Too Much Information & The Death of Consensus

John W.E. Cremin *†

Aix-Marseille University, CNRS, and AMSE

November 19, 2025

Latest Version

Abstract

Modern society is increasingly polarized, even on purely factual questions, despite greater access to information than ever. In a model of sequential social learning, I study the impact of *motivated reasoning* on information aggregation. This is a belief formation process in which agents trade-off accuracy against ideological convenience. I find that even Bayesian agents only learn in very highly connected networks, where agents have arbitrarily large neighborhoods asymptotically. This is driven by the fact that motivated agents sometimes reject information that can be inferred from their neighbors' actions when it refutes their desired beliefs. Observing any finite neighborhood, there is always some probability that all of an agent's neighbors will have disregarded information thus. Moreover, I establish that *consensus*, where all agents eventually choose the same action, is only possible with relatively uninformative private signals and low levels of motivated reasoning. JEL Codes: D72, D83, D85.

Keywords: Social Learning, Motivated Reasoning, Networks, polarization

1 Introduction

Why are we so polarized? On questions of ethics and ideology persistent disagreement is unsurprising, yet polarization extends even to matters of fact: Was the 2020 election swung by mass voter fraud? Is climate change man-made? Do vaccines cause autism? The '*polarization of reality*' observed by Alesina et al. (2020) now seems a fixed feature of American politics;

*Aix-Marseille School of Economics: john-walter-edward.cremin@univ-amu.fr.

†I would like to thank Evan Sadler for his guidance with this project, as well as Navin Kartik and Jacopo Perego for many fruitful discussions and suggestions. I would also like to thank anonymous referees and participants in the Columbia Economic Theory Workshop for their useful comments.

even before the 2020 election, 62% of Trump supporters endorsed the claim that millions of illegal votes were cast in 2016, against 25% amongst Clinton supporters, whilst conversely 9% and 50% believed Russian tampering with vote tallies improved Trump's performance respectively (Nyhan, 2020). That partisan *factual* polarization has grown more severe[1] would seem paradoxical in a world with ever greater access to, and ability to share, information. However, the obvious feature common to all these fields (politics, religion, ethics...) is their emotional salience, and relevance to the very identity of the reasoner.

To study social learning on such questions, we must consider that people may be engaging in *motivated reasoning*.[2] If so, belief formation is no longer directed only at forming accurate beliefs, but at striking a balance between this and the desire to believe that reality is convenient or pleasant. Specifically, I model motivated reasoners as ignoring information that can be inferred from their neighbors' behavior when it provides strong evidence against their preferred state, and using a biased prior otherwise. Precise details can be found in Section 3. Experimentally, Oprea and Yuksel (2022) find evidence of motivated reasoning when subjects form beliefs concerning their own IQ, and Guilbeault et al. (2018) show that increasing the 'salience of partisanship' can significantly damage social learning by provoking motivated reasoning. More generally, Westen et al. (2006) and Moore et al. (2021) use fMRI scanners to show that emotional, not analytical, reactions are triggered by 'threatening' political information.[3]

As people become more polarized on ideology and values, as polling evidence suggests they have (Geiger, 2021), the importance of understanding social learning with motivated reasoning grows. Such value polarization implies stronger 'directional motives' (Little, 2021), i.e. a greater desire to believe the true state of the world is congenial to one's ideological leanings, and thus even more biased belief-formation on purely factual questions. For example, upon reading a study on the effectiveness of mask mandates, an agent growing more and more libertarian will be more likely to conclude that they did not successfully suppress Covid-19. Their ideological opposites, reading the same study, shall form even stronger opinions that

[1]As noted by Rekker (2022), there is a surprisingly sparse literature on the change in factual polarization over time, and thus it is difficult to confidently establish that it has in fact increased. The literature on partisan perceptions of economic performance provides some evidence of persistent and growing factual polarization (Campbell et al., 1980; Gerber and Huber, 2009; Bartels, 2002), and in any case the pervasive interest in 'post-truth' politics suggests possible explanations of such an increase are worth pursuing.

[2]I do not suggest that this is the *only* possible explanation of polarization, but rather a particularly plausible one for such subjects.

[3]Bénabou and Tirole (2016) observe that the emotional nature of some subjects is evident even without fMRI scanners: '*Heat versus light: Finally, in "motivated" [reasoning] there is also emotion. Challenging cherished beliefs directly- like a person's religious identity, morality, or politics- evokes strong emotional and even physical responses of anger, outrage, and disgust. Such pushback is a clear "signature" of protected beliefs: not only would a Bayesian always welcome more data, but so would any naïve boundedly rational thinker.*' They also note the fact that low individual stakes are conducive to motivated reasoning (holding incorrect beliefs about the extent to which climate change is man-made does not cause your own house to catch fire!), a point reinforced by Zimmermann (2020).

they did. This paper offers a resolution of the seemingly paradoxical coincidence of increased factual polarization and the unprecedented ease of access to information we have today. I show that increasing access to information, the connectivity of social networks, the level of value polarization, or any combination of these three can break social consensus, and cause individuals to divide along party lines. Hence the above stylised facts form a perfect storm for factual polarization with motivated reasoners.

A canonical model economists have studied to analyse the spread of information and beliefs is that of sequential social learning, where agents act in turn and observe both an independent and identically distributed private signal, and a social signal: the behavior of some subset of their predecessors. This allows the study of learning with arbitrary social networks, without sacrificing sophisticated belief formation procedures. Agents each observe these signals, before attempting to match their own action to the state, where both are binary. The exact subset of predecessors observed is determined by the network topology: I study the conditions on this and the private signals necessary for and/or sufficient to achieve learning.

I find that motivated reasoning seriously impedes both *learning* (where Bayesians eventually match the state with arbitrarily high probability) and *consensus* (where all agents eventually choose the same action). Firstly, *Consensus* (eventual action unanimity) becomes impossible as the support of private beliefs expands, as this ensures that all social beliefs are either rejected or too weak to command unanimity (Theorem 1). The absence of even complete-network[4] consensus with unbounded beliefs, despite common knowledge of the true model, reflects the comparatively extreme nature of the bias I study, and specifically the non-monotonic response to information it involves. In contrast, Bohren and Hauser (2021) study a wide range of biases, and find in all cases that consensus[5] is necessarily achieved with sufficiently little misspecification.

Beyond consensus, Theorem 2 establishes that in order to achieve learning, we need a much stronger condition than in the standard Bayesian model. In that setting, being indirectly connected to ever-larger sets of agents (*Expanding Observations*) is necessary and sufficient for learning with *unbounded signals* (Acemoglu et al., 2011), where these are signals that can leave a Bayesian arbitrarily close to certain of the state. With motivated reasoners, we can instead only achieve learning if agents are *directly* connected to such sets (*Expanding Sample Sizes*, a novel condition), albeit with *nonstationary* signals (a category that includes unbounded and some bounded signal structures). One intuition for this is that with motivated reasoners information can be 'lost,' since these agents sometimes 'reject' information gleaned from observing their neighbors' actions. This is easier to see in deterministic, simple structures: a tractable example is the line network, in which agents each observe only their immediate predecessor.

[4]The complete network involves each agent observing the actions of all predecessors.

[5]What I call *consensus* corresponds to what they call *learning*.

Such a network structure guarantees learning with unbounded beliefs and Bayesian agents, but fails with motivated agents since it does not satisfy expanding sample sizes. In my model, expanding sample sizes is necessary for learning as upon observing any neighborhood with only M members there is some probability that all observed neighbors rejected their social signals, and an agent cannot achieve a level of accuracy arbitrarily close to one with such information. This problem fades as neighborhood size expands. It is also the case that in those settings where, in Bayesian societies without expanding sample sizes, only finitely many agents would make the wrong decision in equilibrium, introducing an arbitrarily tiny number of motivated reasoners, engaging in an arbitrarily tiny level of motivated reasoning, is enough to ensure that infinitely many Bayesians instead fail to match the state. I demonstrate this in Corollary 2.1, in which any amount of motivated reasoning can be seen to impede the *fast* convergence required for such consensus.

Theorem 3 and Proposition 2 then elucidate in what precise sense more connected and clustered social networks produce more polarization. Defining a *nested neighbor* of agent n as any neighbor whose neighborhood n also certainly observes, Theorem 3 establishes that with nonstationary signal structures, any network topology in which agents observe an ever-growing number of nested neigbours will achieve learning, and the exact pattern of party polarization this produces. Proposition 2 then establishes that even if agents do not asymptotically observe infinitely many nested neighbors, the probability an agent rejects his social signal can be made arbitrarily close to 1 by supposing that they observe enough of them. As I argue in Section 6, nested neighborhoods may be a reasonable representation of online social networks, particularly given the level of connectivity and clustering they exhibit, and the sequential structure of the game. Polarization is a complex phenomenon that will result from many factors, such as misspecification (Bohren and Hauser, 2021), selective news sharing (Bowen et al., 2023), and echo chambers (Levy and Razin, 2019; Acemoglu et al., 2024) to name a few. The results of this paper illustrate how motivated reasoning can exacerbate it, and especially so in the world the internet has built.

The paper is organised as follows. After reviewing the literature in Section 2, in Section 3 I describe the set-up of the model, how motivated agents form beliefs and what information they have. Section 4 defines the solution concept, my notions of learning and consensus, and my partition over information structures, as well as presenting the decision rules of agents and establishing equilibrium existence. Consensus is discussed in both Sections 5 and 6, with Theorem 1 setting out the main implications of motivated reasoning for this, and Corollary 2.1 demonstrating the fragility of consensus in Bayesian societies without expanding sample sizes. My main result on learning is Theorem 2, in Section 6, on the role of expanding sample sizes. This section also contains a simple example illustrating the problems motivated reasoners present for learning, and presents a sufficient condition for learning with nonstationary signal structures in Theorem 3. Discussion of extensions is contained in section 7, and section 8 concludes.

2 Literature Review

There is a large literature on sequential social learning, going back to classic articles by Bikhchandani et al. (1992) and Banerjee (1992). These and much of the early literature are developed by Smith and Sorensen (2000), who provide a general analysis of sequential social learning on a complete network, in which all agents observe all those who came before, and were the first to discover the distinction between bounded and unbounded beliefs. Following on from this initial benchmark, there are two clear literatures developing this in two distinct directions. The first concerns the exploration of social learning in general network topologies, and the second investigates behavioral biases and/or misspecification.

Acemoglu et al. (2011) extend Smith and Sorensen (2000) in allowing for arbitrary social network structures, with the caveat that agents' neighborhoods are independent of each other. In this setting, they find that *Expanding Observations*- a minimal connectivity condition- is necessary and sufficient for asymptotic learning with unbounded beliefs.[6] A small literature following Acemoglu et al. (2011) has developed, for example containing Lobel and Sadler (2015, 2016) and Lomys (2020). The first of these removes the neighborhood independence assumption of Acemoglu et al. (2011), unlike the model I present here,[7] and the second studying learning in a setting where agents have different preferences over the two actions. This turns out to be sufficient to break learning in general networks, and some of the learning problems in my model are reminiscent of it. All four papers establish learning in different settings using either an *Improvement Principle* or a *Large Sample Principle*, I further discuss these and compare their uses to my setting in Section 6. Papers such as Çelen and Kariv (2004), Acemoglu et al. (2009), Dasaratha and He (2024) and Rosenberg and Vieille (2019) consider the speed of learning, and whether or not convergence in such general networks is almost sure or in probability. I have notions of learning and consensus defined in both modes of convergence, and in particular use the example of fast convergence in a line network from Rosenberg and Vieille (2019) to establish the fragility of Bayesian correct consensus to motivated reasoning

[6]In addition to these articles studying general network topologies with the Acemoglu et al. (2011) framework, there are of course articles such as that by Çelen and Kariv (2004) that study specific non-complete network topologies such as the line network. I follow the Acemoglu et al. (2011) approach, as real world networks are inevitably going to contain all sorts of arbitrary patterns, making results on general network topologies of much more use in studying social learning.

[7]Note that an implication of this it that motivated reasoning can damage social learning even *without* type homophily and echo chambers.

in my Corollary 2.1. Separately, the results of Dasaratha and He (2024) concern the speed of learning with 'confounding' (when agents do not observe the common neighbors of their own neighbors, and cannot correct for the resulting correlation in observed actions) in a model with generations of agents, and is thus related to my Proposition 2 in which I show that observing enough *nested neighbors* (neighbors for which there is no such 'confounding', c.f. Condition 3) guarantees a minimum probability of success. They also speak of information 'loss', but in a different sense to the loss I discuss in Section 6.1: improvement reasoning still holds in their model, so whereas my loss reflects agents performing worse than their predecessors, the loss of Dasaratha and He (2024) concerns them improving on said neighbors very slowly.

I am by no means the first to introduce behavioral biases into social learning, though articles pursuing this extension do so largely on the complete network (Bohren, 2016; Bohren and Hauser, 2021; Eyster and Rabin, 2009; Arieli et al., 2023). This fact ensures that the analyst can study 'the' social belief at period n, and analyze asymptotic learning outcomes by characterising the properties of this stochastic process. Intuitively, it means that agents always have access to all the social information their predecessors did, and this information cannot be 'lost'. As I explain in section 6.1, this possible loss of information is a major problem for learning in more general network topologies. Beyond this, the particular bias I study is quite different to those studied by other articles. As mentioned in the Introduction, Bohren and Hauser (2021) alone cover an array of behavioral biases where agents also hold misspecified beliefs over the fractions of agents with each bias, but find that without misspecification agents always achieve consensus with unbounded beliefs (Bohren and Hauser, 2021, Theorem 6). In contrast, agents in my model fail to achieve consensus with unbounded beliefs on the complete network, even in the absence of misspecification. This is because for each bias they consider, in the absence of misspecification, agents all agree which actions are optimal in each state, regardless of their type, and respond to strong evidence that the true state is θ by choosing that action. Conversely, my model involves a more extreme bias, with agents engaging in non-Bayesian updating and sometimes exhibiting a non-monotonic response to social information. In this sense it resembles other models with non-Bayesian updating in the non-sequential literature, such as the classic DeGroot (1974), but the non-Bayesian updating rule here is far less mechanical. Molavi et al. (2018) use an axiomatic approach to study a generalization of DeGroot (1974), which does involve agents using Bayesian reasoning to incorporate private information as here. However, their monotonicity axiom explicitly rules out the motivated reasoning procedure I model.

3 Model

An infinite sequence of agents each labelled by $n \in \mathbb{N}$ must choose one of two available actions $x_n \in \{0,1\}$, and seek to match the binary state: $\theta \in \{0,1\}$. Their utility function is the following:

$$u_n(x_n, \theta) = \begin{cases} 1 & \text{if } x_n = \theta, \\ 0 & \text{if } x_n \neq \theta, \end{cases}$$

All agents have a common prior, and nature draws the state at the beginning of the game according to this prior. To simplify notation, I assume that $\mathbb{P}(\theta = 1) = \frac{1}{2}$, though the results generalise. Each agent receives an independent and identically distributed private signal $\varsigma_n \sim \mathbb{F}_\theta$, where $(\mathbb{F}_0, \mathbb{F}_1)$ makes up the information structure of the game. These private signal distributions are assumed to be absolutely continuous with respect to each other, i.e. there are no perfectly informative signals, and informative (their Radon-Nikodym derivative is not almost surely 1). Every private signal ς_n implies a private belief p_n, and the distributions of these private signals induce private belief distributions: $(\mathbb{G}_0, \mathbb{G}_1)$. The unconditional private belief distribution is $\mathbb{G} = \frac{1}{2}\mathbb{G}_0 + \frac{1}{2}\mathbb{G}_1$.

In addition to their private signal, each agent observes the actions of some ordered subset of those agents before them, their neighborhood $B(n)$, and the indexes of those agents. I.e. they observe $\{x_k : k \in B(n)\}$. These neighborhoods are drawn at the beginning of the game, and are independent across n. The distribution of agent n's neighborhood, $\mathbb{Q}_n$, is common knowledge. Upon observing this social information, agents form their (motivated) social belief. For each social signal one can define the *Bayesian social belief* as the probability a Bayesian would assign to $\theta = 1$ upon observing it (given their understanding of the behavioral behavior of their neighbors). As in standard models, these are then combined into an overall belief, and the order in which an agent forms social and private beliefs does not matter (Lemma 2, which reproduces Proposition 2 from Acemoglu et al. (2011), makes this clear, by showing that the decision of the agent depends on the sum of their private and social beliefs).

Agents can be of type 1,0 or U[8], where τ_n denotes the type of agent n; type 1 agents are biased towards believing that $\theta = 1$, type 0 agents that it is $\theta = 0$, and type U agents are Bayesian (showing no bias either way). I assume types are independent and identically distributed. Let the probability a given agent is Bayesian be β. For notational simplicity I assume there is equal chance of an agent being type 0 or 1: $\frac{1}{2}(1-\beta)$.[9] The specific reasoning procedure I assume for the motivated social belief is adapted from Little (2021), which provides a tractable and elegant representation of an agent trading-off a desire for accuracy against a

[8]For *unbiased*, though I shall frequently refer to them as simply Bayesian.

[9]All results generalise to any distribution with non-zero measure on type 1 and type 0 agents, though the fraction of agents choosing one action or the other asymptotically will depend on this.

desire to believe the state of the world takes particular values.[10] A type 0 agent observing a social signal in favor of $\theta = 1$ is an agent *of non-congenial type*, and similarly in reverse for type 1 agents. For any agent, their *Bayesian equivalent* is a hypothetical agent with identical information, but no bias. This agent's hypothetical action is $\chi_n \in \{0, 1\}$, and the unconditional probability with which it matches θ is their *Bayesian accuracy*: $\alpha_n := \mathbb{P}(\chi_n = \theta)$.

