Synthese (2024) 203:46
https://doi.org/10.1007/s11229-024-04491-5

ORIGINAL RESEARCH

Rational factionalization for agents with probabilistically related beliefs

David Peter Wallis Freeborn[1]

Received: 9 September 2023 / Accepted: 5 January 2024 / Published online: 29 January 2024

Abstract
General epistemic polarization arises when the beliefs of a population grow further apart, in particular when all agents update on the same evidence. Epistemic factionalization arises when the beliefs grow further apart, but different beliefs also become correlated across the population. I present a model of how factionalization can emerge in a population of ideally rational agents. This kind of factionalization is driven by probabilistic relations between beliefs, with background beliefs shaping how the agents' beliefs evolve in the light of new evidence. Moreover, I show that in such a model, the only possible outcomes from updating on identical evidence are general convergence or factionalization. Beliefs cannot spread out in all directions: if the beliefs overall polarize, then it must result in factionalization.

Keywords Polarization · Factionalization · Bayesian networks · Network epistemology · Social epistemology · Philosophy of science

1 Introduction

Epistemic polarization arises when a population's beliefs about some hypothesis grow further apart. This is sometimes operationalized as an increase in the spread or dispersion of the belief across the population (for example, see Bramson et al., 2017; DiMaggio et al., 1996; Freeborn, 2023, 2024a, 2024b; Madsen et al., 2018; Pallavicini et al., 2021). For example, suppose that most of a population are very unsure about the safety of vaccines. If this belief polarizes, then more people might become very sure

✉ David Peter Wallis Freeborn
dfreebor@uci.edu

[1] Department of Philosophy, Northeastern University London, Devon House, 58 St Katharine's Way, London E1W 1LP, UK

that vaccines are safe, more people might become very sure that vaccines are unsafe, and fewer people may be left highly unsure.[1]

However, we are often interested in agents who hold many different beliefs, and in how those beliefs might be related. For instance, different polarized beliefs might also become more closely correlated. Epistemic factionalization arises when *multiple*, different beliefs become correlated in a population of agents (see Bramson et al., 2017; Kawakatsu et al., 2021; Levin et al., 2021; Weatherall & O'Connor, 2021). For example, suppose that some population's beliefs about vaccination efficacy and anthropogenic climate change have both polarized. However, perhaps the same people who are skeptical about vaccine efficacy also tend to be skeptical about anthropogenic climate change, whilst those who strongly believe that vaccines are effective also tend to believe in anthropogenic climate change. Then, if I know that someone is highly skeptical about anthropogenic climate change, this could give some degree of evidence that they might also be skeptical of vaccines.[2] This would be a case of factionalization.

Perhaps such factionalization could be driven by the relationships between different beliefs. Consider that proposed correlation between skepticism about anthropogenic climate change and skepticism about vaccines. At first glance, these might seem like unrelated beliefs, pertaining to two very different fields, climate science and medicine. However, these beliefs might be related by an underlying belief, perhaps regarding the trustworthiness of scientists or scientific institutions. If someone regards scientific institutions as generally reliable, this could drive them to accept scientific results about both anthropogenic climate change and vaccines. On the other hand, if someone regards scientific institutions as generally unreliable, this could drive skepticism about both anthropogenic climate change and vaccines.

Previous research has already shown how underlying background beliefs can drive rational polarization of individual beliefs (see Freeborn, 2023, 2024a, 2024b; Jern et al., 2014). In this paper, I demonstrate how factionalization can arise even for populations of ideally rational agents who have probabilistic relations between their beliefs.

To do this, I will assume that the agents are as similar as possible, sharing the same probabilistic relationships between their beliefs, and updating on the same evidence, differing only in their initial degrees of belief about various hypotheses. I show how patterns of factionalization spontaneously emerge due to the probabilistic relations between beliefs themselves. One can think of this model as explicating one particular kind of factionalization—arising due to certain underlying background beliefs, worldviews or ideologies shaping how the agents' beliefs evolve in the light of new evidence.

The paper is structured as follows. In Sect. 2, I outline a general model for representing a population of agents with multiple beliefs, which could undergo factionalization. I also outline some of the formalism that I will use throughout the rest of the paper. In Sect. 3, I suggest three different approaches for operationalizing "factionalization", "convergence" and "general divergence" within this model. In Sect. 4, I present three

[1] For one recent empirical study with similar findings to this, see Lee and Sibley (2020).

[2] Indeed, some studies suggest that beliefs about vaccines and climate change may in fact be correlated within the U.S. population (Hamilton et al., 2015; Latkin et al., 2022).

simple examples of belief networks, one that leads to convergence and two that lead to factionalization. I explain whether and how convergence, polarization and factionalization arise in each case. In Sect. 5, I explain why factionalization must arise when agents' overall beliefs polarize: general divergence never arises.

2 General model

To talk about factionalization more concretely, it will help to have a basic model of a population in mind. This model will include only certain minimal necessary features for factionalization to emerge.[3] My aim is to distill one particular form of factionalization that emerges due to the relationships between beliefs.

This model is highly idealized, but it will be helpful to have a concrete real-world picture in mind. The model might represent a population, accumulating exactly the same evidence about some particular hypotheses, and updating their beliefs about many other hypotheses on this basis. For instance, we might imagine a subset of the general public reading a series of newspaper articles about the a particular Covid-19 vaccine. From this evidence, each population member might update many other (more or less closely related) beliefs: about the efficacy of vaccines in general, about the reliability of scientists, or about whether humans cause anthropogenic climate change, and so forth.

I assume a finite population of agents. I assume that there is a set of hypotheses or propositions describing the world or some system within it, each of which can be true or false, represented by discrete, binary random variables.[4] Each agent holds a degree of belief, a probability, about each hypothesis. The agents can have conditional probabilities relating pairs of different beliefs. However, I assume that all the agents agree about each of the conditional relations between beliefs: any disagreement comes down to disagreements about the hypotheses themselves.

To represent relations between beliefs, I use the formalism of Bayesian networks (see Sect. 2.1). A Bayesian network specifies a set of variables, representing hypotheses or propositions, and the conditional relationships between variables. Implicit in this model is that the agents are rational: all of their beliefs must be probabilistically consistent at each time, and upon learning any evidence, their beliefs are updated in a dynamically coherent way.[5]

[3] This simple model also allows for a very direct comparison with other recent models looking at polarization (Freeborn, 2023, 2024a, 2024b; Jern et al., 2014) as well as the formation of scientific paradigms (Grim et al., 2022a).

[4] This is for simplicity only, the analysis extends straightforwardly to discrete random variables more generally. However, requiring the variables to be discrete allows it to keep the analysis in Sect. 3.3 significantly simpler (see Lazo & Rathie, 1978).

[5] Recent work in philosophy of science has used Bayesian networks as tools to explicate webs of interconnected beliefs, paradigms, or scientific hypotheses (Dizadji-Bahmani et al., 2011; Grim et al., 2022a, 2022b; Hartmann & Bovens, 2002; Sprenger, 2017). Other research has already used Bayesian networks as a tool to study belief polarization (Freeborn, 2023, 2024a, 2024b; Jern et al., 2014).

2.1 Formalism of Bayesian networks

More formally, a Bayesian network is a graphical model that aims to capture some subset of the independence relationships given by a joint probability distribution (Pearl, 2009). Let $\mathcal{X} = \{X_1, X_2, \ldots X_N\}$ be a set of N random variables, defined on a probability space. Then, a joint probability distribution $P(X_1, X_2, \ldots X_N)$ gives the probability that each of $X_1, X_2, \ldots X_N$ falls within some range or a discrete set of values specified for that variable. A factorization of a joint probability distribution makes a choice about how variables depend upon others. Given some particular ordering of variables 1 to N, a factorized representation $P(X_1, X_2 \ldots X_N)$ takes the form,

$$P(X_1, \ldots X_N) = P(X_1 \mid X_2, \ldots, X_N) \times P(X_2 \mid X_3, \ldots, X_n) \ldots P(X_n). \quad (1)$$

$$P(X_1, \ldots X_N) = \prod_{i=1}^{N} P(X_i \mid X_1, \ldots X_{i-1}). \quad (2)$$

Each of the $N!$ factorizations of a joint probability distribution will correspond to a different Bayesian network. Let $\mathcal{G} = (\mathcal{V}, \mathcal{D})$ be a directed, acyclic graph, where V is a set of vertices (or "nodes"), and D is a set of directed edges, pointing from one vertex to another. In a directed, acyclic graph, these directed edges can never form a closed cycle. Nodes are associated with unique variables, and edges represent the conditional relations between different variables. A directed edge (X_a, X_b) exists in the network if $P(X_b, X_a)$ is a factor in the joint probability distribution. If there is a directed edge from node A to node B, we call A the "parent" and B the "child". Bayesian networks encode a series of *local Markov independence assumptions*. If the joint probability distribution factorizes with respect to a directed graph $\mathcal{G}$, then each variable in the joint probability distribution, associated with some node in the graph, is probabilistically independent of its non-descendants, given its parents (Geiger & Pearl, 1993; Pearl, 2009). So, we can fully specify a Bayesian network by a set of nodes, $\mathcal{V}$, directed edges, $\mathcal{D}$, random variables, $\mathcal{X}$, where there is a 1–1 map between the random variables and the nodes (I will often use the two interchangeably), and conditional probability distributions $P(X_i \mid X_{\text{par}_i})$, where X_{par_i} are the variables associated with the parents of X_i.

