UNIVERSITY OF CALIFORNIA,
IRVINE

Polarization and Factionalization for Agents with Multiple, Related Beliefs

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

David Peter Wallis Freeborn

Dissertation Committee:
Professor James Owen Weatherall, Co-chair
Professor Cailin O'Connor, Co-chair
Professor Simon Huttegger
Distinguished Professor Brian Skyrms

2023

TABLE OF CONTENTS

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

This work behind this dissertation has been a collaborative endeavor, one made possible by the support, guidance, and insights of numerous individuals to whom I owe my sincere gratitude.

First, I would like to thank my co-chairs, Jim Weatherall and Cailin O'Connor, whose insights and mentorship have been invaluable. They contributed significantly to the ideas behind this project, and offered continual, meticulous feedback. Likewise, I would like to extend my gratitude to Brian Skyrms and Simon Huttegger, who have generously lent their expertise offered invaluable critique that has significantly augmented my research. I would also like to express my gratitude Michael Lee for his judicious thoughts and advice as a member of my advancement committee. The late Louis Narens also provided exceptionally percipient comments on several of the presentations and research that ultimately contributed to this dissertation.

The enriching dialogues and feedback during the NSF research group meetings have been pivotal in polishing my understanding and presentation of the subject. The members of the Social Dynamics group at UCI have also been a constant source of intellectual stimulation and guidance. Their probing discussions and constructive feedback have immensely contributed to shaping this dissertation. I am grateful to Adam Chin, Benjamin Genta, Daniel Hermann, Jack VanDrunen, Jingyi Wu, Matthew Coates, Nathan Gabriel, and Saira Khan for comments on draft papers that contributed towards this body of work.

I am thankful for the opportunities to present early parts of this work at various conferences and symposia. The invaluable input and engaging discussions at the Politics, Philosophy, and Economics Society Annual Meeting in New Orleans (2022), the Philosophy of Science Association in Pittsburgh (2022), and the Formal Epistemology Workshop in Irvine (2022) have helped shape and refine my research. My deepest appreciation goes to workshop participants including Kevin Dorst, Jiin Jung, Haixin Dang, Manuel Almagro, Aydin Mohseni, and Gerard Rothfus, each of whom have offered comments and suggestions that have contributed towards the corpus of work that developed into this dissertation.

Finally, this dissertation is also made possible through generous support I received through the National Science Foundation (NSF) Grant 1922424, as well as the John Templeton Foundation Grant 61048. I am very grateful for the financial support that I have received.

VITA

David Peter Wallis Freeborn

Ph.D. in Philosophy **2023**
University of California, Irvine *Irvine, CA, USA*

M.Sc. in Philosophy, Logic and Scientific Method **2017**
London School of Economics and Political Science *London, UK*

Ph.D. in High Energy Physics **2016**
University College London *London, UK*

Master of Physics (Hons.) **2012**
University of Oxford *Oxford, UK*

SELECTED PUBLICATIONS

The Invention of New Strategies in Bargaining Games **2022**
Philosophy of Science

BOOK REVIEWS

Review of Wayne Myrvold's *Beyond Chance and Credence* (with Daniel Hermann) **2022**
Philosophica Mathematica

SELECTED PRESENTATIONS

Rational belief polarization: origins and responses **2022**
Politics, Philosophy and Economics Society, New Orleans

Belief polarization in agents with Bayesian Belief Networks **2022**
Philosophy of Science Association, Pittsburgh

ABSTRACT OF THE DISSERTATION

Polarization and Factionalization for Agents with Multiple, Related Beliefs

By

David Peter Wallis Freeborn

Doctor of Philosophy in Philosophy

University of California, Irvine, 2023

Professor James Owen Weatherall, Co-chair
Professor Cailin O'Connor, Co-chair

Epistemic polarization arises when the statistical dispersion of a population's beliefs increases, especially when all agents update on exactly the same evidence. Factionalization arises when not just one belief, but many different beliefs become correlated across a population. Polarization and factionalization have generally been viewed as examples of human irrationality.

In this dissertation I study the phenomena of epistemic convergence, polarization and factionalization for ideally rational agents, with multiple, probabilistically related beliefs. I demonstrate that rational belief polarization arises very naturally for such agents, even when they update on identical evidence. I demonstrate that probabilistic relations between beliefs can drive various kinds of belief convergence, polarization, and factionalization. Importantly, polarization and factionalization arise generically, without needing specific initial conditions. Under certain circumstances, polarization can even be rationally anticipated in advance. Furthermore, I show that a population of rational agents, should always expect their beliefs to either converge or factionalize under certain conditions.

Introduction

Epistemic polarization arises when the statistical dispersion of a population's beliefs increases, especially when all agents update on exactly the same evidence. Factionalization arises when not just one belief, but many different beliefs become correlated across a population. Polarization and factionalization have generally been viewed as examples of human irrationality.

In this dissertation I study the phenomena of epistemic convergence, polarization and factionalization for ideally rational agents, with multiple, probabilistically related beliefs. I demonstrate that rational belief polarization arises very naturally for such agents, even when they update on identical evidence. I demonstrate that probabilistic relations between beliefs can drive various kinds of belief convergence, polarization, and factionalization. Importantly, polarization and factionalization arise generically, without needing specific initial conditions. Under certain circumstances, polarization can even be rationally anticipated in advance. Furthermore, I show that a population of rational agents, should always expect their beliefs to either converge or factionalize under certain conditions. Thus, beliefs cannot spread out uniformly: if belief polarization takes place, it must also result in epistemic factionalization.

Throughout, I use the toolkit of Bayesian networks, developed by Pearl (see Pearl 1985 and Pearl 2009). Bayesian networks provide a rich and convenient model for representing

and studying belief structures involving complex, probabilistic inter dependencies and their dynamics.

In chapter 1, I systematize the kinds of updating that can arise when two rational agents update a single belief on the same evidence. I then investigate how probabilistic relations between beliefs can drive various kinds of belief convergence, polarization, and related phenomena. I show that these phenomena arise generically, even for ideally rational agents, without the need for specific initial conditions.

In chapter 2 I address the question of whether and how such polarization can be epistemically "rational". Traditionally, belief polarization has often been viewed as an example of irrationality (Baron, 2008; Gerber and Green, 1999). Merging results, such as those of Blackwell and Dubins (1962), Huttegger (2015) and Nielsen (2018) suggest that, under certain plausible assumptions for a "learning scenario", the beliefs of rational agents will converge in the limit, as they update on the same information. I argue that rational belief polarizes arises under natural and general conditions for agents who have multiple, probabilistically connected beliefs, compatible with standard Bayesian assumptions. This epistemic polarization is driven by differences in the background beliefs held by agents, which cause them to update in different ways on the same information. I also show that certain types of polarization can be rationally "expected" in advance.

In Chapter 3, I turn to the phenomenon of epistemic factionalization. This arises when not just one belief, but many different beliefs become correlated across a population. I present a model of how factionalization can emerge in a population of ideally rational agents. This is driven by probabilistic relations between beliefs, with background beliefs shaping how the agents' beliefs evolve in the light of new evidence. I show that in such a model, the only possible outcomes from updating on identical evidence are convergence or factionalization. Beliefs cannot spread out uniformly: if polarization takes place, it must result in factionalization.

Chapter 1

Convergence and Polarization for Agents with Bayesian Belief Networks

1.1 Introduction

Epistemic polarization is widely seen as a growing social, political and scientific problem. Traditionally, such polarization has been viewed as an "irrational" phenomenon(Baron, 2008; Gerber and Green, 1999; Munro and Ditto, 1997a). However, many recent studies have demonstrated a variety of different mechanisms of *rational* belief polarization (Kelly, 2008; Bramson et al., 2017; Jung et al., 2019; Dorst, 2022; Almagro, 2022). Notably, Jern et al. (2014) have shown that one type of polarization can arise under very natural and general conditions for agents who have multiple, logically connected beliefs, compatible with standard Bayesian assumptions [1]. If polarization can arise naturally even for pairs of ideally rational agents, simply as a result of the relations between their prior beliefs, without any appeal

[1] A number of prominent psychological studies point to cases of belief polarization that seem to be driven by the relations between beliefs (Batson, 1975; Lord et al., 1979; Liberman and Chaiken, 1992; McHoskey, 1995; Munro and Ditto, 1997b; Taber and Lodge, 2006; Taber et al., 2009; Plous, 1991).

to cognitive or information biases, network effects or social phenomena, then plausibly, it might arise at least as easily, by similar mechanisms, in populations of more psychologically realistic agents.

In this chapter, I provide a systematic and comprehensive analysis of the effects of belief structure on polarization and related phenomena, between two agents with multiple, probabilistically connected beliefs. Specifically, I investigate how the degrees of belief of two rational agents with the same conditional relations between their beliefs but different priors can converge, diverge, update in the same or opposite directions, and so forth , when they encounter the same evidence. I study the necessary conditions for different updating phenomena to arise, and relate these conditions to the possible belief structures of the agents. I show that these phenomena arise generically, without the need for specific initial conditions. Finally, I explore the relationships between the belief structure and the probabilities that these different phenomena arise.

A key assumption throughout this chapter is that there are two rational agents, with the same conditional probabilistic relations between their beliefs, updating on the same information. However, I allow the agents to hold different priors. The restriction to rational agents is deliberate– if belief polarization arises by these mechanisms for ideally rational agents, then plausibly, it might arise even more naturally for more psychologically realistic agents, with various limitations and cognitive biases. I restrict the focus to just two agents for simplicity and to provide certain minimal results under which polarization can arise.

The chapter is structured as follows. In section 1.2, I introduce Bayesian networks and explain how they can be used to represent and analyze belief structures. In section 1.3 I use a simple example to illustrate how polarization arises between agents with multiple, probabilistically related beliefs. In section 1.4, I systematize the kinds of updating that can arise when two rational agents update their beliefs on the same evidence. These phenomena include different kinds of polarization, as well as related phenomena. In section 1.5, I

investigate some necessary independence conditions on the relations between beliefs under which certain kinds of polarization can arise. I relate these conditions to certain structural conditions on Bayesian belief networks. In section 1.6, I use computer simulations to study the propensity towards polarization among agents with randomly generated Bayesian belief networks. Importantly, I show that these phenomena arise generically, without the need for specific initial conditions, and relate them to certain structural conditions on Bayesian belief networks.

1.2 Bayesian Networks as a Tool to Study Belief Polarization

Bayesian networks provide an especially useful tool for studying probabilistically connected beliefs, and the relations between them. A Bayesian network provides a natural way to represent which beliefs are (conditionally) independent of each other, given other beliefs, as well as how updating one belief can affect other beliefs. Of course, a Bayesian network is not a necessary tool for this; however, it provides a convenient formalism by which to study such relations. Recent work in philosophy of science has begun to study Bayesian networks as models of webs of interconnected beliefs, hypotheses or theories (Hartmann and Bovens 2002, Dizadji-Bahmani et al. 2011, Sprenger 2017, Grim et al. 2021).

In many real world contexts of polarization, we might be interested in understanding how updating one belief can affect another particular belief of interest. Perhaps, changing beliefs about the trustworthiness of climate scientists can increase, or decrease polarization with regards to economic impact of a transition to renewable energy. Or perhaps encountering new evidence about a vaccine can increase, or decrease polarization with regards to the reliability of scientific institutions.

1.2.1 Bayesian Networks

More, formally, a Bayesian network (Pearl, 2009) is a graphical model of a factorized representation of a joint probability distribution. A joint probability distribution, $P(v_1, v_2, \ldots)$, for random variables $v_1, v_2, \ldots$, defined on a probability space, is a probability distribution that gives the probability that each of $v_1, v_2, \ldots$ falls in any particular range or discrete set of values specified for that variable.Given some particular ordering of variables 1 to N, a factorized representation $P(X_1, X_2 \ldots X_N)$ takes the form,

$$P(X_1, \ldots X_N) = P(X_1 \mid X_2, \ldots, X_N) \times P(X_2 \mid X_3, \ldots, X_n) \ldots P(X_n). \tag{1.1}$$

$$= \prod_{i=1}^{N} P(X_i \mid X_1, \ldots X_{i-1}). \tag{1.2}$$

Each of the $N!$ factorizations of a joint probability distribution will correspond to a different Bayesian network.

Let $H = (V, D)$ be a directed, acyclic graph, where V is a set of nodes or vertices, and D a set of directed edges between pairs of vertices in V. Let $X = (X_v)$, $v \in V$ be a set of random variables. Then, X is a Bayesian network with respect to G if its joint probability density function can be written as a product of the individual density functions, conditional on their parent variables. Each node is conditionally independent of any subset of the nodes that are not descendants of itself, given its parents.[2] The probability function of any node takes as input the possible values from the parent nodes' variables, and gives as output the probability, or probability distribution, of the variable associated with the node. Thus we can fully specify a Bayesian network, $G(V, D)$, by a set of nodes, V, directed edges, D,

[2]This is known as the Causal Markov assumption, see Geiger and Pearl (1993); Pearl (2009)

random variables, X with an isomorphism to$V$, and conditional probability distributions $p(x_i|x_{\text{par}_i})$, where x_{par_i} are the variables associated with the parents of i.

Each node of a Bayesian network is associated with a unique random variable, representing the agent's credence in a corresponding hypothesis. Directed edges show how the agents' beliefs are related. A directed edge (v_a, v_b) exists in the network if $P(v_b, v_a)$ 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".

The values of the variables of the different nodes should be consistent with the axioms of probability theory. If a probability distribution can be represented by a particular graph, we say it is "compatible" with the graph. In generating Bayesian belief networks initial probabilities may be given for nodes without parents; subsequently, probabilities for the children are generated by a downwards propagation algorithm. Upon learning new evidence, beliefs may be updated using Bayes' rule, and propagated through the network by an upwards propagation algorithm, such that each proposition eventually will be assigned a measure consistent with the axioms of probability theory. This process of propagation is 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.

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 Jeffrey 1983; Diaconis and Zabell 1982; Bradley 2005) [3]). The belief propagation process is 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.

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

1.3 Polarization between two agents

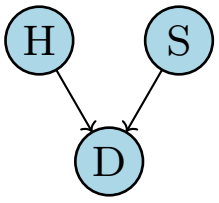
Before moving on, let us look at an example of how belief polarization can arise between agents who have multiple, probabilistically related beliefs, and how this can be represented using Bayesian belief networks. Suppose that we have two rational agents, whose beliefs can be represented by a Bayesian network. Let us further suppose that both agents share the same network structure (roughly implying that both agents share certain beliefs are conditionally independent of others (Pearl, 2009)), and furthermore, let us assume that their belief networks share the same conditional probabilities for child variables conditional on the parent variables (in other words, agents agree on how they should update beliefs about one variable given knowledge of the parent variables). For agents with consistent beliefs, the differences can be fully specified by giving the beliefs about exogeneous variables (those without any parents), and calculating other beliefs by downwards propagation. As such, the agents only have room to disagree about their prior beliefs.

Jern et al. (2014) demonstrate belief polarization can arise between two such agents. To see how this can happen, consider a network with three discrete, two-valued nodes, as in figure 1.1. We have a hypothesis node, H, a data node, D, and a third node, which we can label a "switch" node, S, which affects how conditional probabilities about D depend upon H, and vice versa. Updating on the data node D, leads to divergence in the agents' beliefs about the switch node, S. In effect, beliefs that are not settled (S) determine how agents update belief (H) on the basis of the common evidence (D). Agents with different, unsettled beliefs, may update some beliefs in opposite directions, leading to belief polarization. We could imagine that the "switch" node represents some underlying worldview or belief influencing how an agent responds to data.

For example, suppose two people, Olivia and Peter, are deciding about the effectiveness of a medication. They start with different prior beliefs: Olivia believes that scientific journals

are generally reliable sources of evidence, whereas Peter believes that scientific journals systemically and reliably lie, deceive or are mistaken in their reportings of findings. H could represent the belief that a medication is an effective treatment. S could represent the belief that scientific journals are systemically reliable. Suppose that they receive the same evidence: they both read the same scientific journal, which says that the medication is an effective treatment (D). Then their beliefs about H may update in opposite directions upon reading the same piece of evidence, D, driven by their different beliefs about S [4].

Let us now consider a quantitative example. Suppose that Olivia and Peter share the network and conditional probability table given in figure 1.1. Let us suppose that each of the variables are two valued, H can take values h or $\neg h$, S can take values s or $\neg s$ and D can take values d or $\neg d$. Let us suppose that Olivia starts with priors, $P(H = h) = 0.5$, $P(s) = 0.9$ and so $P(D = d) = 0.66$. Let us suppose that Peter starts with priors, $Q(H = h) = 0.5$, $Q(S = s) = 0.1$ and so $Q(D = d) = 0.43$. Now suppose both agents learn the same evidence, that d is in fact true. This evidence is enough to settle D, but not their beliefs about S, which will influence how beliefs about H is updated. Updating by Bayesian conditionalization, we find their posteriors will be, $P(H = h) \approx 0.62$ and $Q(H = h) \approx 0.21$. Starting with the same initial credence in $H = h$, the agents' beliefs about H have polarized, moving in opposite directions, and growing further apart.