3.1 Motivated Reasoning

I represent motivated reasoning with two parameters $(R, s) \in (\frac{1}{2}, 1) \times (0, 1)$, and illustrate this process in Figure 1. 'R' denotes a threshold for '*signal rejection*,'[11] and 's' is a parameter governing '*prior-shifting*.' If agent n is of type $\tau_n = 0$ (he), he rejects any Bayesian social belief strictly greater than R and adopts motivated social belief $\frac{1}{2}$ in its place. This (relatively stark) assumption is easy to work with, but it should be noted that all results hold for any belief rejection process satisfying Assumption 1:

Assumption 1. *There exist $\epsilon > 0$ and $\delta > 0$ such that for any Bayesian social belief $\mu \in (1-\epsilon, 1]$, a type $\tau_n = 0$ agent will instead form social belief y with probability z where $y \leq 1-\epsilon$ and $z \geq \delta$. Type $\tau_n = 1$ agents will act symmetrically for Bayesian social beliefs in $[0, \epsilon)$.*

In other words, we simply need some strictly positive probability that agents will act with a social belief bounded away from 1 or 0. Our agents do not reject Bayesian social beliefs below R, but instead form a motivated social belief by updating an adjusted prior $(1-s) \times \frac{1}{2} + s \times (0) = \frac{1}{2}(1-s)$ that places weight s on their type and $1-s$ on the true prior. Conversely, an agent of type $\tau_n = 1$ (she) rejects Bayesian social beliefs below $1 - R$, and interprets other social signals according to the prior $(1-s) \times \frac{1}{2} + s \times (1) = \frac{1}{2}(1+s)$. Notice the asymmetry here: private beliefs are still processed in a Bayesian fashion with the correct prior of $\frac{1}{2}$: agents always form exactly the same private beliefs a Bayesian would, whatever their type. I model belief formation in this way to reflect experimental evidence that people are distinctly irrational when asked to process social information (Conlon et al., 2022; Oprea and

[10] There are various models of Motivated Reasoning in the literature, notable alternatives are studied by Bénabou (2015), but Little's has the advantages of using a straightforward alteration of Bayesian reasoning, and thus great tractability.

[11] Epley and Gilovich (2016) cite the example of an admirer of the actor Jimmy Stewart who would refuse to read profiles of him if ever she glimpsed negative keywords such as 'womanizer' and 'FBI informant', instead slapping the magazine shut. My agents can be thought of as behaving similarly, logging into their social media accounts and scanning to see roughly what their friends are saying on a given topic, before logging off quickly and suppressing the memory if it looks like bad news. Rejection in this model is also similar to information avoidance and forgetting in other models of motivated reasoning, Bénabou (2015) presents a number of models in which agents can forget unpleasant signals, and the experimental literature provides evidence both for information avoidance (Oster et al., 2013; Ganguly and Tasoff, 2017) and forgetting (Zimmermann, 2020; Saucet and Villeval, 2019)

Yuksel, 2022; Weizsäcker, 2010; Conlon et al., 2021).[12,13,14] Moreover, the processing of social signals here involves a severe discontinuity at R and $1-R$ for $\tau_n = 0$ and $\tau_n = 1$ type agents respectively. An extension considered in Section 7 smooths out this discontinuity, without changing any results, as can be done with any rejection procedure satisfying Assumption 1.

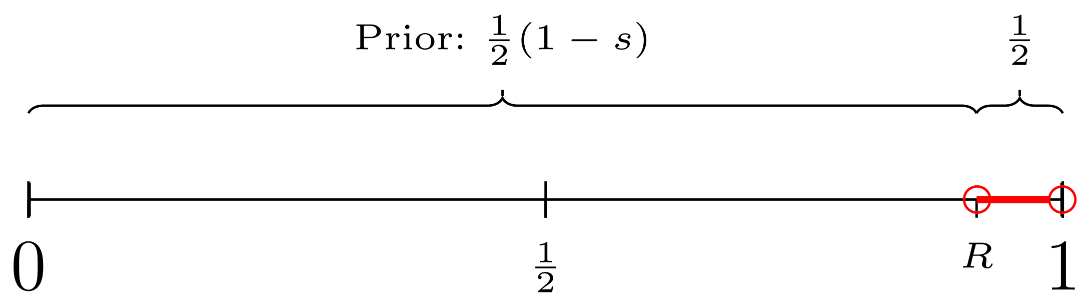


Figure 1: A $\tau_n = 0$ agent forms a motivated social belief. If the social belief a Bayesian would form is in the red region, they adopt a motivated social belief of $\frac{1}{2}$. Otherwise, the form the belief a Bayesian would form, if said Bayesian had prior belief $\frac{1}{2}(1-s)$.

4 Equilibrium Strategies and Outcomes

A convenient representation of each agent's decision making procedure is outlined in Lemma 1. This reasoning procedure of course involves a non-monotonic response to social information: an agent will increase their belief in response to stronger evidence up to a point, before snapping and discounting this information entirely.[15] Given the above and a convenient manipulation of Bayes' Rule, each agent's motivated decision rule can be represented as in the following lemma:

[12]Motivated Reasoning can be considered as consisting of two distinct parts: (1) evidence recruitment and (2) evaluation. Epley and Gilovich (2016) note this distinction in particular. Given the structure of this model (where evidence is not specifically recruited, but simply observed), I focus on the latter.

[13]Appendix F.2 considers an alternative model in which agents process all information in a motivated fashion, finding substantially similar results.

[14]One oddity of this specification occurs with a sufficiently extreme s parameter, and Bayesian social belief just above R (taking type $\tau_n = 0$ agents without loss of generality). In this case, it is possible that whilst rejecting the social signal produces motivated social belief $\frac{1}{2}$, accepting it and processing it with the biased prior will produce a preferable motivated social belief below $\frac{1}{2}$. Since belief rejection represents the most extreme psychological defense measure, this is odd. However, if instead we suppose that agents reject social signals beyond $\overline{R} = \max\{R, \frac{1}{2}(1+s)\}$, this is no longer the case. All of the results I prove for this model also hold if we assume agents use $\overline{R}$ as their rejection threshold. For simplicity I will not incorporate $\overline{R}$ in my analysis, but one could without any important change.

[15]This can be extended to a less abrupt procedure, as discussed in Section 7.

Lemma 1. *Consider an agent, n, who forms (motivated) social belief 'SB'. They will choose $x_n = 1$ if the following condition is satisfied:*

$$SB + \mathbb{P}_\varsigma(\theta = 1|\varsigma_n) \geq 1$$

If agent n is a Bayesian ($\tau_n = B$), they form $SB = \mathbb{P}_\varsigma(\theta = 1|B(n))$. If they are type 0 ($\tau_n = 0$), they form SB according to:

$$SB = \begin{cases} \frac{(1-s)\mathbb{P}_\varsigma(\theta=1|B(n))}{(1+s)-2s\mathbb{P}\sigma(\theta=1|B(n))} & \textit{if } \mathbb{P}_\varsigma(\theta = 1|B(n)) \leq R \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

Finally, if they are type 1 ($\tau_n = 1$), SB is formed according to:

$$SB = \begin{cases} \frac{(1+s)\mathbb{P}_\varsigma(\theta=1|B(n))}{\left(2s\mathbb{P}_\varsigma(\theta=1|B(n))+(1-s)\right)} & \textit{if } \mathbb{P}_\varsigma(\theta = 1|B(n)) \geq (1-R) \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

Otherwise they will choose $x_n = 0$.

Proof. See Appendix B. □

Having defined the belief formation process, the solution concept I use is Perfect Bayesian Equilibrium, except of course that the agents do not use Bayes' Rule to form their beliefs, but rather the relevant motivated distortion of it.

Definition 1. *(Motivated Equilibrium) A strategy profile ς is a Motivated Equilibrium if:*

(i) Sequential Rationality: Every agent's strategy is an optimal response to their belief given the strategies of other agents ς_{-n}.

(ii) (In)consistency: Agents' beliefs are updated according to the Motivated Reasoning Procedure outlined above, applying Bayes' Rule with the distortion implied by their type, and (R, s).

Denote the set of all equilibria as Σ, with typical member ς.

With this definition, equilibrium existence is immediate, as is standard in models of sequential social learning:

Proposition 1. *A Motivated Equilibrium exists.*

Proof. The Motivated Reasoning procedure always uniquely defines a belief for every history and private signal (ς_n, h_n). The set of optimal actions is non-empty for every belief (and is not a function of the strategies of other agents except through this belief), thus recursively applying these facts gives equilibrium existence. □

We are interested in whether or not agents ever reach consensus, and whether this consensus is correct. Failing this, it is of particular interest whether Bayesian agents make the correct decision asymptotically, as this reflects the extent to which the presence of motivated agents damages information aggregation within a given network. I define these two outcomes formally as follows, with *complete Bayesian learning* and *consensus*. The most extreme form of polarization would involve all motivated reasoners choosing their own type, and I call this *tribalism*. Tribalism and consensus are mutually exclusive by definition. For each of these three outcomes, one can speak about the outcome occurring within the whole population $\mathbb{N}$ or some subset of it S. As I discuss in Section 6, this allows me to explain a larger variety of patterns of polarization.

Definition 2. *(Learning, Consensus & Tribalism)* ***Complete Bayesian learning*** *obtains if* χ_n *converges to* θ *in probability (according to measure* $\mathbb{P}_\varsigma$*) in all equilibria* $\varsigma \in \Sigma$*, i.e.* $\lim_{n\to\infty} \alpha_n = 1\ \forall\theta$*. We say that* ***Almost-sure Bayesian learning*** *obtains if this converges almost surely.* ***Consenus*** *obtains if* x_n *converges to* ω *for any* $\omega \in \{0, 1\}$ *almost surely in all equilibria* $\varsigma \in \Sigma$*, i.e.* $\mathbb{P}_\varsigma(\lim_{n\to\infty} x_n = \omega) = 1$*. Beyond this,* ***Correct Consenus*** *obtains if* x_n *converges to* θ *almost surely in all equilibria* $\varsigma \in \Sigma$*, i.e.* $\mathbb{P}_\varsigma(\lim_{n\to\infty} x_n = \omega) = 1\ \forall\theta$*.* ***Tribalism*** *obtains if, for motivated reasoners,* x_n *converges to* τ_n *almost surely in all equilibria* $\varsigma \in \Sigma$*, i.e.* $\mathbb{P}_\varsigma(\lim_{n\to\infty} x_n = \tau_n) = 1\ \forall\tau_n \in \{0, 1\}$*. Tribalism does not restrict the asymptotic actions of Bayesians. Any of these outcomes occurs within subset* $S \subseteq \mathbb{N}$ *if the limit probability of agents within* S *converges to 1.*

I shall sometimes have cause to refer to tribalism and consensus *in probability*, which simply switches the notion of convergence in these definitions, though it is primarily the almost-sure notions above that will be of interest. Almost-sure Bayesian learning of course implies consensus on the true state when no motivated reasoning is present, and complete Bayesian learning analogously would imply consensus in probability.

4.1 Categorizing Information Structures

Generally, an information structure is said to be bounded if the support of the private beliefs it can induce is $[\underline{B}, \overline{B}]$ where $\underline{B} > 0$ and $\overline{B} < 1$; for example, normally distributed signals are unbounded, but finite signal structures are all bounded. This is usually an important distinction, and only with unbounded beliefs are there no Bayesian social beliefs for which, should they hold the motivated social belief that corresponds to it, an agent's action carries no information at all about their private signal. Here however, motivated reasoning reduces the pertinence of this distinction. To supplement it, I divide information structures between those that are *nonstationary* and those that are *stationary*. Let $\psi(\lambda, \theta)$ denote the probability that an agent n of unknown type chooses $x_n = 1$ if they observe a social signal producing Bayesian social belief λ and the state of the world is θ. In order to define this partition, I first define the notions of *cascade* and *stationary* beliefs:

Definition 3 (Cascade and Stationary Beliefs)**.** *A Bayesian social belief λ is a cascade belief for type τ, if an agent of type τ with Bayesian social belief λ takes the same action $x_n = x$ for any private signal. λ is a stationary belief if it is a cascade belief for every type, implying that $\psi(\lambda, 0) = \psi(\lambda, 1) \in \{0, 1\}$.*

Suppose a Bayesian agent observes a single neighbor, $n-1$, of unknown type. An agent cannot precisely know the Bayesian social belief of their neighbor without observing their neighborhood, but suppose for the sake of this definition that we simply tell them this value. If λ is a nonstationary belief, they will form a social belief *strictly* above λ if $x_{n-1} = 1$ and *strictly* below λ if $x_{n-1} = 0$. If instead λ is stationary, their action communicates no new information and our agent will adopt λ as their own Bayesian social belief, since it reflects all social information observed by the neighbor. Stationary beliefs are those that cannot be rejected or overturned by any private signal, whatever the type of the agent. Nonstationary beliefs are those that can be, for at least one type of agent. Armed with these definitions, we can define our partition:

Definition 4 (Nonstationary & Stationary Signal Structures)**.** *A signal structure* $(\mathbb{F}_0, \mathbb{F}_1)$ *is nonstationary for a given parameter tuple (R, s, β), if the set of stationary beliefs is empty. Otherwise, it is stationary for* (R, s, β).

In the standard Bayesian model, this definition of nonstationary information structures is equivalent to the definition of unbounded signals, and unbounded signal structures are necessarily nonstationary here. Bounded signal structures may be either stationary or nonstationary however, and this fact is illustrated by Figure 2 and the following discussion.

Let us consider a bounded signal structure. The presence of prior-shifting first expands the region of Bayesian social beliefs that can be overturned (the orange region in Figure 2b for example), since type 0 agents will form motivated social beliefs more in favor of $\theta = 0$, and type 1 agents in favor of $\theta = 1$. The region of social beliefs which are not cascade beliefs for Bayesians is $[1 - \overline{B}, 1 - \underline{B}]$. Let us define $[1 - \overline{b_0}, 1 - \underline{b_0}]$ and $[1 - \overline{b_1}, 1 - \underline{b_1}]$ analogously for the motivated types. To compute these values, we set $SB + \mathbb{P}_\varsigma(\theta = 1 | \varsigma_n) = 1$ in Lemma 1, substitute $\underline{B}$ or $\overline{B}$ for the private belief, and use the first case of the piecewise function defining SB. As we increase s, the interval of over-turnable beliefs for each motivated type moves towards the extreme corresponding to their type. In Figure 2b the region of over-turnable Bayesian social beliefs grows from $[1-\overline{B}, \overline{B}]$ to $[1-\overline{b_1}, 1-\underline{b_0}]$. Secondly, signal rejection entails that for Bayesian social beliefs in the regions $[0, 1-R)$ and $(R, 1]$, the agent will choose each action with strictly positive probability. This follows from the fact that if they are of non-congenial type they will simply reject this belief entirely. As we increase the maximum signal strength of our private signal $\overline{B}$ or reduce the threshold for signal rejection R, we will eventually ensure that these two regions meet. Given this, for any parameters (R, s), sufficiently informative bounded signals instead behave like unbounded signals. Increasing s alone can also turn a stationary signal structure into a nonstationary one, though for some values of $\{\overline{B}, \underline{B}, R\}$ it will not, as I discuss next.

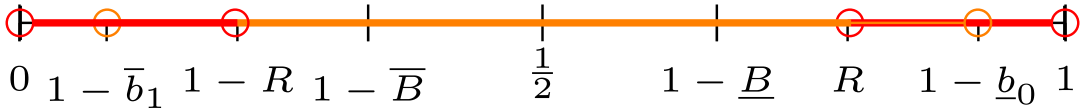
Focusing our attention on stationary signal structures, there are two important properties that they can exhibit: they can be *conducive to tribalism*, *conducive to consensus*, or both.

Definition 5 (Conducive Signal Structures)**.** *A signal structure* $(\mathbb{F}_0, \mathbb{F}_1)$ *is conducive to consensus for a given parameter tuple* (R, s, β) *if there is a non-empty set of stationary social beliefs for which agents of all types choose the same action with probability 1. It is conducive to tribalism if there are stationary beliefs for which motivated reasoners of each type match their action to their type with probability 1.*

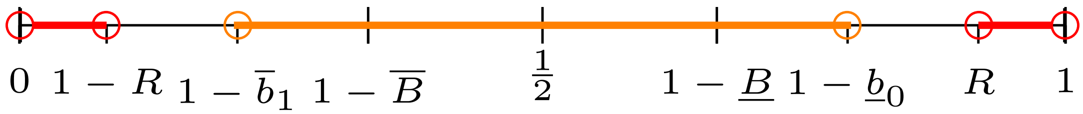
Increasing the extent of motivated reasoning (by increasing s or decreasing R), and increasing access to information (the support of the private beliefs) will eventually ensure that the signal structure is not conducive to consensus, and beyond that nonstationary. The conditions that make a signal structure conducive to tribalism are more counterintuitive: increasing motivated reasoning via s helps to make a signal structure conducive to tribalism (since it pushes the interval of over-turnable beliefs for each motivated type away from the center), but increasing it by decreasing R does not achieve this. Decreasing R works against both tribalism and consensus. The same holds for increasing the support of private beliefs: both consensus and tribalism are eventually rendered impossible as we expand $[\underline{B}, \overline{B}]$.

5 Consensus with Motivated Reasoning

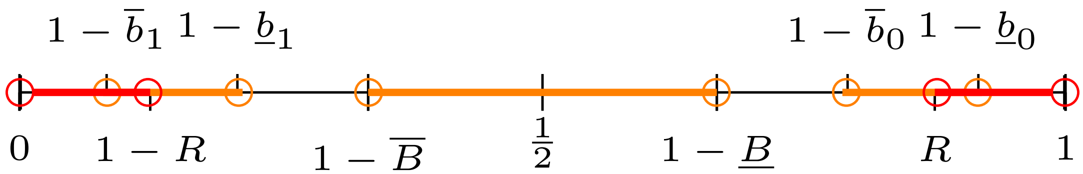
First and foremost, it can be easily seen that complete Bayesian learning and consensus are incompatible with this model of motivated reasoning. The former implies extreme Bayesian social beliefs, which are necessarily rejected by agents of non-congenial type. This in turn gives that nonstationary beliefs preclude learning, as either complete Bayesian learning obtains, or all Bayesian social beliefs can be overturned with positive probability asymptotically. In either case, clearly no action commands unanimity. This is not to imply that consensus is impossible, however, and the third part of this theorem observes that with stationary beliefs there are some network topologies and parameter values that do produce it (the complete network can



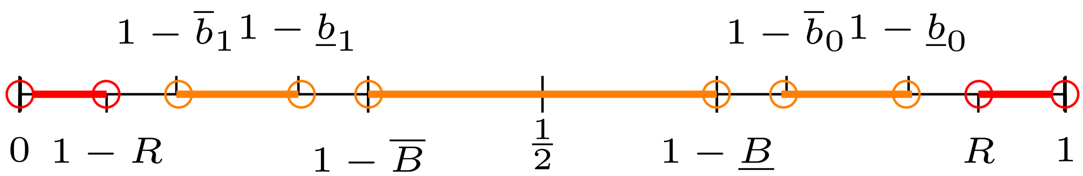
(a) Nonstationary Signal Structure



(b) Conducive to Consensus but not Tribalism



(c) Conducive to Tribalism but not Consensus



(d) Conducive to Both

Figure 2: *Nonstationary vs Stationary* and *Conducive to Tribalism/Consensus*: Any Bayesian social belief in the red region may be rejected by a non-congenial motivated reasoner, and any in the orange region can be overturned. Bayesian social beliefs in neither are stationary beliefs, and it is the absence of these that defines nonstationary information structures. The final three panels all exhibit stationary information structures, and the first is nonstationary. In panel 2b, agents will all take action 0 in the left region of stationary beliefs, and 1 in the right. In panel 2c motivated agents will choose their own type in the stationary regions; Bayesians will choose 0 on the left and 1 on the right. In panel 2d Bayesians will choose 0 for all stationary beliefs on the left, and 1 for all stationary beliefs on the right. In the two inner regions of stationary beliefs motivated agents will choose their own type, in the outer regions we will have consensus. On the left this consensus will be on $x_n = 0$, and on the right it will be on $x_n = 1$.

provide such a case).