Bayesian networks can be updated on new evidence using upwards and downwards propagation procedures, such that the updated Bayesian network remains consistent with the axioms of probability theory. Downwards propagation involves a simple application of the specified conditional probabilities, upwards propagation involves a Bayesian inference procedure. In practice this requires a particular algorithm; in this case I use successive *variable elimination* (see Darwiche, 2009 for a comprehensive overview). Successive updating makes use of the *rigidity* assumption, that conditional probabilities of the form $P(X_i \mid X_j)$ do not change when X_j is updated (see Bradley, 2005; Diaconis & Zabell, 1982; Jeffrey, 1983).[6] The belief propagation process is

[6] Probability kinematics is a generalization of Bayesian updating for uncertain evidence in which the updating still obeys the rigidity condition.

governed by probability functions for each node which take as input the possible values of the parent nodes, and give as output the probability, or probability distribution, of the variable associated with the node.

2.2 Specification of the evidence

In this model, the agents update their beliefs based on accumulating evidence over time. So, I assume that the agents begin at some timestep 0, and the population evolves through T discrete timesteps. All agents receive *the same* evidence at each timestep, and then updates all of their beliefs in their belief network on the basis of this evidence.[7] I will assume that all the evidence, at every timestep, pertains to just one single belief, corresponding to one single node, let us call it the "data node".[8] However, the effects of updating this single belief will propagate through the network to other beliefs.

In order to explore the evolution of beliefs over time, I will look at successive updating on uncertain evidence.[9] Rather than the evidence determining that one of the hypotheses is definitely true or false (with probability 1 or 0), I will specify this as fixed likelihood evidence.

What does it mean for agents to receive the same likelihood evidence? In this case, I will represent that as receiving evidence with the same likelihood ratio. Following, Mrad et al. (2015), I define likelihood evidence η on a variable H of a Bayesian network, as evidence given by a likelihood ratio,

$$L(H = h_1) : \ldots : L(H = h_n) = P(\eta \mid H = h_1) : \ldots : P(\eta \mid H = h_n), \qquad (3)$$

where the $L(H = h_i)$ are likelihoods, representing the probability of the observed evidence, given that H is in the state h_i. This is a natural standard of "sameness" of evidence for several reasons. First, it allows the updating procedure to be commutative (see Field, 1978; Huttegger, 2015; Jeffrey, 1988; Wagner, 2002 for a philosophical discussion; see also Diaconis & Zabell, 1982; Mrad et al., 2015 for some mathematical considerations about the explication of uncertain evidence relevant to Bayesian networks). Second, the same likelihood evidence of this kind can also be thought of

[7] For reasons of simplicity, I do not consider network effects or information sharing in this paper. Every agent has access to exactly the same data. However, the interaction of network effects and belief networks suggests a promising avenue for further study.

[8] In this sense, the evidence that the agents obtain will be "incomplete" (see Freeborn, 2024b for a discussion of this point). The results in this paper do generalize to evidence received on multiple different beliefs. However, the other assumptions of this paper satisfy the Blackwell-Dubins assumptions about Bayesian merging (see Blackwell & Dubins, 1962; Huttegger, 2015; Kalai & Lehrer, 1994; Nielsen, 2018; Schervish & Seidenfeld, 1990), so were the agents receive the same sufficient evidence to settle *all* of their beliefs, then the agents' beliefs should converge. The kind of factionalization results I will discuss here are most relevant to the case where the information is insufficient to settle every belief—see Freeborn (2024b) for an argument that this is a reasonable assumption under a broad range of conditions.

[9] However, nothing in this analysis will depend on the use of uncertain evidence: the results also apply to the special case of agents updating on certain evidence. I focus on uncertain evidence because it is a more general case than certain evidence, and because it will generally yield more gradual changes in the agents' beliefs than certain evidence. It is easier to observe the evolution of the population's beliefs when they change more gradually.

as exactly the same hard "virtual evidence" in an augmented Bayesian network (Chan & Darwiche, 2005; Jacobs, 2018; Pearl, 1988).[10]

2.3 Agreement between agents

Summarizing, I assume that the agents agree about *almost* everything.

- The agents will form beliefs about the same set of propositions, X.
- The agents will agree about which beliefs are dependent or independent of others (i.e. the agents will share the same belief network structure G).
- The agents will agree about the conditional relations between beliefs (i.e. the agents will share the same conditional probability distributions between parent and child beliefs).
- Each agent will receive the same likelihood evidence η_t, at each timestep t.

The agents will only disagree about one thing: the initial probabilities that they assign to each proposition. Given the Bayesian network structure, and the rationality constraints on the agents, this disagreement can entirely summarized by their beliefs about the *exogenous variables*: those with no parents. Beliefs about these variables are in some sense prior to other beliefs: we could imagine as basic background beliefs held by the agents. Any polarization or factionalization that arises must be driven entirely by these disagreements about those exogenous variables. I will assume that the exogenous beliefs of our population are drawn from a random distribution (more precisely, that the degrees of belief are drawn from a uniform distribution between 0 and 1). As such, the exogenous variables will be statistically independent of each other, at least at the initial timestep, t_0.

2.4 Limitations of the model

This idealized model is not intended to fully capture the complexity of real-world factionalization, which is likely to arise from multiple factors. A sophisticated understanding of real-world factionalization should also consider other potential sources, which may include social trust, political alliance-building or underlying psychological attitudes (for example, see Lakoff, 2010; Weatherall & O'Connor, 2021). None of these play a role in the model presented here.

However, this model may still provide insight of one plausible mechanism that drives factionalization. It seems likely that the principles driving factionalization in this idealized model could also be at work within the multifaceted models that better represent the complexities of real-world factionalization.

[10] To represent evidence about some variable, H, we augment the original Bayesian network with a virtual node, η, which has no children and whose only parent is the node corresponding to variable H. We can represent uncertain evidence pertaining to H as certain evidence about this virtual node, and update H by Bayes' rule. The uncertainty regarding evidence on H is now specified by the likelihoods given the virtual evidence η, i.e. $P(\eta \mid H = h_i)$. Therefore if different agents obtain evidence from virtual nodes with the same conditional probabilities, this represents evidence with the same likelihoods for each agent. If the reader is still uncomfortable with this notion of sameness of uncertain evidence, they can at least be reassured that the results in this paper will apply to cases of certain evidence, as a straightforward limiting case.

Furthermore, this model does demonstrate how epistemic factionalization, a phenomenon that one might intuitive suppose to be a result of "irrationality", can arise for a population of rational agents, who are all updating on the same evidence in highly idealized circumstances. This insight challenges the notion that factionalization is solely a product of cognitive biases or misinformation, suggesting instead that it can be a natural outcome of rational interrelations among beliefs. Therefore, addressing factionalization is not as straightforward as correcting cognitive biases or rectifying skewed information sources; it demands a deeper understanding of the inherent dynamics between beliefs.

2.5 Related models

With this model in hand, it is worth considering how it relates to, and differs from certain other models. Weatherall and O'Connor (2021) demonstrate how factionalization can arise in networks of agents. These agents adopt a heuristic for evaluating the reliability of evidence—they discount evidence from other agents as a function of the overall differences between their beliefs. This model deliberately avoids appealing to background beliefs, worldview or ideologies. Indeed each of the agents' beliefs are assumed to be independent (except insofar as they depend on the agents beliefs about other agents). Nonetheless, the beliefs systematically become correlated as the population updates its beliefs. As such, they explicate a form of factionalization that emerges solely "from trust grounded in shared belief".

The approach taken here is importantly different: the factionalization does not arise from network effects or social trust *between* agents. Indeed, in the model presented here, all agents have access exactly the same evidence. Rather, it arises from relationships between the beliefs of agents. As such, whilst Weatherall and O'Connor (2021) treat beliefs as independent, in the model presented here, the beliefs are explicitly probabilistically related.

Grim et al. (2022a) also create a model with some similarities to the one presented in this paper. In their model, individual agents with multiple, probabilistically related beliefs exhibit patterns of stable beliefs and punctuated equilibria, which they suggest might resemble patterns of paradigm shifts. However, these equilibria arise under different conditions, and by a different mechanism from the factions that I study in this paper. In the Grim et al. (2022a) model, agents receive an "evidence barrage" of continually surprising evidence, of different likelihoods. As such, this does not represent a "learning scenario" (see Huttegger, 2015) in which the agents cumulatively learn the state of the world. Stable belief patterns arise when the agents' credences become resistant to change as a result of nearing either 0 or 1. By contrast, I will study a population of many agents who receive an increasing (but incomplete) set of information about the world. Most of the time, most of the agents' credences never become close to 0 or 1.