H	S	$P(D = d)$
$\neg h$	$\neg s$	0.5
$\neg h$	s	0.5
h	$\neg s$	0.1
h	s	0.9

Figure 1.1: A Bayesian network with three, two-valued variables, and an associated conditional probability table for D. This network can lead to polarization in beliefs about H if agents with different priors update on the same evidence about D.

[4]See Jern et al. (2014); Cook and Lewandowsky (2016) for similar examples of belief polarization, or Batson (1975); Lord et al. (1979); Liberman and Chaiken (1992); McHoskey (1995); Munro and Ditto (1997b); Taber and Lodge (2006); Taber et al. (2009); Plous (1991) for some empirical studies.

1.4 Epistemic Updating Phenomena

Now we should systemically define belief polarization, and the other kinds of phenomena that can arise when agents update their beliefs on the same evidence in these models. Belief polarization has been operationalized in a number of different ways in different contexts (for an overview, see Bramson et al., 2017). When considering populations of multiple agents, one such choice (see Bramson et al., 2017, Pallavicini et al., 2021 and Madsen et al., 2018) is to use the statistical dispersion, quantified by the standard deviation, of the agents' probabilities about a hypothesis as a measure of the polarization. In the two-agent context, an increase in the statistical dispersion would correspond to the degrees of belief of the agents about the hypothesis growing further apart. I will call this criterion "belief divergence".

Jern et al. (2014) proposes a stricter criterion of belief polarization, widely adopted elsewhere (for example, see Nielsen and Stewart, 2021, page 56 and implicitly Dorst, 2022, page 2). According to this criterion, two agents polarize about hypothesis H if the agent who starts with a higher degree belief in H increases this degree of belief, and the agent who starts with lower degree of belief in H decreases this degree of belief [5]. I will call this criterion "belief radicalization"– the intuition is that both agents move to become more extreme in their views.

However, when considering how updating one belief can affect other beliefs, there are a number of possible behaviors that one might wish to consider. Beliefs may grow apart, or grow closer together. The degrees of belief may both increase or both decrease, or each move in opposite directions. Or the beliefs may cross over, an option excluded by the radicalization definition. These phenomena are not obviously relevant to the case of updating a single belief, or unrelated beliefs on an infinite set of data. However, updating one belief can have a number of intuitively surprising effects on other probabilistically related beliefs in a network. They

[5] Or if two agents start with identical degrees of belief, they move apart in opposite directions

may be important when considering polarization in agents with many connected beliefs, on data that does not settle every belief at either one or zero. For example, the criterion of radicalization requires belief divergence, but also excludes cases in which the agents' beliefs move in opposite directions and do not cross over.

Let us try to categorize these behaviors more precisely. Let there be two agents, who update on a single datapoint (corresponding to updating their beliefs about some variable, D to either one or zero), and in turn update their beliefs about some other hypothesis, H. We might be interested in several different aspects with regards to how their beliefs might change. Let us label the two agents 1 and 2, and label their prior and posterior degrees of belief in H, prior_1, prior_2, posterior_1, posterior_2 respectively. Let us consider three possible behaviors of the belief updating, each of which could occur in two ways. This leads to $2^3 = 8$ possible combinations, which I label cases $A - H$, as schematized in figure 1.2.

Do the agents' degrees of belief in H become closer or drift further apart upon updating?

- **Convergent updating:** the agents's beliefs become closer upon updating: $|\text{posterior}_2 - \text{posterior}_1| \leq |\text{prior}_2 - \text{prior}_1|$.
- **Divergent updating:** the agents's beliefs grow apart upon updating: $|\text{posterior}_2 - \text{posterior}_1| > |\text{prior}_2 - \text{prior}_1|$.

Do the agents degrees of belief in H update in the same direction? That is, do they both increase or both decrease their degree of belief in the hypothesis upon updating?

- **Co-directional updating:** the agents both update in the same direction: $(\text{posterior}_2 - \text{prior}_2) \times (\text{posterior}_1 - \text{prior}_1) \geq 0$.
- **Contra-directional updating:** the agents update in different directions: $(\text{posterior}_2 - \text{prior}_2) \times (\text{posterior}_1 - \text{prior}_1) < 0$.

Do the agents degrees of belief in H cross over upon updating? That is, does the agent with a higher initial degree of belief in H end up as the agent with a lower degree of belief in H?

- **Cisvergent updating:** the agents' beliefs do not cross over:
  $(\text{posterior}_2 - \text{posterior}_1) \times (\text{prior}_2 - \text{prior}_1) \geq 0$.
- **Transvergent updating:** the agents' beliefs cross over:
  $(\text{posterior}_2 - \text{posterior}_1) \times (\text{prior}_2 - \text{prior}_1) < 0$.

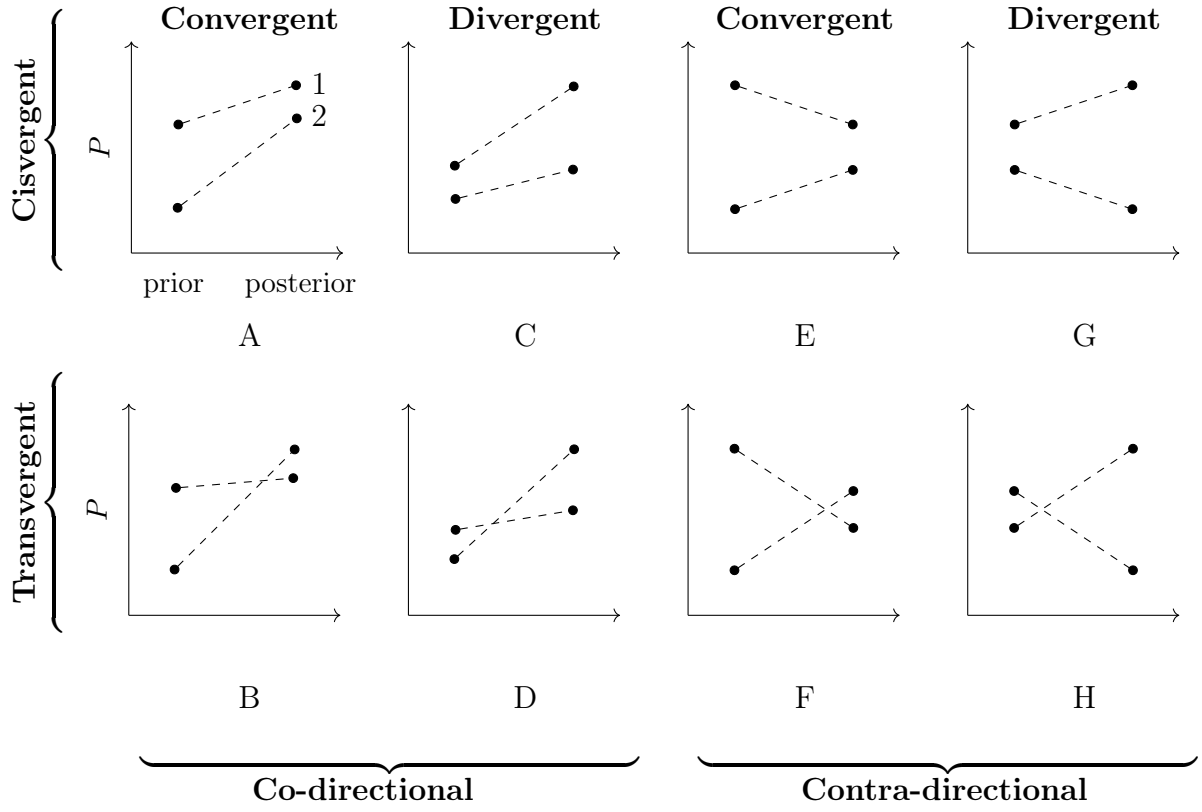


Figure 1.2: Schematics of the eight possible cases ($A - H$) of two generic agents, 1 and 2, some probability, P, (priors on the left, posteriors on the right).

With this in place, we can consider various criteria for two-agent "belief polarization", some stronger, some weaker. Often we are concerned with cases in which the overall difference in agents' beliefs increases: this is belief divergence (cases C, D, G and H). However, situations in which the distance increases, but both agents update in the same direction (cases C and D)

may be less relevant than cases in which the agents update their beliefs in opposite directions (cases G and H). In some situations, the direction of belief change is more relevant than whether the agents beliefs converge or diverge. So we could consider a criterion of contra-directional updating, including cases E, F, G and H. However, this includes cases like E and F in which the agents' beliefs grow closer together. We could combine both conditions and require diverging, contra-directional updating, restricting polarization to cases G and H. I will call this condition "diverging contra-directional updating".

Whilst belief "transvergence" is not in itself a form of belief polarization, it is a separate, intriguing phenomenon that arises when we consider the effects of updating one belief on other beliefs as it propagates through the network. Furthermore, it is important to consider because certain definitions of polarization implicitly *exclude* it. The condition of "belief radicalization", used by Jern et al. (2014) and Nielsen and Stewart (2021), is a particularly strict condition, including only case G, and excluding cases in which the beliefs cross over (case H).

Clearly, the appropriate criterion to use will be sensitive to the context and precise question we are trying to answer. Therefore, rather than advocating for one particular definition, I will consider the collection of criteria together. I show some relations between the criteria in figure 1.3.

1.5 Polarization Conditions for Two Agents with Multiple, Connected Beliefs

Under what conditions can these different updating phenomena arise? The structure of a Bayesian network places important constraints on what types of belief updating are possible. In particular, it turns out that both contra-directional updating, and transvergent updating

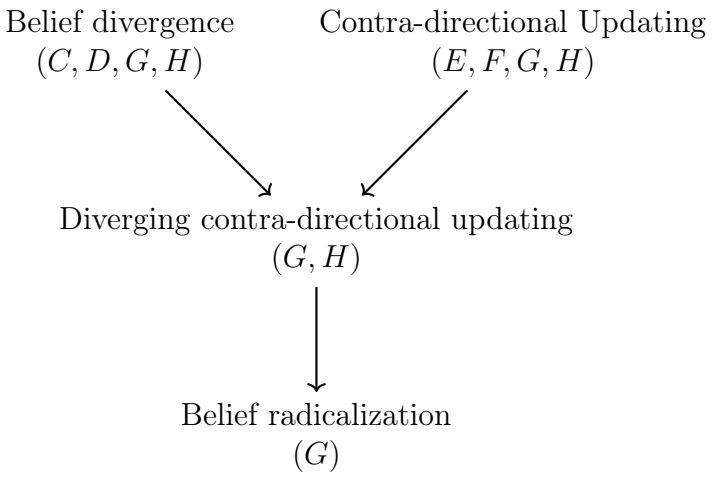


Figure 1.3: Some possible criteria for belief polarization in the case of two-agent updating and their relations. Arrows point from weaker to stronger conditions. Each condition is labelled with the cases from figure 1.2 that it includes.

condition (and so any of the stronger conditions such as divergent contrary updating and belief radicalization) can only occur if a necessary independence condition is satisfied For proofs, see appendix C. This condition generalizes a result by Jern et al. (2014) (see also Jern et al., 2009 for additional details), who demonstrate this condition is necessary for the case of cisvergent, contra-directional updating [6].

> **Independence Condition.** Suppose that we have two agents, with an identical Bayesian network structure, G. Let D and H be two nodes within G and let $\mathbf{V}$ be the set of all exogeneous nodes, that is nodes with no parents. We assume that the associated variables, H and D, can only take two values, 1 or 0. Let the two agents have identical conditional probabilities, but may differ in the initial probabilities associated with the nodes in V. Let β be a virtual node, that is parent to all the nodes in V and has no other edges connecting to the nodes in G. Then contra-directional updating and transvergent updating with regards to H as a result of updating D is impossible unless both of these requirements are satisfied:

[6] Jern et al. (2014) implicitly assumes only cisvergent updating in the definition of contrary updating provided; however, the result straightforwardly generalizes to the general case of contra-directional updating.

1. D and β are conditionally dependent given H.

2. D and H are conditionally dependent given β.

The virtual node β can be understood as encoding the different beliefs of the two agents. Changing the value of β switches us between a network representing the beliefs of agent 1 and a network representing the beliefs of agent 2. Examples of networks with the β node included and excluded are shown in figure 1.4.

We can relate this to the phenomenon of d-separation in Bayesian networks (for details on d-separation, see appendix B or Pearl, 2009). Let G be a directed, acyclic graph, and X, Y, Z be three sets of nodes. Then, for almost all distributions compatible with G, if X is d-connected to Y, given Z, then the corresponding variables will be conditionally dependent. Furthermore, for all distributions compatible with G, then if X is d-separated from Y, given Z, then the corresponding variables will be conditionally independent. Thus, for any belief network, the independence condition will hold only when D and β are d-connected given H and D and H are d-connected given β. As discussed in section B, d-separation is a structural feature of the networks. Therefore, we can give a necessary condition on the structure of networks for which contra-directional updating and transvergent updating can take place, without needing to consider the numerical values of the beliefs.

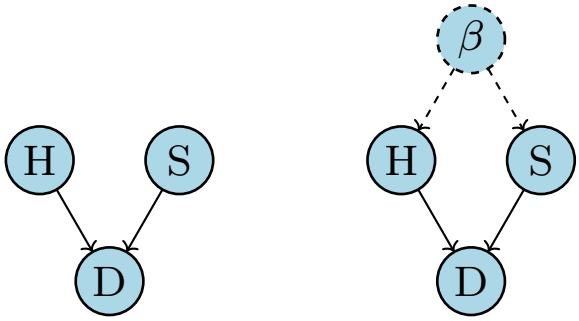
Structural Condition. Suppose that the same assumptions for the independence condition apply. Then contra-directional updating and transvergent updating with regards to H as a result of updating D cannot occur for almost all distributions compatible with G unless both of these two requirements is satisfied:

1. D and β are d-connected given H.

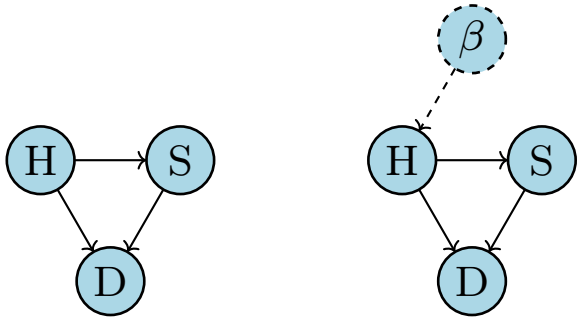
2. D and H are d-connected given β.

Loosely, the first condition, that D and β are d-connected given H, states that the initial beliefs can provide extra information about H, once the D is known. Intuitively, then the initial beliefs of the agents can further influence the direction in which H is shifted, once the data is accounted for. Loosely, the second requirement, that D and H are not d-separated given β, states that the data node can give any additional information about hypothesis, once the priors are known. Intuitively, unless this condition is met, then D and H are separated from influence, except via the initial beliefs. I show example networks that do and do not satisfy the structural condition in figure 1.4.

Notably, graphs of fewer than three nodes will never satisfy the structural condition. Such belief networks could represent simple cases of updating a single based on direct evidence, or updating one belief based on changing another belief, without any confounding beliefs. Therefore, structural contra-directional and transvergent updating will never occur for agents with such networks, under the assumptions of the condition. More generally, structural contra-directional and transvergent updating will not occur if the relevant beliefs are independent.



(a) Left: A Bayesian network that satisfies the structural condition, shown without the β node. Right: When the β node is included (dashed lines), then D and β are *not* d-separated by H. Therefore two agents with this belief network can, in principle, exhibit structural contra-directional and transvergent updating under the assumptions of the structural condition.



(b) Left: A Bayesian network that does not satisfy the structural condition, shown without the β node. Right: When the β node is included (dashed lines), then D and β are d-separated by H. Therefore two agents with this belief network cannot exhibit structural contra-directional and transvergent updating under the assumptions of the structural condition.

Figure 1.4

1.6 Propensity towards Polarization in Bayesian Networks

It would be of interest to better understand the propensity of networks with different characteristics towards belief polarization for at least two reasons. First, it would be useful to understand whether polarization is a fairly generic phenomenon or whether it occurs only in specially chosen Bayesian networks. Second, doing so may help us better understand some of the characteristics of Bayesian networks that tend to lead to polarization. For example, does polarization occur more frequently in Bayesian networks with more nodes, or with more edges for any given number of nodes?