Theorem 1.

(i) Complete Bayesian learning implies that neither consensus nor tribalism obtain.

(ii) Thus consensus and tribalism cannot obtain with nonstationary beliefs.

(iii) Consensus and tribalism can occur with stationary signal structures. There exists at least a network topology and stationary signal structure that produce consensus, and the same is true for tribalism.

Proof. Let $\theta = 0$,

(i) Complete Bayesian Learning implies the Bayesian social belief converges to (probability) 1, which implies agents of type $\tau_n = 1$ are rejecting their social signals, and instead using $\mathbb{P}(\theta = 1|\varsigma_n) > \frac{1}{2}$.

(ii) With nonstationary beliefs therefore, one of two things happens. The Bayesian social belief may not converge to certainty, in which case all agents' Bayesian social beliefs can either be overturned or rejected. Otherwise it does, and we have complete Bayesian learning: then Part (i) implies we do not have consensus.

(iii) This can be seen by taking the example of the complete network (which is particularly convenient, since it allows the almost direct application of Smith and Sorensen (2000)), with bounded beliefs: Figure 2b illustrates the set of all possible Bayesian social beliefs if we have an information structure conducive to consensus but not tribalism. A complete proof with details is relegated to Appendix B, but the intuition is as follows. The assumption of bounded beliefs implies that there is a minimum Bayesian social belief that can possibly be overturned by a private signal for each type ($1 - b_1 < \underline{B} < 1 - b_0$), and similarly maximum Bayesian social beliefs that can be overturned by a weak signal ($b_1 < \overline{B}, b_0$). The exact levels of these are given by the type-parameters, and implied by proposition 1. Thus if the Bayesian social belief is below $1 - b_1$ or above b_0, all future actions become completely uninformative (as they do not reflect private signals) unless the Bayesian social belief is low (high) enough to be below $1 - R$ (above R). If R is picked to be high enough, however, this will be impossible. The Bayesian social belief will then get stuck in the set $[1 - R, 1 - b_1] \cup [b_0, R]$, and learning will stop with all agents choosing the same action. Taking an information structure conducive to tribalism (i.e. the signal structures in Figure 2c or Figure 2d), the same argument establishes that tribalism obtains.

□

Whereas stationary signal structures maintain the possibility of consensus, the level of polarization in networks exhibiting expanding nested samples with a nonstationary information structure sizes can be pinned-down exactly: if the true state is $\theta = 1$ agents will choose $x_n = 1$ with asymptotic probability $\beta + \frac{1}{2}(1-\beta)(2 - \mathbb{G}_1(\frac{1}{2}))$.

For fixed (R, s), increasing the availability of information (assuming this translates to an increase in the most informative available signal $\overline{B}$) could create a discontinuous jump in polarization if it engenders a shift from a stationary to a nonstationary information structure, or indeed from a stationary signal structure conducive to only consensus, to one conducive to tribalism. Hence a sudden and marked increase in polarization is unsurprising with the expansion of the internet over the past 20 years: once a tipping point has been reached the character of the information environment changes completely. Secondly, if one accepts that in democratic societies it is not plausible (or at all desirable more generally!) to restrict the accessibility of information to the citizenry, it follows that one could only recover the 'stationary' information structure in managing to increase the R parameter (or its distribution as in the aforementioned extension) above its original value enough to compensate for the increased availability of information. This fits well with the findings of Guilbeault et al. (2018), who show experimentally that increasing the salience of partisanship produces motivated reasoning: in the context of this model increasing the salience of partisanship would simply correspond to decreasing and increasing the values of R and s respectively.[16]

If we begin with an environment that permits and is producing consensus, expanding the support of private beliefs and the level of motivated reasoning will eventually produce a nonstationary signal structure. Depending on the relative rates as which s, R and $[\underline{B}, \overline{B}]$ are changing, the signal structure may become conducive to tribalism en route. If an information structure is stationary for $s = 0$, there is some value of s that will make this structure conducive to tribalism.

Beyond what we have seen in Theorem 1, the forthcoming results on learning have further implications on consensus and the circumstances in which we should expect it. Corollary 2.1, a corollary to the main theorem of this section, Theorem 2, establishes that correct consensus in a purely Bayesian society is fragile to the introduction of any non-zero probability β of motivated reasoning. With enough *nested neighbors* (c.f. Condition 3), I also show that one can ensure that non-congenial agents reject social signals with arbitrarily high probability, whilst congenial type agents form arbitrarily strong beliefs in favor of the true state.

[16] They note that increasing the salience of partisanship reduces social learning, though *social learning* as defined in their setting corresponds to *consensus* here; clearly the observed action/ stated belief in an experimental setting corresponds to x_n, not χ_n.

6 Learning and the absence of Sparse Correct Consensus

In searching for sufficient conditions for complete Bayesian learning, we might first guess that expanding observations (originally from Acemoglu et al. (2011), set out below in Condition 1) might be enough.

Condition 1 (Expanding Observations)**.** *A network topology $\mathbb{Q}$ satisfies expanding observations if for all $K \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{Q}_n(\max_{b\in B(n)} b < K) = 0$*

Expanding observations is certainly a necessary condition for learning, since if it fails there is some finite K such that infinitely many agents' decisions are based on at most K private signals, but is no longer sufficient in this setting. This insufficiency follows from Theorem 2, which establishes that a much stronger (and novel) condition '*Expanding Sample Sizes*' is necessary for complete Bayesian learning to obtain. Notice that this condition implies expanding observations.

Condition 2 (Expanding Sample Sizes)**.** *A network topology has expanding sample sizes if for all $K \in \mathbb{N}$, we have*

$$\lim_{n\to\infty} \mathbb{Q}_n\Big(|B(n)| < K\Big) = 0$$

If the network topology does not satisfy this property, it has non-expanding sample sizes. If for some subset $S \subseteq \mathbb{N}$ we have $\lim_{n\to\infty} \mathbb{Q}_n(|B(n) \cap S| < K) = 0$, then we have expanding sample sizes for S.

This condition is of course weaker than requiring a network topology to be complete, but it is still very strong, since it requires that agents eventually observe arbitrarily large neighborhoods with very high probability. Whereas expanding observations ensures that agents indirectly observe an ever-increasing number of neighbors, expanding sample sizes requires that they do so directly. This condition, as I now establish in Theorem 2, is necessary for complete Bayesian learning. Beyond this, as I then demonstrate in Corollary 2.1, in its absence[17] consensus is impossible with a nonstationary signal structure, unlike in the purely Bayesian model.

Theorem 2. *Complete Bayesian learning obtains only if the network topology satisfies expanding sample sizes.*

Proof. See Appendix B. □

[17] It is this absence that I reference with 'sparse', though networks can in truth be extremely densely connected without satisfying expanding sample sizes.

The proof is relegated to the appendix (and an alternative proof by contradiction provided in Appendix D), but the rough intuition behind it is as follows. Any neighbor is either acting on the basis of a weak Bayesian social belief, and is thus not very informative to observe, or a very strong one, in which case they will be rejecting this belief if of non-congenial type. For any finite neighborhood, the probability that all neighbors happen to be of non-congenial type is bounded away from zero. Any Bayesian agent may therefore be acting on the basis of an entire neighborhood of neighbors rejecting their social signals.

This intuition, that the fact of signal rejection causes information to be 'lost' (or possibly lost), is easier to see in deterministic, simple networks. A tractable example is the line network, in which agents each observe only their immediate predecessor. This network structure guarantees learning with unbounded beliefs and Bayesian agents, but of course fails with motivated agents since it does not satisfy expanding sample sizes.

Note that Theorem 2 establishes that expanding sample sizes is a necessary condition for learning for any proportion of Bayesians $\beta \in (0, 1)$. In so far as we care about whether or not complete Bayesian learning obtains in a given network topology therefore, this provides a sense in which the traditional Bayesian model is fragile. Any small fraction of motivated reasoners is sufficient to move from a model in which even the line network suffices for learning, to one in which agents instead need arbitrarily large neighborhoods. However, this fragility is only in terms of the exact achievement of complete learning, and does not say that a very small β/very large R must produce a *large* drop in asymptotic accuracy. Given that our main concern in this paper is with consensus however, it is more important that we do find a substantial fragility with this outcome as set out in Corollary 2.1.

Corollary 2.1. *There exist network topologies and signal structures that produce correct consensus when $\beta = 1$, but fail to when $\beta = 1 - \epsilon$ for any $\epsilon > 0$. Hence, in some cases introducing an arbitrarily small probability of motivated reasoning switches an equilibrium in which only finitely many Bayesian agents fail to match the state for one in which infinitely many of them do, almost surely in each case.*

Proof. See Appendix B □

In games that would produce almost sure learning in the Bayesian model therefore, and in which all agents will match the state after some finite time, an arbitrarily tiny proportion of motivated reasoners engaging in an arbitrarily small amount of motivated reasoning ($R \approx 1$, perhaps reverting to $1 - \epsilon$ or ϵ as per Appendix F.1) is enough to thwart almost sure learning whenever we do not have expanding sample sizes, and ensure infinitely many Bayesians (and by extension, agents of any type) fail to match the state where only finitely many would do so in the Bayesian model. A line network with sufficiently informative beliefs[18] is an example of such a setting, and it is in such high information, sparse network settings in which the introduction of motivated reasoning will cause such problems. Though Theorem 1 establishes that in the complete network, the densest possible, motivated reasoning will eventually kill consensus as we introduce more and more information. It is also interesting to note that motivated reasoning breaks the equivalence between the line and complete networks that Rosenberg and Vieille (2019) find holds with sufficiently informative private signals. In their purely Bayesian setting, the outcome is the same in both ('efficiency'- which is stronger than almost sure learning) whereas here there is still almost sure learning in the complete network, but not in the line network.

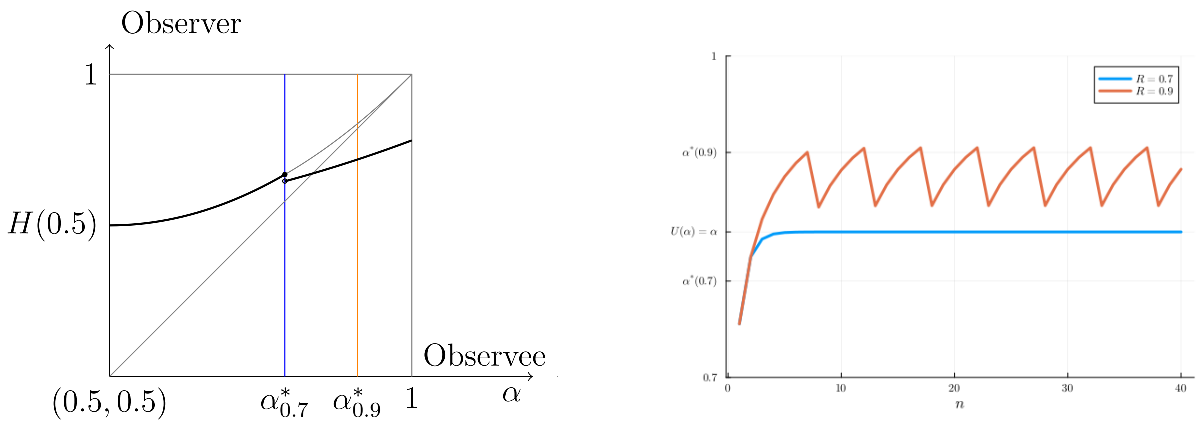
6.1 Information Loss & Uninformative Actions

Whether motivated reasoning helps or harms information aggregation, measured by the attainment of complete Bayesian learning, varies according to the network topology. Two extreme deterministic cases demonstrate this. The line network shows how motivated reasoning can cause the loss of information, and the complete network shows how it can ensure information continues to accumulate. Signal rejection, specifically when applied to the social belief, can engender the repeated loss of all accumulated information. Learning results derived using improvement principles are thus inapplicable, as the neighbor upon whom a given agent is improving may be acting on the basis of only his own private information. In the line network this problem is at its most extreme: there will be infinitely many signal rejections, and agents will never be sure exactly how many agents back the last rejection was. I work through a simple line network example in Appendix C, but its conclusions are captured by Figure 3, in which I plot the case with $R = 0.7$ in blue and $R = 0.9$ in orange. The black function in Figure 3a gives the Bayesian accuracy of an agent (he) as a function of the Bayesian accuracy of his neighbor (she). Upon observing his neighbor, if her Bayesian accuracy is sufficiently high, our agent must account for the fact that she will be rejecting her social signal if of non-congenial type, hence the discontinuity. That this black function never reaches a value of 1 reflects the fact that our agent can never achieve this level of accuracy, however accurate his neighbor. Plotting α_n against n in Figure 3b, we can see that the Bayesian accuracy of agents never approaches 1 in either case. Incidentally, that it need not converge at all[19] (e.g. as is the case for $R = 0.9$) provides another point of distinction from the Bayesian model, where for a very

[18] I refer to a condition in Rosenberg and Vieille (2019) here, see the proof for details.

[19] If α_R^* is above the intersection of the black function with the 45° line in figure 3a, the accuracy of agents repeatedly climbs up this curve only to drop back down below α_R^* when some agent exceeds it. This produces the sawtooth patten for $R = 0.9$ in figure 3b. Otherwise, the process keeps climbing after reaching the second section of the black curve, converging to its fixed point as with $R = 0.7$.

general class of network topologies (including the line network) it is known that Bayesian accuracy must converge (Kartik et al., 2022).



(a) Observer accuracy against observee accuracy in a line network

(b) α_n against n

Figure 3: $R \in \{0.7, 0.9\}$, $s = 0$, $\beta = 0$, $(f_0, f_1) = (2(1 - \varsigma), 2\varsigma)$

In the standard Bayesian model, the assumption of non-empty neighborhoods and expanding observations are sufficient to ensure that as $n \to \infty$ more and more information is continually introduced, whilst agents have indirect access to the signals of all agents in the chain before them; the depth of each agent's 'information path' converges to infinity. With motivated agents, however, periodic signal rejection ensures that the actions of agents before a given signal rejection occurs are completely independent of those after. Thus even with expanding observations, an agent only ever (in a line network) observes a social signal reflecting a finite number of signals.

This breakdown in information monotonicity is reminiscent of that seen in Lobel and Sadler (2016), where agents have heterogeneous preferences over the two actions. However, it is even worse in the present model, as in their paper one would at least be able to extract the information of an observed agent upon learning their type (and due to this they can achieve learning with homophily in their Proposition 3). In this model, however, knowing the type of a neighbor does not fully fix the issue. If you observe the type of your most accurate neighbor, and also observe (by observing their neighborhood) that they are of congenial type, this tells you that x_n is highly likely to be equal to χ_n (I say only 'highly' likely since prior-shifting can still produce a difference even in the absence of signal rejection). However, if they are of non-congenial type, their decision may in no way reflect their social information, and evening knowing that they are of non-congenial type does not allow you to recover that information. To salvage an improvement principle, we would need to impose that agents were always of congenial type, which is of course inconsistent with the independence of type and neighborhood.

On the other hand, in the complete network, agents do not depend on their immediate predecessors to learn about the history of the game. The only value in observing one's immediate predecessor is in the fact that observing their action communicates information about that agent's private signal. If all agents are Bayesian and beliefs are bounded, as is well known, complete Bayesian learning will not obtain. With motivated reasoners however, if the (R, s) parameters are sufficient to produce a nonstationary information structure, the action of each individual in a complete network necessarily communicates information about their private signal. As I set out in Lemma 4, this ensures that complete Bayesian learning obtains in the complete network.

When agents depend upon a small number of observations to learn about the entire history of the game, information loss will prevent learning. When all agents eventually arrive at Bayesian cascade beliefs and take uninformative actions in the Bayesian model, motivated reasoning can rescue learning by making the signal structure nonstationary.

6.2 Sufficient Conditions for Learning

Theorem 2 establishes that learning obtains only under highly connected and large networks. Though the complete network is the most obvious example of a network topology that satisfies expanding sample sizes, it is not the only one under which we can achieve learning in this setting. One sufficient (though not necessary, as is demonstrated by Example 1 in Appendix D) condition for learning within any (infinite) subset S of agents is that it satisfies *expanding nested samples*:

Condition 3 (Expanding Nested Samples)**.** *For agent n, let $B^n(n) \subseteq B(n)$ be the set $\{m \in B(n) : \mathbb{Q}\big(B(m) \subset B(n)\big) = 1\}$. A network topology has expanding nested samples for S if for all $K \in \mathbb{N}$, we have* $\lim_{n\to\infty} \mathbb{Q}_n(|B^n(n) \cap S| < K) = 0$.

In words, agent m (he) is a *nested neighbor* of n (she) if she observes both his action and the actions of his entire neighborhood, and does so with probability 1. Expanding nested samples then requires that agents observe ever greater numbers of such agents. Clearly a network topology that satisfies expanding nested samples also satisfies expanding sample sizes, but two examples illustrate that the reverse does not hold: i.e. that expanding nested samples is a stronger condition than expanding sample sizes. Firstly, if we consider the network topology defined by each agent observing the $\lceil ln(n) \rceil$ agents before him, we have an example of a deterministic network topology that satisfies expanding sample sizes but not expanding nested samples, since whilst $|B(n)|$ converges to infinity, agents have at most 1 nested neighbor, and most commonly none. Secondly, amongst stochastic networks, consider a network topology in which each agent observes with equal probability either all of their even predecessors or all of their odd ones (with agent 2 certainly observing agent 1). In such a network, no agent $n > 2$ observes any nested neighbors, since for any $m < n$ we have $\mathbb{Q}\big(B(m) \subset B(n)\big) = 0.5 \neq 1$.

Though a very demanding condition on the network topology, the requirement that agents observe a large number of nested neighbors is arguably not so unreasonable when studying social networks.[20] With the advent of social media, they are increasingly connected, and also display large amounts of clustering- where agents are disproportionately likely to be connected to any agents their friends are connected to. Since the neighborhood of agent n represents the set of agents n is connected to who act before them in the game, and social media provides a public record of all friends' comments, it is not necessarily unreasonable to suppose that agents at least observe some nested neighbors.