3 Convergence, polarization and factionalization

Recall the model in mind from Sect. 2. What should we expect to happen to the population's beliefs as they update on the successive datapoints? We might distinguish three ways in which the population's beliefs could evolve: *convergence*, *general divergence* and *factionalization*. In this section, I will suggest three different ways to explicate convergence, general divergence and factionalization within this model.[11]

3.1 Intuitive idea

To begin with, let us consider an informal first pass, meant to capture the intuitive ideas of convergence, general divergence and factionalization. We can understand these possibilities as follows.

- **Convergence** The beliefs of the population members will grow closer together as they gain evidence.
- **General Divergence** The beliefs of the population members will grow further apart in all directions as they gain evidence.
- **Factionalization** The beliefs of the population members spread out, but not uniformly. Instead, different beliefs become more correlated.

Convergence would be perhaps the least surprising of these possible outcomes. After all, it is well known that Bayesian agents will often converge when they update on the same information (as indicated by the famous results of Blackwell & Dubins, 1962; Huttegger, 2015; Nielsen, 2018; Schervish & Seidenfeld, 1990; see Freeborn, 2024b for a discussion of these results in the context of agents with a Bayesian belief network).[12] However, it is well known that Bayesian agents can polarize in single beliefs when they update on evidence (see Freeborn, 2024a; Jern et al., 2014). General divergence and factionalization would be more surprising outcomes: in some sense the agents would be polarizing not just in one belief, but in their overall beliefs.

I will suggest some more precise definitions in Sects. 3.2 and 3.3, but it will be useful to keep this intuitive picture in mind. I represent an example of each of these cases for an imaginary population in Fig. 1.

3.2 Variance explication

We can use the statistical variance to measure the spread of a single belief is across the population. A high variance in a population's beliefs about hypothesis X suggests that the agents' beliefs are spread out, whilst a low variance suggests that the agents' beliefs are closely clustered together. We can use the absolute covariance to give one measure of the degree to which one belief gives us information about another. If the absolute covariance between X and Y is large, then knowing an agent's belief about X

[11] However, note that different authors have used these terms in a wide variety of different ways—see Bramson et al. (2017) for an overview.

[12] We may not see belief merging if the evidence is not complete, in the sense of being enough to settle every belief that the agents hold (see Freeborn, 2024b).

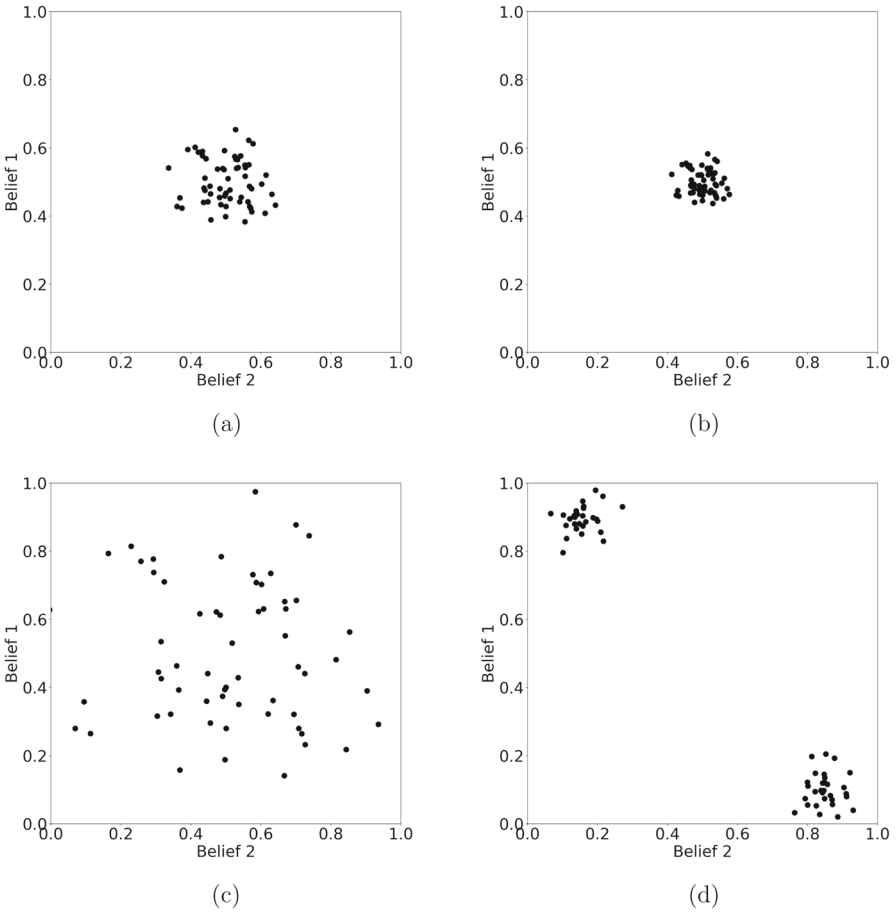


Fig. 1 A schematic representation of an imaginary population of 60 agents, with two different beliefs, 1 and 2, represented by probabilities. The beliefs are shown at a starting timestep, and three hypothetical evolutions of this population at a later timestep. **a** A starting distribution of beliefs for the population. **b** A possible evolution from (**a**) in which the both beliefs have grown closer together. This is a case of **convergence**. **c** A possible evolution from (**a**) in which both beliefs have grown apart. This is a case of **general divergence**. **d** A possible evolution from (**a**) in which both beliefs have grown apart, but not uniformly: the two beliefs have become correlated. This is a case of **factionalization**

allows us to predict something about their belief in Y.[13] We can define these quantities for our population as follows,

$$\textbf{Variance}: \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 \tag{4}$$

[13] More precisely, it tells us the linear joint variability. I use the absolute variances and covariances in particular, rather than correlation coefficients, because we are not interested in the direction of the relationship between two variables, only the degree to which one variable tells us about the other.

$$\textbf{Absolute Covariance}: |\sigma_{X,Y}| = \frac{1}{N}\sum_{i=1}^{N}|(x_i - \mu_x)(y_i - \mu_y)|, \tag{5}$$

where X, Y are binary random variables representing two propositions, x_i and y_i are the probabilities assigned to propositions X or Y being true by agent i, μ_x and μ_y are the corresponding average degree of beliefs across the population, σ_X and σ_Y are the corresponding standard deviations across the population.

With this in hand, we can give a new explication the concepts of convergence, general divergence and factionalization.

- **Convergence** The average variance of the population's beliefs decreases as the agents gain evidence.
- **General Divergence** The average variance of the population's beliefs increases, and the average absolute covariance increases or remains the same, as the agents gain evidence.
- **Factionalization** The average variance of the population's beliefs increases, but the average absolute covariance decreases, as the agents gain evidence.

3.3 Information-theoretic explication

Finally, we are ready to develop a more general explication of convergence, general divergence and factionalization. To do this, we will deploy several concepts from information theory (see Appendix A for definitions and a brief discussion; see Cover and Thomas (2006) for further detail).

Suppose that we have two joint probability distributions with the same support, $P(X_2, X_2 \ldots X_N)$ and $Q(X_2, X_2 \ldots X_N)$. The Jensen–Shannon (JS) divergence $D_{JS}(P \mid Q)$ gives one natural way to measure the overall relatedness between two joint probabilistic distributions. It is given by,

$$D_{JS}(P \mid Q) = \frac{1}{2}D_{KL}\left(P \;\middle|\; \frac{P+Q}{2}\right) + \frac{1}{2}D_{KL}\left(Q \;\middle|\; \frac{P+Q}{2}\right). \tag{6}$$

where D_{KL} is the Kullback–Leibler divergence, given by,

$$D_{KL}(P \mid Q) = -\sum_{\substack{x_1 \in \mathcal{X}_1, \\ \ldots, \\ x_N \in \mathcal{X}_N}} P(x_1, \ldots x_N)\log\frac{P(x_1, \ldots x_N)}{Q(x_1, \ldots x_N)}. \tag{7}$$

The Jensen–Shannon entropy effectively gives a measure of the symmetrized joint information between two such distributions. It has the advantage of measuring the overall information that one distribution gives us about another, whereas the absolute covariance is only sensitive to linear relations.

For each joint probability distribution, $P(X_1, X_2, \ldots X_N)$, we can define a corresponding product of marginal probabilities, $P^m = P(X_1)P(X_2)\ldots P(X_N)$. In effect,

the marginal probabilities product tells us what the probability distribution of the random variables would be if they were all independent. If we regard each of the $P(X_i)$ as telling us the agent's credence about some *salient hypothesis* of interest, X_i, then we could interpret the marginal probabilities product as telling us the agent's credences about each individual *salient hypothesis*, whilst neglecting beliefs about how those salient hypotheses are related.