To explore the propensity towards polarization, I consider a model similar to that of Jern et al. (2014). I randomly generate Bayesian networks. For simplicity, I consider only Bayesian networks with discrete variables that can take two values (1 or 0, i.e. true or false). For each Bayesian network, I consider two copies, representing the beliefs of two agents, with identical nodes and edges, and identical conditional probabilities, except that they differ in the the probabilities associated with the exogeneous variables (i.e. the probabilities assigned to nodes without parents). I select one node to be a data node, and one node to be a hypothesis node. Then, both agents update their beliefs associated with the data node to the same value (1 or 0, chosen at random with equal probability). I propagate these beliefs through the Bayesian networks and test the effect of this updating on the hypothesis. I compare the beliefs of the two agents before and after updating, to test whether the beliefs of the two agents have polarized, according to various possible criteria of "polarization".

1.6.1 The Random Generation of Bayesian Networks

In order to randomly draw Bayesian networks, we must first specify a sample space (a set of all the possible Bayesian networks we are considering), an event space of possible outcomes

(the sets of Bayesian networks we could draw), and a probability measure (which assigns each outcome a probability between 0 and 1). In the results below, for each possible number of nodes, I draw each possible Bayesian network with approximately equal probability, according to the Ide and Cozman (2002) algorithm (see appendix A). Of course, Bayesian belief networks chosen in this manner will not necessarily be representative of the types of networks we might expect to be associated with realistic agents of interest in any particular setting.

For each network, I select two variables: a variable upon which the agents update their beliefs, and a second variable, which we use to test the belief polarization. I call the node associated with the former the "data node" and the node associated with the latter the "hypothesis node". I select a data node at random from all possible nodes, and then select a hypothesis node from all nodes that are not the data node, but which are d-connected to it.

1.6.2 Simulation Results

Here I present the results from running simulations of 100,000 pairs of agents for each node number from 2 to 10. In figure 1.5, I show the percentage of simulations that pass the conditions discussed in section 1.5. Heatmaps for the simulations passing the polarization conditions for various node and edge numbers are plotted in figure 1.6 [7]. Finally, the relative proportions of simulations that correspond to each of the cases represented in figure 1.2. Stacked population charts are shown in figure 1.7. There are several key observations to make.

First, polarization seems to be a fairly generic phenomenon: it does not require specially chosen initial conditions in order to arise. Belief divergence arises even in the simple case of

[7]Note that not all possible edge numbers will be represented equally. The algorithm for generating directed acyclic graphs generates all possible graphs of a given node number with approximately equal probability. There are generally more possible graphs of medium edge density than very high or low edge density.

two-node graphs, in almost 40% of cases– involving only instances of case C in figure 1.2, i.e. updating that is divergent but cisvergent and co-directional. Other kinds of polarization, such as contra-directional updating, diverging contra-directional updating and radicalization arise generically for graphs with at least three nodes. This is expected: the structural condition is only satisfiable for graphs with at least three nodes.

Second, larger graphs (i.e. graphs with more nodes) are more likely to experience the intuitively surprising phenomena. Belief divergence, contra-directional updating and transvergent updating all increase with node number. In the case of contra-directional updating and transvergent updating, this can be partly explained by the fact that larger graphs are more likely to pass the structural condition. However, restricting only to graphs pass the structural condition, these phenomenon still increase [8]. In other words, all of these surprising phenomena arise more often in larger graphs [9].

Third, when we look at graph density (i.e. the relative number of edges pre nodes), the picture is more complicated. Graph density is the ratio of edges to some theoretical maximum number of edges. A directed, acyclic graph with n nodes must have at least $n-1$ edges and a maximum of $\frac{n(n-1)}{2}$ edges. I define the density for a directed acyclic graph by

$$\Delta_{DAG} = \frac{2e}{n(n-1)}. \tag{1.3}$$

[8]The increase in contra-directional updating with node number cannot be fully accounted for by the increase in simulations passing the structural condition (see figure 1.5b). However, almost all of the increase of diverging contra-directional updating and radicalization with node number can be accounted for by the increase in contra-directional updating with node number. Among those graphs that do experience contrary updating, the relative proportions of cases E, F, G and H do not seem to significantly change (see figures 1.5c and 1.7b). For node numbers greater than 2, essentially all of the increase in the transvergent updating comes from the increase in the co-directional cases B and D, not the contra-directional cases, F and H. Likewise, for node numbers greater than 2, essentially all of the increase in divergent updating comes from the increase in the co-directional cases C and D, not the contra-directional cases, G and H (see figures 1.7a and 1.7b).

[9]However, of those graphs experiencing contra-directional updating, the proportion experiencing belief radicalization is approximately constant, corresponding to around half of cases. We would expect precisely this: Jern et al. (2009, section 3) provides a proof that "belief divergence" will account for half of all cases for which contra-directional updating is structurally possible; however, this result implicitly neglects the case of transvergent belief divergence and so only corresponds to case G (radicalization). Thus radicalization should account for half of cases for which contra-directional updating is structurally possible.

The least dense graphs for node number n will have an edge number equal to or only a little greater than $n-1$, whilst most dense graphs will have numbers of edges close to $\frac{n(n-1)}{2}$.

The proportion of graphs exhibiting belief divergence seems to increase straightforwardly with graph density. However, the proportion of graphs passing the structural condition, as well as stronger conditions that depend on this such as contra-directional and transvergent updating seem to arise most often in graphs of intermediate density. Plausibly, when there are more edges, nodes are more likely to be causally connected. However, there is a smaller decrease in the proportion passing the independence condition for very high density graphs as well. I hypothesize that this may relate to an observation by Borboudakis et al. (2012): nodes are less likely to be d-connected in very dense graphs. [10].

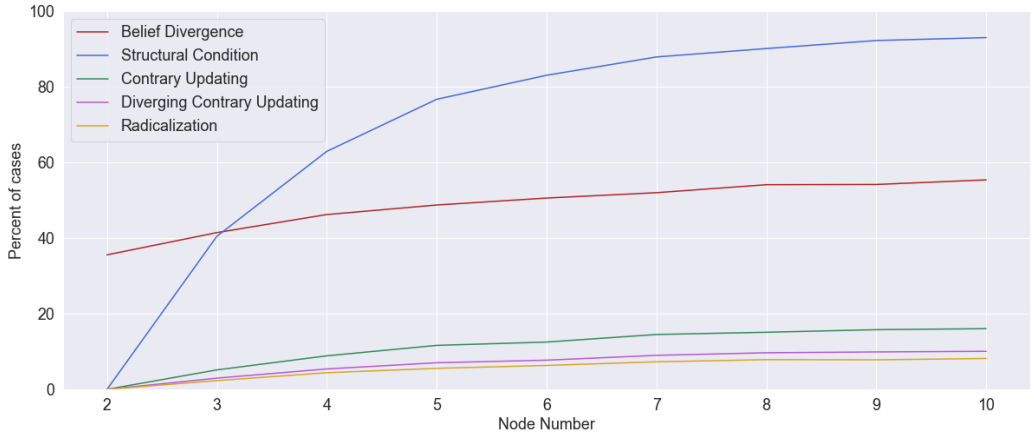
1.7 Conclusions

We have seen that the belief structures of agents can drive various updating phenomena, including various kinds of belief polarization. This kind of polarization arises even when the agents have the same belief network structure and the same conditional relations, and update on exactly the same evidence. In particular, when the evidence is not complete enough to settle all of the agents beliefs, those beliefs that are not settled can influence how the agents update other beliefs. This kind of belief polarization arises fairly generically: it does not require specially chosen initial conditions, but arises often in randomly generated networks. I have argued that Bayesian networks provide a valuable tool for studying belief polarization in this case. Importantly, this polarization can only take place if certain independence

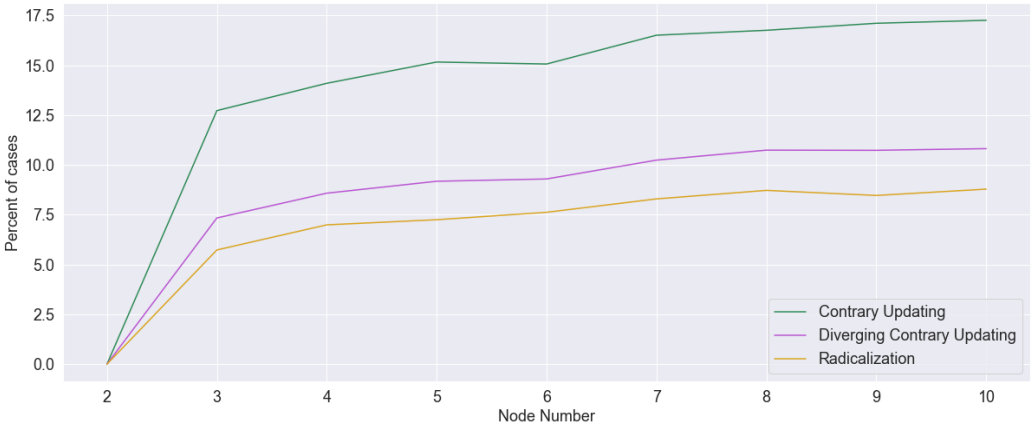
[10]Intuitively, we can see this as follows intermediate nodes have more paths by which to "screen-off" a d-connection in denser graphs. In our case, dense graphs have fewer exogeneous variables on average, as nodes are more likely to have parents. When there are fewer exogeneous variables, then the virtual β node representing the agents' initial conditions relatively more *disconnected* from the rest of the graph: it will have fewer children (the exogeneous) to which it connects. It is therefore easier for the initial conditions to be screened off by the hypothesis node from other parts of the graph. As a result, it is more likely that β and D will be conditionally independent, given H.

conditions are met between the beliefs. These independence conditions relate to requirements on the structure of the Bayesian networks themselves. Furthermore, we have seen that there are certain kinds of expectable polarization that can arise; however expectable contra-directional updating cannot arise.

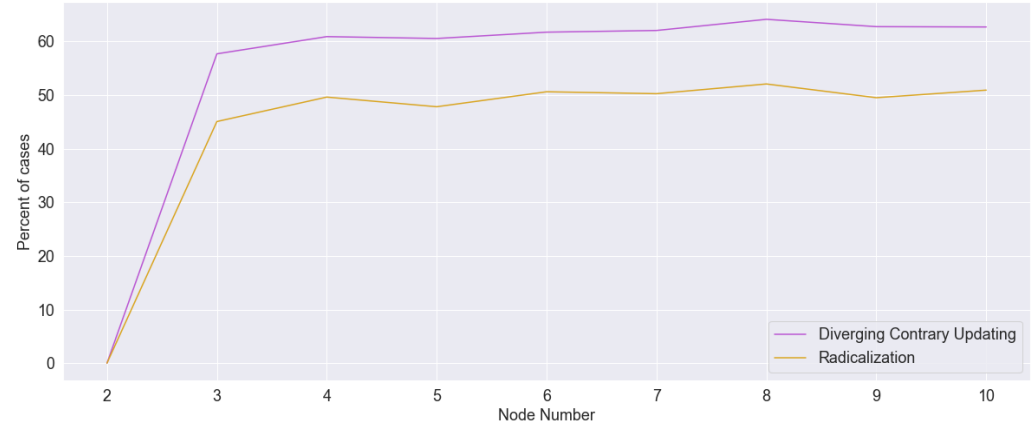
It would be of great interest to better understand more about the networks that can lead to the various kinds of belief networks, beyond the conditions demonstrated here. In addition, what other conditions, on either the network structures, the values of variables lead to belief polarization? It would also be of great interest to use the tools of Bayesian networks to study polarization in more psychologically realistic agents than those studied here. For example, we can adapt Bayesian networks to study agents who are not perfectly rational. It would also be interesting to belief networks better motivated by real world agents, such as those studied by psychologists (Powell et al., 2018; Cook and Lewandowsky, 2016). Above all, real world polarization tends to arise in social settings, such as the scientific, economic or political communities. Significant research has studied polarization in these settings for agents with indepedent beliefs (Axelrod, 1997; Hegselmann and Krause, 2002; Macy et al., 2003; Baldassarri and Bearman, 2007; Deffuant, 2006; Deffuant et al., 2002; O'Connor and Weatherall, 2018). A promising avenue for future research would be to study belief polarization between agents with multiple, connected beliefs in a social setting.



(a) Percentage of all simulations that pass each of the proposed polarization conditions.

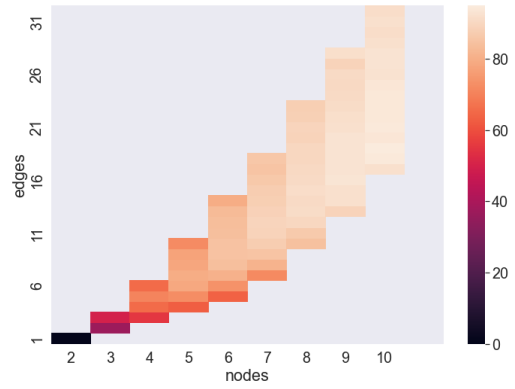


(b) Percentage of those simulations passing the structural condition that also pass each of the proposed polarization conditions.

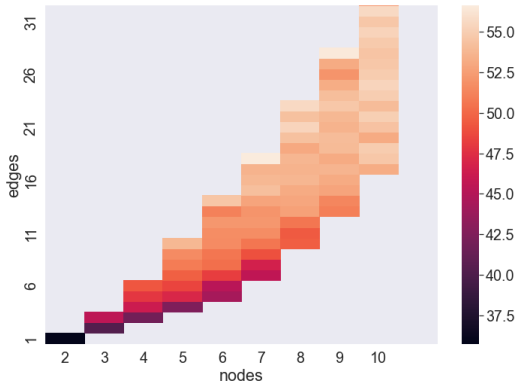


(c) Percentage of those simulations passing the contra-directional updating condition that also pass each of the proposed polarization conditions.

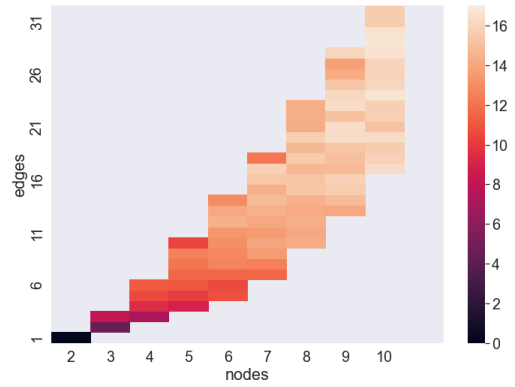
Figure 1.5



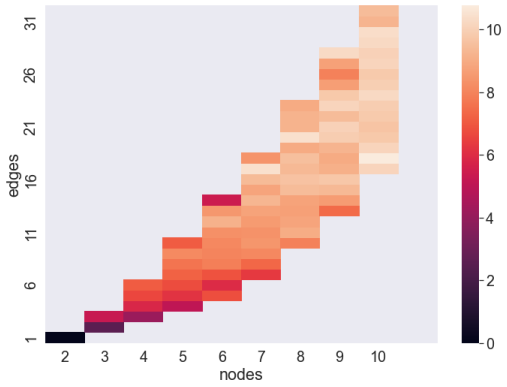
(a) Percentage of simulations that pass the structural condition for each node and edge number.



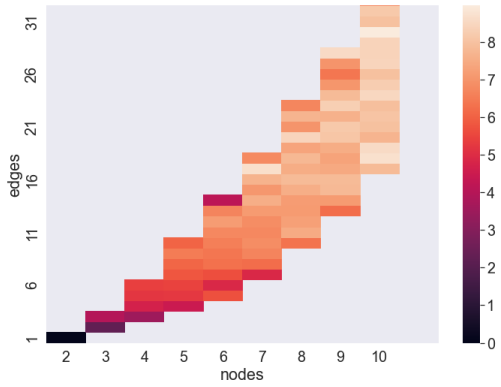
(b) Percentage of simulations that pass the belief divergence condition for each node and edge number.



(c) Percentage of simulations that pass the contra-directional updating condition for each node and edge number.

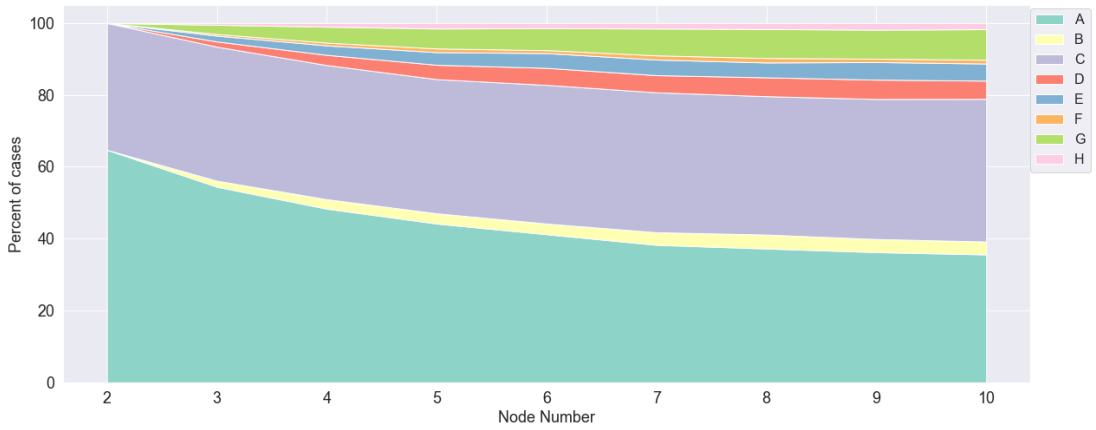


(d) Percentage of simulations that pass the diverging contra-directional updating condition for each node and edge number.