Theorem 3. *If a network topology has expanding nested samples for S, and the information structure is nonstationary, complete Bayesian learning obtains within S.*

Proof. See Appendix B □

One implication of this is that any network topology that satisfies expanding nested samples for $\mathbb{N}$ exhibits complete Bayesian learning. This re-proves complete Bayesian learning for the complete network, but also extends it to other nested topologies. For example, if all agents observe $B(n) = \{m \in \mathbb{N} : m < n\} \cap \big\{\{1, 100\} \cup \{i \times 100 + 50, ..., i \times 100 + 99, \forall i \in \mathbb{N}\}\big\}$ then the network topology satisfies expanding nested samples. When establishing that complete Bayesian learning obtains within proper subsets of $\mathbb{N}$, we can demonstrate that even network topologies that do not satisfy expanding sample sizes can nonetheless achieve *similar* outcomes. Specifically, we can see networks in which the asymptotic fraction of agents who correctly match the state with probability greater than $1 - \epsilon$ for any $\epsilon > 0$ is one. Take, for example, the network topology in which all agents in the set $S' = \{10^m : m \in \mathbb{N} \cup \{0\}\}$ observe only their immediate predecessor in S', and any agent $n \in \mathbb{N} \setminus S'$ has $B(n) \supseteq S' \cap \{1, ..., n-1\}$. This network topology satisfies expanding nested samples for $\mathbb{N} \setminus S'$, since the set S' has infinitely many members, and the neighborhood of every agent within S' is contained within S' (apart from the very first agent). When discussing polarization in politics, it is normally the overall fraction of agents taking the incorrect action that is of concern. Thus, even network topologies that do not achieve complete Bayesian learning or consensus may nonetheless result in similar levels of polarization as those that do if a sufficiently large subset of agents achieve these outcomes. Had I instead defined $S' = \{3m : m \in \mathbb{N}\} \cup \{1\}$, then learning in the group $\mathbb{N} \setminus S'$ would no longer pin down the exact asymptotic fraction of consensus as that

[20] Note that this assumption is not strong enough to allow the use of standard martingale convergence arguments (as in Smith and Sorensen (2000); Goeree et al. (2006) and others). Instead I argue that one can define a specific Bernoulli trial such that observing M neighbors is at least as good as observing M copies of said trial. Since this M converges to infinity, we can then appeal to the Law of Large Numbers for the result.

achieved by complete Bayesian learning, but it would still explain a large proportion of the polarization that resulted, as two thirds of all agents are within a set that satisfies expanding nested samples. This set S' resembles the notion of a *Royal Family* in Bala and Goyal (1998), where this is a *small*[21] set of agents who are observed by all other agents. In their model, in which all agents are Bayesians, the presence of a royal family can *prevent* learning. Here it is permitting learning within $\mathbb{N} \setminus S'$, but preventing consensus. In reality, this royal family might represent a population of influencers.

As I discussed earlier, motivated reasoning can help learning in the complete network when the signal structure is bounded. The same holds for any network topology involving expanding nested samples, the above example in which a subpopulation satisfies this condition, and also in Example 1 in Appendix D. This establishes that though expanding nested samples for S is a sufficient condition for complete Bayesian learning within S, it is not a necessary condition. It also illustrates that motivated reasoning can help by reintroducing a certain amount of independence between agents' actions, roughly speaking.

These sufficient conditions are more powerful when considered in the context of the extensions considered in Section 7, since the assumption of nonstationary information structures becomes less demanding. In one of these, agents do not all have the same (R, s) parameters, but have these independently drawn for them from some distribution. In such a setting, all that is required for a signal structure to be nonstationary is for each agent to have a sufficiently low R with *some* non-zero probability. Thus whilst unbounded beliefs might itself seem a relatively strong assumption, here what matters is only that we have a sufficiently informative (possibly bounded) signal given the rejection thresholds of those agents in the population most prone to rejecting social information. Whereas this assumption of unbounded beliefs is essential for many results in the Bayesian model (most notably Acemoglu et al's sufficiency of expanding observations for complete learning), here the unbounded-bounded distinction is much less important.

Even if we are not quite willing to suppose that a network topology satisfies expanding nested samples, we can nonetheless show that non-congenial type agents will reject their social signals with probability $1 - \epsilon$ for arbitrary $\epsilon > 0$ if they have neighborhoods with at least M nested neighbors[22] for some sufficiently large M. This is the main mechanism behind polarization, and produces the sharp difference in stated beliefs between different political factions. As discussed in Section 6, high clustering and connectivity of online social networks gives cause to hope that this may hold. Condition 4 mirrors Condition 3, except that nested neighbor samples need not explode to infinity.

[21] In Bala and Goyal (1998) this can be small as in finite and small, here it is small in a proportional sense.
[22] neighbors whose neighborhoods they also certainly observe.

Condition 4 (M-Nested neighbor Samples)**.** *For agent n, let $B^n(n) \subseteq B(n)$ be the set of n's nested neighbors. A network topology has M-nested neighbor samples for S if we have*

$$\lim_{n\to\infty} \mathbb{Q}_n\Bigg(|B^n(n) \cap S| < M\Bigg) = 0$$

Proposition 2. *For any $\epsilon > 0$, there is an $M \in \mathbb{N}$ such that, if agents within set S have M-nested neighbor samples and the information structure is nonstationary, then if sb_n is the Bayesian social belief of n,*

$$\liminf_{n\in S\cap\mathbb{N}} \mathbb{P}_\varsigma(sb_n > R|\theta = 1) \geq 1 - \epsilon$$
$$\liminf_{n\in S\cap\mathbb{N}} \mathbb{P}_\varsigma(sb_n < 1 - R|\theta = 0) \geq 1 - \epsilon$$

asymptotically agents of non-congenial type each reject their Bayesian social beliefs with probability at least $1 - \epsilon$.

Proof. See Appendix B. □

Hence whilst complete learning in this environment results from agents observing ever more information, if instead they simply observe *a lot* of information, that can suffice for the widespread social signal rejection that drives polarization. It is in this sense that increasing the extent to which a social network is 'connected' and highly clustered can increase polarization. If for a given network topology the asymptotic distribution of Bayesian social beliefs is not extreme enough to produce widespread social signal rejection, this can be 'rectified' by increasing the number of nested neighbors agents observe asymptotically. The asymptotic probability that agents' beliefs are in the region regions can be pushed arbitrarily close to 1 by ensuring they observe a large enough number of nested neighbors.

None of this, it should also be noted, establishes that stationary beliefs preclude learning. Much as with Acemoglu et al. (2011, Theorem 4), network topologies with infinitely many sacrificial lambs within a larger core can produce learning. The rise of ideological polarization and growing access to information suggest that nonstationary information structures are most relevant to investigating modern polarization, but I include some discussion of stationary beliefs in Appendix E for completeness.

7 Discussion

There are a number of simple extensions to which we can easily extend the results in this paper. As mentioned in Section 3, any rejection process that satisfies Assumption 1 will not interfere with our results. The details of the more involved extensions are relegated to

Appendix F. Firstly, the assumption that all agents have the same rejection threshold and prior-shifting parameter seems unrealistic. However, the proofs of several of results in this paper largely depend upon the existence of social belief regions in which non-congenial types will possibly reject the Bayesian social belief. Thus the results extend to the case in which we draw an R and s parameter individually for each agent from some distribution, so long as the support of the R parameter includes values strictly less than 1. In other words, if the convex hull of this support is $[1-\overline{R}, \overline{R}]$, we must have $\overline{R} < 1$. Specifically, this condition is sufficient to reproduce the proof of Theorem 2, the reasoning within the line network example, and the first two parts of Theorem 1. The third part requires an additional assumption, namely that the convex hull of the support of the R distribution does not extend all the way down to $\frac{1}{2}$: $\underline{R} > \frac{1}{2}$. The definitions of nonstationary and stationary information structures (Definition 5) still hold in this case, except that the tuple of parameters that determine if an information structure is one or the other are now $(\overline{R}, s, \beta)$ instead of (R, s, β). In a similar vein, the fact that rejection operates with a sharp threshold here creates a discontinuity that itself seems a poor representation of real behavior, but this too can be resolved without loss. Instead, we can define each agent as having a rejection function, giving the probability with which they reject beliefs of a given strength. Such a function, assigning higher probabilities of rejection to more extreme Bayesian social beliefs, could be defined to continuously increase the rejection probability for non-congenial Bayesian social beliefs so long as it assigns strictly positive probability to the rejection of some Bayesian social beliefs. To be precise, we can simply endow each agent with two parameters $R_{min} > \frac{1}{2}$, $R_{max} < 1$ and a rejection function $\mathcal{R} : [\frac{1}{2}, 1] \to \{0, 1\}$ taking value 0 below R_{min}, 1 above R_{max}, and $f(\cdot)$ between where $f(z)$ is any function from $[0, 1]$ to $[0, 1]$ (though of course a strictly increasing function makes the most sense intuitively). Much as before, Theorem 2, the line network example and the first two parts of Theorem 1 carry through so long as agents reject some sufficiently extreme non-degenerate Bayesian social beliefs with some strictly positive probability. There are no stationary information structures without the condition $R_{min} > \frac{1}{2}$, much as $\overline{R} > \frac{1}{2}$ was necessary for this in the previous extension.

A more substantial change would be to apply motivated reasoning to the entire 'combined' signal, instead of only the social signal. As I discuss early in the paper, I choose to model agents as engaging in social motivated reasoning as there is some experimental evidence to suggest that this reflects how people process information in real life. Having said that, such experiments involve a clearly defined private signal that unambiguously carries independent information, following a clearly defined distribution. Perhaps the signals to which agents have access in real life are less clear, in which case it is not inconceivable that they would be both treated equivalently. The result that expanding sample sizes is a necessary condition for social learning will still hold in this case, but the sufficiency of expanding nested samples is more difficult to establish. This follows from the fact that Combined Motivated Reasoning re-introduces a form of confounded learning, where, in a complete network for example, the game could reach a point in which agent n is exactly as likely to choose $x_n = 1$ in either state of the world. Once this point has been reached in a complete network, learning stops. In more general network topologies satisfying expanding sample sizes, it will rarely be the case that the game can get irreversibly stuck like this (this will require specific nested network topologies), but the same problem nonetheless prevents us bounding the minimal informativeness of any given action. This particular extension is discussed further in Appendix F.2, but in short the fundamental obstacles that motivated reasoning poses to learning remain.

8 Conclusion

Social learning, and under what conditions it should be expected, has been extensively studied. Despite this, the implications of motivated reasoning on the necessary and sufficient conditions for learning have not previously been investigated in this literature. Given the pertinence of motivated reasoning to learning à propos of political and ethical questions, and the increasing interest in explaining polarization in society, this omission is in need of correction. This paper not only fills that gap, but establishes that the perfect storm of increasing value-polarization, ever greater access to information, and the advent of more connected and clustered social networks with social media can explain the increase in fact-polarization we observe widely in modern life, particularly in politics.

The difficulty of learning from neighbors who reject social information they dislike strongly ensures that much more demanding conditions on the network topology are needed: we move from the requirement that asymptotically an agent indirectly observes infinitely many agents (expanding observations) to that they do so directly (expanding sample sizes). Whilst learning is thus severely obstructed by the presence of social learning, the ability of societies to settle on consensus is not spared either. If a society can achieve consensus in one setting, it can always be broken by increasing value-polarization or access to information. Furthermore, in those settings in which purely Bayesian societies can achieve correct consensus, an arbitrarily small amount of motivated reasoning suffices to kill it.

Arguably, however, this shift in network topology is descriptive of the change that the internet has brought about. When paired with the shift from stationary to nonstationary information structures that increased value polarization and information access can entail, this provides an additional mechanism explaining the increased fact-polarization of modern political discourse. The change is not all bad, as there are some network topologies that can produce learning with bounded beliefs when a similar network of Bayesians would not. Even in this instance, however, motivated reasoning produces more and more polarization as agents become more motivated to hold beliefs conducive to their ideology.

That increasing the level of information available to agents can make consensus harder to achieve, as demonstrated most concretely in the complete network, is counterintuitive but seems a compelling explanation of the increasing polarization we are currently observing. The clear practical lesson for reestablishing consensus, since reducing either our access to information or the greater density of networks are clearly not feasible (in addition to being alarming policy objectives in general!), is that reducing the political charge of important issues is essential. Though this is in and of itself a very tough nut to crack, it is the only clear route my model suggests to reducing polarization with motivated agents in our ever more connected and informationally-overwhelmed societies.

A Useful Lemmas

The first two lemmas here serve only to reproduce important results in Acemoglu et al. (2011) that I use in the analysis in this paper. The first is Proposition 2 in their paper, and characterises the decision rule of Bayesian agents. Lemma 3 similarly reproduces Theorem B.1 from Smith and Sorensen (2000), which is the central theorem describing the fixed points of Markov-Martingale processes.

Lemma 2 (Acemoglu et al. (2011) Proposition 2)**.** *If they are Bayesian, agent n will choose $x_n = 1$ upon observing neighborhood $B(n)$ and private signal s_n if:*

$$\mathbb{P}\big(\theta = 1 | B(n)\big) + \mathbb{P}(\theta = 1 | s_n) > 1$$

Proof. See Acemoglu et al. (2011, Proposition 2). □

This next lemma is Acemoglu et al. (2011, Lemma 1), and gives several useful properties of the belief distributions.

Lemma 3 (Acemoglu et al. (2011) Lemma 1)**.** *The private belief distributions, $\mathbb{G}_0$ and $\mathbb{G}_1$, satisfy the following properties:*

(a) For all $r \in (0, 1)$, $d\mathbb{G}_0(r)/d\mathbb{G}_1 = (1 - r)/r$

(b) For all $0 < z < r < 1$, $\mathbb{G}_0(r) \geq ((1 - r)/r)\mathbb{G}_1(r) + ((r - z)/2)\mathbb{G}_1(z)$

(c) For all $0 < r < w < 1$, $1 - \mathbb{G}_1(r) \geq (r/(1 - r))(1 - \mathbb{G}_0(r)) + ((w - r)/2)(1 - \mathbb{G}_0(z))$

(d) *The term* $\mathbb{G}_0(r)/\mathbb{G}_1(r)$ *is non-increasing in* r *and is strictly larger than 1 for all* $r \in (\underline{\beta}, \overline{\beta})$

Proof. See Acemoglu et al. (2011, Lemma 1). □

Lemma 4. *Suppose without loss of generality that the true state is 0. With a 'single rational type' (in the context of this paper, this only reflects that all agents have the same utility function) and 'unbounded' (here nonstationary suffices) beliefs, the likelihood ratio,* $l_n = \frac{(1-q_n)}{q_n}$*, of the Bayesian social belief,* q_n*, that* $\theta = 1$ *of Bayesian agents in a complete network converges to zero almost surely* $l_n \to 0$.

Proof. See Smith and Sorensen (2000, Theorem 1b). □

Lemma 5 (Bayesian Social Belief Distribution Relationship)**.** *If the Bayesian social belief of agent* n *in state* θ *has PMF* $h^n_\theta(\cdot)$*, they obey the following relation:*

$$h^n_1(SB_n)(1 - SB_n) = h^n_0(SB_n)SB_n$$

Proof. This follows almost exactly the proof of Acemoglu et al. (2011, Lemma A1 (a))- adjusted in necessary ways. By then definition of a Bayesian social belief, we have for any $sb_n \in (0, 1)$:

$$\mathbb{P}(\theta = 1|SS_n) = \mathbb{P}(\theta = 1|SB_n)$$

Using Bayes' Rule, it follows that:

$$SB_n = \mathbb{P}_\varsigma(\theta = 1|SB_n) = \frac{\mathbb{P}_\varsigma(SB_n|\theta = 1)\mathbb{P}_\varsigma(\theta = 1)}{\sum_{j=0}^{1} \mathbb{P}_\varsigma(SB_n|\theta = j)\mathbb{P}_\varsigma(\theta = j)}$$

(*Note this differs from the analogous expression in Acemoglu et al. (2011) since there are only a finite number of possible Bayesian social beliefs at any point.)

$$SB_n = \frac{\mathbb{P}_\varsigma(SB_n|\theta = 1)}{\mathbb{P}_\varsigma(SB_n|\theta = 0) + \mathbb{P}_\varsigma(SB_n|\theta = 1)}$$
$$\mathbb{P}_\varsigma(SB_n|\theta = 1) = [\mathbb{P}_\varsigma(SB_n|\theta = 0) + \mathbb{P}_\varsigma(SB_n|\theta = 1)]SB_n$$

Using the notation that h^n_θ is the probability mass function for the Bayesian social beliefs of agent n in state θ:

$$h^n_1(SB_n)(1 - SB_n) = h^n_0(SB_n)SB_n$$

□

Lemma 6. *If we take a set A containing N agents, each of whom has an ex-ante success probability of $\alpha_n > 1 - \frac{\delta/2}{N}(1 - \mathbb{G}_0(R))$, the ex-ante probability that their Bayesian social beliefs are all within a rejection region is at least:*

$$\mathbb{P}(sb_n \in [0, 1-R) \cup (R, 1] \ \forall n \in A) > 1 - \delta$$

Proof. For any agent n, there are a finite number of possible social signals, and the Bayesian social belief induced by each of them is deterministic. These beliefs can be listed in order: $\{sb_n^1, sb_n^2, ..., sb_n^E\}$.

$$\begin{aligned}
\alpha_n &= \frac{1}{2}\mathbb{P}(p_n + SB_n < 1|\theta = 0) + \frac{1}{2}\mathbb{P}(p_n + SB_n > 1|\theta = 1) \\
2\alpha_n - 1 &\leq \mathbb{P}(p_n + SB_n < 1|\theta = 0) - 0 \\
&= \sum_{k=1}^{E} \mathbb{P}(SB_n = sb_n^k|\theta = 0)[\mathbb{G}_0(1 - sb_n^k)] \\
&= \sum_{k=1}^{E} h_0^n(sb_n^k)\mathbb{G}_0(1 - sb_n^k)
\end{aligned}$$

The more mass we place on lower Bayesian social beliefs, the higher this gets. Thus any mass placed on values above $1 - R$ would be better placed on $1 - R$:. Similarly, any mass placed on values between $1 - R$ and the minimum Bayesian social belief $sb_n^{\min}$ is best placed on $sb_n^{\min}$.

$$2\alpha_n - 1 \leq (1 - \mathbb{H}_0^n(1 - R))\mathbb{G}_0(R) + \mathbb{H}_0^n(1 - R)\mathbb{G}_0(1 - sb_n^{\min})$$

The Bayesian social belief that maximises this is $sb_n^{\min} = 0$, thus we have the upper bound:

$$\begin{aligned}
2\alpha_n - 1 &\leq (1 - \mathbb{H}_0^n(1 - R))\mathbb{G}_0(R) + \mathbb{H}_0^n(1 - R)\mathbb{G}_0(1 - 0) \\
2\alpha_n - 1 &\leq (1 - \mathbb{H}_0^n(1 - R))\mathbb{G}_0(R) + \mathbb{H}_0^n(1 - R)
\end{aligned}$$

This highest α_n can possibly get with mass ρ on Bayesian social beliefs that are not rejected cannot be higher than:

$$\begin{aligned}
2\alpha_n - 1 &\leq \rho\mathbb{G}_0(R) + (1 - \rho) \\
&\leq 1 - \rho(1 - \mathbb{G}_0(R))
\end{aligned}$$

Thus if we have probability ρ that agent n will observe a signal that is not in a rejection region, α_n is (extremely loosely) bounded above by:

$$\alpha_n \leq 1 - \frac{\rho}{2}(1 - \mathbb{G}_0(R))$$

Thus if α_n is strictly greater than this threshold, the ex ante probability that the social signal is not in a rejection region is less than ρ. Boole's inequality tells us that, for events $\{A_i\}_{i \in \{1,...,N\}}$, $\mathbb{P}(\cup_{i=1}^{N} A_i) \leq \sum_{i=1}^{N} \mathbb{P}(A_i)$. Thus the probability that at least one agent does not have a rejection region signal: $\mathbb{P}(\cup_{i=1}^{N}(sb_i$ not in Rejection Region) is less than or equal to $\sum_{i=1}^{N} \rho$. Let us call the probability that at least one agent does not have a rejection region signal P. It follows that $P \leq N\rho$. The probability that all agents receive a rejection region signal is $1 - P$.

