the aligned strategy did better than the gullible strategy - especially in the presence of disinformation.

There remains much more work to be done investigating the interactions between levels and types of misinformation on the one hand, and ways of accommodating higher-order evidence about it on the other. A systematic investigation of these interactions within networks of different shapes (or topologies) might be pursued, for example. And other metrics of performance might be explored - for instance, not whether a correct consensus is achieved within the community, and in how many steps, but whether e.g. a two thirds majority is achieved, and how long that takes; or whether the most central nodes have been convinced, and in how many steps. But our initial studies have already shown how misinformation can adversely affect public opinion even within a community of rational agents, and that it is far from obvious what the best way of individually responding to its presence is.

In particular, Zollman (2007) found that, when it comes to the communicative structure of a group of agents engaged in rational inquiry, there is a trade-off between accuracy and efficiency: more connected (denser) networks are quicker to arrive at the truth (i.e.

more efficient) than less connected (sparser) ones, but less likely to do so at all (i.e. less accurate). Our results suggest that there is another aspect to this trade-off: when confronted with higher-order evidence of misinformation, a more cautious, ‘aligned’ information-processing strategy will be more accurate than a more ‘gullible’ one (and so, in this sense, will produce fewer errors), but it will be less efficient (and so be less successful in grasping the truth within a reasonable time frame). Indeed, William James (1896) once drew to our attention a fundamental issue in the ethics of belief: whether to employ sceptical belief forming practices so as to avoid error, or to be less cautious in the hope of grasping the truth in a timely manner.²⁸ This problem, it seems, is still with us.

Indeed, if we are not mistaken, this fact has a broad bearing on the epistemological debate surrounding higher-order evidence: for it seems that, even if we resolve to assess what rationality requires of us in broadly consequentialist terms, we will still need to choose a dimension on which to assess the epistemic consequences of the adoption of the different strategies. Should we seek to arrive at the truth (relatively quickly), or to avoid error? But our results also appear to shed some light on the problem of misinformation: for, if they are indicative, no matter what rational strategy individual agents pursue for coping with the presence of misinformation in their environments, they and their communities will suffer adverse epistemic consequences of one sort or another.

Nevertheless, not all the news is bad. Turning to the large (1005 node) real-world (email communication) network we investigated, we found that (when pursuing the gullible strategy) the community took longer to increase its credence in the correct opinion, not only as the level of misinformants distributed (uniformly) at random within it increased, but even when just a small number (10) of central nodes were unreliable purveyors of mis- and disinformation. While this reinforces the concern that degradation of the informational environment may have serious negative social epistemological consequences, it also hints at the possibility that structurally targeted, community level (global) responses (of a sort that could be pursued by large tech platforms) may prove effective in tackling that concern.

Data availability

The source code for this study is available in the [GitHub repository](#) for the PolyGraphs project.

Received: 28 March 2024; Accepted: 16 September 2024;

Published online: 28 September 2024

Notes

- There is some dispute over how best to think of the epistemological issues in this general vicinity. Dorst (2019), for instance, argues in favour of reframing the discussion: instead of distinguishing first-order and higher-order evidence, he claims, we should instead consider how our total evidence bears on our first-order and higher-order opinions. In introducing the discussion as we have done, we do not intend to take sides on this controversial issue. Indeed, following Christensen (2010) we are happy enough to identify the target phenomena through examples, as below, and trust that they bear sufficient similarity to those under discussion elsewhere in the literature to be of interest. What is crucial from our point of view, however, is that the type of situation we are considering bears on the question of what it is rational for agents to do in response to the general (‘higher-order’) evidence that is available to them in the situations we model.
- Avnur (2020) and Levy (2023) also discuss higher-order evidence in relation to misinformation, though they do not attempt to shed light on the higher-order evidence debate itself, as we shall do here, instead focussing on whether agents in echo chambers have higher-order evidence that rationalizes their beliefs. (For what it

is worth, our models treat individual agents as rational - at least, according to a certain conception - and so align perhaps more closely with Levy’s view here than Avnur’s. But this modelling assumption is, first and foremost, an idealization which allows us to explore what would happen to communities of agents that were rational in this way, whether or not they are as a matter of fact.)