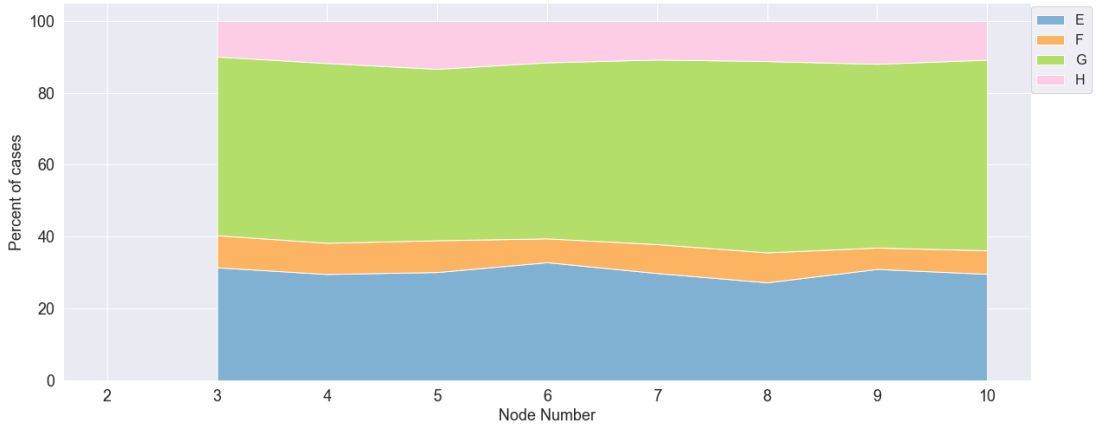


(e) Percentage of simulations that pass the radicalization condition for each node and edge number.

Figure 1.6: Heatmaps showing the percentages of simulations passing various conditions. The scales are varied for each graph for easier visibility. To improve the legibility, node-edge combinations with fewer than 50 simulations are excluded.



(a) Percentage of all simulations that update according to each of the eight cases represented in figure 1.2.



(b) Percentage of simulations passing the contra-directional updating condition that update according to each of the eight cases represented in figure 1.2.

Figure 1.7

Chapter 2

Rational Polarization for Agents with Multiple, Related Beliefs

2.1 Introduction

Imagine two people, Oliver and Pauline. They share the same initial credence about the effectiveness of vaccines as a preventative measure against COVID-19. They both carefully read exactly the same results from a scientific study concluding that vaccines are effective. However, they have different background assumptions. Oliver is optimistic about the scientists who wrote the paper–he believes that they are truthful and highly reliable. Upon reading the results, Oliver now thinks it is more likely that vaccines are effective. However, Pauline is pessimistic about the scientists who wrote the paper–she believes that, more often that not, scientists present false results. As a result, Pauline updates her beliefs in the opposite direction. She now thinks that it is less likely that vaccines are effective. Oliver and Pauline's beliefs have *polarized*; they have grown further apart, after updating on the same information. Cases of polarization arising when two people receive exactly the same

evidence, have been observed in a number of studies (Cook and Lewandowsky, 2016; Lord et al., 1979; Batson, 1975).

Traditionally, belief polarization has often been viewed as an example of irrationality (Baron, 2008; Gerber and Green, 1999). Merging results, such as those of Blackwell and Dubins (1962), Huttegger (2015) and Nielsen (2018) suggest that, under certain plausible assumptions for a "learning scenario", the beliefs of rational agents will converge in the limit, as they update on the same information. However, recent work, across a number of fields has revealed circumstances under which belief polarization can arise even for rational agents (Nielsen and Stewart, 2021; Kelly, 2008; Dorst, 2022; Almagro, 2022; Freeborn, 2023a; Jern et al., 2014). Such work has primarily focused on relaxing various assumptions of an ideal learning scenario. Notably, Dorst (2022) argues that *predictable* polarization cannot arise under standard Bayesian assumptions in an idealized learning situation.

The kind of polarization that Oliver and Pauline experience arises from a different mechanism, one compatible with standard Bayesian assumptions. Updating on evidence about one belief (the results of the scientific paper) drives polarization in another belief (the effectiveness of vaccines), due to the influence of a third belief (their trust in the claims of scientists). Background beliefs, worldviews and ideologies can shape how different agents update on the same information. Recent work in psychology and cognitive science has studied just such situations (Jern et al., 2014). However, this kind of polarization has received little attention in the philosophical discourse.

In this chapter, I argue that rational belief polarizes arises under natural and general conditions for agents who have multiple, logically connected beliefs, compatible with standard Bayesian assumptions. This epistemic polarization is driven by differences in the background beliefs held by agents, which cause them to update in different ways on the same information. Importantly, I do not assume that the information that the agents receive is complete enough to settle every belief that the agents hold, although it may settle some beliefs. I

show that this polarization arises without any need to select for special initial conditions. I also show that certain kinds of expectable polarization can arise.

2.2 Definition of Epistemic Polarization

Here, we are interested in the polarization of a single belief, when two agents update on exactly the same evidence. Let A be some event in an event space $\mathcal{F}$. Then, following the same terminology as chapter 1, we can distinguish two definitions of what it might mean for a *single* belief to become polarized between two agents, with prior probability measures P_1 and Q_1 and posterior probability measures P_2 and Q_2.

- **Divergent updating:** the agents's beliefs grow apart upon updating: $\mid P_2 - Q_1 \mid > \mid P_1 - Q_1 \mid$.
- **Contra-directional updating:** the agents update in opposite directions: $(P_2 - P_1) \times (Q_2 - Q_1) < 0$.

2.3 Background: Bayesian Merging and Rational Polarization

Bayesians often appeal to merging results such as the Blackwell-Dubins theorem to argue that rational disagreement is transient, disappearing once enough evidence is collected (Blackwell and Dubins, 1962; Lehrer and Smorodinsky, 1996; Huttegger, 2015). Informally, the theorem shows that if two agents have sufficiently similar probability distributions, learn from the same, increasing and complete series of evidence, and both update their beliefs by Bayesian

conditioning, then each should be almost certain that their beliefs will converge in the limit. Such convergence stands in contrast to the above notion of "belief polarization".

More precisely, first define a probability space, $(\Omega, \mathcal{F}, P)$. This could represent a setup such as tossing a coin twice, with various possible outcomes and associated probabilities. Here, Ω is a sample space, a set of elementary events, for example in the case of tossing a coin twice, we could have $\Omega = \{HH, HT, TH, TT\}$. We use $\mathcal{F}$ to represent a sigma algebra on Ω (i.e. a non-empty collection of subsets of Ω, closed under complement, countable unions and countable intersections), with elements $A \in \mathcal{F}$ representing events. In the example of tossing a coin twice, particular events could be each possible outcome, $\{HH\}$, $\{HT\}$, $\{TH\}$, $\{TT\}$, or particular combinations of outcomes, such as the event of getting at least one head, $\{HH, HT, TH\}$, or at least one tail, $\{HT, TH, TT\}$, the empty set, $\emptyset$, or the set of all possible outcomes, Ω. We use P to represent a probability measure, a function on $\mathcal{F}$ to the real unit interval, $P : \mathcal{F} \to [0, 1]$, satisfying countable additivity, and with $P(\emptyset) = 0$ and $P(\Omega) = 1$. The probability measure assigns a probability between 0 and 1 to each event. We call $(\Omega, \mathcal{F})$ without a probability measure, an "event space".

We characterize a particular state of evidence by a partition, ϵ, on the sample space, Ω. For example, we could partition the sample space into the event of getting at least one head, and its complement. This represents exactly sufficient evidence to determine whether at least one head obtains, or not. An increasing sequence of evidence is represented by a sequence of finite partitions, $\epsilon_1, \ldots \epsilon_n$, $n \in \mathbb{N}$, such that ϵ_{i+1} is always a finer partition than ϵ_i. For example, in our example of tossing a coin twice, a a sequence of increasing evidence could be: $\epsilon_1 = \{\{HH, HT, TH, TT\}\}$, $\epsilon_2 = \{\{HH, HT\}, \{TH, TT\}\}$, $\epsilon_3 = \{\{HH\}, \{HT\}, \{TH\}, \{TT\}\}$. Each partition represents the information an agent may get about the world at a particular timestep; the increasing fineness implies that the information increases successively each timestep.

Let $\mathcal{F}_i$ be the smallest sigma-algebra that contains the every set in the partition, ϵ_i. We call $\mathcal{F}_i$ the sigma algebra generated by ϵ_i. For example, using ϵ_2 above, $F_2 = \{\emptyset, \{HH, HT\}, \{TH, TT\}\, \Omega\}$. Then the refinement condition on the sequence of evidence implies that $\mathcal{F}_i \subset i+1$ for all $i = 1, 2, \ldots$. The sequence $\{\mathcal{F}_1, \ldots \mathcal{F}_n\}$ is known as a filtration. We say that the information is "complete" if it settles every event in $\mathcal{F}$, i.e. the sigma field generated by all the $\mathcal{F}_i$ equals $\mathcal{F}$. In the example of tossing two coins above, we reach complete information at $\epsilon_3 = \{\{HH\}, \{HT\}, \{TH\}, \{TT\}\}$: this partition is fine enough to specify each individual event. If an agent knows which element of the partition obtains, then they know precisely which outcome arises from the experiment of tossing a coin twice.

Suppose that P and Q are two probability measures on the event space, $(\Omega, \mathcal{F})$. Then the "total variational distance" is the largest possible difference between the probabilities that the two probability distributions can assign to the same event, defined by,

$$d(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|. \tag{2.1}$$

Let P_i be the probability measure we arrive at by updating P on some element of the ith partition, ϵ_i, i.e. for some $\omega \in \Omega$, $P_i(\omega) = P(\cdot | E \in \epsilon_i(\omega))$. Note that this updating does not need to be Bayesian conditionalization, for now we will leave the updating rule unspecified. We will say that "P merges with Q" if,

$$d\big(P_n(\omega), Q_n(\omega)\big) \to 0, \tag{2.2}$$

as $n \to \infty$, for almost every $\omega \in \Omega$. Using a convenient shorthand, let us call P an agent with beliefs corresponding to P and Q an agent with beliefs corresponding to Q. Then if P merges to Q, agent P believes with probability one that their conditional degrees of belief about all events $A \in \mathcal{F}$ will get arbitrarily close to those of Q.

P is absolutely continuous with respect to Q if, whenever $Q(A) = 0$, then $P(A) = 0$, for all $A \in \mathcal{F}$. That is, any event which Q thinks has a probability of zero, P also thinks has a probability of zero. P and Q are mutually absolutely continuous if P is absolutely continuous with respect to Q and Q is absolutely continuous with respect to P. Now we can state the theorem.

> **Theorem 2.1** (Blackwell and Dubins, 1962)**.** Let P and Q be two, mutually absolutely continuous probability measures over the same event space, $(\Omega, \mathcal{F})$. Let both P and Q both update on the same sequence of increasing and complete evidence by Bayesian conditionalization. Then P expects to merge with Q (and Q expects to merge with P).

Under the assumptions of the theorem, long-run divergent and contra-directional updating are not possible when agents update on increasing and complete evidence. This raises the question as to how belief polarization can arise in rational agents. Let us briefly review several paths taken in the literature that relax the various conditions of the Blackwell-Dubins theorem.

2.3.1 Irrational Agents

Many successful models of polarizing agent have used agents that are not explicitly rational. Often these have focused on polarization arising in a social setting, in which agents learn from each other, but in which some measure of similarity of beliefs determines the degree to which one agent influences another (Hegselmann and Krause, 2002; Deffuant et al., 2002; O'Connor and Weatherall, 2018). However, these models do not answer the question above, of whether polarization can emerge in rational agents.

2.3.2 Relaxing Mutual Continuity

Perhaps the next most natural condition to relax is mutual continuity (agreement about which events to assign zero probability). Nielsen and Stewart (2021) demonstrate that when this assumption is relaxed, rational agents should sometimes expect their beliefs to diverge. Some Bayesians have convincingly argued that probability measures should be regular, i.e. they should assign positive probability to all non-empty events (Skyrms, 1995). However, in infinite probability spaces, there is no measure to which all other probability measures will be absolutely continuous.

2.3.3 Relaxing Dynamic Coherence

Dorst (2022) studies polarization arising by relaxing a dynamic coherence condition (see also Huttegger, 2015). Dorst shows that predictable, persistent polarization can arise amongst rational agents if we relax a property of *No Self Doubt*. This is the condition that rational opinions are always sure what rational opinions are. Dorst assumes a value of evidence requirement, that a sequence of updates is rational if and only if each belief state values its successor. Together, the value of evidence requirement and *No Self Doubt* entail a principle of reflection: "my prior rational confidence must equal my rational expectation of my future, more-informed rational confidence". With these conditions loosened, the beliefs of agents can move apart and the agents can anticipate in advance the directions in which their opposing beliefs will polarize.

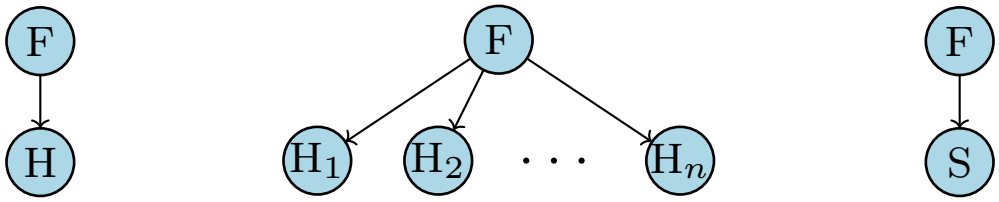
2.3.4 Relaxing Completeness

Finally, we might relax the requirement that agents can access a complete set of evidence, one that can settle every event in $\mathcal{F}$. The finite evidence case has been studied by Jern et al.

(2014), who create a model in which contra-directional arises for rational agents with multiple, probabilistically related beliefs, represented by a Bayesian belief network. In chapter 1 (see also Freeborn (2023a)), I generalize this model generalizes this model to other belief phenomena, and relates it to certain structural features of the belief network. We will discuss further details in the section 2.5.

2.4 Representing multiple beliefs

As in section 1.2, Bayesian networks provide one useful model to represent agents with multiple probabilistic beliefs. For example, suppose we will toss a coin once, which may land heads, H, or tails, T. And suppose that the coin is either fair, F, or biased 60% towards heads, B. Then the sample space could be $\Omega = \{FH, FT, BH, BT\}$. If an agent (reasonably) believes that the fairness of the coin influences whether or not it lands heads: then we can use the network structure in figure 2.1a to represent the agent's beliefs.



(a) (b) (c)

Figure 2.1: (a) A Bayesian network structure with two variables, representing an agent's beliefs about tossing a possibly biased coin once. F represents their beliefs about whether the coin is fair, H represents their beliefs about whether it lands heads in a single coin toss. (b) A similar network, now representing beliefs about tossing a coin finitely many, n, times. (c) A network representing tossing a coin finitely many times, now with a single variable, S, used to represent the outcomes of the entire sequence.

What sort of information might the agent plausibly acquire? After the coin is tossed, they learn how the coin lands. We can represent this state of information by the partition $\{\{FH, BH\}, \{FT, BT\}\}$, sufficient to determine whether the coin has landed heads or tails. However, the information received by the agent is incomplete in two ways. First, it is

not enough to fully specify both outcomes. Our agent should update their beliefs about the fairness of the coin by Bayes' theorem, but does not, in general, have enough information to settle this belief at either 1 or 0. Second, the evidence does not fix all of the conditional probabilities.

Note that updating on a single datapoint is more general than it first appears: conditionalization on any finite string of evidence, $E_1, \ldots E_n$ can be characterized as updating on a single datapoint, $E = \cup_{i=1}^{n} E_i$ (Nielsen and Stewart, 2021, page 56 makes a similar point). For example, suppose that we wanted to represent the agent above experimenting with many coin tosses, rather than just one. A possible Bayesian network to represent such a scenario is shown in figure 2.1b. However, we could also use a single variable to represent the entire sequence of coin tosses, as in figure 2.1c.

2.5 How Relations Between Beliefs Drive Polarization

When agents hold multiple, related beliefs, contra-directional updating can arise easily. Jern et al. (2014) demonstrate that different underlying beliefs can drive such polarization, and in chapter 1, I relate this to certain structural features of the belief networks. In fact, such polarization is possible, even when the agents agree about almost everything. Suppose that we have two rational agents, whose beliefs can be represented by the same Bayesian network structure, with the same conditional probabilities for child variables conditional on the parent variables (in other words, agents agree on how they should update beliefs about one variable given knowledge of the parent variables).