Therefore if we choose $\alpha_n > 1 - \frac{\epsilon/2}{N}(1 - \mathbb{G}_0(R))$, the probability that all agents receive signals in the rejection region is at least $1 - N(\frac{\epsilon}{N}) = 1 - \epsilon$. □

Lemma 7. *If a set of $n \in \{1, ..., N\}$ Bernoulli-Blackwell (in other words simple binary) experiments all have parameters (p_n^0, p_n^1) such that $p_n^1 - p_n^0 > \underline{\Delta}$, $p_n^0 \leq \overline{p}^0$, and $p_n^1 \geq \underline{p}^1$ for all n; all dominate an informative lower bound experiment in the confidence order of Weber (2010). This implies that they give a higher value in the decision problem of this paper.*

Proof. Consider an experiment $(\mathcal{X} = \{0, 1\}, 2^{\{0,1\}}, p^\theta : \theta \in \Theta = \{0, 1\})$, where $p^1 - p^0 > \underline{\Delta}$, $p_n^0 \leq \overline{p}^0$, and $p_n^1 \geq \underline{p}^1$. The 'confidence parameters'[23] of this experiment are:

$$\kappa = \frac{p_n^1}{p_n^0} \qquad\qquad \mathcal{L} = \frac{1 - p_n^0}{1 - p_n^1}$$

One experiment dominates another in the confidence order if both its κ and $\mathcal{L}$ are higher, and an experiment is informative if $(\kappa, \mathcal{L}) > (1, 1)$. For any experiment satisfying the conditions of this lemma, observe that:

$$\kappa = \frac{p_n^1}{p_n^0} \geq \frac{p_n^0 + \underline{\Delta}}{p_n^0} \geq 1 + \frac{\underline{\Delta}}{\overline{p}^0} > 1$$
$$\mathcal{L} = \frac{1 - p_n^0}{1 - p_n^1} \geq \frac{1 - p_n^1 + \underline{\Delta}}{1 - p_n^1} \geq 1 + \frac{\Delta}{1 - p_n^1} \geq 1 + \frac{\Delta}{1 - \underline{p}^1} > 1$$

Thus if we choose any simple binary experiment with confidence parameters $\left(1 + \frac{\underline{\Delta}}{\overline{p}^0}, 1 + \frac{\Delta}{1 - \underline{p}^1}\right)$, we have an informative lower bound experiment dominated in the confidence order by any experiment in our set of n experiments. An experiment that satisfies this is that with

[23] Weber uses κ and λ, but since I am already using λ elsewhere I replace it with $\mathcal{L}$.

success parameters: $\left(\frac{1}{2}\left(1+\frac{1}{2\overline{p}^0/\underline{\Delta}+1}\right), \frac{1}{2}\left(1-\frac{1}{2\overline{p}^0/\underline{\Delta}+1}\right)\right)$. The value of the increment $\frac{1}{2\overline{p}^0/\underline{\Delta}}$ is clearly strictly positive but less than $\frac{1}{2}$, so this is a well-defined Blackwell experiment.

Weber establishes that if one experiment A dominates experiment B in the confidence order, A gives a higher *decision value* in any *standard decision problem*, which he describes in section 3.1. The decision problem of this paper is clearly a standard decision problem, therefore our lower bound experiment gives a lower decision value than any of the n experiments in our set. □

Lemma 8. *If a set of $n \in \{1, ..., N\}$ simple binary experiments all have parameters (p_n^0, p_n^1) such that $p_n^1 < \overline{p}^1$ and $p_n^0 > \underline{p}^0$, then there exists upper bound experiment that dominates them all in the confidence order or Weber (2010). This implies that they give a lower value in the decision problem of this paper.*

Proof. Consider an upper bound experiment with parameters $(\overline{p}^1, \underline{p}^0)$, for this experiment we have confidence parameters:

$$\kappa = \frac{\overline{p}^1}{\underline{p}^0} \qquad \mathcal{L} = \frac{1-\underline{p}^0}{1-\overline{p}^1}$$

Clearly any experiment with $p^1 \leq \overline{p}^1$ and $p^0 \geq \underline{p}^0$ will have lower $(\kappa, \mathcal{L})$ than this upper bound experiment, and be dominated in the confidence order. As per the equivalent reasoning in the previous lemma, they must therefore give a lower value in the decision problem each agent faces. What's more, this upper bound experiment also Blackwell dominates our set of experiments, since it has lower Type 1 and Type 2 error in both states of the world. □

For Theorem 3 I need an equivalent of Blackwell's product experiment dominance result (Blackwell, 1951, Theorem 12), which says that if experiments E_1 and E_2 dominate F_1 and F_2 respectively, then the product experiment $E_1 \bigotimes E_2$ dominates $F_1 \bigotimes F_2$. Since I am using the confidence order of Weber, I cannot apply Blackwell's Theorem 12 directly. Let us define $\succeq_c$ as the confidence order, which applies to simple binary experiments (experiments with two states and two signals). Let us further define $\succeq$ as the binary relation that orders experiments according to the decision value they achieve. Weber gives us that $\succeq_c = \succeq$ over the space of simple binary experiments for our decision problem

Lemma 9. *Suppose we have M experiments $E_1, ..., E_M$, and M experiments $F_1, ..., F_M$*

such that E_i dominates F_i in the confidence order for each $i \in 1, ..., M$. Then the decision value upon observing $\{E_1, ..., E_M\}$ is at least as high as upon observing $\{F_1, ..., F_M\}$.

Proof. First, observe that the product experiment $E_1 \bigotimes E_2$ (in which E_1 and E_2 are independent of each other, conditional on the state of the world) must dominate the the product experiment $E_1 \bigotimes F_2$. To see this, consider the agent observing first E_1 and then the second experiment (for Bayesian updating, it does not matter in what order the agent observes them, it is 'divisible' to adopt Cripps (2018) terminology). Whatever posterior they form upon observing E_1, call this the interim belief, it serves as their prior for observing E_2 or F_2. $E_2 \succeq_c F_2$ implies that the decision value upon observing E_2 is weakly higher than that for F_2 for *any* prior. Thus $E_1 \bigotimes E_2 \succeq E_1 \bigotimes F_2$. By the same reasoning, we can also deduce that $E_1 \bigotimes E_2 \succeq F_1 \bigotimes E_2$, and that both $F_1 \bigotimes E_2 \succeq F_1 \bigotimes F_2$ and $E_1 \bigotimes F_2 \succeq F_1 \bigotimes F_2$. Since this preference relation has a utility representation (that of the expected value of the decision problem), it is transitive. Therefore $E_1 \bigotimes E_2 \succeq F_1 \bigotimes F_2$.

Now let us adopt the labels $E^k := E_1 \bigotimes ... \bigotimes E_k$ and $F^k := F_1 \bigotimes ... \bigotimes F_k$. $E^{k+1} = E^k \bigotimes E_{k+1}$ and $F^{k+1} = F^k \bigotimes F_{k+1}$. Applying once more exactly the same reasoning above, we can see that if for some $k \in \mathbb{N}$ $E^k \succeq F^k$, it follows that $E^{k+1} \succeq F^{k+1}$. We have established that $E^2 \succeq F^2$ (and of course we assumed that $E^1 = E_1 \succeq F_1 = F^1$), thus by induction it follows that for any M, $E^M \succeq F^M$ □

B Omitted Proofs

Proof of Lemma 1. The fact that the private and (motivated) social beliefs can be summed and compared to 1 in order to choose an action can be seen by manipulating Bayes' Rule. In this case, the fact that my agents are assumed to combine their private and motivated social beliefs as a Bayesian, as if they had both been formed in a Bayesian fashion, is essential. The sum representation is then simply an application of Acemoglu et al. (2011, Proposition 2).

The exact form of the motivated social belief (MSB) if $\mathbb{P}_\varsigma(\theta = 1|B(n)) < (1 - R)$, is also a straightforward consequence of the motivated reasoning procedure.

The form of MSB when $\mathbb{P}_\varsigma(\theta = 1|B(n)) \geq (1 - R)$ results from comparing the belief formed by a Bayesian with prior $\frac{1}{2}$, and that formed by a Bayesian with prior $\frac{1}{2}(1 + s)$. Let us call the former $\mathbb{P}$ and the latter $\tilde{\mathbb{P}}$:

$$(1 + s)\mathbb{P} = \frac{(1 + s)\mathbb{P}(B(n)|\theta = 1)}{\mathbb{P}(B(n)|\theta = 1) + \mathbb{P}(B(n)|\theta = 0)} \tag{B.1}$$

$$\tilde{\mathbb{P}} = \frac{\mathbb{P}(B(n)|\theta = 1)(1 + s)}{\mathbb{P}(B(n)|\theta = 1)(1 + s) + \mathbb{P}(B(n)|\theta = 0)(1 - s)} \tag{B.2}$$

Dividing B.1 by B.2, and applying Bayes' Rule (using the true prior) twice on the Right-hand side produces:

$$\begin{aligned}(1+s)\frac{\mathbb{P}}{\tilde{\mathbb{P}}} &= 1+s\big(\mathbb{P}(\theta=1|B(n))-\mathbb{P}(\theta=0|B(n))\big)\\ (1+s)\frac{\mathbb{P}}{\tilde{\mathbb{P}}} &= 1-s+2s\mathbb{P}\\ \tilde{\mathbb{P}} &= \frac{(1+s)\mathbb{P}}{2s\mathbb{P}+(1-s)}\end{aligned}$$

Equivalently for type 0 agents one can find $\tilde{\mathbb{P}} = \frac{(1-s)\mathbb{P}}{(1+s)-2s\mathbb{P}}$. □

I now present the proof of Theorem 2, this proof is by contradiction, though I present an alternative that assumes the signal structure is nonstationary Appendix E.1.

Proof of Theorem 2. Let's first prove that if there is a finite upper bound on each agent's neighborhood size ($|B(n)| < M_1$ for all n), and a restriction such that $\arg\min_k\{B(n)\} \geq n - M_2$ (where of course $M_2 \geq M_1$), the statement holds.

Suppose for contradiction that $\alpha_n \to 1$ as $n \to \infty$. For any $\epsilon > 0$, there is some $N_\epsilon \in \mathbb{N}$ such that for all $n > N_\epsilon$, $\alpha_n > 1-\epsilon$. Define $\alpha_{IOR}^{M_1-1} < 1$ as the accuracy of a Bayesian with $M_1 - 1$ rejecting neighbors, who is additionally informed that all his neighbors have rejected their social signals. Choose any $\delta \in (0,1)$, and ϵ such that:

$$\epsilon < \min\{\big(\frac{1}{2}(1-\beta)\big)^{M_1-1}(1-\delta)\big(1-\alpha_{UCD}^{M_1-1}\big), \frac{\delta/2}{M_1-1}(1-\mathbb{G}_0(R))\}$$

This first term is reverse engineered to provide a contradiction, this second allows us to use Lemma 6. Consider an agent m such that $m > N_\epsilon + M_2$, and without loss of generality let $|B(m)| = M_1 - 1$. By our Lemma, the probability m has an all-rejection-region neighborhood (all of his neighbors receive a social signal in $[0, 1-R) \cup (R, 1]$) is some $P_m > 1-\delta$. The probability that they all reject is then the probability that each neighbor is of noncongenial type multiplied by P_m, which is $\Big(\frac{1}{2}(1-\beta)\Big)^{M_1-1} P_m$.

Thus we have

$$\alpha_m < \big(1-\big(\frac{1}{2}(1-\beta)\big)^{M_1-1}P_1\big)\cdot(1) + \big(\frac{1}{2}(1-\beta)\big)^{M_1-1}P_1\cdot\alpha_{UCD}^{M_1-1}$$

Our selection of ϵ then implies

$$1-\epsilon < \alpha_m < 1-\big(\frac{1}{2}(1-\beta)\big)^{M_1-1}(1-\delta)\big(1-\alpha_{UCD}^{M_1-1}\big)$$

Given our definition of ϵ, this is a contradiction! We have proved our theorem subject to the additional condition. □

Now we must prove the theorem without the added restriction. The logic of the proof is the same, except that without this restriction a neighborhood could always contain agents from arbitrarily far back, preventing the use of Lemma 6 on the entire neighborhood.

Dropping the requirement that $\arg\min_k\{B(n)\} \geq n - M_2$, we can take the same δ as before, and choose ϵ such that:

$$\epsilon < \min\{\epsilon^*, \Big(\frac{1}{2}(1-\beta)\Big)^{M_1-1}(1-\delta)\Big(1-\alpha_{UCD}^{M_1-1}\Big), \frac{\delta/2}{M_1-1}(1-\mathbb{G}_0(R)) \tag{B.3}$$

where ϵ^* is defined later in the proof.

Let us partition the neighborhood into agents before N_ϵ and agents after N_ϵ (taking N_ϵ now to be that N_ϵ corresponding to this newly selected (smaller) ϵ). Call $B_1(m) = B(m) \cap \{k : k \leq N_\epsilon\}$ and $B_2(m) = B(m) \setminus B_1(m)$. There must be an upper bound on the accuracy of agents in $B_1(m)$ that is strictly less than 1, since only a finite amount of information has been generated by any finite time in the game, including by N_ϵ. Call this upper bound $\overline{\alpha}_1$.

The information contained in the actions of $B_1(m)$ must be at most the information that would be contained in the Blackwell supremum experiment of the set of all possible Blackwell experiments with $|B_1(m)|$ binary signals and arbitrary correlation structure (such a thing exists when $|\Theta| = 2$ by Bertschinger and Rauh (2014)) with each of the $|B_1(m)|$ agents each correct with probability $\overline{\alpha}_1$. Consider m acting with the benefit of $|B_1(m)|$ such signals and $|B_2(m)|$ signals of agents who rejected the social signal, call the accuracy that would result from this $\alpha^*(|B_2(m)|)$. This must be strictly less than 1 and no outcome of this is possible in one state of the world, and impossible in the other.

The probability that at least one agent in $B_2(m)$ rejects their signal is now $|B_2(m)| \times \delta \leq N\delta$. Thus since we have chosen $\alpha_n \geq 1 - \frac{\delta}{|B_1(m)|+|B_2(m)|}(1 - \mathbb{G}_0(R))$, the probability that all agents receive signals in the rejection region is at least $\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)|+|B_2(m)|}\Big)$. Let the probability that all neighbors in $|B_2(m)|$ are in their rejection regions be P_2.

We must have that:

$$\begin{aligned}
\alpha_m &< \big(1 - (\frac{1}{2})^{M_1-1-|B_1(m)|}P_2\big) \cdot 1 + (\frac{1}{2})^{M_1-1-|B_1(m)|}P_2\alpha^*(|B_2(m)|) \\
\alpha_m &< 1 - (\frac{1}{2})^{M_1-|B_1(m)|}P_2(1 - \alpha^*(|B_2(m)|)) \\
\alpha_m &< 1 - (\frac{1}{2})^{|B_2(m)|}\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)| + |B_2(m)|}\big)(1 - \alpha^*(|B_2(m)|))
\end{aligned}$$

Now let us define:

$$\epsilon^* = min_{B_2(m)}\big((\frac{1}{2})^{|B_2(m)|}(1 - |B_2(m)|\frac{\delta}{|B_1(m)| + |B_2(m)|})(1 - \alpha^*(|B_2(m)|))\big)$$

This must exist since $B_2(m) \in \{0, ..., M_1\}$. Thus we have from our definition of ϵ that:

$$1 - \epsilon < \alpha_m < 1 - (\frac{1}{2})^{|B_2(m)|}\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)| + |B_2(m)|}\big)(1 - \alpha^*(|B_2(m)|))$$
$$1 - \epsilon < 1 - (\frac{1}{2})^{|B_2(m)|}\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)| + |B_2(m)|}\big)(1 - \alpha^*(|B_2(m)|))$$

This is a contradiction. If $|B_1(m)| = M_1 - 1$ for all agents after N_ϵ, then the expanding observations condition is not even satisfied, and the agents are acting on the basis of a finite amount of information in perpetuity, and of course complete Bayesian learning does not obtain. □

Proof of Corollary 2.1. First, I prove that correct consensus does not obtain when we do not have expanding sample sizes. Secondly I prove that there exist settings with expanding sample sizes in which there is correct consensus in a model with only Bayesians ($\beta = 1$).

On this first point: by Theorem 2, in the presence of motivated reasoning there is no complete Bayesian learning without expanding sample sizes so almost sure Bayesian learning necessarily does not obtain (it is stronger than and implies complete Bayesian learning). If the private beliefs under consideration are nonstationary, there cannot be consensus (as any interior social belief can be overturned or rejected with strictly positive probability with nonstationary beliefs). Part 2 of Theorem 1 also implies this.