Suppose that our population of A agents holds the set of joint probability distributions, $P_1, P_2, \ldots, P_A$, with corresponding marginal probabilities products, $P_1^m, P_2^m, \ldots, P_A^m$. Then the average JS divergence between the joint distributions across the population, $\langle D_{JS}^{\text{joint}} \rangle$, gives one way to measure the overall relatedness of the joint probability distributions. On the other hand, the average JS divergence between the marginal probabilities products across the population, $\langle D_{JS}^{\text{marginal}} \rangle$, gives one way to measure the overall closeness of the agents' beliefs about the propositions, ignoring any correlations between these beliefs.

Now we have the tools in place for a plausible information-theoretic explication of convergence, general divergence and factionalization.

- **Convergence** $\langle D_{JS}^{\text{marginal}} \rangle$ decreases as the as the agents gain evidence.
- **General Divergence** $\langle D_{JS}^{\text{marginal}} \rangle$ increases and $\langle D_{JS}^{\text{joint}} \rangle$ increases or stays the same as the agents gain evidence.
- **Factionalization** $\langle D_{JS}^{\text{marginal}} \rangle$ increases and $\langle D_{JS}^{\text{joint}} \rangle$ decreases as the agents gain evidence.

Seen this way, there is one sense in which factionalization can be understood as a form of epistemic divergence, but another in which it can be thought of as a form of epistemic convergence. Factionalization is a form of divergence in the sense that the agents' beliefs about the key, salient hypotheses grow further apart overall, $\langle D_{JS}^{\text{marginal}} \rangle$ increases. However, it is a form of convergence, in the sense that, when the dependencies between beliefs are taken into account, the overall joint probability distributions grow closer together, $\langle D_{JS}^{\text{joint}} \rangle$ decreases.

From hereon, I will primarily use the information-theoretic approach, which has the advantage of being sensitive to any statistical relation between the variables across the population, linear or not. However, at times it will be convenient to consider the variances of variables and the covariances or correlations between variables.

4 Simple examples

To get a better grasp on convergence and factionalization, it will be helpful to investigate some relatively simple examples. These should allow us to see how an actual belief network might drive convergence or factionalization. I will not provide an example of general divergence, for reasons that I will explain in Sect. 5.

In each example, we will follow the model assumptions set out in Sect. 2. I will also simulate a randomly generated population in each case, and demonstrate how its

beliefs evolve. In each case I will assume that the agents' degrees of belief about the exogenous hypotheses are uniformly distributed between 0 and 1.[14]

4.1 Example 1: Convergence

Let us suppose that agents have beliefs about two distinct hypotheses, H_1 and H_2, and agree that H_2 probabilistically depends on H_1 as in Fig. 2. However, the agents do not agree about the probabilities that they assign to the two hypotheses, H_1 and H_2: let us assume beliefs about H_1 are uniformly distributed across the population.[15] Perhaps, H_1 represents the proposition, "The air pressure is low today", and H_2 represents the proposition, "It will rain today". All agree that learning that it is raining today (H_2 is true) provides the same degree of evidence that the air pressure is low today (H_1 is true), and vice versa. Therefore, we should not expect any polarization to take place.

If agents receive the same evidence, then their beliefs will all update in the same direction, as shown in Fig. 3. The variance in their beliefs about H_2 will decrease, and this in turn may drive a decrease in the variance of their beliefs about H_1. Overall, epistemic convergence takes place. The joint probability distributions, $P(H_1)P(H_2 \mid H_1)$, and marginal probabilities products, $P(H_1)P(H_2)$, will move closer together.[16]

4.2 Example 2: Factionalization

Now, let us allow the agents to have a slightly more complex network of beliefs, one that allows them to update particular beliefs in opposite directions. Let the population hold beliefs about three related hypotheses, H_1, H_2 and H_3. It is already well known that Bayesian networks of this form can drive the polarization of individual beliefs (see Freeborn, 2023, 2024a, 2024b; Jern et al., 2014 for similar examples).[17]

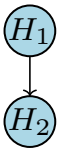
Once again, suppose that the agents start with uniformly distributed degrees of belief between 0 and 1, now about each of the exogenous variables, H_1 and H_3. Suppose

[14] Figures 3, 5 and 7 show results for simulated populations. However, I draw the exogenous variables from a quasi-random 3-dimensional Halton sequence, with prime-numbered bases 2, 3 and 5, rather than from a true random uniform distribution. This is for purely demonstrative purposes: the Halton sequence exhibits low mathematical discrepancy. As such the sequence is generally more evenly spaced than a sequence generated by random draws (see Halton & Smith, 1964; Kocis & Whiten, 1997).

[15] As a result of agreeing about the conditional relations, the agents will agree more about H_2 than H_1. In general, for a population who share a chain belief network, in which all nodes have at most one parent, the variance of the children variables across the population will be always be less than or equal to the variance of the parents. For instance, suppose that 2-valued variable B depends *only* on 2-valued variable A, through a linear conditional probability distribution. We can write $P(A = \text{true}) = aP(B = \text{true}) + bP(B = \text{false}) = cP(B = \text{true}) + b$, for some $a, b \in [0, 1]$, $c = a - b$. Then $\text{var}(B) = c^2\text{var}(A) \leq \text{var}(A)$.

[16] Note that the beliefs in H_1 and H_2 across the population both begin and end perfectly correlated. There are no external sources of information that can serve to change the perfect correlation: H_2 depends entirely on H_1 However, the slope of the relation between H_1 and H_2 has changed. In accordance with the rigidity assumption, the probability $p(H_1 \mid H_2)$ does not change, but the probability $p(H_2 \mid H_1)$ can change for each agent. One way to see this is that not every probability can change by the same amount in light of the same evidence, as the probabilities are fixed between 0 and 1.

[17] More precisely, they allow for what (Freeborn, 2024a) terms the contra-directional updating of individual beliefs. The network in example 1 already allows for a different kind of polarization, belief divergence, in which the difference between particular beliefs increases.



(a)

H_1	$P(H_2 = \text{true})$
True	0.9
False	0.1

(b)

Fig. 2 **a** A Bayesian network structure with two variables, corresponding to degrees of belief about hypotheses H_1 and H_2. I assume that all agents agree about this structure. **b** The conditional probabilistic relations between H_1 and H_2

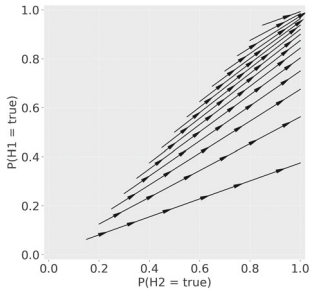
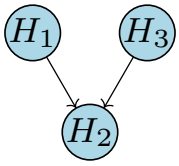


Fig. 3 Belief trajectories for a population of 15 agents, with regards to two related hypotheses, H_1 and H_2 as in Fig. 2b. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives all agents to update in the same, positive direction about H_1. Arrow are indicative, showing only the directions in which their degrees of belief change

(a)



(b)

H_1	H_2	$P(H_3 = \text{true})$
False	False	0.9
False	True	0.1
True	False	0.1
True	True	0.9

Fig. 4 **a** A Bayesian network structure with three variables, corresponding to degrees of belief about hypotheses H_1, H_2 and H_3. I assume that all agents agree about this structure. **b** The conditional probabilistic relations between H_1, H_2 and H_3

that all agents agree that these beliefs are related: H_2 probabilistically depends on H_1 (as in Fig. 4). Perhaps H_1 represents the proposition "The air pressure is low today", H_3 represents "My barometer will give the correct reading" and H_2 represents "*My barometer states that* the air pressure is low today". All agree about the same conditional relationships between these hypotheses. However, their different beliefs regarding H_3 will partly determine how agents update their expectations about what the barometer will say. If I believe that the barometer is a systematically reliable instrument, then a low air pressure reading should increase my degree of belief that the air pressure really is low. On the other hand, if I believe the barometer systematically gives incorrect readings, then a low air pressure reading should decrease my degree of belief that the air pressure is low.