For example, consider two belief networks, representing two absolutely mutually continuous probability measures over some event space. Let us assume that there are three discrete, two-valued nodes, a hypothesis node, H, a data node, D, and a third node, which we can

intuitively think of as a "switch" node, S, which in affects how conditional probabilities about D depend upon H (see figure 1.1). Updating on the data node D, can lead to divergent and contra-directional updating in the agents' beliefs about the hypothesis node, H. In effect, beliefs that are not settled (S) determine how agents update belief (H) on the basis of the common evidence (D). Agents with different, unsettled beliefs, may update some beliefs in opposite directions, leading to belief polarization. We could imagine that the "switch" node represents some underlying worldview or belief influencing how an agent responds to data.

This might offer a good model of the beliefs of Oliver and Pauline. H can represent the hypothesis that vaccines are an effective treatment; D the evidence that the scientific paper *claims* that vaccines are an effective treatment, and S the belief that scientists give reliable indications of true results. Pauline distrusts scientists (her credence in S is low). Oliver trusts scientists (his credence in S is high). As a result, both update their beliefs about H in opposite directions upon receiving the evidence of D. I show one possible schematic in figure 2.2.

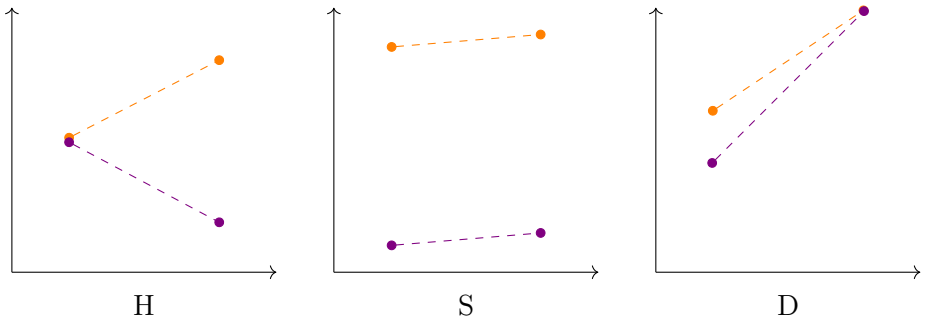


Figure 2.2: One possible way that Oliver (orange) and Pauline (purple) might update their beliefs about H, S and D, upon receiving the same evidence that D is true. Different beliefs in S drive contra-directional updating in their beliefs about H. In this example, their beliefs about S change only slightly.

The kind of polarization studied here can take place for standard Bayesian argents in what seem to be ideal learning situations, in which knowledge is accumulating. Furthermore, both agents have consistent, mutually absolutely continuous probability measures over the same

event space, represented by their Bayesian belief networks. How then are these examples of belief polarization compatible with merging results of Blackwell-Dubins?

As I hinted in section 2.3.4, the key is that the information is not complete in the sense required by the Blackwell-Dubins theorem. Even if the evidence is complete enough to settle one belief (for example hard or certain evidence about node D), this is not sufficient to settle *all* of the agents' other beliefs, namely those about H and S. Nothing here violates the Blackwell-Dubins theorem.

It is likely that in real world settings, agents never have access to sufficient evidence to settle every belief. For example, consider two scientists testing the hypothesis that a treatment is effective. They test the treatment on numerous patients, and examine the outcomes with regards to certain disease symptoms. Suppose that we could even idealize to an infinite sample population of patients. The results of testing alone *still* may not completely settle the scientists' beliefs about the treatment. Perhaps the scientists disagree about how to interpret the data: one believes that the symptoms tested are sufficient to judge the health outcomes of the treatment, whereas the other thinks this data excludes certain important symptoms. In practice, background disagreements may be a persistent feature in scientific disagreements.

We might think that this case is rather unusual: plausibly, people might disagree about the reliability of evidence, but rarely update in opposite directions altogether. Yet for contra-directional updating to take place, Pauline and Oliver need to actively update in opposite directions based on the same data. Pauline believes that scientific papers are not just unreliable indicators of true results, but that they are reliable indicators of *false* results.

However, recall from section 2.4, that the single data node (D) can represent compound data. For example, perhaps the datapoint represents not just a single study, but two different sources of information that come to opposite conclusions. If Pauline and Oliver disagree

about which study is the most reliable, then they might plausibly update in opposite directions. For example, perhaps D represents the outcomes of a scientific journal article *and* and a popular social media account. Oliver and Pauline may disagree about which source they consider to be most reliable. [1]

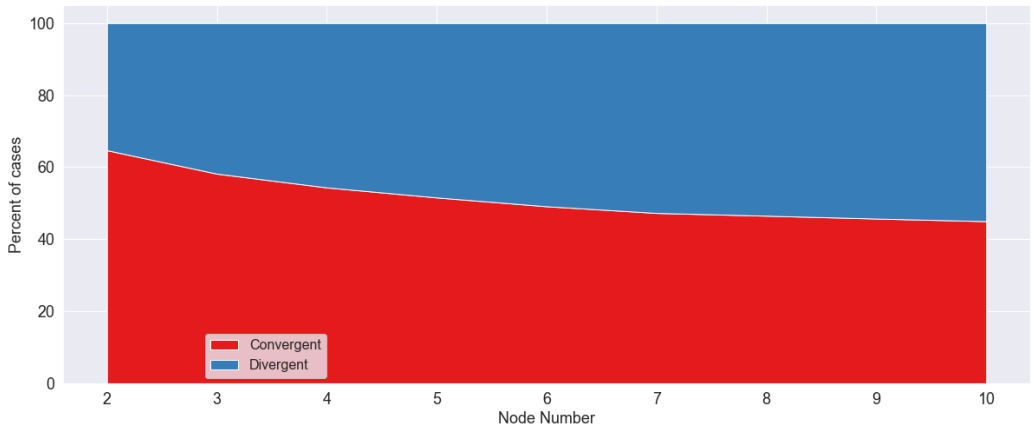
2.6 Sensitivity to Intitial Conditions

We have seen that belief polarization can arise when we allow for probabilistic relations between beliefs, driven by differences in underlying beliefs between agents. This raises the question: how naturally or generically does this polarization arise?

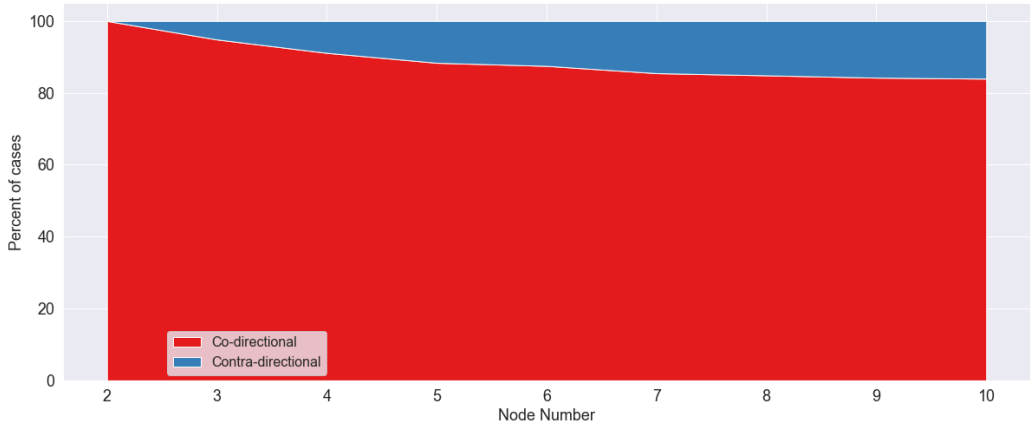
We can test this is with simulations under randomly generated initial conditions. Following a similar model to chapter 1 and Jern et al. (2014); Freeborn (2023a), I randomly generate pairs of identical Bayesian networks with different initial beliefs, and test whether the pairs exhibit divergent or contra-directional updating on randomly generated data. [2]

Polarization arises naturally and generically without any need for specially chosen initial conditions. Figure 2.3 shows the results from running 100,000 simulations for each node number between two and ten.Both contra-directional and divergent updating are surprisingly common, and arise moer often for more complex networks, with more nodes. Divergent updating is fairly common for all network types. Contra-directional updating arises only in networks of three or more nodes, but even for three-node networks arises in around 5% of cases.

[1] Indeed, the study of (Lord et al., 1979) use just such compound data to study polarization empircally.
[2] See appendix D for the simulation details.



(a) Percentage of all simulations that show convergent or divergent updating.



(b) Percentage of all simulations that show co-directional or contra-directional updating.

Figure 2.3

2.7 Expectable Polarization

It is natural to ask whether agents can rationally *expect* belief polarization to arise. Expectable contra-directional updating arises when agents can predict that they will update in opposite directions in the light of certain evidence, *and* can successfully predict the directions in which they will update their beliefs. Expectable belief divergence arises when agents can rationally predict that their beliefs will grow apart in advance of receiving evidence.

Recall from section 2.3.3 that Dorst argues that expectable polarization (in the sense of contra-directional updating) is not compatible with standard Bayesianism, as it violates the principle of reflection Consistent with this, we will see that, under plausible assumptions, expectable contra-directional updating cannot arise; however, expectable belief divergence can arise.

Suppose two agents have prior joint probability distributions O and P, and posterior distributions O' and P', over some set X of binary hypotheses, including H and D. Let the agents share the same conditional relationships between the beliefs in X. We will further supplement O and P with additional beliefs, about their own beliefs and the beliefs of the other agent. [3] Let $\mathbb{E}_P(A)$ refer to the expectation of A with respect to P. Let us further assume that the agents obey a standard principle of reflection, i.e. their expectations of their future beliefs about a hypothesis must equal their current degree of belief in that hypothesis. In this context, we can state the principle of reflection as,

Principle of Reflection:

$$O(H) = \mathbb{E}_O(O'(H)) = O(D)O(H|D) + O(\neg D)O(H|\neg D) \tag{2.3}$$

$$P(H) = \mathbb{E}_P(P'(H)) = P(D)P(H|D) + P(\neg D)\big]P(H|\neg D). \tag{2.4}$$

[3] See appendix E for a discussion of the corresponding Bayesian networks.

Let us also assume a related mutual knowledge condition. Both agents know the other agents' probability distributions and know they will update by Bayesian conditioning,

Mutual Knowledge Condition:

$$\mathbb{E}_P(O'(H)) = P(D)O(H|D) + 1 - P(\neg D)O(H|\neg D) \tag{2.5}$$

$$\mathbb{E}_O(P'(H)) = O(D)P(H|D) + 1 - O(\neg D)P(H|\neg D) \tag{2.6}$$

The agents have expectations of contra-directional updating if one expects that their credence in some hypothesis will drop, whilst the other believes that their credence will increase,

Expectations of contra-directional updating:

$$\mathbb{E}_P(O'(H)) < O(H) \tag{2.7}$$

$$\mathbb{E}_O(P'(H)) > P(H). \tag{2.8}$$

whilst actual contra-directional updating arises if,

Actual contra-directional updating:

$$O'(H) = O(H|D) < O(H) \tag{2.9}$$

$$P'(H) = P(H|D) > P(H). \tag{2.10}$$

Expectable contra-directional updating arises when the agents have expectations of contra-directional updating *and* actual contra-directional updating arises, with the agents updating in the same directions as expected. Then,

Expectable Polarization Incompatibility Condition. Suppose that O and P are prior joint probability distributions over some set X of binary hypotheses, including H and D, with O' and P' as posterior probability distributions. Suppose the agents share the same conditional relationships between the beliefs in X. Suppose that they obey the reflection and mutual knowledge conditions. Then expectable contra-directional updating cannot arise. [4]

Note that both agents O and P might expect contra-directional updating, and make a prediction about the directions in which the other agent will update. However, they cannot both be right. They cannot both rationally predict the directions in which contra-directional updating will take place if they hold beliefs that will in fact lead to contra-directional updating. Thus, although agents *can* have rational expectations of contra-directional updating, expectable contra-directional updating still cannot arise, in the sense meant by Dorst.

Nonetheless, other kinds of expectable polarization can arise. For instance, both agents can correctly predict that their beliefs will grow apart, if they do not know the directions in which they will diverge. Let us define the following,

Expectations of belief divergence:

$$\mathbb{E}_P|P'(H) - O'(H)| > \mathbb{E}_P|P(H) - O(H)| \tag{2.11}$$

$$\mathbb{E}_O|P'(H) - O'(H)| > \mathbb{E}_O|P(H) - O(H)|. \tag{2.12}$$

[4]This is proven in appendix F, and I present an example in appendix G.

Actual belief divergence:

$$|P'(H) - O'(H)| = |P(H|D) - O(H|D)| > |P(H) - O(H)| \tag{2.13}$$

$$|P'(H) - O'(H)| = |P(H|D) - O(H|D)| > |P(H) - O(H)|. \tag{2.14}$$

We have expectable belief polarization if both of these are satisfied, i.e. the agents expect beliefs to diverge, and they actually do. This kind of expectable polarize can arise together for O and P, even when both agents satisfy reflection and the mutual knowledge condition. [5]

The results here are compatible with those presented by Dorst; however, they demonstrate that one must pay attention to the kinds of expectable polarization in question.

2.8 Conclusions

Far from being an irrational or unusual phenomenon, for agents with multiple related beliefs, belief polarization arises very naturally, driven by differences in the background beliefs. The key is that the evidence must be incomplete, i.e. insufficient to settle all of these differing background beliefs. However, this is a plausible assumption for modeling many realistic scenarios. Agents can even rationally expect that certain kinds of polarization will take place. These results are not inconsistent with the arguments of Dorst; however, they demonstrate the significance of distinguishing between different kinds of polarization.

[5] I present an example in appendix G..

Chapter 3

Rational Factionalization for Agents with Probabilistically Related Beliefs

3.1 Introduction

As we have seen in chapters 1 and 2, epistemic polarization arises when a population's beliefs about some issue grow further apart. For this chapter, I will take polarization to arise when the statistical dispersion of a population's beliefs increases regarding *one* particular statement, proposition or hypothesis (for examples of this use, see Bramson et al., 2017; Pallavicini et al., 2021; Madsen et al., 2018; DiMaggio et al., 1996; Freeborn, 2023a,b). To take one example, the New Zealand public's beliefs about the safety of childhood vaccination may have become more polarized over time, with more people becoming increasingly confident that vaccines are either safe or unsafe, and fewer people being unsure (Lee and Sibley, 2020).

However, we are often interested in many different beliefs, and the relationships between them. For example, different polarized beliefs might become closely correlated: people may

cluster into groups who all share not just one, but many, similar beliefs. Epistemic factionalization arises when *multiple*, different beliefs become correlated in a population of agents (see Weatherall and O'Connor, 2021; Bramson et al., 2017; Levin et al., 2021; Kawakatsu et al., 2021). For example, if a U.S. citizen is skeptical about anthropogenic climate change, this gives some degree of evidence that they might also be skeptical of vaccines as a response to the COVID-19 pandemic (Latkin et al., 2022; Hamilton et al., 2015).

Perhaps, such factionalization could be driven by the relationships between different beliefs. Consider again the correlation between skepticism about anthropogenic climate change and vaccine-skepticism. At first glance, these might seem like unrelated beliefs, one pertaining to climate science, the other to medicine. However, these beliefs could be related by an underlying belief. One possibility could be a belief about the trustworthiness of scientists or scientific institutions. If one regards scientific institutions as generally unreliable, this could drive skepticism about both anthropogenic climate change and vaccines.

Previous studies have already shown how underlying background beliefs can drive rational polarization in particular beliefs (see Freeborn, 2023a,b; Jern et al., 2014). In this chapter, I show how factionalization can arise even for populations of ideally rational agents who have probabilistic relations between their beliefs.

I 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. Rather than appealing to trust between agents, 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, and shaping how the agents' beliefs evolve in the light of new evidence.

The chapter is structured as follows. In section 3.2, 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 chapter. In section 3.3, I suggest three different approaches for operationalizing "factionalization", "convergence" and "general divergence" within this model. In section 3.4, I present three simple examples of belief networks, one that leads to convergence and two that lead to factionalization. I explain how factionalization arises in each case. In section 3.5, I explain why factionalization must arise when beliefs polarize: general divergence never arises.

3.2 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 [1]. 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 might update many other (more or less closely related) beliefs: about the efficacy of vaccines in general, about the reliability of scientists, or whether humans cause anthropogenic climate change.

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

I assume a finite population of A 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 [2]. 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 a disagreement about the hypotheses themselves.

To represent relations between beliefs, I use the formalism of Bayesian networks (see section 1.2).

3.2.1 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 [3]. 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" [4]. However, the effects of updating this single belief will propagate through the network to other beliefs.

[2]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 section 3.3.3 significantly simpler (see Lazo and Rathie, 1978).

[3]For reasons of simplicity, I do not consider network effects or information sharing in this chapter. 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.