With stationary information structures, either there are interior stationary beliefs in favour of each state in which infinitely many agents can be stuck with positive probability, or this is only true on one side in which case in at least one state there will not be consensus on the correct state, and thus overall no correct consensus.

Moving onto the second point: Rosenberg and Vieille (2019, Theorem 3) establishes that such Bayesian networks do exist. Assuming that private signals are not only unbounded but also very informative in the sense that $\int_0^1 \frac{1}{\mathbb{G}(p)}dp < +\infty$ and $\int_0^1 \frac{1}{1-\mathbb{G}(1-p)}dp < +\infty$, and imposing two technical assumptions, they prove that learning is 'efficient' (the expected number of wrong actions is finite) and therefore that there is almost sure learning and consensus (on the true state) in the line network. Thus, there exist networks of Bayesian agents that do not satisfy expanding sample sizes but nonetheless exhibit correct consensus.

Finally addressing the fragility element of the result: Theorem 2 holds for any $\beta \in (0, 1)$, any level of $R < 1$ and any $s \in (0, 1)$. As discussed in Appendix F.1, it even holds if agents do not snap back to 1/2 but $\{\epsilon, 1-\epsilon\}$. Hence it does not matter how little motivated reasoning is introduced, unbounded beliefs producing correct consensus in the Bayesian-only setting will not continue to do so with motivated reasoners. □

Proof of Theorem 3. Take an agent n with M nested neighbors. Using Lemma 7, we can observe that each nested neighbor action must be more informative than the lower bound Blackwell experiment. By Lemma 9, it therefore follows that a Bayesian agent observing M nested neighbors must manage to match the state with greater probability than one observing M lower bound experiments. As M converges to infinity, this lower bound probability converges to 1. Therefore an agent observing M nested neighbor experiments must have Bayesian accuracy converging to 1 as M converges to infinity. It follows that the agents within S must have Bayesian accuracy converging towards 1, since for any $M \in \mathbb{N}$ the probability that they observe fewer than M nested neighbors converges to zero. □

Proof of Theorem 1 Part 3. To recall the assumptions already mentioned in the main text, assume we have: (i) A complete network and (ii) Bounded beliefs: $[1-\overline{B}, \overline{B}]$.

For any Bayesian social belief, λ, define $\Delta_a(\lambda)$ as the (positive) distance between λ_n and λ_{n+1} upon observing $x_n = a$ with $a \in \{0, 1\}$. Δ_a is well-defined (since the posterior depends only upon the prior (the Bayesian social belief here) and likelihood) and a continuous function of λ for all a, by the properties of Bayesian updating. The set $[1-b_1, b_0]$ is compact, so the maxima $\overline{\Delta}_a := \max_{\lambda \in [1-b_1, b_0]}\{\Delta_a(\lambda)\}$ exist by the Weierstrass extreme value theorem.

Suppose that $R > b_0 + \overline{\Delta}_1$, and $1 - R < (1 - b_1) - \overline{\Delta}_0$; this ensures that for no Bayesian social belief will any observation be able to push the updated Bayesian social belief into a rejection in which some types will reject it. These definitions establish that the Bayesian social belief cannot enter a rejection region, so to complete the proof we must simply establish that the Bayesian social belief will necessarily enter the region $(1 - b_1, b_0)^c$ at some point, and that if it does so, agents will all take the same action.

This second point is easy to see; if, for example, the Bayesian social belief drops below $1 - b_1$, it is sufficiently low that no signal can satisfy the decision rule in Proposition 1 and lead the agent to choose $x_n = 1$. Since the belief is not in the rejection region, this is sufficient to establish that type 1 agents will choose $x_n = 1$. The other type agents have equivalent thresholds $1 - b_0$ and $1 - \overline{B}$ that are both higher than $1 - b_1$, so all agents are choosing $x_n = 0$ in this region. To prove the first point, we can simply deploy Smith and Sorensen (2000, Theorem B.1); the Martingale-Markov Bayesian social belief must eventually settle at some fixed-point (converging almost surely to it). Since the convergence in this theorem of Smith and Sorensen is almost sure, it establishes consensus (/tribalism) as defined in Definition 2. □

Proof of Proposition 2. Suppose, without loss of generality, that $\theta = 0$, a symmetric argument holds for $\theta = 1$. Firstly, we can establish that when experiment A dominates experiment B in the confidence order, it engenders beliefs stronger than R with at least weakly greater probability.

To see this, in place of every agent, consider an agent who observes exactly the same information, but chooses actions to solve a different 'standard' decision problem (where again I mean 'standard' in the sense of Weber (2010)). Specifically, suppose they solve a binary problem where action $x_n = 0$ is optimal if they have a belief less than $1 - R$ that $\theta = 1$ and $x_n = 1$ is optimal otherwise.[24] Whether an agent's decision value is higher or lower in such a problem corresponds exactly to the probability they have a belief less than $1-R$ in each state of the world. Since this is a standard decision problem, Lemma's 7 and 9 apply, and we can see that observing M nested neighbors dominates observing M lower bound experiments for any $M \in \mathbb{N}$, i.e. produces a higher probability of having a posterior lower than R. Since $x_n = 0$ is chosen whenever an agent has belief greater than $1 - R$, they choose $x_n = 0$ whenever their log-likelihood ratio is less than $log\frac{1-R}{R}$.

Following Moscarini and Smith (2002) and Tamuz (2022), and calling the posterior upon observing m copies of the lower bound experiment q_m, we can observe that the probability of forming a belief above $1 - R$ after m can be expressed:

$$\mathbb{P}(q_m \leq 1 - R|\theta = 0) = \exp\Big(- m\rho^0_{\underline{\Delta}} + o(n)\Big)$$

where $\rho^0_{\underline{\Delta}} = \min tK^0_{\underline{\Delta}}(t)$ and $K^0_{\underline{\Delta}}$ is the cumulant generating function of the log-likelihood ratio of the lower bound experiment $\underline{\Delta}$ conditional on $\theta = 0$. Thus, for any $\epsilon > 0$ there is some M big enough such that $\mathbb{P}(q_m \leq 1 - R|\theta = 0) > 1 - \epsilon$ for all $m > M$. □

References

Acemoglu, Daron, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar (2009): "Rate of Convergence of Learning in Social Networks," .

——— (2011): "Bayesian learning in social networks," *The Review of Economic Studies*, 78, 1201–1236.

[24]An example of such a problem can be produced by adjusting the decision problem of Cremin (2025). In this paper, agents choose both a binary action $x_n \in \{0, 1\}$ and whether their action is visible $v_n \in \{0, 1\}$, where they want x_n to match the state, but are rewarded (punished) more for visible actions if they are correct (incorrect). If we strip them of the ability to choose $x_n = 1$, relabel $(x_n = 1) := (x_n = 1, v_n = 0)$ and $(x_n = 0) := (x_n = 0, v_n = 1)$, and choose the value of their 'confidence' parameter appropriately, we have such a problem.

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2024): "A model of online misinformation," *Review of Economic Studies*, 91 (6), 3117–3150.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2020): "The Polarization of Reality," *AEA Papers and Proceedings*, 110, 324–28.

Arieli, Itai, Yakov Babichenko, Stephan Müller, Farzad Pourbabaee, and Omer Tamuz (2023): "The Hazards and Benefits of Condescension in Social Learning," *arXiv preprint arXiv:2301.11237*.

Bala, Venkatesh and Sanjeev Goyal (1998): "Learning from neighbours," *The review of economic studies*, 65 (3), 595–621.

Banerjee, Abhijit V. (1992): "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, (August).

Bartels, Larry M (2002): "Beyond the running tally: Partisan bias in political perceptions," *Political behavior*, 24 (2), 117–150.

Bénabou, Roland (2015): "The economics of motivated beliefs," *Revue d'économie politique*, 125 (5), 665–685.

Bénabou, Roland and Jean Tirole (2016): "Mindful economics: The production, consumption, and value of beliefs," *Journal of Economic Perspectives*, 30 (3), 141–64.

Bertschinger, Nils and Johannes Rauh (2014): "The Blackwell relation defines no lattice," in *2014 IEEE International Symposium on Information Theory*, 2479–2483.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992): "A Theory of Fads , Fashion , Custom , and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100 (5), 992–1026.

Blackwell, David (1951): "Comparison of experiments," in *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, University of California Press, vol. 2, 93–103.

Bohren, J. Aislinn (2016): "Informational herding with model misspecification," *Journal of Economic Theory*, 163, 222–247.

Bohren, J. Aislinn and Daniel N. Hauser (2021): "Learning With Heterogeneous Misspecified Models: Characterization and Robustness," *Econometrica*, 89 (6), 3025–3077.

Bowen, T Renee, Danil Dmitriev, and Simone Galperti (2023): "Learning from shared news: When abundant information leads to belief polarization," *The Quarterly Journal of Economics*, 138 (2), 955–1000.

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes (1980): *The american voter*, University of Chicago Press, originally published in 1960.

Çelen, Boğaçhan and Shachar Kariv (2004): "Observational learning under imperfect information," *Games and Economic Behavior*, 47 (1), 72–86.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach (2021): "Learning in the Household," Tech. rep., National Bureau of Economic Research.

——— (2022): "Not Learning from Others," Working Paper 30378, National Bureau of Economic Research.

Cremin, J.W.E (2025): "Bot Got Your Tongue? Social Learning with Timidity and Noise," .

Cripps, Martin W (2018): "Divisible updating," *Manuscript, UCL*.

Dasaratha, Krishna and Kevin He (2024): "Aggregative Efficiency of Bayesian Learning in Networks," *arXiv preprint arXiv:1911.10116 (2019)*.

DeGroot, Morris H (1974): "Reaching a consensus," *Journal of the American Statistical association*, 69 (345), 118–121.

Epley, Nicholas and Thomas Gilovich (2016): "The mechanics of motivated reasoning," *Journal of Economic perspectives*, 30 (3), 133–40.

Eyster, Erik and Matthew Rabin (2009): *Rational and naive herding*, Centre for Economic Policy Research.

Ganguly, Ananda and Joshua Tasoff (2017): "Fantasy and dread: The demand for information and the consumption utility of the future," *Management Science*, 63 (12), 4037–4060.

Geiger, Abigail (2021): "Political polarization in the American public," .

Gerber, Alan S and Gregory A Huber (2009): "Partisanship and economic behavior: Do partisan differences in economic forecasts predict real economic behavior?" *American Political Science Review*, 103 (3), 407–426.

Goeree, Jacob, Thomas Palfrey, and Brian Rogers (2006): "Social learning with private and common values," *Economic Theory*, 28, 245–264.

Guilbeault, Douglas, Joshua Becker, and Damon Centola (2018): "Social learning and partisan bias in the interpretation of climate trends," *Proceedings of the National Academy of Sciences*, 115 (39), 9714–9719.

Hamilton, James Douglas (2020): *Time series analysis*, Princeton university press.

Kartik, Navin, S Lee, Tianhao Liu, and Daniel Rappoport (2022): "Beyond Unbounded Beliefs: How Preferences and Information Interplay in Social Learning," Tech. rep.

Levy, Gilat and Ronny Razin (2019): "Echo Chambers and Their Effects on Economic and Political Outcomes," *Annual Review of Economics*, 11 (1), 303–328.

Little, Andrew T (2021): "Detecting Motivated Reasoning," *URL: osf. io/b8tvk*.

Lobel, Ilan and Evan Sadler (2015): "Information diffusion in networks through social learning," *Theoretical Economics*.

——— (2016): "Preferences, Homophily and Social Learning," *Operations Research*.

Lomys, Niccolo (2020): "Collective search in networks," *Available at SSRN 3197244*.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie (2018): "A Theory of Non-Bayesian Social Learning," *Econometrica*, 86 (2), 445–490.

Moore, Adam, Sujin Hong, and Laura Cram (2021): "Trust in information, political identity and the brain: an interdisciplinary fMRI study," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376 (1822), 20200140.

Moscarini, Giuseppe and Lones Smith (2002): "The Law of Large Demand for Information," *Econometrica*, 70 (6), 2351–2366.

Nyhan, Brendan (2020): "Facts and Myths about Misperceptions," *Journal of Economic Perspectives*, 34 (3), 220–36.

Oprea, Ryan and Sevgi Yuksel (2022): "Social exchange of motivated beliefs," *Journal of the European Economic Association*, 20 (2), 667–699.

Oster, Emily, Ira Shoulson, and E Ray Dorsey (2013): "Optimal expectations and limited medical testing: Evidence from Huntington disease," *American Economic Review*, 103 (2), 804–830.

Rekker, Roderik (2022): "Political polarization over factual beliefs," in *Knowledge resistance in high-choice information environments*, Routledge, 222–236.

Rosenberg, Dinah and Nicolas Vieille (2019): "On the Efficiency of Social Learning," *Econometrica*, 87 (6), 2141–2168.

Saucet, Charlotte and Marie Claire Villeval (2019): "Motivated memory in dictator games," *Games and economic Behavior*, 117, 250–275.

Smith, Lones and Peter Sorensen (2000): "Pathological Outcomes of Observational Learning," *Econometrica*, 68 (2), 371–398.

Tamuz, Omer (2022): "The Value and Costs of Information," *Mini Course Lecture Notes.*

Weber, Thomas A (2010): "Simple methods for evaluating and comparing binary experiments," *Theory and decision*, 69, 257–288.

Weizsäcker, Georg (2010): "Do we follow others when we should? A simple test of rational expectations," *American Economic Review*, 100 (5), 2340–2360.

Westen, Drew, Pavel Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann (2006): "Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgment in the 2004 U.S. Presidential Election," *Journal of Cognitive Neuroscience*, 18 (11), 1947–1958.

Zimmermann, Florian (2020): "The Dynamics of Motivated Beliefs," *American Economic Review*, 110 (2), 337–61.

C Online Appendix: Line Network Example Working

C.1 A Line Network Example & Information Loss

The line network is a particularly convenient topology to study, especially with some specific parameter assumptions (though these assumptions serve to create a convenient example, the basic intuition is general, and applies to any line network with a strictly positive measure of motivated reasoners and $1/2 < R < 1$). In it, we can straightforwardly express the Bayesian accuracy, α_n, of an agent n, who observes a Bayesian predecessor, as a function of the Bayesian accuracy of said predecessor: α_{n-1}. Even if we suppose that there are in fact no Bayesians at all, how they would behave reflects the extent to which information is successfully aggregated through the observation network as the game proceeds. In this simplified example I assume there is no prior-shifting ($s = 0$), $R = 0.7$, $\beta = 0$, and the private signals have density functions: $(f_0(\varsigma), f_1(\varsigma)) = (2(1-\varsigma), 2\varsigma)$. These signal distributions produce identical belief distributions: $f_\theta(\cdot) = g_\theta(\cdot)$.

It follows that the CDFs of the belief distributions take the following forms:

- $\mathbb{G}_0(x) = \mathbb{F}_0(x) = x(2-x)$
- $\mathbb{G}_0(1-x) = (1-x)(1+x)$
- $\mathbb{G}_1(x) = x^2$
- $\mathbb{G}_1(1-x) = (1-x)^2$

Claim 1. *The relationship between the Bayesian accuracy and possible Bayesian social beliefs of n is given by:*

$$2\alpha_n - 1 = \sum_{k=1}^{E} h_1^n(sb_n^k)\frac{(1-sb_n^k)}{sb_n^k}\mathbb{G}_0(1-sb_n^k) - \sum_{k=1}^{E} h_1^n(sb_n^k)\mathbb{G}_1(1-sb_n^k)$$

Proof. There are a finite number of possible social signals, and the Bayesian social belief induced by each of them is deterministic. These beliefs can be listed in order: $\{sb_n^1, sb_n^2, ..., sb_n^E\}$.

$$\begin{aligned}
\alpha_n &= \frac{1}{2}\mathbb{P}(p_n + SB_n < 1|\theta = 0) + \frac{1}{2}\mathbb{P}(p_n + SB_n > 1|\theta = 1) \\
\alpha_n &= \frac{1}{2}\mathbb{P}(p_n + SB_n < 1|\theta = 0) + \frac{1}{2}\big(1 - \mathbb{P}(p_n + SB_n \leq 1|\theta = 1)\big) \\
2\alpha_n - 1 &= \mathbb{P}(p_n + SB_n < 1|\theta = 0) - \mathbb{P}(p_n + SB_n \leq 1|\theta = 1) \\
&= \sum_{k=1}^{E}\mathbb{P}(SB_n = sb_n^k|\theta = 0)[\mathbb{G}_0(1 - sb_n^k)] - \sum_{k=1}^{E}\mathbb{P}(SB = sb_n^k|\theta = 1)\mathbb{G}_1(1 - sb_n^k) \\
&= \sum_{k=1}^{E} h_0^n(sb_n^k)\mathbb{G}_0(1 - sb_n^k) - \sum_{k=1}^{E} h_1^n(sb_n^k)\mathbb{G}_1(1 - sb_n^k)
\end{aligned}$$

Using Lemma 5 that $h_0^n(SB_n) = h_1^n(SB_n)\frac{(1-SB_n)}{SB_n}$, we get the above expression. □

Let us call $h_1^n(sb_n^0) := q_0^n$, since there are only two possible signals for each agent in the line network it follows that $h_1^n(sb_n^1) := 1 - q_0^n$.

$$\begin{aligned}
2\alpha_n - 1 = q_0^n &\left[\frac{(1 - sb_n^0)}{sb_n^0}\mathbb{G}_0(1 - sb_n^0) - \mathbb{G}_1(1 - sb_n^0)\right] \\
&+ (1 - q_0^n)\left[\frac{(1 - sb_n^1)}{sb_n^1}\mathbb{G}_0(1 - sb_n^1) - \mathbb{G}_1(1 - sb_n^1)\right]
\end{aligned}$$

The symmetry of my assumptions also gives us that $sb_n^0 = 1 - sb_n^1$, so we can simplify this further to:

$$\begin{aligned}
&= q_0^n(sb_n^1)^2\left[\frac{1}{1 - sb_n^1}\right] \\
&+ (1 - sb_n^1)^2\left[\frac{1}{sb_n^1}\right] - q_0^n(1 - sb_n^1)^2\left[\frac{1}{sb_n^1}\right]
\end{aligned}$$

This symmetry also means that $h_1^n(sb_n^1) = h_0^n(sb_n^0)$ and $h_1^n(sb_n^0) = h_0^n(sb_n^1)$, which with lemma 5 implies that $h_1^n(sb_n^1) = sb_n^1$. Using this, the above becomes:

$$\alpha_n = (q_0^n)^2 - q_0^n + 1 = \alpha_{n-1} - \alpha_{n-1} + 1$$

since of course q_0^n is simply α_{n-1} in the line network. Thus upon observing a binary signal that matches the correct state with probability α, a Bayesian agent with belief distributions $(\mathbb{G}_0(\cdot), \mathbb{G}_1(\cdot))$ will match the true state with probability $H(\alpha)$ given by equation C.1. Given the particular belief distributions specified above, this simplifies to C.2.

$$H(\alpha) := \frac{\alpha}{2}\big[\mathbb{G}_0(\alpha) + 1 - \mathbb{G}_1(1-\alpha)\big] + \frac{(1-\alpha)}{2}\big[\mathbb{G}_0(1-\alpha) + 1 - \mathbb{G}_1(\alpha)\big] \tag{C.1}$$

$$H(\alpha) = \alpha^2 - \alpha + 1 \tag{C.2}$$

α^* is then simply defined as the value $H(R)$. Beyond this, with 50% probability n is observing a congenial type agent who matches the state with probability α_{n-1}, and otherwise he is observing an agent with success probability $H(0.5) = 0.75$, since they revert to the prior of $\frac{1}{2}$.