- Of course, the manner in which we do so is not the only possible one - we need to make some auxiliary assumptions in order to make progress. But we or others can hope to refine those assumptions in due course, in part in light of the results we obtain here with their aid.
- One additional benefit of the approach taken here is that it may highlight strengths and weaknesses of this broadly Bayesian conception of rationality. For discussion of this cluster of views, and what is perhaps its main rival, on which belief does not come in degrees, see e.g. Sturgeon (2020).
- This is in line with the subjective Bayesian perspective that any prior probability distribution is a rational credence, so long as it is internally coherent. In future work we hope to explore the effects of initialization.
- As a quick reminder, Bayes’ rule is as follows (where P_f is the ‘final’ or ‘posterior’ and P_i the ‘initial’ or ‘prior’ probability):

$$P_f(h) = P_i(h|e) = \frac{P_i(e|h)P_i(h)}{P_i(e)} \quad (1)$$

- Jeffrey’s rule is a generalization of Bayes’ rule (see previous note), which states that:

$$P_f(h) = P_f(e)P_i(h|e) + P_f(\neg e)P_i(h|\neg e) \quad (2)$$

- Specifically, their no anti-updating rule for determining the final probability of the evidence based on the distance d between the beliefs of the agents involved is as follows (where m is a ‘mistrust multiplier’):

$$P_f(e) = 1 - \min(1, d \cdot m)(1 - P_i(e)). \quad (3)$$

- Of course, real misinformants might have plenty of unconscious biases. We do not regard the present modeling assumption of neutrality as definitional of mis- as opposed to disinformation. But we need to begin our investigations somewhere, and this assumption is not unmotivated. Future work may explore alternative approaches.
- Weatherall et al. (2020), for example, develop models on which ‘propagandists’ practice selective reporting of the (otherwise accurate) results that are available to them.
- Kelly (2005) advocates a view along these lines in relation to the question of what the rationally appropriate response to peer disagreement is.
- Of course, if evidence is factive (cf. Williamson (2000), this suggestion is mistaken for those agents who are in fact unreliable. In future work we plan to investigate a refinement of the present model that accommodates this point.
- After all, should we discount what you say, simply because we know that some people are liars?
- In particular, where r is level of reliability in the network (i.e. the probability that a given agent offers reliable testimony), we can implement Jeffrey’s rule by setting

$$P_f(e) = r. \quad (4)$$

- Our simulation framework contains a large number of hyperparameters, some of which are pertinent to the simulations themselves, others to such computational questions as what data gets stored in memory, and where. Amongst the former, some are only pertinent once the values of others are set in particular ways. In the present context, the key choices concern: network kind, and size; the way in which credences are initialized; what the stopping conditions are for a simulation - all nodes with credence above 0.99, or below 0.5, or the maximum number of simulation steps is reached; what the size of the bias epsilon is; how many trials agents conduct when they think it worth experimenting; what operation is performed at each step - which includes whether there are any unreliable agents, and if so what the level of reliability is in the network (i.e. what the probability is that each node is reliable), and what kind of sampler (binomial or negative epsilon) is used by any unreliable agents, as well as how the agents respond to the evidence they receive, i.e. whether they update using Bayes’ rule, or Jeffrey’s rule operationalized by alignment. Obviously, this leads to a vast parameter space - but thankfully, we can selectively explore it in a useful manner by drawing on previous work (e.g. Rosenstock et al. (2017)) to identify a likely region where interesting results may be found.
- We took a consensus that B is better to have been reached if all agents had a credence above 0.99, and a consensus that B is not better but A is if all agents had a credence below 0.5, so that no new evidence was generated.
- Note that this suggests that, amongst the simulations that reached 20,000 steps without converging to A or B, the proportion that would have gone on to converge to A is likely to have been higher than in the sample of those that converged in less than 20,000 steps. We do not think this to be of huge importance in relation to the analysis we conduct below, but mention it here as interesting in its own right.
- Obviously, since we are restricting attention to cases in which an answer was obtained, in the remaining cases an error will have been made.