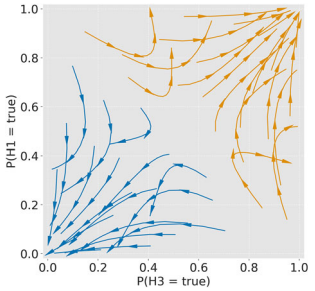


Fig. 5 Belief trajectories for a population of 40 agents, with the belief network shown in Fig. 4. Only two beliefs, H_1 and H_3 are shown. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives the agents to polarize in their beliefs about H_1 and H_3. Observe that the agents beliefs about H_1 and H_3 become correlated as they coalesce into two clusters. Arrow are indicative, showing only the directions in which their degrees of belief change. Colors indicate whether the belief pair ($P(H_1 = \text{true})$, $P(H_2 = \text{true})$) ends closest to (0,0) (blue) or (1,1) (orange) at the final timestep, as measured by the Euclidean distance. (Color figure online)

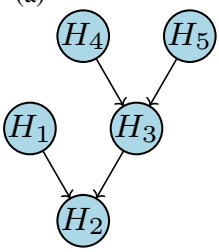
As before, all of the agents receive the same evidence about H_2. Now the agents' beliefs about H_1 and H_3 may be drawn in one of two different directions: either they increase their credence in H_1 being true, and decrease it in H_3 or vice versa, as in Fig. 5. Different degrees of belief in H_3 drive polarization of beliefs H_1, upon updating beliefs about H_2. Likewise, different degrees of belief in H_1 drive polarization of beliefs about H_3. Indeed, the marginal probabilities products, $P(H_1)P(H_2)P(H_3)$ may grow further apart. However, when we look at both beliefs, about H_1 and H_3 together, we see that the beliefs that started independent become correlated. As a result of these correlations, the joint probability distributions, $P(H_1)P(H_3)P(H_2 \mid H_1, H_3)$ grow closer together. The population's beliefs factionalize.

Why do the beliefs factionalize, rather than diverging in all directions, without correlations forming? One way to understand this is in terms of the independencies between the variables. Belief polarization arises here because the agents' beliefs about the H_1 and H_3 can both provide independent information about how to update the other, given some value of H_2.[18] As a result, unlike in the previous example, the correlations between variables can vary after updating H_2. In fact, the correlations *must* vary if H_2 is updated to a new value: given some agreed value of H_2, then knowing the beliefs about H_3 provides new information to us about the beliefs about H_1.

We can draw a more general lesson from examples like this. Whenever updating one variable in a Bayesian population leads to the polarization of another variable, then at least some fully or partly independent variables must experience changes in their correlations. In Appendix B, I explain why this is the case. This realization is very suggestive: if at least some variables must become more correlated, does polarization always lead to factionalization, rather than general divergence? I will return to this question in Sect. 5.

[18] In fact, all that is required is that H_1 and H_3 are fully or partly independence sources of information, conditional on the value of H_2, i.e. $P(H_1 \mid H_2) \neq P(H_1 \mid H_3, H_2)$ (and so likewise, $P(H_3 \mid H_2) \neq P(H_3 \mid H_1, H_2)$)—see Jern et al. (2014) and Freeborn (2024a).

(a)



(b)

H_1	H_3	$P(H_2 = \text{true})$
False	False	0.9
False	True	0.1
True	False	0.1
True	True	0.9

(c)

H_4	H_5	$P(H_3 = \text{true})$
False	False	0.9
False	True	0.1
True	False	0.1
True	True	0.9

Fig. 6 **a** A Bayesian network structure with five variables, corresponding to degrees of belief about hypotheses H_1, H_2, H_3, H_4 and H_5. I assume that all agents agree about this structure. **b** The conditional probabilistic relations between H_1, H_2 and H_3. **c** The conditional probabilistic relations between H_3, H_4 and H_5

4.3 Example 3: Multiple factions

Let us augment the previous example once more, to see how this process can lead to the population dividing into many different factions, rather than just two. A simple way to do this is to add a second polarizing node.

Let the population hold beliefs about *five* related hypotheses, H_1, H_2, H_3, H_4, and H_5. Suppose that all agents agree that these beliefs are related, with H_3 depending on H_4 and H_5, and with H_2 depending on H_1 and H_3, as in Fig. 6. Perhaps H_1 represents the proposition "The air pressure is low today", H_3 represents "My barometer will give the correct reading", H_2 represents "My barometer states that the air pressure is low today", H_4 represents "The barometer is aneroid" and H_5 represents "aneroid barometers give systematically reliable results". Now, different beliefs about H_5 will drive polarization in H_4 (and vice versa) given updated beliefs about H_1. But the updated beliefs about H_1 are themselves already polarized by the different beliefs about H_3, given evidence about H_2. As a result, rather than dividing into two factions as in the previous example, the beliefs about H_4 and H_5 now divide into four distinct factions, as shown in Fig. 7. In general, augmenting networks in this way, by adding more polarizing nodes can increase the number of factions that may form.

5 Why do populations factionalize?

The examples in Sect. 4 illustrate how convergence and factionalization both arise, but not general divergence. In fact, given the definitions in Sect. 3.3, then agents should *never* rationally expect their population to exhibit general divergence upon learning the value of some variable, under the assumptions of our general model, and assuming that they know the population is rational. We can state this as a general condition.

No General Divergence Condition

Suppose that we have two rational agents, with beliefs specified by joint probability distributions $P(X, Y, \ldots Z, D)$ and $Q(X, Y, \ldots Z, D)$ over the same set of discrete, binary variables, $\mathcal{X} = \{X, Y, \ldots D\}$. Let us suppose that the two agents share the same conditional relationships, $P(Y|X) = Q(Y|X)$, for all $X, Y, \in \mathcal{X}$. Let us suppose that

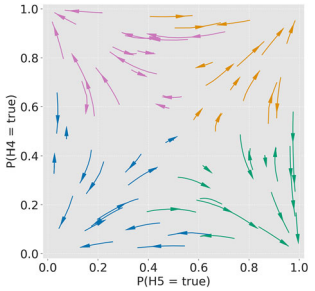


Fig. 7 Belief trajectories for a population of 60 agents, with the belief network shown in Fig. 6. Only two beliefs, H_4 and H_5 are shown. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives the agents to polarize in their beliefs H_1, in turn leading to four-way factionalization in their beliefs about H_4 and H_5. Arrow are indicative, showing only the directions in which their degrees of belief change. Colors indicate whether the belief pair ($P(H_4 = \text{true})$, $P(H_5 = \text{true})$) ends closest to (0,0) (blue), (0,1) (purple), (1,0) (green) or (1,1) (orange) at the final timestep, as measured by the Euclidean distance. (Color figure online)

at least one agent is not certain about the value of D. Then, $D_{JS}(P(X, Y, \ldots Z, D \mid D) \mid (P(X, Y, \ldots Z, D \mid D)) < D_{JS}(P(X, Y, \ldots Z, D) \mid (P))$.

Proof From the Kullback–Leibler divergence chain rule (Eq. 18) and the positivity of Kullback–Leibler entropy, it immediately follows that,

$$\begin{aligned} &D_{KL}(P(X, Y, \ldots Z \mid D) \mid (P(X, Y, \ldots Z \mid D)) \\ &\quad < D_{KL}(P(X, Y, \ldots D) \mid (P(X, Y, \ldots D)). \end{aligned} \tag{8}$$

Furthermore,

$$D_{KL}(P(X, Y, \ldots Z \mid D) = D_{KL}(P(X, Y, \ldots Z, D \mid D). \tag{9}$$

Then,

$$\begin{aligned} &D_{KL}(P(X, Y, \ldots Z, D \mid D) \mid (Q(X, Y, \ldots Z, D \mid D)) \\ &\quad < D_{KL}(P(X, Y, \ldots Z, D) \mid Q(X, Y, \ldots Z, D)). \end{aligned} \tag{10}$$

The result for Jensen–Shannon divergences follows immediately.

Therefore, if the agents' overall beliefs grow further apart, then agents should always expect factionalization, not general divergence.[19] We can understand this as a *cumulativity of information* condition. If all of the rational agents in some sense acquire the same information, then in some sense their beliefs should move closer together. This

[19] However, this does not immediately rule our general divergence as a possibility altogether. As I explain in Appendix A, conditional Kullback–Leibler divergences are the expectations of the Kullback–Leibler divergences of the conditional probabilities relative to the current probability distributions. Thus whilst no agent should rationally expect the Kullback–Leibler divergences to increase upon learning the same information, this does not mean that surprising results could not happen, in which upon learning new information, the actual Kullback–Leibler divergences could increase.

does not mean that beliefs cannot polarize, but rather, if polarization generally takes place across all of their beliefs (i.e. their beliefs about the salient hypotheses become more spread out; D_{JS}^{marginal} increases) then the beliefs across the population must *factionalize*, or become more correlated (i.e. their beliefs about the salient hypotheses become more spread out; D_{JS}^{joint} must decrease). Whilst the population's marginal beliefs about all the hypotheses individually can diverge, if we look at the the joint probabilities, then the population's beliefs must nonetheless grow closer together. Another way to think of this is that, in one sense Bayesian learning is genuinely taking place in such a population. Alternatively, one might say that the population's beliefs are becoming more orderly or predictable, even as the agents' individual beliefs diverge.

Certain kinds of Bayesian belief polarization can only arise given certain structural or independence relations between the variables (see Appendix B).[20] In fact, we can understand these as conditions on the dependence between variables: polarization can only take place if the salient variables are dependent in precisely such a way that they must become more generally correlated after polarization. In other words, they can be viewed as conditions that exclude general divergence but allow for factionalization, consistent with our cumulativity of information approach above. I discuss this further in Appendix C.