The first panel of figure 4, subfigure 4a, illustrates the fact that a line network of Bayesian agents gives learning, since observing an agent with any level of accuracy α gives some $H(\alpha) > \alpha$, the following agent then achieves an accuracy of $H(H(\alpha)) > H(\alpha)$ and so on. The accuracy eventually converges to 1 where $H(1) = 1$, reflecting the fact that complete Bayesian learning does obtain in a line network of exclusively Bayesian agents. Our agents, however, are not Bayesians. Beyond a certain level of accuracy, α^*, an agent must be observing a social signal generating a Bayesian social belief in the region $[0, 1-R) \cup (R, 1]$ in order to do better. Therefore, observing an agent whose Bayesian equivalent matches the state with probability α^* or higher implies observing an agent who will reject their social signal if they are of the non-congenial type. Thus, the action agent n observes when $\alpha_{n-1} > \alpha$ matches the state with probability α_{n-1} if $n-1$ is of congenial type and with probability $\alpha_{Rej} = \alpha_1$ otherwise. Their Bayesian accuracy becomes $H(\frac{1}{2}\alpha_{n-1} + \frac{1}{2}\alpha_{Rej})$, where α_{Rej} is the accuracy of a Bayesian agent whose neighbor will have rejected their social signal if of non-congenial type.[25]

Figure 4b shows $U(\alpha) := H(\frac{1}{2}\alpha + \frac{1}{2}\alpha_{Rej})$ function this bound graphed alongside our $H(\alpha)$ function. Since $H(\alpha)$ is relevant below α^*, and $U(\alpha)$ above it, we can see that the function in figure 4c gives the Bayesian accuracy of any agent observing a neighbor of any level of accuracy. Since the range of this function, unlike its Bayesian cousin $H(\alpha)$, is bounded away from 1, no agent can ever approach perfect accuracy.

[25] The antisymmetric distributions of this example guarantee that there is a single α^* above which non-congenial types reject, and below which they do not. Similar arguments still allow us to bound the maximal accuracy away from 1 without this property.

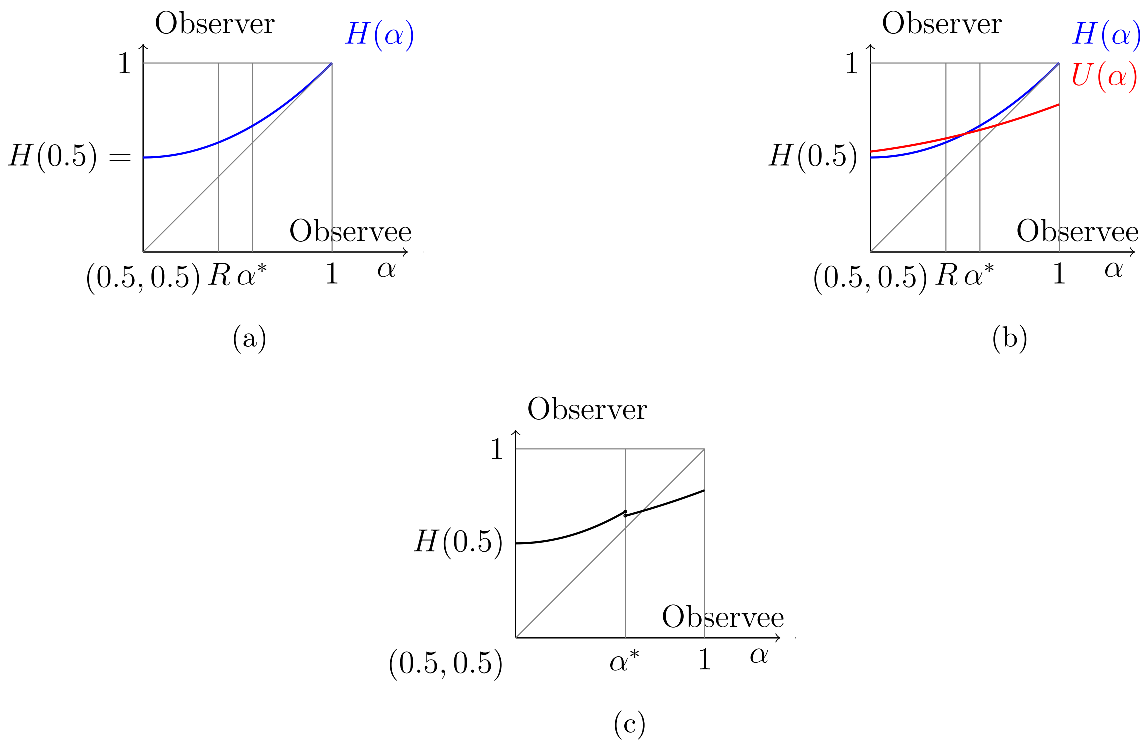
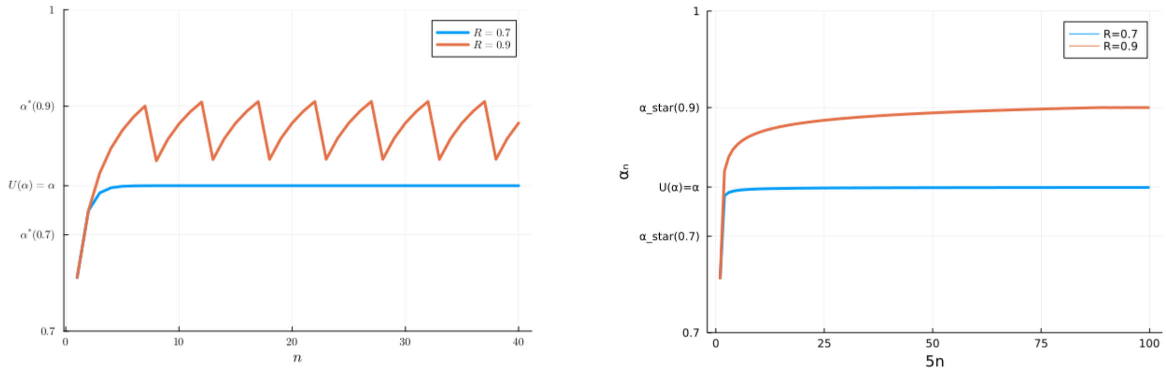


Figure 4

Whatever the values of the above parameters (and implied form of the function $H(\cdot)$), an agent must either be observing an neighbor with $\alpha_{n-1} \leq \alpha^* < 1$- in which case they must have an accuracy below $H(\alpha^*)$- or a neighbor with accuracy $\alpha_{n-1} > \alpha^*$, in which case their accuracy is bounded above by $\sup\{U(\alpha) : \alpha \in (\alpha^*, 1]\} = H(\frac{1}{2}(1-\beta)\alpha_{Rej} + \frac{1}{2}(1+\beta)) < 1$.

$$\forall n \;\; \alpha_n \leq \max\{H(\alpha^*), H(\frac{1}{2}(1-\beta)\alpha_{Rej} + \frac{1}{2}(1+\beta))\} < 1$$

Plotting this as in Figure 5a, we can see that the Bayesian accuracy of agents does not necessarily converge at all, let alone to 1. If the α^* implied by R is above the intersection of $U(\alpha)$ with the 45° line, the accuracy of agents repeatedly climbs the $H(\alpha)$ curve only to drop back down below α^* when an agent achieves an accuracy of above α^*. Otherwise, the process keeps climbing upon reaching the $U(\alpha)$ curve, converging to this $U(\alpha) = \alpha$ fixed point. Figure 5b shows that a network in which each agent draws a neighbor uniformly from all predecessors is similar in that any R implying an α^* value below the fixed point of $U(\alpha)$ produces an asymptotic accuracy of exactly this fixed point. In contrast, however, for higher values of R it achieves smooth convergence to α^*, if it achieves this very slowly.



(a) The line network

(b) Uniform Sampling

Figure 5: The path of α_n against n, with $R \in \{0.7, 0.9\}$, $s = 0$, $\beta = 0$, $(f_0, f_1) = (2(1-\varsigma), 2\varsigma)$.

This failure of complete Bayesian learning clearly demonstrates that the Improvement Principle arguments no longer hold in this setting. The Bayesian equivalent of an agent n observing the action of an agent $n-1$ with accuracy $\alpha_{n-1} > \alpha^*$ can no longer improve upon α_{n-1}, as this is the accuracy of a latent variable. They can only improve upon the probability with which the observed action matches the state. Learning could perhaps be salvaged from this breakdown in information monotonicity if the correlation between x_n and χ_n were converging to 1 as $\alpha_n \to 1$, and the probability $\mathbb{P}(x_n \neq \chi_n)$ converged to zero fast enough, but something closer to the opposite occurs here. The higher α_n, the higher the probability with which non-congenial agents form rejection-region Bayesian social beliefs. Since the probability with which a given agent is of the non-congenial type is $\frac{1}{2}(1-\beta)$, this implies that $\mathbb{P}(x_n \neq \chi_n) \to \frac{1}{2}(1-\beta)$ as $\alpha_n \to 1$.

C.2 Efficiency and Speed of Convergence

Incidentally, this is an example of assumptions that do not give the efficiency of Rosenberg and Vieille (2019) (finitely many wrong actions in expectation) even without motivated reasoners though they would give complete Bayesian learning in that setting, since the signal structure is not efficient ($\int_0^1 \frac{1}{s} = \infty$). In fact Acemoglu et al. (2009) also allow use to give the speed of convergence here. Acemoglu et al. (2009, Proposition 2) states that in this network topology, with private belief distributions with polynomial tails (as here), $\mathbb{P}(x_n \neq \theta) = O(n^{\frac{-1}{K+1}})$ where there exist C and K that satisfy:

$$\min\{\mathbb{G}_1(\frac{1-\alpha}{2}), 1-\mathbb{G}_0(\frac{1+\alpha}{2})\} \geq C(1-\alpha)^K$$

With our private belief distributions, $C = 1$ and $K = 1$, so $\mathbb{P}(x_n \neq \theta) = O(n^{\frac{-1}{2}})$.

D Online Appendix: Example 1 and an Alternative Proof for Theorem 2

D.1 Learning without Expanding Nested Samples

In the main article, I establish that expanding nested samples is a sufficient condition for learning with a nonstationary information structure. Here I show that it is not a necessary condition with a counterexample.

Example 1 (Expanding nested samples is not necessary for complete Bayesian learning within S)**.** *Consider again a set S' containing agents with indices in $\{10^{(m-1)} : m \in \mathbb{N}\}$ in which all agents bar agent 1 observe their immediate predecessor in S', and take the parameter values of the example represented by the upper curve (orange) in Figure 5a (which are $\beta = 0$, $s = 0$, $R = 0.9$, $(f_0, f_1) = (2(1-\varsigma), 2\varsigma)$). The 8th agent in S' is the first to have a lower α_n than their immediate predecessor, and every fifth agent in S' after them (i.e. the 8th, 13th, 18th, 23rd and so on agents within S') is in a similar position, let us refer to this subset of S' as S''. Suppose also that agents in the set $\mathbb{N} \setminus S'$ have neighborhoods $B(n) = S'' \cap \{1, ..., n-1\}$. Expanding nested samples does not hold for $N \setminus S'$, but there is complete Bayesian learning within this set.*

Proof. A covariance stationary process $\{Y_t\}_{t=1}^{\infty}$ with mean μ is defined by the following three properties (Hamilton, 2020):

$$\mathbb{E}(Y_t) = \mu \qquad \forall t$$

$$\mathbb{E}(Y_t - \mu)(Y_{t-j} - \mu) = \gamma_j \qquad \forall t$$

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

Let us use the notation that the m^{th} agent within S'' has index $n(m)$. Thanks to the symmetry of this example, the fact that there is no prior-shifting, and the fact that agents within S'' never reject their signals, we have that agent $n(m)$ will choose $x_{n(m)} = 1$ with probability $\alpha_{n(m)} > 0.5$ if $\theta = 1$, and $1 - \alpha_{n(m)} < 0.5$ if $\theta = 0$. Although we do not have that $\alpha_{n(m)} = \alpha_{n(m+j)}$ for all j, we will have that for some small $\epsilon > 0$ $|\alpha_{n(m)} - \alpha_{n(m+j)}| < \epsilon$.

Let us define the process $\{Y_t\}_{t=1}^{\infty}$ where $Y_t = x_{n(t)}$, and suppose without loss of generality that $\theta = 1$; it follows that $\mathbb{E}(Y_t) = \alpha_{n(t)}$. The covariances will also vary with t, but crucially form an absolutely convergent sequence, as I establish next.

Let the indicator variable Z_t reflect whether or not the possibly-rejecting agent following $n(t)$ actually did reject their social signal. This happens with 50% probability in both states of the world, and is completely independent of Y_t. We can then use the conditional covariance formula to establish that the covariances are absolutely convergent:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y_tY_{t+j}]] &= \mathbb{P}(Z_t=0)\mathbb{E}[Y_tY_{t+1}|Z_t=0] + \mathbb{P}(Z_t=1)\mathbb{E}[Y_tY_{t+1}|Z_t=1] \\ &= \frac{1}{2}\mathbb{E}(Y_tY_{t+j}|Z_t=0) + \frac{1}{2}\mathbb{E}[Y_t]\mathbb{E}[Y_{t+1}|Z_t=1]\end{aligned}$$

In this second line, the final conditional expectation can be split into a product conditional expectation since the Y variables are independent conditional on $Z_t = 1$. Y_t is then independent of Z_t as already mentioned, so the conditional expectation can be replaced with an unconditional one.

$$\begin{aligned}cov(Y_t, Y_{t+1}) &= \frac{1}{2}\mathbb{E}(Y_tY_{t+j}|Z_t=0) + \frac{1}{2}\mathbb{E}[Y_t]\mathbb{E}[Y_{t+1}|Z_t=1] \\ &\quad - \mathbb{E}[Y_t]\big(\frac{1}{2}\mathbb{E}[Y_{t+1}|Z_t=1] + \frac{1}{2}\mathbb{E}[Y_{t+1}|Z_t=0]\big) \\ &= \frac{1}{2}\mathbb{E}(Y_tY_{t+j}|Z_t=0) - \frac{1}{2}\mathbb{E}[Y_t]\mathbb{E}[Y_{t+1}|Z_t=0]\end{aligned}$$

The covariance between these two terms is necessarily positive, and all of the expectations in this expression are positive and strictly less than 1, so it follows that $|cov(Y_t, Y_{t+1})| < \frac{1}{2}$. Similarly, we can consider the covariance between Y_t and Y_{t+2}, and find that:

$$\begin{aligned}cov(Y_t, Y_{t+2}) &= \frac{1}{4}\mathbb{E}[Y_tY_{t+2}|Z_t=0, Z_{t+1}=0] + \frac{3}{4}\mathbb{E}[Y_t]\mathbb{E}[Y_{t+2}|Z_t=1 \text{ or } Z_{t+1}=1] \\ &\quad - \mathbb{E}[Y_t]\big(\frac{3}{4}\mathbb{E}[Y_{t+2}|Z_t=1 \text{ or } Z_{t+1}=1] + \frac{1}{4}\mathbb{E}[Y_{t+2}|Z_t=0, Z_{t+1}=0]\big) \\ &= \frac{1}{4}\mathbb{E}[Y_tY_{t+2}|Z_t=0, Z_{t+1}=0] - \frac{1}{4}\mathbb{E}[Y_t]\mathbb{E}[Y_{t+2}|Z_t=0, Z_{t+1}=0]\end{aligned}$$

This in turn gives that $|cov(Y_t, Y_{t+1})| < \frac{1}{4}$, and similar reasoning will establish that $|cov(Y_t, Y_{t+j})| < \frac{1}{2^j}$ for any j. Thus the covariances are absolutely convergent.

Using this, we can establish, according to the standard line of argument (as in Hamilton (2020, Chapter 7.2, pg. 186)) that:

$$\mathbb{E}(\overline{Y}_T - \overline{\alpha}_T)^2 < (\frac{1}{T})\big\{1 + 2(T-1)\frac{1}{2} + 2(\frac{1}{2^2})(T-2)/T + ... + [1/T]2\big(\frac{1}{2^{T-1}}\big)\big\}$$

which converges to 0 as the sample size grows. If α_T were a fixed value for all T, this would establish mean square convergence to it, but here it simply establishes that $\overline{Y}_T$ will enter the region $[\liminf_{m\in\mathbb{N}} \alpha_n(m), \limsup_{m\in\mathbb{N}} \alpha_n(m)]$. In probability, this average will eventually enter this region if the state of the world is $\theta = 1$, and it will enter (and remain in) an analogous region below 0.5 if $\theta = 0$. Thus for the agents in $N \setminus S''$ $\lim_{n\in N\setminus S''} \alpha_n = 1$, and we have learning. □

The intuition behind this example is that the fact of signal rejection within set S' ensures that to observe S'' is to observe an almost covariance stationary process (it does not quite fit the definition of covariance stationarity, but is close enough to allow us to analyse it in a very similar fashion, as can be seen in the proof). Given this, the time average of the actions of agents in S'' will converge in mean-square to one quantity, strictly greater than 0.5, if $\theta = 1$, and another, strictly less than 0.5, if $\theta = 0$. Those observing these agents thus learn the true state simply by observing this time average, despite never observing the neighborhood of any agent they observe, and we have learning on the basis of a *Large Sample Principle* (cf. Golub & Sadler 2017, who split learning results into improvement and large sample principle results).