- 19 It is worth noting that there may be some bias in the sample: given that there is a (significant) difference in the number of steps required to converge to A vs B, the fact that we have looked only at the simulations that converged one way or another within 20,000 steps may mean that the proportions in the long run would be different if no simulations were cut short.
- 20 We consider a finding significant when $p < 0.05$.
- 21 We are aware that talk of 'real-world' networks may be regarded as overly simplified. Those who harbour this concern can imagine scare-quotes on the phrase 'real-world' throughout - though we note that in the present instance, email communications were in fact sent as indicated in the network graph, so that we have an accurate (if partial) representation of the communications in the underlying social network/community.
- 22 We logged the state of the network every hundred steps in our simulations. As can be seen, with reliability at its maximum level (1.0), the minimum number of logged steps needed to reach the threshold average credence of 0.8 was 2400, the median was 3200, and the maximum was 3600. To reach an average credence of 0.9, the minimum logged number of steps needed was 8400, the median 10,200, and the maximum 11,700. Similar observations substantiate the points that follow.
- 23 The distributions of credences over time across the various reliability settings were found to be (pairwise) significantly ($p < 0.05$) different using the Kolmogorov-Smirnov test.
- 24 Thus, by way of indication, we observed: (i) an increase of at least a 50% in the number of steps taken to reach each average credence threshold when reliability decreased from 1 to 0.75; (ii) increases of at least 2 times, and as many as 5 times, needed to reach the 0.6 and 0.7 thresholds when comparing even lower levels of reliability with the (full reliability) Bala-Goyal model; and (iii) the typical (median) simulation did not reach the 0.9 threshold (within 25,000 steps) when reliability was 0.5, and did not even reach the 0.8 threshold when reliability was 0.25 - despite doing so in just over 10,000 steps in the absence of misinformation. This last point is perhaps of particular interest, given that these high thresholds are often considered necessary for belief (and therefore action) - at least when it comes to individual agents. In a case of societal interest - such as forming a belief on a matter needed to take urgent climate action - such indefinite delays could prove catastrophic.
- 25 As the EU Email Core network is directed - like networks in which edges represent channels of testimonial communication more generally - not all pairs of nodes are connected (in a given direction) by a path. Accordingly, we cannot determine average shortest path length, nor therefore ascertain whether this network of email communications has the so-called 'small-world' property Watts & Strogatz (1998). This said, the average clustering coefficient (0.366) for the network is considerably greater than the density (0.025). Similarly, we have not sought to determine whether the network is 'scale free' as Barabasi and Albert (1999) understand this notion, with its (out) degree distribution following a power-law; but as can be seen in Fig. 6, it does have a heavy tail (i.e. many nodes with relatively low degree centrality), which has been claimed by e.g. Holme (2019) to be the more pertinent network feature. In these respects, then, our chosen real-world communication network is typical of social networks more generally.
- 26 Again, using the Kolmogorov-Smirnov test, we found the differences between configurations in the distributions of credences over time to be significant ($p < 0.05$).
- 27 For example, we have heard it suggested that social media companies cannot be legally regarded as publishers of the contents posted to their sites by their users, on the grounds that having editorial oversight of this volume of material in real time is simply not feasible. And yet perhaps they could be regarded as publishers of the contents posted by their most influential users, setting terms and conditions of use accordingly. Our (admittedly very preliminary) findings suggest that if such changes were effective in eliminating unreliability amongst the influential nodes, this might well have large social epistemic benefits.
- 28 One of James' examples involved the question of whether to propose marriage: and he suggested that waiting indefinitely in the hope of accumulating more decisive evidence in favour would certainly result in a missed opportunity; the plausibility of the case is due, of course, to the time-sensitivity of the issue.