6 Conclusions

Epistemic factionalization arises very naturally, even for ideally rational agents, who update on exactly the same evidence. This factionalization is driven by probabilistic relations between different beliefs. Different background beliefs drive polarization when the agents update beliefs on the same evidence in different ways: the same evidence can cause some agents to increase their confidence, whilst others decrease theirs. However, this same process tends to lead to different beliefs becoming correlated across a population. Factions emerge, in which agents tend to hold not just one, but many similar beliefs. This process often, but not always, corresponds to the coalescence of distinct clusters of agents, who hold many very similar beliefs, different from the agents in other clusters.

This kind of factionalization is an epistemically rational process. Indeed, it arises precisely because the agents are all rationally learning from the same evidence. There are two perspectives through which we might view factionalization. From one perspective, factionalization might look like a kind of convergence, whereas from another viewpoint, factionalization might look like a particularly severe form of polarization. Fully understanding factionalization requires us to study the phenomenon stereoscopically, using both of these lenses.

In the first sense, factionalization corresponds to the agents' beliefs genuinely moving closer together: the agents' overall joint probability distributions become more similar, as measured by the Kullback–Leibler divergences or Jensen–Shannon entropies. As a population factionalizes, the agents' beliefs line up into two or more

[20] Freeborn (2024a) denotes the types of polarization that can only happen under these conditinos as "contra-directional updating".

opposing camps, each of whom agree about many different beliefs. We can see factionalization as a process in which the populations beliefs become more orderly or predictable, as correlations develop or strengthen between the different agents' beliefs.

In the second sense, factionalization can be understood as a form of multi-belief polarization. The key is whether we consider the joint probability distributions or marginal probabilities products more relevant to the task at hand. If we are primarily concerned with the beliefs about the individual hypotheses themselves, then factionalization may represent a particularly severe kind of polarization. After all, factionalization indicates that the agents have grown further apart in their beliefs about each distinct hypotheses, even as their conditional probabilities may have grown closer together. Recall our original example, a population factionalizing over the issues of anthropogenic climate change and Covid-19 vaccines, perhaps driven by an underlying belief in the trustworthiness of scientists. If the agents grow apart on both of these issues, and their beliefs become more correlated, then this seems to correspond to a severe kind of polarization, even as the agents' joint probabilities grow closer together.

Perhaps one way to put this is that a purely formal epistemologist might feel reassured by factionalization. After all, it is the factionalization process that allows a population's overall beliefs (as represented by the joint probability distributions) to converge, even when individual beliefs are polarizing. By contrast, a social epistemologist or social scientist might find factionalization more concerning. After all, factionalization indicates that the population's beliefs about each individual hypotheses are moving further apart; in such a way that the population is dividing into factions that disagree about not just one belief, but many.

Moreover, no matter how rational the process, this kind of regimentation of beliefs into distinct factions might often be problematic for real populations. For instance, it is well-known that trust tends to decrease between people with very different beliefs (Kitcher, 1995; Rogers, 1983). It is plausible that factionalization across many different beliefs might exacerbate the general problems with social epistemic polarization (Kawakatsu et al., 2021; Levin et al., 2021). In a real world population, processes mechanically similar to this might plausibly contribute towards populations dividing into distinct worldviews, ideologies or paradigms. The fact that the beliefs of agents in each such faction might be internally consistent may discourage convergence or learning from agents in other factions.

Ultimately, the model presented here explains only one kind of factionalization. A more complete model of social factionalization would need to include many other factors, not limited to cognitive biases of agents, differential access to information between agents, and biased sources of information. However, the type of model studied here suggests that, even fixing all such biases would not, in itself, be sufficient to eradicate factionalization.

As Freeborn (2024b) points out, this type of rational polarization could potentially be resolved with the right kind of evidence. If rational agents are able to acquire the same sufficient evidence to settle all their beliefs, then such agents should expect their beliefs to merge. However, in practice, we do not generally have such complete evidence. Bridging the gap between such ideological factions could be challenging. The beliefs of each opposing faction are rationally held, and mutually self-supporting, on the basis of the same evidence. As a result, the epistemic factions that so form

could be difficult to remove through a process of convergence. Simply acquiring more evidence pertaining to just one belief could plausibly drive further factionalization.

Acknowledgements Parts of this research were conducted under the National Science Foundation (NSF) Grant 1922424, on Consensus, Democracy, and the Public Understanding of Science. I would particularly like to thank Cailin O'Connor, Jim Weatherall, Brian Skyrms and Simon Huttegger for their detailed and thorough feedback throughout the development of this project, as well as the two anonymous reviewers for their insights and thoughtful contributions. I would also like to thank the helpful contributions from audiences and panelists at a number of talks, workshops and symposia where this work has been discussed. Of particular note were the invaluable discussions at the Formal Epistemology Workshop in Irvine (2022), Politics, Philosophy, and Economics Society Annual Meeting in New Orleans (2022) and the Philosophy of Science Association Conference in Pittsburgh (2022). I am also grateful for the feedback at recent talks at several discussion groups at the University of California, Irvine, the Social Epistemology Research Group (SERG) at the London School of Economics (2023), and the Cognitive Science Research Seminars and the AI and Information Ethics Working Group, both at Northeastern University, London (2023).

Funding Open access funding provided by Northeastern University Library

Declarations

Conflict of interest The author declares that they have no conflicts of interest concerning this work.

Appendix A: Information-theoretic quantities for discrete variables

Here, I outline some of the key information-theoretic quantities that I use (see Cover & Thomas, 2006 for a more detailed overview). For simplicity, I define these only for discrete variables. These concepts can all apply to joint probability distributions of many variables; however, for clarity I will present them as probability distributions over just one variable here unless the multi-variable case is of particular importance. I leave the logarithmic bases unspecified.[21] Figure 8 gives a visualization of some of the quantities of information and their relations.

Information entropy is a measure of the uncertainty of a random variable. If we learn something about the value of a random variable (i.e gain information), then its information entropy will fall. The total information entropy of a random variable tells us how much information we would need to learn its exact state. If X is a discrete random variable, with possible values $x, \ldots \in \mathcal{X}$, then the entropy is defined by,

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x), \qquad \textbf{(Entropy)} \qquad (11)$$

[21] Choose your favorite logarithmic base. Any will do, as long as it is used consistently.

where $P(x)$ is the probability of X taking value x. The entropy of a probability distribution is always greater than or equal to zero, $H(X) \geq 0$; an entropy of zero corresponds to a variable about whose value we are certain. Likewise, if we have a joint probability distribution over N random variables, $X_1, \ldots X_N$ with supports $\mathcal{X}_1 \ldots \mathcal{X}_N$, then the joint entropy is given by,

$$H(X_1, \ldots X_N) = -\sum_{\substack{x_1 \in \mathcal{X}_1, \\ \ldots, \\ x_N \in \mathcal{X}_N}} P(x_1, \ldots x_N) \log P(x_1, \ldots x_N). \quad \textbf{(Joint Entropy)} \tag{12}$$

The joint entropy tells us how much uncertainty is associated with the set of random N random variables. The conditional entropy $H(Y \mid X)$ tells us what entropy we should expect for variable Y after learning X, *on average*, given our current joint probability distribution over X and Y. It is defined by,

$$H(Y \mid X) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)}. \quad \textbf{(Conditional Entropy)} \tag{13}$$

Loosely, we can think of conditional entropy $H(Y \mid X)$ as the expected posterior entropy upon learning X, and the original entropy of X as the prior entropy. It is not symmetric: $H(Y \mid X) \neq H(X \mid Y)$; however, Bayes' rule for entropy tells us how to relate these quantities:

$$H(Y \mid X) = H(X \mid Y) - H(X) + H(Y). \quad \textbf{(Bayes' Rule for Entropy)} \tag{14}$$

This is an additive analogue for Bayes' rule for probabilities. The conditional entropy always greater than or equal to zero, and always less than the marginal entropy: $0 \leq H(Y \mid X) \leq H(Y)$. In other words, upon learning the true value of a variable that we did not previously know (actually, more generally, upon reducing the entropy of one variable), the posterior entropy of our joint probability distribution should increase (on average, according to our probability measure). One can think of this as a cumulativity of information condition. Roughly speaking, one should expect a net gain in information from learning something new.

Suppose that we have a joint probability, $P(X_1, \ldots, X_N)$ over N random variables. Then the joint entropy is can be calculated by the conditional entropies using the chain rule for entropy.

$$H(X_1, \ldots X_N) = \sum_{i=1}^{N} H(X_i \mid X_1, \ldots X_{i-1}). \quad \textbf{(Chain Rule for Entropy)} \tag{15}$$

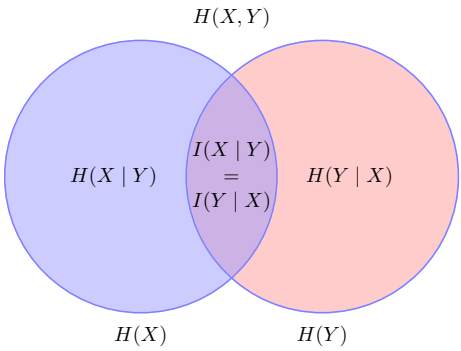


Fig. 8 A Venn diagram relating various quantities of information for two variables, X and Y in a joint probability distribution

This is an additive analogue to the chain rule for probability (see Eq. 2).