The large sample result we have here is arguably more true to the moniker 'large sample' than other results that take this title. Acemoglu et al. (2011, Theorem 4) and Lomys (2020, Theorem 2) both provide large sample results that establish learning when the network contains infinitely many sacrificial lambs who observe small enough neighborhoods that their action always reflects their private signal. Both, however, depend on a 'core' of agents who are either sacrificial lambs with some small and vanishing probability, or observe all preceding agents within this core. This allows the use of martingale convergence arguments, and intuitively acts as 'storage' for all this information. Expanding observations with respect to this set of agents then gives learning for all agents. In Example 1 such a group is not needed: the large samples gives learning alone, without any need for a core facilitating martingale arguments. In this sense, and very specific set of circumstances, the presence of motivated reasoners helps learning. If the private beliefs are bounded but not severely so, then the above network topology would not achieve learning with Bayesian agents, the line network agents in S' would simply copy each other after a certain number of initial actions, and almost all the neighbors observed by agents within $\mathbb{N} \setminus S'$ would be completely uninformative.

E Online Appendix: Learning with Stationary Signal Structures

As is established by part 3 of theorem 1, expanding nested samples is clearly not a sufficient condition for learning when the signal structure is stationary: the complete network exhibits expanding sample sizes, and learning does not occur here with stationary beliefs. However, it does not follow that one cannot achieve complete Bayesian learning with stationary beliefs; in Acemoglu et al. (2011), one of their most surprising results (since their article was written as an extension to Smith and Sorensen (2000), in which bounded beliefs preclude learning) is that there are some network topologies that achieve complete learning even with bounded beliefs. The rough intuition for this is that whereas all agents eventually ignore their signals in the complete network, more generally it is possible to concoct network topologies in which there are infinitely many 'sacrificial lambs' whose neighborhoods will certainly produce a Bayesian social belief weak enough that the agent will choose $x_n = 0$ for some non-null private signals, and $x_n = 1$ for others. A similar trick will allow us to establish the possibility of complete Bayesian learning for stationary beliefs here.

The specific class of network topologies used in Acemoglu et al. (2011, Theorem 4) crucially involves a subset of agents S, whose elements each observe the entire history of the network with some probability bounded away from zero, but of whom infinitely many also act partially on the basis of their private signal (they have a 'non-persuasive neighborhood'- defined below). This set crucially allows both the use of martingale convergence (since the notion of 'the' social belief is well-defined as the belief of each agent in the event they see the entire history) *and* guarantees that each agent relies on their private signal with non-zero probability. A martingale convergence argument gives learning for this subset of agents, and all other agents that are not contained within S are then assumed to have expanding observations with respect to S, and an improvement principle argument gives learning overall.

Definition 6 (Non-Persuasive neighborhoods)**.** *A finite set $B \subset \mathbb{N}$ is a non-persuasive neighborhood in equilibrium $\varsigma \in \Sigma$ if*

$$\mathbb{P}_\varsigma(\theta = 1 | x_k = y_k \text{ for all } k \in B) \in (\underline{B}, \overline{B})$$

for any set of values $y_k \in \{0, 1\}$ for each k. The set of all non-persuasive neighborhoods is $\mathcal{U}_\varsigma$.

The same reasoning clearly cannot apply exactly here, as we have already noted that improvement principles break down in this motivated reasoning setting, as the action-choice to be improved upon is a latent variable, whose correlation with the observed action of an agent does not converge to 1 as $\alpha_n \to 1$. Thus for one set of agents to have expanding observations with respect to a distinct set of agents that learn asymptotically will no longer cause this first set to learn.

Theorem 4. *Let $(\mathbb{F}_0, \mathbb{F}_1)$ be an arbitrary signal structure and let $S \subseteq \mathbb{N}$. Assume the network topology has a lower bound on the probability of observing the entire history of actions along S i.e. there exists some $\underline{\epsilon} > 0$ such that*

$$\mathbb{Q}_n(B(n) = \{1, ..., n-1\}) \geq \underline{\epsilon} \qquad \text{for all } n \in S$$

Assume further that for some positive integer M and non-persuasive neighborhoods $C_1, ..., C_M$ i.e. $C_i \in \mathcal{U}_\varsigma$ for all $i = 1, ..., M$, we have

$$\sum_{n \in S} \sum_{i=1}^{M} \mathbb{Q}_n(B(n) = C_i) = \infty$$

Then complete Bayesian learning occurs in equilibrium ς if the network topology $\{\mathbb{Q}_n\}_{n \in \mathbb{N}}$ has expanding nested neighborhood samples with respect to S.

Proof. Learning within the set S holds on exactly the same basis as in Acemoglu et al. (2011), and learning outside follows from the same argument used to establish the sufficiency of expanding nested neighborhood samples for learning with nonstationary beliefs in Theorem 3. □

Thus learning is certainly still possible with stationary signals, but requires much stronger assumptions on the exact structure of the network. Expanding nested neighborhood samples is already quite an extreme condition (even the weaker expanding sample sizes is quite dramatic), but no longer suffices. As per Acemoglu et al. (2011, Propositions 3 & 4), one can contrive specific network topologies in which the first K agents are necessarily non-persuasive (these constructions of course still produce non-persuasive neighborhoods here), though it seems implausible that a real-life social network would exhibit such a specific structure.

E.1 Alternative Proof for Theorem 1

I have proved Theorem 1 with a proof by contradiction in Appendix B. Here I present a difference proof that assumes the signal structure is nonstationary (and thus does not fully prove the statement) but is more constructive, and thus valuable in providing further intuition.

Proof. Suppose at first that $\theta = 1$. For any collection of M agents, with probability $(\frac{1}{2}(1-\beta))^M$ they are all of non-congenial type (i.e. their type and the state their social beliefs support are opposite). For each of these agents with a social belief stronger than R in favour of either state, they will reject this social belief. Conditional on rejection, there is then probability $\mathbb{G}_1(\frac{1}{2})$ they will take action $x = 0$. There is probability $1 - \mathbb{G}_1(\frac{1}{2})$ they will take action one.

If they do not reject their social belief it is because it was not strong enough (since they are of non-congenial type). Hence a private belief stronger than the social belief corresponding to a Bayesian social belief of R in favour of the state not supported by their social belief will lead them to act according to their private belief. There is a probability of at least $\min\{\mathbb{G}_1(1-\frac{(1-s)R}{(1+s)-2sR}), 1-\mathbb{G}_1(\frac{(1+s)(1-R)}{s-2sR+1})\}$ that this will occur, where these expressions are computed using Lemma 1.

Hence for each individual agent, the probability of them taking action $x=1$ conditional on their being of non-congenial type is at least $\min\{1-\mathbb{G}_1(\frac{(1+s)(1-R)}{s-2sR+1}), 1-\mathbb{G}_1(\frac{1}{2})\}$ (either they are in the rejection region, or they are not and the first probability is strictly positive since the signal structure is nonstationary), and the analogous probability for $x=0$ is $\min\{\mathbb{G}_1(1-\frac{(1-s)R}{(1+s)-2sR}), \mathbb{G}_1(\frac{1}{2})\}$. Let us call these probabilities $\underline{p}(1,1)$ and $\underline{p}(1,0)$. Note that both probabilities are strictly greater than 0 and strictly less than 1 at any. Note also that these events concern only private signals, and are thus conditionally independent across agents. The types of agents are independent as well (not even just conditionally independent).

Any social signal is a sequence of 0s and 1s, and so for each action observed in a given neighborhood we can note that the realisation in question was set to occur with at least the relevant minimum probability derived above. Since the private-signal events discussed are conditionally independent, they probability of observing social signal SS (whatever it is) conditional on all neighbors being of non-congenial type is simply the product of these minimum probabilities. Let us call this $\underline{\mathbb{P}}_1(SS|NC)$. A lower bound on the probability of observing this (not conditional on their all being non-congenial) is therefore $0 < (\frac{1}{2}(1-\beta))^M \times \underline{\mathbb{P}}_1(SS|NC) < 1$. Since the probabilities of observing all possible social signals must sum to 1, and each social signal occurs with probability bounded away from zero, this implies that we can also bound the probability of observing any social signal away from 1.

Note that we can go through exactly the same line of reasoning supposing $\theta = 0$. Hence for every possible social signal, the probability of observing it in either state is bounded away from 1 and 0. Hence, the likelihood ratio of any signal is bounded away from 0 and ∞, and the Bayesian beliefs resulting from them are bounded away from 0 and 1. This in turn implies that their accuracy must also be bounded away from one.

Without expanding sample sizes, as per the hypothesis of this theorem, there is a sequence of infinitely many agents with probability of observing fewer than M agents for some $M \in \mathbb{N}$ that does not converge to 0. Thus there accuracy is at most this probability multiplied by the bound implied by the reasoning above for M and one minus this probability multiplied by 1. This is strictly less than 1. Hence without expanding sample sizes complete Bayesian learning does not obtain.

□

F Online Appendix: Extensions

F.1 Applying Prior-Shifting to Rejected Signals

In the main model of this paper, I assume that agents who reject social signals adopt the social belief $\frac{1}{2}$ in its place. However, there are other assumptions one could make. First of all, the assumption that directly follows Little (2021) is that type 0 agents use $\frac{1}{2}(1-s)$ and type 1 agents $\frac{1}{2}(1+s)$. It might seem extreme that agents who reject social signals in one direction replace them with a belief in favor of the opposite state, but some subjects in Oprea and Yuksel (2022, Result 4) do in fact do this. Alternatively, agents could reject signals but replace them with less extreme beliefs in the same direction. For example, type 0 agents could replace rejected beliefs with $\frac{1}{2}(1+s)$ and type 1 agents with $\frac{1}{2}(1-s)$. The main change this brings about is to ensure that the orange and red regions of Figure 2 meeting no longer guarantees an information structure is nonstationary. If the signals are bounded, and s extreme enough, it can be that (in the first case, where type 0 adopt $\frac{1}{2}(1-s)$) all type 0 agents that reject necessarily choose $x_n = 0$. This would ensure that, on the complete network for example, agents polarise exactly by type and we have tribalism. We would have neither learning nor consensus when $R < b_0$ and $1 - R > 1 - b_1$. Hence the Theorem 1 result on consensus collapse still holds, but tribalism is more common, since motivated reasoners who reject extreme information then choose their own type. With s not sufficiently extreme, as for example is true for all s with unbounded signals, the red and orange regions overlapping is once again sufficient for nonstationarity. On the complete network, therefore, we would have learning in such a setting, but with a larger and larger proportion of non-congenial types choosing their own type as we increase s (subject to the constraint that this not be increased so much as render the information structure stationary). Hence with unbounded beliefs, increasing s to 1 can lead to complete tribalism in the complete network, where all motivated agents choose their own type, and Bayesian agents the correct state.

Theorem 2 holds as long as we assume that all social beliefs $sb \in [0, 1]$ are mapped to some interval $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$. This guarantees that agents rejecting beliefs can only ever be so informative, and the proof of Theorem 2 goes through. Theorem 3 still holds as well, though as above requiring conditions to be nonstationary becomes more demanding.

In the model where type 0 agents replace rejected beliefs with $\frac{1}{2}(1+s)$, we have the perverse result on the complete network that (for strong enough s) all type 0 agents will choose $x_n = 1$ and all type 1 agents $x_n = 0$.

Corollary 2.1 still holds as well, as long again as all social beliefs $sb \in [0, 1]$ are mapped to some interval $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$, and in fact even agents only perform rejection (on beliefs in the rejection region) with some probability $\delta > 0$. This adds further explanation to my suggestion that not much motivated reasoning is needed in order to break correct consensus that obtains in some sparse Bayesian networks with sufficiently informative private signals. Given the prevalence of conspiracy theories, the possibility that some will find a reason to ignore information in the social history does not seem too strong an assumption for political learning.

F.2 Combined Motivated Reasoning

This paper assumes, as discussed in the Model section, that motivated reasoning occurs only with the social belief. This modelling decision reflects the experimental evidence that whilst people do manage to process private signals rationally, they can be widely irrational in their treatment of social signals. Oprea & Yuksel 2022 in particular show this in the context of motivated reasoning, finding patterns of behavior that can be explained by the model of motivated reasoning I used.

However, both of these experiments give agents access to precise private signals that are clearly mathematically defined objects. Perhaps agents process these signals rationally, but extend motivated reasoning to all their information when private signals are less well-defined. Perhaps private signals simple represent an agent's judgement on a question, or the fruit of their own investigations on the internet. If so, a model of motivated reasoning that applies the motivated process to the overall signal might be interesting to study. I refer to this as 'Combined Motivated Reasoning', and outline here the ways in which this model diverges from or resembles the main model of this paper.

The necessity of expanding sample sizes is easier to show with combined motivated reasoning; since it is the overall belief that now may be rejected, we can observe that for every Bayesian social belief $\lambda \in (0, 1)$ either there is some non-zero probability that the agent will form a combined belief within $[0, 1 - R) \cup (R, 1]$ or we can bound their overall belief away from 0 and 1. Constructing a contradiction argument as in my proof of Theorem 2, for any neighborhood of size m, the probability that all neighbors are of non-congenial type is $(\frac{1}{2})^m$, and multiplying this by the m strictly positive probabilities of rejection-region beliefs gives a non-zero probability that all neighbors have rejected their beliefs (taking agents observed sufficiently recently, and thus with positive probability of forming rejection region beliefs). Agents who reject their beliefs (with combined motivated reasoning) simply choose their type as their action, and no information at all about θ is communicated by this fact.

Combined motivated reasoning resurrects the possibility of confounded learning in the complete network, since as the Bayesian social belief moves to one extreme, the probability that non-congenial types reject their beliefs increases (in which case they choose $x_n = \tau_n$). Thus, expanding nested samples is no longer a sufficient condition for learning.

In this setting the results of Theorem 1 all hold. Complete Bayesian learning implies the combined belief is within a rejection region (that corresponding to the true state), and thus that non-congenial types all choose the incorrect state of the world. Part two holds on the basis of exactly the same reasoning as proves it in the main paper. For part 3, with (1) $\overline{B} > \frac{1}{2}$ sufficiently low, (2) $\underline{B} < \frac{1}{2}$ sufficiently high, and (3) very high $R < 1$; the social belief can converge to a stationary point, where even the strongest private signal does not produce a combined belief in a rejection region.

With stationary beliefs, there will be some region of Bayesian social beliefs around $\frac{1}{2}$ such that, even combined with the most extreme possible private signals, they cannot lead to a rejection region combined relief, thus consensus will be achieved in a complete network. Hence, though this paper is primarily concerned with social motivated reasoning, which seems to best reflect the experimental evidence we have on how people behave, the key message that polarization becomes inevitable in a world of increasing informational access and ideological disagreement still holds.

F.3 Reject mild signals, not extreme ones

An alternative approach to motivated reasoning is suggested by Epley and Gilovich (2016), and their view that: "*When considering propositions they would prefer to be true, people tend to ask themselves something like "Can I believe this?" This evidentiary standard is rather easy to meet; after all, **some** evidence can usually be found even for highly dubious propositions... In contrast, when considering propositions they would prefer not to be true, people tend to ask themselves something like "**Must** I believe this?"*" Given this, one might be interested in studying motivated reasoners who instead reject information in the set $(0.5, R]$ if they are type 0, or $[1-R, 0.5)$ if they are type 1, rather than $(R, 1]$ and $[0, 1-R)$ as in my specification.

A first point is to note that if we adopt a probabilistic interpretation of this, the results of this article still all hold. By 'probabilistic interpretation', I mean a model in which agents only ever reject social information with some probability decreasing in the strength of the evidence. For example, if type 0 agents rejected evidence as in figure 6, rejected Bayesian social beliefs above 0.5 with probability $\frac{1}{2}(1-SB)^2 + \epsilon$ for some $\epsilon > 0$.

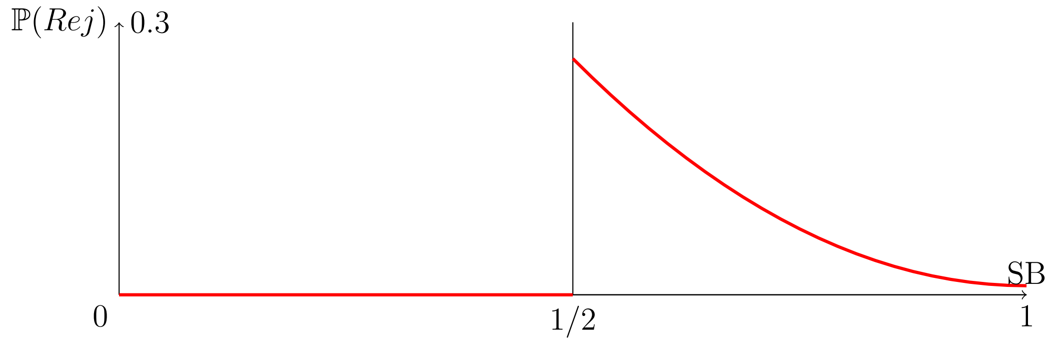


Figure 6: $(1-x)^2 + 0.01$ rejection probability beyond $\frac{1}{2}$

Such a rejection procedure preserves the nonstationary property that underlies many of my results, thus preserving the results themselves. In fact, it preserves this property for any information structure (the distinction I draw in the main article between stationary and nonstationary becomes moot- all signal structures are nonstationary). The major difference with this model, compared to the model I study, is that the magnitude of asymptotic dissensus with expanding sample sizes will be much smaller (since as Bayesian social beliefs converge to 1, the fraction of agents rejecting their Bayesian social beliefs converges to $\frac{1}{2}\epsilon$.

However, if we do not take this specification, and study a model in which agents reject sufficiently weak Bayesian social beliefs with probability 1 and otherwise with probability zero, this is no longer the case. If we first consider the line network, and suppose that agents do not engage in prior-shifting, we can see that belief rejection is no longer sufficient to make expanding sample sizes a necessary condition.

- Agent 1 starts with 0.5, and clearly does not reject.
- Agent 2 rejects with probability 0.5.
- Therefore the probability with which his action matches the state is $H(0.5\alpha_1 + 0.5\alpha_{Rej})$.
- Similarly, all agents $n \in \{3, ...\}$ have accuracy $H(0.5\alpha_{n-1} + 0.5\alpha_{Rej})$.

From this it follows that if $H(0.5 + 0.5\alpha_{Rej}) < \alpha^*$ (recall that α^* is that success probability that implies the agent has observed a social signal outside the region $[1-R, R]$), agents will never achieve a Bayesian social belief high enough to exit the rejection region and we will not have learning. On the other hand, if $H(\alpha_{Rej}) > \alpha^*$, then the agents leave the rejection region immediately and complete Bayesian learning (and in this model, consensus) occurs. Thus with such a specification it is still the case that expanding observations is not a sufficient condition for learning, but also that expanding sample sizes in no longer necessary.