[4]In this sense, the evidence that the agents obtain will be "incomplete" (see chapter 2 for a discussion of this point). The results in this chapter 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 and Dubins, 1962; Nielsen, 2018; Huttegger, 2015; Kalai and Lehrer, 1994; Schervish and Seidenfeld, 1990), so were the agents receive the same sufficient evidence to settle *all* of their beliefs, then the agents' beliefs should all 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 (2023b) for an argument that this is a reasonable assumption under a broad range of conditions.

In order to explore the evolution of beliefs over time, I will look at successive updating on uncertain evidence [5] 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,

$$L(H = h_1) : \ldots : L(H = h_n) = P(\eta \mid H = h_1) : \ldots : P(\eta : H = h_n), \tag{3.1}$$

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 Wagner, 2002; Jeffrey, 1988; Field, 1978, and Huttegger, 2015 for a philosophical discussion; see also Mrad et al., 2015; Diaconis and Zabell, 1982 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 as exactly the same hard "virtual evidence" in an augmented Bayesian network (Pearl, 1988; Jacobs, 2018; Chan and Darwiche, 2005) [6]

[5] Nothing in this analyses depends on the use of uncertain 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.

[6] 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 chapter will apply to cases of certain evidence, as a straightforward limiting case.

3.2.2 Agreement Between Agents

Summarizing, 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 affect 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 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 *exogeneous variables*: those with no parents. Beliefs about these variables are in some sense prior to other beliefs: we can imagine as basic background beliefs held by the agents. Any polarization or factionalization that arises must be driven entirely by these disagreements about those exogeneous variables. I will assume that the exogeneous 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.

3.2.3 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

actual factionalization should also consider other potential sources, which may include social trust, political alliance-building or underlying psychological attitudes (for example, see Weatherall and O'Connor, 2021; Lakoff, 2010). None of these play can 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.

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.

3.2.4 Related Models

Before proceeding, it is worth considering how the model presented here relates to, and differs from certain similar 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, there are no information differentials between agents. 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 explicited probabilistically related.

Grim et al. (2022) create a model in which single agents with multiple, probabilistically related beliefs exhibit patterns of stable beliefs and punctuated equilibria, which they suggest might resemble paradigm shifts. Such paradigm shifts are importantly different from the factions that I study here, and arise from a different mechanism. In the Grim et al. (2022) 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 gradually learn the state of the world. Stable belief patterns arise when the agent's 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.

3.3 Convergence, Polarization and Factionalization

Recall the model in mind from section 3.2. 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. [7]

3.3.1 Intuitive Idea

First, 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 will grow closer together as they gain evidence.
- **General Divergence:** The beliefs of the population will grow further apart in all directions as they gain evidence.
- **Factionalization:** The beliefs of the population 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 and Dubins, 1962; Nielsen, 2018; Huttegger, 2015; Schervish and Seidenfeld, 1990; see Freeborn, 2023b for a discussion of these results in the context of agents with a Bayesian belief network) [8]. However, it is well known that Bayesian agents can polarize in single beliefs when they update on evidence (see Freeborn, 2023a; Jern et al., 2014. General divergence and factionalization would be more

[7] However, note that these terms have been defined in a wide variety of different ways– see Bramson et al. (2017) for an overview.

[8] 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, 2023b)

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 sections 3.3.2 and 3.3.3, 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 figure 3.1.

3.3.2 Variance Explication

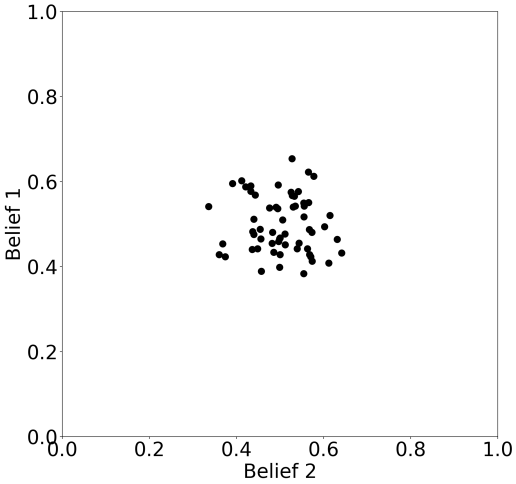
We can use the variance to measure the spread of a single belief is across the population. A high variance in a population's beliefs about hypothesis X suggests that the agents' beliefs are spread out, a low variance suggests that the agents' beliefs are closely clustered together, on average. We can use the absolute covariance to measure 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 allows us to predict something about their belief in Y [9]. We can define these quantities for our population as follows,

$$\textbf{Variance:} \qquad \sigma_X^2 \quad = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 \tag{3.2}$$

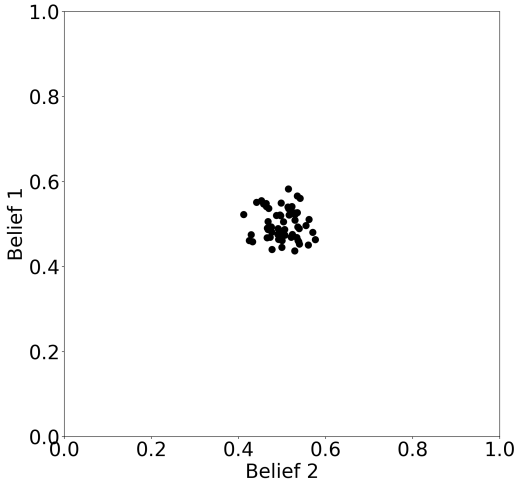
$$\textbf{Absolute Covariance:} \qquad \mid \sigma_{X,Y} \mid = \frac{1}{N}\sum_{i=1}^{N} \mid (x_i - \mu_x)(y_i - \mu_y) \mid, \tag{3.3}$$

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 corre-

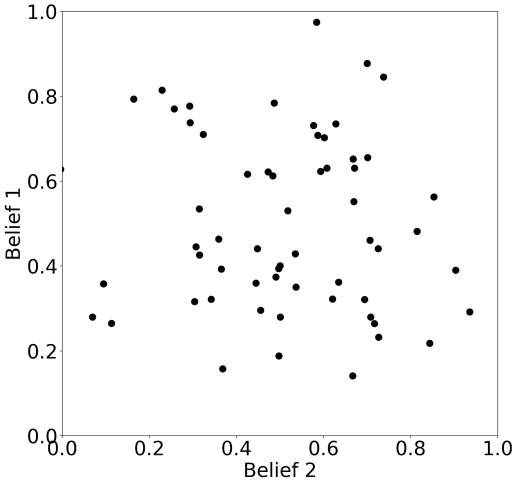
[9] More precisely, it tells us the linear joint variability. I use the absolute variance rather than the variance or Pearson correlation coefficient because we are not interested in the direction of the relationship between two variables, only the degree to which one variable tells us about the other.



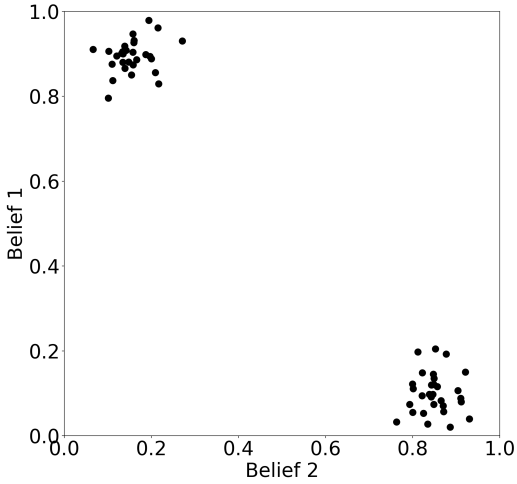
(a) A starting distribution of beliefs for the population.



(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 the both beliefs have grown apart, but not uniformly: the two beliefs have become correlated. This is a case of **factionalization**.

Figure 3.1: 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.

sponding 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

3.3.3 Information-Theoretic Explication

Finally, we are ready to develop a general explication of convergence, general divergence and factionalization. To do this, we will deploy several concepts from information theory (see appendix H 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,

$$D_{JS}(P \mid Q) = \frac{1}{2} D_{KL}\left(P \,\middle|\, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q \,\middle|\, \frac{P+Q}{2}\right). \tag{3.4}$$

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

$$D_{KL}(P \mid Q) = -\sum_{\substack{x_1 \in \mathcal{X}_1, \\ \dots, \\ x_N \in \mathcal{X}_N}} P(x_1, \dots x_N) \log \frac{P(x_1, \dots x_N)}{Q(x_1, \dots x_N)}. \tag{3.5}$$

These 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, $P(X_1, X_2, \dots X_N)$, we can define a corresponding marginal distribution, $P^m = P(X_1)P(X_2)\dots P(X_N)$. In effect, the marginal probability distribution tells us the probability distribution of the random variables if they were all independent.

Suppose that our population of A agents holds the set of joint probability distributions, $P_1, P_2, \dots, P_A$, with corresponding marginal probability distributions, $P_1^m, P_2^m, \dots, P_A^m$. Then the average JS divergence between the joint distributions, $\langle D_{JS}^{\text{joint}} \rangle$, gives one way to measure the overall relatedness of the joint probability distributions. On the other hand, the average JS divergence between the marginal distributions, $\langle D_{JS}^{\text{marginal}} \rangle$, 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 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.

3.4 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 section 3.5.

In each example, we will follow the model assumptions set out in section 3.2. 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 exogeneous hypotheses are uniformly distributed between 0 and 1 [10].

3.4.1 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 figure 3.2. 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. [11]. 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 figure 3.3. The variance in their beliefs about H_2 will decrease, and this in turn drives a decrease in the variance of their beliefs about H_1. Overall then, epistemic convergence takes place. The joint probability distributions, $P(H_1)P(H_2 \mid H_1)$, and marginal probability distributions, $P(H_1)P(H_2)$, will move closer together [12].

[10] Figures 3.3, 3.5 and 3.7 show results for simulated populations. However, I draw the exogeneous 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 Kocis and Whiten, 1997; Halton and Smith, 1964).

[11] 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 $P(A = \text{true}) = aP(B = \text{true}) + bP(B = \text{true}) = cP(B = \text{true}) + b$, for some $a, b \in [0, 1]$, $c = a - b$. Then $\text{var}(B) = c^2\text{var}(A) \leq \text{var}(A)$.

[12] 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

3.4.2 Example 2: Factionalization

Now, let us allow the agents to have a slightly more complex network of beliefs, one that allows for the polarization of particular beliefs. 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 Jern et al., 2014; Freeborn, 2023a,b for similar examples).

Once again, suppose that the agents start with uniformly distributed degrees of belief between 0 and 1, now about each of the exogeneous variables, H_1 and H_3. Suppose that all agents agree that these beliefs are related: H_2 probabilistically depends on H_1 (as in figure 3.4). Perhaps, H_1 represents the proposition, "*My barometer states that* the air pressure is low today", H_2 represents "It will rain today" and H_3 represents "My barometer will give the correct reading." All agree about the same conditional relationships between these hypotheses. However, H_3 will determine how agents update their expectations about what the barometer will say. If I believe that the barometer is a reliable instrument, then the presence of rain should increase my degree of belief that it will state that the air pressure is low. On the other hand, if I believe the barometer systematically gives the wrong readings, then the presence of rain should decrease my degree of belief that it will state that the air pressure is low.

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 figure 3.5. 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

on H_1 However, the slope 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 $p(H_2 \mid H_1)$ for each agent can change. 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.

marginal probability distributions, $P(H_1)P(H_2)P(H_3)$ 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 have 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 [13]. 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, as I demonstrated in section 1.5. This result 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 section 3.5, after building up a more precise machinery for discussing factionalization.

[13] In fact, all that is required is that H_1 and H_3 are fully or partly independence sources of information, given 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, $P(H_3 \mid H_2) \neq P(H_3 \mid H_1, H_2)$).

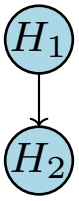
3.4.3 Example 3: Multiple Factions

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 and H_3. Suppose that all agents agree that these beliefs are related: H_2 probabilistically depends on H_1 (as in figure 3.4). Perhaps, H_1 represents the proposition, "*My barometer states that* the air pressure is low today", H_2 represents "It will rain today", H_3 represents "My barometer will give the correct reading," 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 figure 3.7. In general, augmenting networks in this way, by adding more polarizing nodes can increase the number of factions that may form.

3.5 Why do Populations Factionalize?

The examples in section 3.4 show how convergence and factionalization both arise, but not general divergence. In fact, given the definitions in section 3.3.3, 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.



(a)

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

(b)

Figure 3.2: (a) A Bayesian network structure with two variables, represented as 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_2 and H_1.

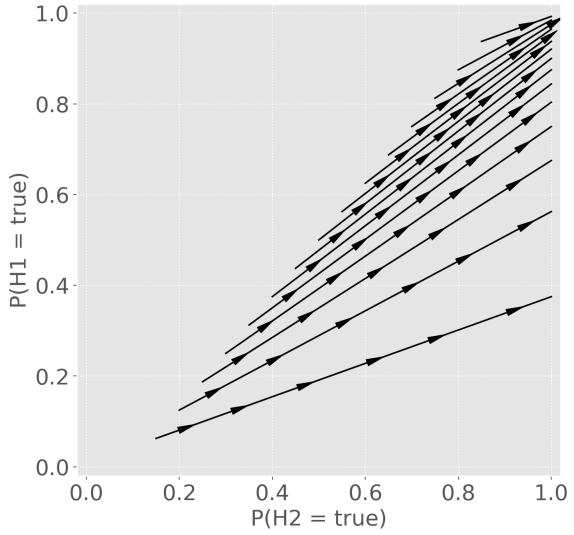
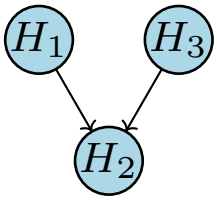


Figure 3.3: Belief trajectories for a population of 15 agents, with regards to two related hypotheses, H_1 and H_2 as in figure 3.2b. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives all agents to update in the same, positive direction about H_1. Arrow are indicative, showing only the directions in which their degrees of belief change.



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

Figure 3.4: (a) A Bayesian network structure with three variables, represented as 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_3, H_2 and H_1.

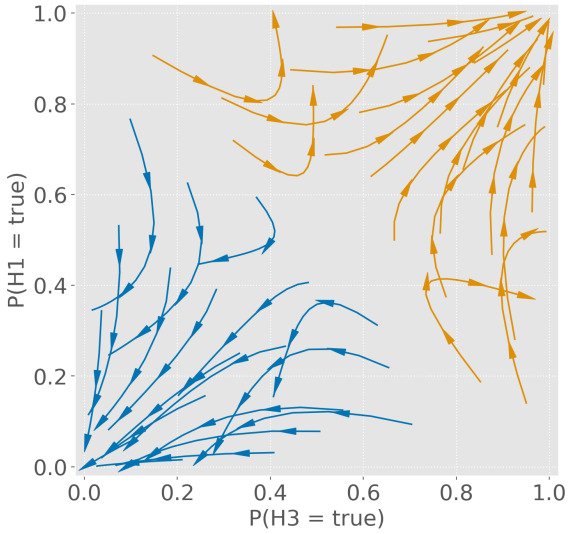
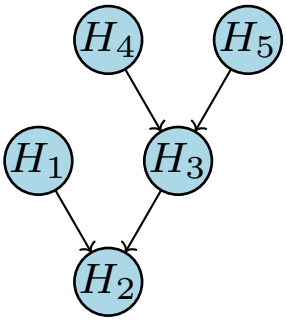


Figure 3.5: Belief trajectories for a population of 40 agents, as in figure 3.4. Only two beliefs, H_1 and H_3 are shown. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives the agents to polarize in their beliefs about H_1 and H_3. Observe that the agents 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 .



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

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

Figure 3.6: (a) A Bayesian network structure with five variables, represented as 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_3, H_2 and H_1. (c) The conditional probabilistic relations between H_5, H_4 and H_1.

]

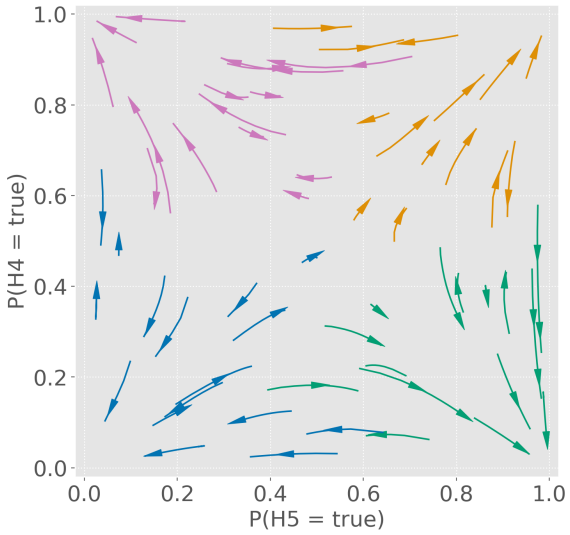


Figure 3.7: Belief trajectories for a population of 60 agents, as in figure 3.6. Only two beliefs, H_4 and H_5 are shown. The agents all update on 20 datapoints about H_2, each with a likelihood ratio of 0.65. This drives the agents to polarize in their beliefs H_1, in turn leading to four-way factionalization 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.