The mutual information gives us the amount of information we expect to gain about Y upon learning X, given our current joint probability distribution over X and Y. It equals the difference between the original entropy of Y and the conditional entropy of Y upon learning X.

$$\begin{aligned} I(X \mid Y) = & \quad - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x) P(y)} \\ = & \quad H(Y) - H(Y \mid X). \qquad \textbf{(Mutual Information)} \end{aligned} \tag{16}$$

The mutual information is symmetric: $I(X \mid Y) = I(Y \mid X)$. Another way to think of the mutual information is that it tells us about the independence of variables. If X and Y are independent, then the mutual information is zero, $I(X \mid Y) = 0$: in other words, neither independent variable provides us with any information about the other (this corresponds to $H(X)$ and $H(Y)$ having no overlap in Fig. 8). On the other hand, if X and Y are perfectly correlated, then $I(X \mid Y) = H(X) = H(Y)$ (this corresponds to $H(X)$ and $H(Y)$ having total overlap in Fig. 8) . In general, the mutual information is bounded between these two quantities, $0 \leq I(X \mid Y) \leq H(X), H(Y)$. The mutual information gives us a more general way to measure the dependencies between variables than the correlation or covariance (Eq. 5), in particular one more suited to handling nonlinear dependencies.

One can think of the mutual information, between a joint probability distribution $P(X, Y)$ and a marginal probabilities product $P(X)P(Y)$, as a special case of the Kullback–Leibler divergence. The Kullback–Leibler (KL) divergence between two joint probability distributions on the same support is given by,

$$D_{KL}(P \mid Q) = - \sum_{\substack{x_1 \in \mathcal{X}_1, \\ \ldots, \\ x_N \in \mathcal{X}_N}} P(x_1, \ldots x_N) \log \frac{P(x_1, \ldots x_N)}{Q(x_1, \ldots x_N)}, \quad \textbf{(KL Divergence)} \tag{17}$$

where P and Q are two joint probability distributions with support X. The Kullback–Leibler divergence gives a measure of the information-theoretic difference between two distributions between two distributions, according to the probabilities of one distribution or the other. As such, the Kullback–Leibler divergence is not generally symmetric, unlike the mutual information: $D_{KL}(P \mid Q) \neq D_{KL}(Q \mid P)$. Kullback–Leibler divergences also obey an additive chain rule,

$$D_{KL}(P(x, y) \mid Q(x, y)) = D_{KL}(P(x) \mid Q(x)) + D_{KL}(P(x \mid y) \mid Q(x \mid y)), \quad \textbf{(KL Divergence Chain Rule)} \tag{18}$$

where the conditional Kullback–Leibler divergences are shorthands for the expectations of the Kullback–Leibler divergences of the conditional probability distributions, relative to the former probability distribution, $D_{KL}(P(x \mid y) \mid Q(x \mid y)) = \mathbb{E}_P[D_{KL}(P(x \mid y) \mid Q(x \mid y))]$.

Unlike the mutual information, the Kullback–Leibler divergence is generally unbounded. For example, if one agent is certain about a variable, (say $P(X = x) = 1$), in a way that contradicts another ($Q(X = x) \neq 0$), then the Kullback–Leibler divergence $D_{KL}(P \mid Q)$ will be infinite for probability P. In other words, no finite quantity of information can be sufficient to shift distribution P to Q.

For these reasons, it is often more convenient to use the Jensen–Shannon (JS) divergence to measure the information-distance between two joint probability distributions. This is given by,

$$D_{JS}(P \mid Q) = \frac{1}{2} D_{KL}\left(P \,\middle|\, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q \,\middle|\, \frac{P+Q}{2}\right). \quad \textbf{(JS Divergence)} \tag{19}$$

The Jensen–Shannon divergence can be understood as a smoothed and symmetrized version of the Kullback–Leibler divergence. If the probability distributions of two agents move generally closer together, then the JS divergence will decrease. If the probability distributions of two agents move generally further apart, then the JS divergence will increase. For instance, if the probability distributions are identical, $P = Q$, then $D_{JS}(P \mid Q) = 0$. On the other hand, if the probability distributions are as different as they can be, for a set of N variables, e.g. $P(X_i) = 1$, $Q(X_i) = 0$, for all binary variables $X_i \in \mathcal{X}$, then the JS divergence will take its maximum possible value, $(P \mid Q) = \frac{N}{2}\log(2)$.

There are many other possible different measures of the similarity of joint probability distributions, known as f-divergences (see Ali & Silvey, 1966; Csisz'ar, 1964; Morimoto, 1963; Rényi, 1961). However, the Jensen–Shannon entropy has some desirable properties. One can think of the Jensen–Shannon entropy as giving an "information radius" between two joint probability distributions (see Nielsen, 2021). It has many convenient properties that make it suitable to measure the information-distance between two joint probability distributions. Furthermore, it is symmetric,

$D_{JS}(P \mid Q) = D_{JS}(Q \mid P)$. The square root of the Jensen–Shannon divergence is a metric distance (Endres & Schindelin, 2003; Fuglede & Topsoe, 2004).

One way to think of these quantities is as follows. The correlation and covariance both give a measure of the statistical linear relatedness of two variables. The mutual information gives a way to measure the overall statistical relatedness of two variables, regardless of the linearity of the relation. The KL divergence and JS divergence extend this, giving a measure of the overall relatedness of two joint probability distributions. The KL gives this measure relative to one or the other probability distribution, whereas the JS divergence gives a way to average this for both probability distributions.

Appendix B: Contra-directional updating for Bayesian agents

Jern et al. (2014) and Freeborn (2024a) define contra-directional (or contrary) updating as updating in which one agent increases their degree of belief in some hypothesis, whilst another agent decreases their degree of belief:

$$(\text{posterior}_2 - \text{prior}_2) \times (\text{posterior}_1 - \text{prior}_1) < 0. \tag{20}$$

Suppose that there are two agents, with an identical Bayesian belief networks, G, with discrete variables, including at least two binary variables, D and H. Let $\mathbf{V}$ be the set of all exogenous variables. Let the two agents have identical conditional probability distributions for all children conditional on their parents, but may differ in the probabilities associated with each variable. Let β be a virtual node, with no parents, whose children are the set of the exogenous nodes $\mathbf{V}$ (see Fig. 9). Given that the only differences between the beliefs of the agents can be traced to differences about the exogenous variables, we can understand the virtual node β as encoding all of the differences between the beliefs of the two agents.

Under these assumptions, Jern et al. (2014) prove that only certain kinds of Bayesian belief networks can exhibit contra-directional updating. Following the terminology of Freeborn (2024a), we can express this either through an independence condition or a structural condition.

Independence Condition *Contra-directional updating and transvergent updating with regards to H as a result of updating D is only possible if the belief network satisfies these criteria:*

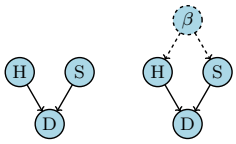


Fig. 9 Left: An example Bayesian network without the virtual node β included. Right: The same network with the virtual node β included. Observe that it is parent to the exogenous variables, and only the exogenous variables

1. D and β are conditionally dependent given H.
2. D and H are conditionally dependent given β.

The independence condition states that contra-directional updating and transvergent updating with regards to node H as a result of updating node D can only occur if two requirements are met: (1) D and the virtual node β are conditionally dependent given H and (2) D and H are conditionally dependent given β. β represents the differing beliefs of two agents with the same Bayesian network structure, G, and variables that can only take on values of 1 or 0.

The structural condition expresses this in terms of d-separation, a graphical or structural property of Bayesian networks (i.e. one pertaining to the nodes and edges only, rather than the numerical values of variables). Loosely, d-separation tests the connectedness of the two variables (Pearl, 2009, pp. 16–19). Roughly, speaking, two sets of nodes are conditionally dependent if they are d-connected given a third set of nodes and conditionally independent if they are d-separated given a third set of nodes.

Structural Condition *Then contra-directional updating and transvergent updating with regards to H as a result of updating D cannot occur for almost all distributions compatible with G unless both of these two requirements is satisfied:*

1. D and β are d-connected given H.
2. D and H are d-connected given β.

The structural condition states that almost all distributions compatible with G, contra-directional updating and transvergent updating with regards to H can only occur if (1) D and β are d-connected given H and (2) D and H are d-connected given β. The first requirement means that the initial beliefs of the agents can provide additional information about H once D is known, and the second requirement means that the data node D can give additional information about the hypothesis node H given the initial beliefs of the agents.

These independence conditions demonstrate that the polarization of one variable leads to changes in the correlations of other variables. To see this, observe that the independence condition implies the following relations (see Jern et al., 2014):

1. $P(\beta \mid D) \neq P(\beta \mid HD)$,
2. $P(H \mid D) \neq P(H \mid \beta D)$.