References

- Avnur Y (2020) What's wrong with the online echo chamber: A motivated reasoning account. *J Appl Philos* 37(4):578–593
- Bala V, Goyal S (1998) Learning from neighbours. *Rev economic Stud* 65(3):595–621
- Ball, B., Koliouisis, A., Mohanan, A., and Peacey, M. (forthcoming). Ignorance in social networks: discounting delays and shape matters. In Arnold, M., Herrman, M., Kaminski, A., Resch, M., and Wiengarn, J. (Eds.), *Trust and Disinformation*, Springer
- Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512
- Burge T (1993) Content preservation. *Philos R review* 102(4):457–488
- Cairncross, F. *The Cairncross Review: a Sustainable Future for Journalism*. Department for Digital, Culture, Media, and Sport. Available at: <https://www.gov.uk/government/publications/the-cairncross-review-a-sustainable-future-for-journalism> (2019)
- Christensen D (2010) Higher-Order Evidence. *Philos Phenomenological Res* 81(1):185–215
- Coady, C.A.J. *Testimony: A philosophical study*. Oxford: Clarendon Press (1992)
- Dorst, K. Higher-order uncertainty. In M. Skipper & A. Steglich-Petersen (Eds.), *Higher-order evidence: New essays*, Oxford University Press, 35–61 (2019)
- Feldman R (2005) Respecting the evidence. *Philos Perspect* 19:95–119
- Fitelson B (2012) Evidence of evidence is not (necessarily) evidence. *Analysis* 72(1):85–88
- Holme P (2019) Rare and everywhere: Perspectives on scale-free networks. *Nat Commun* 10(1):1016
- Horowitz, S. Higher-Order Evidence. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition. Available at: <https://plato.stanford.edu/archives/fall2022/entries/higher-order-evidence/> (2022)
- James W (1896) The will to believe. *N. World* 5:327–347
- Kelly T (2005) The epistemic significance of disagreement. *Oxf Stud Epistemology* 1:167–196
- Leskovec, J. & Mcauley, J. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25 (2012)
- Levy N (2023) Echoes of covid misinformation. *Philos Psychol* 36(5):931–948
- O'Connor C, Weatherall JO (2018) Scientific polarization. *Eur J Philos Sci* 8(3):855–875
- Reid, T. *Inquiry and Essays*. Beanblossom, R.E. and Lehrer, K. (Eds.), Indianapolis: Hackett (1983)
- Rosenstock S, Bruner J, O'Connor C (2017) In epistemic networks, is less really more? *Philos Sci* 84(2):234–252
- Skipper, M. and A. Steglich-Petersen *Higher-order evidence: New essays*. Oxford University Press (2019)
- Sturgeon, S. *The Rational Mind*. Oxford University Press (2020)
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442
- Weatherall, J., C. O'Connor, and J. Bruner How to beat science and influence people: Policymakers and propaganda in epistemic networks. *The British Journal for the Philosophy of Science* (2020)
- Williamson, T. *Knowledge and its limits*. Oxford University Press (2000)
- Zollman KJS (2007) The communication structure of epistemic communities. *Philos Sci* 74(5):574–587

Acknowledgements

This work was supported by the British Academy, the Royal Academy of Engineering, the Royal Society, and the Leverhulme Trust, under the APEX Award scheme, grant number APX\R1\211230.

Author contributions

Brian Ball guided the philosophical approach of this article. Alexandros Koliouisis developed the simulation framework. Amil Mohanan ran simulations and analyzed data. Mike Peacey provided the statistical methodology. All authors made substantial contributions to the conception or design of the work and/or the acquisition, analysis, and interpretation of the data. They contributed to drafting the work or revising it critically for important intellectual content and have given final approval of the version to be published. They agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Ethical approval

The research for this article did not involve any studies with human participants performed by any of the authors.

Informed consent

Informed consent was not required, as no human participants were involved in the research for this article.

Additional information

Correspondence and requests for materials should be addressed to Brian Ball.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024