No General Divergence Condition. Suppose that we have two rational agents, with beliefs specified by joint probability distributions P(X, Y, ... Z, D) and Q(X, Y, ... Z, D) over the same set of discrete, binary variables, $\mathcal{X} = \{X, Y, \dots D\}$. 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 at least one agent is not certain about the value of D. Then, $D_{JS}(P(X, Y, \dots Z, D \mid D) \mid (P(X, Y, \dots Z, D \mid D)) < D_{JS}(P(X, Y, \dots Z, D) \mid (P))$.

Proof. From the Kullback-Leibler divergence chain rule (equation H.8) and the positivity of Kullback-Leibler entropy, it immediately follows that,

$$D_{KL}(P(X, Y, \dots Z \mid D) \mid (P(X, Y, \dots Z \mid D)) < D_{KL}(P(X, Y, \dots D) \mid (P(X, Y, \dots D)). \tag{3.6}$$

Furthermore,

$$D_{KL}(P(X, Y, \dots Z \mid D) = D_{KL}(P(X, Y, \dots Z, D \mid D). \tag{3.7}$$

Then,

$$D_{KL}(P(X, Y, \dots Z, D \mid D) \mid (Q(X, Y, \dots Z, D \mid D)) < D_{KL}(P(X, Y, \dots Z, D) \mid Q(X, Y, \dots Z, D)). \tag{3.8}$$

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 [14]. We can understand this as a *cumulativity of in-*

[14]However, this does not rule our general divergence as a possibility altogether. As I explain in appendix H, conditional Kullback-Leibler divergences are the expectations of the Kullback-Leibler divergences of the

formation 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 does not mean that the 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 relations between the variables (see appendix **??**). In fact, we can understand these conditions 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 I.

3.6 Conclusions

Epistemic factionalization arises very naturally and generically, 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:

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, surprising results could happen, and upon learning new information, the actual Kullback-Leibler divergences could increase.

agents to 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 a 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 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 or marginal probability distributions 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 kind of salient 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 I have explained in chapters 1 and 2, 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.

Appendix A

The Ide-Cozman Algorithm

I generate Bayesian networks according to an algorithm developed by Ide and Cozman (2002). This takes a number of nodes as input, and generates any of the possible directed acyclic graphs with that number of nodes with approximately equal probability. In fact, there is no known algorithm to uniformly sample Bayesian networks for any N in a finite time Ide and Cozman (2002). The algorithm here runs for some number I of iterations. It begins with a chain network, a connected graph in which each node has at most one parent and one child, and then iteratively adds and removes edges. The algorithm is "asymptotic" in that in the limit of an infinite number of iterations, it produces a uniformly distributed sample of Bayesian networks. In practice, to uniformly sample networks with N nodes, the authors recommend running the algorithm for $I = 4N^2$ iterations, a practice that I follow here. Once a directed acyclic graph is generated, I draw the associated conditional probabilities from a flat Dirichlet distribution (in effect, each number is independently drawn from a uniform distribution between 0 and 1). Finally, the initial probabilities associated with each exogeneous variable for agent are independently drawn from uniform a from a uniform distribution between 0 and 1. Ide and Cozman (2002) refer to this particular process, first drawing directed acyclic graphs and second generating the conditional probabilities, as the

"uniform generation of Bayesian networks", but note that this uniformity is only aprpoximate when run for a finite number of iterations.

Appendix B

D-Separation

We say that a set of variables, X, is independent of a second set, Y, given a third set Z when $P(X|Z) = P(X|YZ)$. Conditional independence is closely related to "d-separation", a structural property of causal graphs (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, pages 16-19). We define d-separation as follows. We define d-separation as follows.

> **Definition B.1** (d-separation)**.** Within a graph, G, a set of nodes, X, is d-separated from a second set, Y, given a third set Z if and only if every path between them is blocked by Z. A path p is blocked by a set of nodes Z if and only if one of these two criteria holds,
>
> 1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$, such that the middle node m is in Z, or
>
> 2. p contains an inverted fork (also known as a collider) $i \to m \leftarrow j$, such that the middle node m is not in Z and further, there are no descendants of m in Z.

If two sets of nodes are not d-separated, we say that they are "d-connected". In the case that the given set, Z is empty, we simply say that X and Y are d-separated. Finally, note that, although d-separation as defined is a property of sets of nodes, not individual nodes, in many cases we want to consider singleton sets. In such cases we use a convenient shorthand, whereby an individual node is d-separated from another node, given some third node.

Conditional independence and d-separation are related as follows.

> **Theorem B.1** (Pearl, 2009)**.** If sets X and Y are d-separated by Z in a directed acyclic graph, G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z, then X and Y are dependent conditional on Z in at least one distribution compatible with G.

In fact, we can say something stronger: if X and Y are not d-separated by Z, then they will be conditionally dependent in *almost all* distributions compatible with the graph G (see Verma and Pearl, 2013, Geiger and Pearl, 1993, Pearl, 2009, pages 18-19, Spirtes et al., 1993, pages 43-47).

Appendix C

Proofs of the Independence Conditions

Here, I prove that the independence condition is necessary for both contra-directional and transvergent updating.

Proof for contra-directional updating

Let us label the two agents 1 and 2, with joint probability distributions P_1 and P_2 respectively. We can write their prior beliefs about H as $P_1(H)$ and $P_2(H)$ respectively, and their conditional beliefs about H after updating on D, $P_1(H|D)$ and $P_2(H|D)$. Then recall that contra-directional updating will occur if,

$$\big(P_1(H|D) - P_1(H)\big) \times \big(P_2(H|D) - P_2(H)\big) < 0. \tag{C.1}$$

If we assume H, D are binary variables, this condition implies that,

$$\big(P_1(D|H) - P_1(D|\neg H)\big) \times \big(P_2(D|H) - P_2(D|\neg H)\big) < 0. \tag{C.2}$$

This requires that one of two conditions hold. Either,

$$\big(P_1(D|H) - P_1(D|\neg H)\big) > 1 \text{ and } \big(P_2(D|H) - P_2(D|\neg H)\big) < 1,$$

or

$$\big(P_1(D|H) - P_1(D|\neg H)\big) < 1 \text{ and } \big(P_2(D|H) - P_2(D|\neg H)\big) > 1,$$

We can write these conditions in terms of the likelihood ratios.

$$\frac{P_1(D|H)}{P_1(D|\neg H)} > 1 \text{ and } \frac{P_2(D|H)}{P_2(D|\neg H)} < 1,$$

or

$$\frac{P_1(D|H)}{P_1(D|\neg H)} < 1 \text{ and } \frac{P_2(D|H)}{P_2(D|\neg H)} > 1.$$

Two such Bayesian networks, differing only in the prior probabilities, can be re-expressed using a virtual node, β, with directed edges coming from β to each of the exogeneous nodes. The difference in the two agents' belief distributions can be entirely encoded through β. As such, in the rewritten networks, both agents can be represented using a single probability

distribution, but with different values assigned to β. Without loss of generality, let us assign $\beta = \beta_1$ to agent 1 and $\beta = \beta_2$ to agent 2. Then we can give both agents joint probability distributions differing only in the value of β, $P(\beta = \beta_1)$ and $P(\beta = \beta_2)$ respectively. Then we can rewrite the above conditions,

$$\frac{P(D|H, \beta = \beta_1)}{P(D|\neg H, \beta = \beta_1)} > 1 \text{ and } \frac{P(D|H, \beta = \beta_2)}{P(D|\neg H, \beta = \beta_2)} < 1,$$

or

$$\frac{P(D|H, \beta = \beta_1)}{P(D|\neg H, \beta = \beta_1)} < 1 \text{ and } \frac{P(D|H, \beta = \beta_2)}{P(D|\neg H, \beta = \beta_2)} > 1.$$

This entails that that D and H must not be conditionally independent given β, or that D and β must not be conditionally independent given H.

Proof for transvergent updating

Let us label the two agents 1 and 2, with joint probability distributions P_1 and P_2 respectively. We can write their prior beliefs about H as $P_1(H)$ and $P_2(H)$ respectively, and their conditional beliefs about H after updating on D, $P_1(H|D)$ and $P_2(H|D)$. Then recall that transvergent updating will occur if,

$$\big(P_1(H|D) - P_2(H|D)\big) \times \big(P_1(H) - P_2(H)\big) < 0. \tag{C.3}$$

Without loss of generality, assume that $P_1(H) - P_2(H) < 0$. Then, the condition implies that,

$$\frac{P_1(H|D)}{P_2(H|D)} < 1 \text{ and } \frac{P_1(H)}{P_2(H)} > 1.$$

Once again, rewriting in terms of the β variable,

$$\frac{P(H|D, \beta = \beta_1)}{P(H|D, \beta = \beta_2)} < 1 \text{ and } \frac{P_1(H|\beta = \beta_1)}{P_2(H|\beta = \beta_2)} > 1.$$

This entails that that H and D must not be conditionally independent given β. Furthermore,

$$P(\beta = \beta_1|D, H)P(D, H) < 1 \text{ and } P(\beta = \beta_2|D, H)P(D, H) > 1$$

or

$$P(\beta = \beta_1|D, H)P(D, H) > 1 \text{ and } P(\beta = \beta_2|D, H)P(D, H) < 1.$$

This entails that that β and D must not be conditionally independent given H.

Appendix D

Simulation Details

To investigate belief polarization, I generated random Bayesian networks with discrete variables that can take two values (1 or 0, indicating true or false). Each Bayesian network had two copies, representing the beliefs of two agents. The copies had identical nodes, edges, and conditional probabilities, but differed in the probabilities associated with exogeneous variables (i.e., variables without parents). As in the model of Freeborn (2023a), I "uniformly" generate the Bayesian networks using the Ide and Cozman (2002) algorithm. For each possible number of nodes, the algorithm generates any of the possible directed acyclic graphs with that number of nodes with approximately equal probability. The algorithm is "asymptotic" in that in the limit of an infinite number of iterations, it produces a uniformly distributed sample of Bayesian networks. I run the algorithm for $I = 4N^2$ iterations. Once a directed acyclic graph was generated, I drew the associated probabilities of exogeneous variables and conditional probabilities from a uniform distribution between 0 and 1. Note that these Bayesian belief networks may not represent realistic agents in any particular setting, and other choices could be made for how to generate such networks. However, the choice here is suitable for the task at hand, namely demonstrating that the polarization phenomena do not depend on any specific initial conditions.

For each network, I selected two variables: a variable upon which the agents update their beliefs (the "data node"), and a second variable, which we use to test the belief polarization ("the hypothesis node"). I selected a data node at random from all possible nodes, and then select a hypothesis node from all nodes that are not the data node, but which are d-connected to it. The agents both update on the same, randomly selected datapoint. They do this by updating their beliefs about the variable associated with the data node to either 1 or 0, chosen at random with equal probability (but the same for each agent). I then propagated the beliefs through each of their Bayesian networks and compared the beliefs of the two agents before and after updating.

Appendix E

Beliefs about Beliefs

Bayesian networks are factorized representations of joint probability distributions, and so can represent an agent's beliefs about their own beliefs, or other beliefs. However, as the literature has generally not discussed such Bayesian networks (for a partial exception, see Goodman et al., 2006), it is worth considering an example of what this would look like. Suppose that we begin with a joint probability distribution $P(H, S, D)$ (with the network from figure G.2) which we want to supplement with further beliefs about another (or the same) joint probability distribution, $O(H, S, D)$, to represent beliefs about another agent's (or the agent's own) beliefs.

We can create a supplemented joint probability distribution $P(H, S, D, H', S', D', \epsilon)$. One (though by no means the only) natural way to represent this network would be as in figure E.1. Here, H', S' and D' represent P's beliefs about the values O will believe about H, S and D respectively. ϵ represents the belief that an observation has taken place– and so both agents should expect to *agree* about the values of D i.e. ϵ should be defined such that whenever $\epsilon = 1$, then $P(D) = P(D')$, regardless of the values of the other variables.

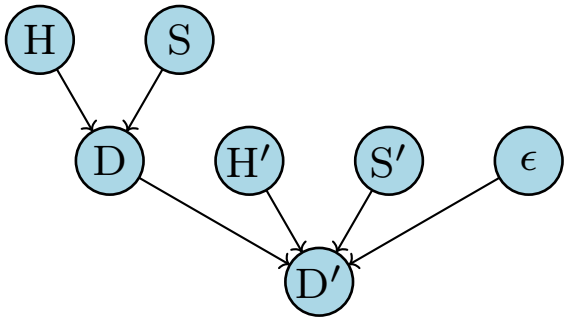


Figure E.1: A Bayesian network representing a factorized joint probability distribution $P(H, S, D, H', S', D', \epsilon)$, containing beliefs about H, S, and D, as well as beliefs about another (or the same) agent's beliefs about H, S and D. ϵ represents beliefs about an act of observing D.

Appendix F

Proof of Expectable Polarization Incompatibility Condition

Let us suppose that we have two agents with prior joint probability distributions O and P over some set X of binary hypotheses, including H and D, let us refer to O' and P' as posterior probability distributions. Let us suppose that the agents share the same conditional relationships between the beliefs in X, but may differ in their prior probabilities. Let us further supplement probability distributions O and P with additional beliefs, about their own beliefs and the beliefs of the other agent.

Suppose both expectable polarization and actual polarization arise for O and P. Applying the principles of reflection (2.3 and 2.4) and mutual knowledge (2.5 and 2.6) to equations 2.7 and 2.8, we have,

$$P(D)\big[O(H|D) - O(H|\neg D)\big] + O(H|\neg D) < O(D)\big[O(H|D) - O(H|\neg D)\big] + O(H|\neg D) \tag{F.1}$$

$$O(D)\big[P(H|D) - P(H|\neg D)\big] + P(H|\neg D) > P(D)\big[P(H|D) - P(H|\neg D)\big] + P(H|\neg D). \tag{F.2}$$

However, by equations 2.9 and 2.10, we have that,

$$O(H|D) - O(H|\neg D) < 0 \tag{F.3}$$

$$P(H|D) - P(H|\neg D) > 0. \tag{F.4}$$

Substituting this into equations F.1 and F.2 yields,

$$P(D) > O(D) \tag{F.5}$$

$$P(D) < O(D), \tag{F.6}$$

which is a contradiction.

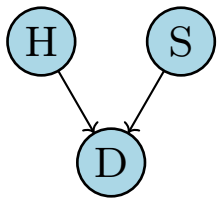
Appendix G

Examples of Expected or Expectable Polarization

G.1 Expected contra-directional updating

Suppose that O and P obey the assumptions of section 2.7, and have beliefs about the three two-valued hypotheses H, D and S as shown in figure G.1. Let $O(H) = P(H) = 0.5$, let $O(S) = 0.9$ and let $P(S) = 0.1$. Here both agents will expect to polarize in their beliefs about H upon updating on D. Agent O expects that P will decrease their degree of belief in H ($\mathbb{E}_O(P'(H)) = 0.44\ldots < 0.5$), whilst agent P expects that O will increase their degree of belief in H ($\mathbb{E}_P(O'(H)) = 0.73\ldots > 0.5$). This polarization is driven by their different beliefs about the probability of the evidence, that D is true or false ($O(D) = 0.86$, whereas $P(D) = 0.54$). Each agent believes that the other has wrong beliefs about D, and so will update their beliefs in a particular direction.

However, there is a subtlety here. Both agents cannot be right: upon updating upon an actual value of D (regardless of whether it is found to be $D = d$ or $D = \neg d$), contra-



H	S	$P(D = d)$
$\neg h$	$\neg s$	0.1
$\neg h$	s	0.9
h	$\neg s$	0.9
h	s	0.9 †

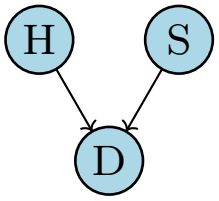
Figure G.1: A Bayesian network relating two-valued variables H, S and D, with the associated conditional probability table for variable D. For clarity, beliefs about the agents' own beliefs, and about the other agents' beliefs are not depicted. Two agents with these beliefs, and satisfying reflection and the mutual knowledge condition, may expect to experience contra-directional updating.

directional updating will not in fact take place. Both agents must update in the same direction. What if we were to tweak the agent's conditional beliefs to force them to contra-directionally update H when they learn the true value D? For instance what if we could lower the value of $P(D = d|HS)$ († in figure G.1) to 0.8? Now the different beliefs about S would now be just enough to guarantee that the agents update in opposite directions. However, this same change means that the agents agree more about the probability that $D = d$, ($O(D) = 0.815$, whereas $P(D) = 0.535$), just enough that they no longer expect to update in opposite directions. For the difference in beliefs about D to be large enough to drive they expectation of polarization, the difference must also be large enough that updating on D must cause the agents to update H in the same direction upon learning the true value of D.