Recall that, under these assumptions, all of the differences between agents can be summarized by the differences in the exogenous variables, which in turn can be entirely represented by the virtual node, β. Thus, these conditions can be understood as stating that, given some data pertaining to D, there are some independent sources of information (captured within β), which vary between agents, and which will affect how the agents update H. In other words, the value of H, upon updating D will vary, given different independent beliefs, β. As such, the correlations between H and other, at least partly independent variables, will change.

Appendix C: Factionalization and the independence conditions

Suppose that we have two joint probability distributions, $P(X, Y, \ldots Z, D)$ and $Q(X, Y, \ldots Z, D)$, where there is some uncertainty about the value of D. The no general divergence condition (Sect. 5) shows that the Kullback–Leibler divergence berween the two joint probability distributions must decrease if we learn the true value of some variable, e.g. D. We can use this to gain a new understanding of the independence conditions in Appendix B.

Recall (see Eq. 15) that we can rewrite the conditional entropy of a joint probability distribution, given some variable as follows,

$$H(X, Y, \ldots Z \mid D) = H(X) + H(X \mid Y) + \cdots H(D|X, Y, \ldots) - H(D). \quad (21)$$

More generally, given some factorization, with a choice of endogenous variables $\mathcal{A}$ and exogenous variables, $\mathcal{B}$, we can write,

$$H(X, Y, \ldots Z \mid D) = \sum_{A \in \mathcal{A}} H(A) + \sum_{B \in \mathcal{B}} H(B \mid \mathcal{A}) - H(D). \quad (22)$$

$$H(X, Y, \ldots Z \mid D) = \sum_{A \in \mathcal{A}} H(A \mid D) + \sum_{B \in \mathcal{B}} H(B \mid \mathcal{A}, D) \quad (23)$$

Let us call the first term the exogenous entropy and the second term the endogenous entropy. Now, if the value of D is not certain, $H(D) \geq H(D \mid X)$ for any variable X. If this is the case then either the exogenous entropy or the endogenous entropy (or both) be expected to fall upon learning D.

Suppose that we satisfy the two independence conditions in Appendix B,

$$P(\beta \mid H) \neq P(\beta \mid DH) P(H \mid \beta) \neq P(H \mid D\beta) \quad (24)$$

Thus, at least two variables must conditionally depend on D. Thus, at least two conditional entropies must change upon learning D. Given the positivity of entropy, these conditional entropies must fall. If P and Q both share the same graph structure, then these same conditional entropies must change in both of these graphs. Given that the Kullback–Leibler divergence must be expected to decrease upon updating on D, both of these entropies must change in the same direction.

One way of understanding this is that the belief structures must carry precisely the conditional relationships to allow for variables to become more correlated, upon updating. In other words, polarization can arise precisely when the independencies between the variables allow for increased dependence between the variables. This allows for the Kullback–Leibler divergence between the joint probability distributions to fall, even when the Kullback–Leibler divergence between the marginal probabilities products increases.

References

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological), 28*(1), 131–142.

Blackwell, D., & Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics, 33*(3), 882–886.

Bradley, R. (2005). Radical probabilism and Bayesian conditioning. *Philosophy of Science, 72*(2), 342–364.

Bramson, A., Grim, P., Singer, D., Berger, W., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science, 84*(1), 115–159.

Chan, H., & Darwiche, A. (2005). On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence, 163*(1), 67–90.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley.

Csisz'ar, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutat'o Int. K"ozl., 8*, 85–108.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press.

Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association, 77*(380), 822–830.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have Americans' social attitudes become more polarized? *American Journal of Sociology, 102*(3), 690–755.

Dizadji-Bahmani, F., Frigg, R., & Hartmann, S. (2011). Confirmation and reduction: A Bayesian account. *Synthese, 179*, 321–338.

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory, 49*(7), 1858–1860.

Field, H. (1978). A note on Jeffrey conditionalization. *Philosophy of Science, 45*(3), 361–367.

Freeborn, D. P. W. (2023). *Polarization and factionalization for agents with multiple, related beliefs*. PhD thesis, University of California, Irvine.

Freeborn, D. P. W. (2024a). Convergence and polarization for agents with Bayesian belief networks. Unpublished manuscript.

Freeborn, D. P. W. (2024b). Rational polarization for agents with multiple, related beliefs. Unpublished manuscript.

Fuglede, B., & Topsoe, F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In *International symposium on information theory, ISIT 2004 proceedings*. IEEE.

Geiger, D., & Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics, 21*(4), 2001–2021.

Grim, P., Seidl, F., McNamara, C., Astor, I. N., & Diaso, C. (2022a). The punctuated equilibrium of scientific change: A Bayesian network model. *Synthese, 200*(4), 1–25.

Grim, P., Seidl, F., McNamara, C., Rago, H., Astor, I., Diaso, C., & Ryner, P. (2022b). Scientific theories as Bayesian nets: Structure and evidence sensitivity. *Philosophy of Science, 89*(1), 42–69.

Halton, J. H., & Smith, G. B. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM, 7*, 701–702.

Hamilton, L. C., Hartter, J., & Saito, K. (2015). Trust in scientists on climate change and vaccines. *SAGE Open, 5*(3), 2158244015602752.

Hartmann, S., & Bovens, L. (2002). Bayesian networks in philosophy. In Benedikt Lowe, Wolfgang Malzkorn & Thoralf Räsch (Eds.), *Foundations of The Formal Sciences II. Applications of Mathematical Logic in Philosophy and Linguistics* [Trends in Logic] (pp. 39–46). Kluwer Academic Publishers.

Huttegger, S. M. (2015). Merging of opinions and probability kinematics. *The Review of Symbolic Logic, 8*(4), 611–648.

Jacobs, B. (2018). A mathematical account of soft evidence, and of Jeffrey's 'destructive' versus Pearl's 'constructive' updating. *CoRR*, arXiv: abs/1807.05609

Jeffrey, R. C. (1983). *The logic of decision*. University of Chicago Press.

Jeffrey, R. C. (1988). *Conditioning, kinematics, and exchangeability*, vol. 1, pp. 221–255. Kluwer.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review, 121*(2), 206–224.

Kalai, E., & Lehrer, E. (1994). Weak and strong merging of opinions. *Journal of Mathematical Economics, 23*, 73–86.

Kawakatsu, M., Lelkes, Y., Levin, S. A., & Tarnita, C. E. (2021). Interindividual cooperation mediated by partisanship complicates Madison's cure for mischiefs of faction. *Proceedings of the National Academy of Sciences, 118*(50), e2102148118.
Kitcher, P. (1995). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press on Demand.
Kocis, L., & Whiten, W. J. (1997). Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software, 23*, 266–294.
Lakoff, G. (2010). *Moral politics: How liberals and conservatives think*. Chicago University Press.
Latkin, C., Dayton, L., Coyle, C., Yi, G., Winiker, A., & German, D. (2022). The association between climate change attitudes and covid-19 attitudes: The link is more than political ideology. *The Journal of Climate Change and Health, 5*, 100099.
Lazo, A., & Rathie, P. (1978). On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory, 24*(1), 120–122.
Lee, C. H., & Sibley, C. G. (2020). Attitudes toward vaccinations are becoming more polarized in New Zealand: Findings from a longitudinal survey. *EClinicalMedicine, 23*, 100387.
Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences, 118*(50), e2116950118.
Madsen, J., Bailey, R., & Pilditch, T. (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports, 8*, 12391.
Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan, 18*(3), 328–331.
Mrad, A. B., Delcroix, V., Piechowiak, S., Leicester, P., & Abid, M. (2015). An explication of uncertain evidence in Bayesian networks: Likelihood evidence and probabilistic evidence. *Applied Intelligence, 43*(4), 802–824.
Nielsen, F. (2021). On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius. *Entropy, 23*(4), 464.
Nielsen, M. (2018). Deterministic convergence and strong regularity. *The British Journal for the Philosophy of Science, 71*, 1461–1491.
Pallavicini, J., Hallsson, B., & Kappel, K. (2021). Polarization in groups of Bayesian agents. *Synthese, 198*, 1–55.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.
Rényi, A. (1961). *On measures of entropy and information*. University of California Press.
Rogers, E. M. (1983). *Diffusion of innovations*. Simon and Schuster.
Schervish, M. J., & Seidenfeld, T. (1990). An approach to consensus and certainty with increasing information. *Journal of Statistical Planning and Inference, 25*, 401–414.
Sprenger, J. (2017). Foundations of a probabilistic theory of causal strength. http://philsci-archive.pitt.edu/14108/
Wagner, C. G. (2002). Probability kinematics and commutativity. *Philosophy of Science, 69*(2), 266–278.
Weatherall, J., & O'Connor, C. (2021). Endogenous epistemic factionalization. *Synthese, 198*, 6179–6200.