We see that the agents can have rational expectations of contra-directional updating to take place. However expectable contra-directional updating, in which the directions of belief revision can be predicted in advance cannot arise.

G.2 Expectable Belief Divergence

Now, suppose that O and P obey the assumptions of section 2.7 and have beliefs about the three two-valued hypotheses H, D and S as shown in figure G.2. Let $O(H) = P(H) = 0.5$, let $O(S) = 0.9$ and let $P(S) = 0.1$. Here, neither each agent thinks it is equally likely for the other agent to change their beliefs in either direction ($\mathbb{E}_O(P'(H)) = 0.5$ and $\mathbb{E}_P(O'(H)) = 0.5$). However, both rationally expect the beliefs to grow further apart upon updating on D ($\mathbb{E}_P|P'(H) - O'(H)| = \mathbb{E}_O|P'(H) - O'(H)| = 0.35\ldots$). Furthermore, upon updating their beliefs on $D = d$ or $D = \neg d$, the agents will experience actual belief divergence in their beliefs about H. The key is that the agents may expect their beliefs to diverge, and indeed may expect their beliefs to update in opposite directions. However, the agents cannot predict in advance which direction their beliefs will diverge in.



H	S	$P(D = d)$
$\neg h$	$\neg s$	0.9
$\neg h$	s	0.1
h	$\neg s$	0.1
h	s	0.9

Figure G.2: A Bayesian network relating two-valued variables H, S and D, with the associated conditional probability table for variable D. For clarity, only a sub-network of the beliefs is shown. Beliefs about the agents' own beliefs, and about the other agents' beliefs are not depicted. Two agents with these beliefs, and satisfying reflection and the mutual knowledge condition, may expect to experience belief divergence.

Appendix H

Information-theoretic Quantities for Discrete Variables

Here, I outline some of the key information-theoretic quantities that I use (see Cover and 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 [1]. Figure H.1 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,

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

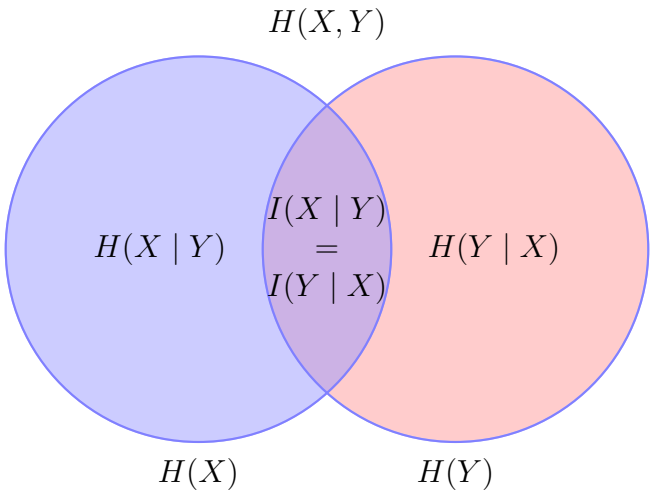


Figure H.1: A Venn diagram relating various quantities of information for two variables, X and Y in a joint probability distribution.

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x), \qquad \textbf{(Entropy)} \quad \text{(H.1)}$$

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,

$$H(X_1, \ldots X_N) = -\sum_{\substack{x_1 \in \mathcal{X}_1, \\ \ldots, \\ x_N \in \mathcal{X}_N}} P(x_1, \ldots x_N) \log P(x_1, \ldots x_N). \qquad \textbf{(Joint Entropy)} \quad \text{(H.2)}$$

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,

$$H(Y \mid X) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)}. \qquad \textbf{(Conditional Entropy)} \quad \text{(H.3)}$$

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:

$$H(Y \mid X) = H(X \mid Y) - H(X) + H(Y). \qquad \textbf{(Bayes' Rule for Entropy)} \quad \text{(H.4)}$$

This is an additive analogue for Bayes' rule for probabilities. The conditional is entropy always greater than or equal to zero, and always less than the marginal entropy: $0 \leq H(Y \mid X) \leq (Y)0$. 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.

$$H(X_1, \ldots X_N) = \sum_{i=1}^{N} H(X_i \mid X_1, \ldots X_{i-1}). \qquad \textbf{(Chain Rule for Entropy)} \quad \text{(H.5)}$$

This is an additive analogue to the chain rule for probability (see equation 1.2).

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.

$$I(X \mid Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} = H(Y) - H(Y \mid X). \quad \textbf{(Mutual Information)} \quad \text{(H.6)}$$

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 figure H.1). 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 figure H.1) . 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 (equation 3.3), 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 distribution $P(x)P(y)$, 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,

$$D_{KL}(P \mid Q) = -\sum_{\substack{x_1 \in \mathcal{X}_1, \\ \cdots, \\ x_N \in \mathcal{X}_N}} P(x_1, \ldots x_N) \log \frac{P(x_1, \ldots x_N)}{Q(x_1, \ldots x_N)}, \qquad \textbf{(KL Divergence)} \quad \text{(H.7)}$$

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,

$$D_{KL}(P(x, y) \mid Q(x, y)) = D_{KL}(P(x) \mid Q(x)) + D_{KL}(P(x \mid y) \mid Q(x \mid y)),$$
$$\textbf{(KL Divergence Chain Rule)} \quad \text{(H.8)}$$

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, $D_{KL}(P(x \mid y) \mid Q(x \mid y)) = \mathbb{E}_P[D_{KL}(P(x \mid y) \mid Q(x \mid y))]$.

Moreover, unlike the mutual information, the Kullback-Leibler divergence 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 two 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,

$$D_{JS}(P \mid Q) = \frac{1}{2} D_{KL}\left(P \,\middle|\, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q \,\middle|\, \frac{P+Q}{2}\right). \quad \textbf{(JS Divergence)} \tag{H.9}$$

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 $(P \mid Q) = 0$. 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 Rényi, 1961; Morimoto, 1963; Csisz'ar, 1964; Ali and Silvey, 1966). 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. Unlike the Kullback-Leibler divergence, it is bounded: for a joint probabilty

distribution over N variables, $0 \leq D_{JS}(P|Q) \leq \frac{N}{2}\log(2))$. 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 and Schindelin, 2003; Fuglede and 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 I

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 (section 3.5) 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 section 1.5.

Recall (see equation H.5) that we can rewrite the conditional entropy of a joint probability distribution, given some variable as follows,

$$H(X, Y, \ldots Z \mid D) = H(X) + H(X \mid Y) + \ldots H(D|X, Y, \ldots) - H(D). \tag{I.1}$$

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

$$H(X, Y, \ldots Z \mid D) = \sum_{A \in \mathcal{A}} H(A) + \sum_{B \in \mathcal{B}} H(B \mid \mathcal{A}) - H(D). \tag{I.2}$$

$$= \sum_{A \in \mathcal{A}} H(A \mid D) + \sum_{B \in \mathcal{B}} H(B \mid \mathcal{A}, D) \tag{I.3}$$

Let us call the first term the exogeneous entropy and the second term the endogeneous 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 exogeneous entropy or the endogeneous entropy (or both) be expected to fall upon learning D.

Suppose that we satisfy the two independence conditions in section 1.5,

$$P(\beta \mid H) \neq P(\beta \mid DH) P(H \mid \beta) \neq P(H \mid D\beta) \tag{I.4}$$

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 probability distributions increases.

Bibliography

Ali, Syed Mumtaz and Samuel D. Silvey (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.

Almagro, Manuel (2022). Political polarization: Radicalism and immune beliefs. *Philosophy and Social Criticism*.

Axelrod, Robert (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, 41 (2): 203–226.

Baldassarri, Delia and Peter Bearman (2007). Dynamics of political polarization. *American sociological review*, 72 (5): 784–811.

Baron, J. (2008). *Thinking and Deciding*. Cambridge University Press, 4th edition.

Batson, C. Daniel (1975). Rational processing or rationalization? the effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, 32: 176–184.

Blackwell, David and Lester Dubins (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3): 882–886.

Borboudakis, Giorgos, Sofia Triantafillou, and Ioannis Tsamardinos (2012). Tools and algorithms for causally interpreting directed edges in maximal ancestral graphs. *Sixth European Workshop on Probabilistic Graphical Models*.

Bradley, Richard (2005). Radical probabilism and bayesian conditioning*. *Philosophy of Science*, 72(2): 342–364.

Bramson, Aaron, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, 84(1): 115–159.

Chan, Hei and Adnan Darwiche (2005). On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163(1): 67–90.

Cook, John and Stephan Lewandowsky (2016). Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science*, 8(1): 160–179.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. Wiley, Hoboken, New Jersey.

Csisz'ar, Imre (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.

Deffuant, Guillaume (2006). Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, 9 (3).

Deffuant, Guillaume, Frederic Amblard, Gerard Weisbuch, and Thierry Faure (2002). How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5 (4).

Diaconis, Persi and Sandy L. Zabell (1982). Updating subjective probability. *Journal of the American Statistical Association*, 77(380): 822–830.

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

Dizadji-Bahmani, Foad, Roman Frigg, and Stephan Hartmann (2011). Confirmation and reduction: A bayesian account. *Synthese*, 179.

Dorst, Kevin (2022). Rational polarization. `https://philarchive.org/archive/DORRP-2v4`.

Endres, Dominik Maria and Johannes E Schindelin (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7): 1858–1860.

Field, Hartry (1978). A note on jeffrey conditionalization. *Philosophy of Science*, 45(3): 361–367.

Freeborn, David P. W. (2023a). Convergence and polarization for agents with bayesian belief networks. unpublished manuscript.

Freeborn, David P. W. (2023b). Rational polarization for agents with multiple, related beliefs. unpublished manuscript.

Fuglede, Bent and Flemming Topsoe (2004). Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory, ISIT 2004 Proceedings*, page 31. IEEE.

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

Gerber, A. and D. Green (1999). Misperceptions about perceptual bias. *Annual Review of Political Science*, 2: 189–210.

Goodman, Noah, Chris Baker, Elizabeth Bonawitz, Vikash Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua Tenenbaum (2006). Intuitive theories of mind: A rational approach to false belief. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society.*

Grim, Patrick, Frank Seidl, Calum McNamara, Isabell N. Astor, and Caroline Diaso (2022). The punctuated equilibrium of scientific change: A bayesian network model. *Synthese*, 200(4): 1–25.

Grim, Patrick, Frank Seidl, Calum McNamara, Hinton Rago, Isabell Astor, Caroline Diaso, and Peter Ryner (forthcoming, 2021). Scientific theories as bayesian nets: Structure and evidence sensitivity. *Philosophy of Science.*

Halton, John H. and G. B. Smith (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7: 701–702.

Hamilton, Lawrence C, Joel Hartter, and Kei Saito (2015). Trust in scientists on climate change and vaccines. *Sage Open*, 5(3): 2158244015602752.

Hartmann, Stephan and Luc Bovens (2002). *Bayesian Networks in Philosophy.* Dordrecht: Kluwer.

Hegselmann, Rainer and Ulrich Krause (2002). Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5.

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

Ide, Jaime and Fabio Cozman (2002). Random generation of bayesian networks. volume 2507, pages 366–375.

Jacobs, Bart (2018). A mathematical account of soft evidence, and of jeffrey's 'destructive' versus pearl's 'constructive' updating. *CoRR*, abs/1807.05609.

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

Jeffrey, Richard C. (1988). *Conditioning, Kinematics, and Exchangeability*, volume 1, pages 221–255. Kluwer, Dordrecht.

Jern, Alan, Kai-Min K Chang, and Charles Kemp" (2009). Bayesian belief polarization (supporting material). In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS).*

Jern, Alan, Kai-Min K Chang, and Charles Kemp (2014). Belief polarization is not always irrational. *Psychological Review, 121(2), 206–224.*

Jung, Jiin, Patrick Grim, Daniel J. Singer, Aaron Bramson, William J. Berger, Bennett Holman, and Karen Kovaka (2019). A multidisciplinary understanding of polarization. *American Psychologist*, 74: 301–314.

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

Kawakatsu, Mari, Yphtach Lelkes, Simon A. Levin, and Corina E. Tarnita (2021). Interindividual cooperation mediated by partisanship complicates madison's cure for "mischiefs of faction". *Proceedings of the National Academy of Sciences*, 118(50): e2102148118.

Kelly, Thomas (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy*, 105(10): 611–633.

Kitcher, Philip (1995). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press on Demand.

Kocis, Ladislav and William J. Whiten (1997). Computational investigations of low-discrepancy sequences. *ACM Trans. Math. Softw.*, 23: 266–294.

Lakoff, G. (2010). *Moral Politics: How Liberals and Conservatives Think*. Chicago University Press, Chicago.

Latkin, Carl, Lauren Dayton, Catelyn Coyle, Grace Yi, Abigail Winiker, and Danielle German (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.V. and P. Rathie (1978). On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, 24(1): 120–122.

Lee, Carol H.J. and Chris G. Sibley (2020). Attitudes toward vaccinations are becoming more polarized in new zealand: Findings from a longitudinal survey. *EClinicalMedicine*, 23: 100387.

Lehrer, Ehud and Rann Smorodinsky (1996). Merging and learning. *Lecture Notes-Monograph Series*, 30: 147–168.

Levin, Simon A., Helen V. Milner, and Charles Perrings (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences*, 118(50): e2116950118.

Liberman, Akiva and Shelly Chaiken (1992). Value conflict and thought-induced attitude polarization. *Journal of Personality and Social Psychology*, 63(4): 618–628.

Lord, Charles G., Lee D. Ross, and Mark R. Lepper (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37: 2098–2109.

Macy, Michael W, James A Kitts, Andreas Flache, and Steve Benard (2003). Polarization in dynamic networks: A hopfield model of emergent structure. *Dynamic social network modeling and analysis*, pages 162–173.

Madsen, Jens, Richard Bailey, and Toby Pilditch (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports*, 8.

McHoskey, John W. (1995). Case closed? on the john/joan case. *Psychology, Public Policy, and Law*, 1(1): 134–142.

Morimoto, Tetsuzo (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3): 328–331.

Mrad, Ali Ben, Véronique Delcroix, Sylvain Piechowiak, Philip Leicester, and Mohamed Abid (2015). An explication of uncertain evidence in bayesian networks: Likelihood evidence and probabilistic evidence. *Applied Intelligence*, 43(4): 802–824.

Munro, G. D. and P. H. Ditto (1997a). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6): 636–653.

Munro, Geoffrey D. and Peter H. Ditto (1997b). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6): 636–653.

Nielsen, Frank (2021). On a variational definition for the jensen-shannon symmetrization of distances based on the information radius. *Entropy*, 23(4): 464.

Nielsen, Michael (2018). Deterministic convergence and strong regularity. *The British Journal for the Philosophy of Science*, 71.

Nielsen, Michael and Rush T. Stewart (2021). Persistent disagreement and polarization in a bayesian setting. *British Journal for the Philosophy of Science*, 72(1): 51–78.

O'Connor, Cailin and James Owen Weatherall (2018). Scientific polarization. *European Journal for Philosophy of Science*, 8(3): 855–875.

Pallavicini, Josefine, Bjorn Hallsson, and Klemens Kappel (2021). Polarization in groups of bayesian agents. *Synthese*, 198.

Pearl, Judea (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. Technical Report CSD-850017, University of California, Los Angeles, Computer Science Department.

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

Pearl, J. (2009). *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press.

Plous, Scott (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, 21: 1058–1082.

Powell, D., Kara Weisman, and E. Markman (2018). Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions. *Cognitive Science*.

Rényi, Alfréd (1961). *On measures of entropy and information.* University of California Press, Berkeley, CA.

Rogers, Everett M (1983). *Diffusion of innovations.* Simon and Schuster.

Schervish, M. J. and T. Seidenfeld (1990). An approach to consensus and certainty with increasing information. *Journal of Statistical Planning and Inference*, 25: 401–414.

Skyrms, Brian (1995). Strict coherence, sigma coherence and the metaphysics of quantity. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 77(1): 39–55.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993). *Causation, Prediction, and Search*, volume 81. Springer Series in Statistics.

Sprenger, Jan (2017). Foundations of a probabilistic theory of causal strength. `http://philsci-archive.pitt.edu/14108/`.

Taber, Charles S., Damon Cann, and Simona Kucsova (2009). The motivated processing of political arguments. *Political Behavior*, 31(2): 137–155.

Taber, Charles S. and Milton Lodge (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3): 755–769.

Verma, Tom and Judea Pearl (2013). Causal networks: Semantics and expressiveness. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, 4.

Wagner, Carl G (2002). Probability kinematics and commutativity. *Philosophy of Science*, 69(2): 266–278.

Weatherall, James and Cailin O'Connor (2021). Endogenous epistemic factionalization. *Synthese*, 198.