JOB MARKET PAPER
Too Much Information & The Death of Political Consensus
Latest Version

J.W.E Cremin*

September 19th 2023

Abstract

Why is modern society so polarized, even on purely factual questions, despite greater access to information than ever? In a model of sequential social learning with a binary state, I study the impact of motivated reasoning, which is a belief formation process in which agents trade-off accuracy against ideological/emotional convenience, on information aggregation. I find that even Bayesian agents only learn under extremely connected network structures (those that satisfy '*Expanding Sample Sizes*'), where agents have arbitrarily large neighborhoods asymptotically. This is driven by the fact that motivated agents sometimes reject their social information, and with any finite neighborhood, there is always some probability that all neighbors of a given agent observes will have rejected their social information. Moreover, I establish that in such a model *consensus*, where asymptotically agents of all types choose the same action, is only possible with relatively uninformative private signals and low levels of motivated reasoning, whatever the network structure.

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, Miano, and Stantcheva (2020) now seems a fixed feature of American politics; even before the 2020 election, 62% of Trump supporters endorsed the claim that millions

*Department of Economics, Columbia University. jwc2166@columbia.edu.

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] seems 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*. 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: maybe my IQ is higher than the evidence reasonably suggests, the need to save less urgent than my financial advisor insists, or the science more equivocal than the climatologist is telling me? Experimentally, Oprea and Yuksel (2022) find evidence of motivated reasoning when subjects form beliefs concerning their own IQ, and Guilbeault, Becker, and Centola (2018) show that increasing the 'salience of partisanship' can significantly damage social learning by provoking motivated reasoning. More generally, Westen, Blagov, Harenski, Kilts, and Hamann (2006); Moore, Hong, and Cram (2021) use fMRI scanners to show that emotional, not analytical, reactions are triggered by 'threatening' political information.[2] As people become more polarised 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 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 stylized facts form a perfect storm for fact polarization with motivated reasoners.

[1]The literature on partisan perceptions of economic performance provides concrete evidence of persistent polarization (Campbell, Converse, Miller, and Stokes, 1980; Gerber and Huber, 2009; Bartels, 2002), and Brady, Ferejohn, and Parker (2022) use Gallup time series data to show that this has grown more severe over time, with the gap in perceptions between Republicans and Democrats doubling between 1999 and 2020.

[2]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).

A canonical set-up economists have investigated to study the spread of information throughout society is the sequential social learning model, with agents acting in turn and observing both an independent and identically distributed private signal, and a social signal: some subset of their predecessors. Adopting this approach allows me to study learning with arbitrary social networks, without resorting to a mechanical belief-updating procedure (e.g. DeGroot (1974)). Faced with a binary state, agents each observe these signals, before attempting to match their own binary action to the state. 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. Conventionally, one partitions all private signals into those that are *unbounded* and *bounded*: unbounded signals are those that can leave a Bayesian arbitrarily close to certain of the state. I instead partition signal structures between the *uniformly informative* and the *severely bounded* (cf. Definition 3). *Severely bounded* signals are those for which some social signals are definitive irregardless of type: the agent will follow their social belief whatever private signal they receive. All unbounded and some bounded information structures are uniformly informative: motivated reasoning brings about the equivalence of unbounded and bounded yet very informative signal structures. To model motivated reasoning, I assume agents process private signals correctly, but treat social signals with a biased prior, and sometimes reject them entirely if they support their disfavored state too strongly.[3]

Motivated reasoning seriously impedes both information aggregation (reflected here by *learning*) and *consensus*, as the results of this model demonstrate. Firstly, Theorem 1 establishes that in order to achieve asymptotic Bayesian learning, where at least Bayesian (non-behavioral) agents manage to learn to true state, we need a much stronger condition than in standard Bayesian model. Whereas in that setting agents, with access to unbounded signals, being indirectly connected to an ever-larger sets of agents (*Expanding Observations*) is necessary and sufficient for learning (Acemoglu, Dahleh, Lobel, and Ozdaglar, 2011), with motivated reasoners we can instead only achieve learning if they are *directly* connected to such sets (*Expanding Sample Sizes*), albeit with uniformly informative (and thus possibly bounded) signals. In this 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. Furthermore, *consensus*, where all agents asymptotically choose the same action, is impossible once private signals become too informative (Theorem 2), since there are no social beliefs at which the system can settle such that all agents choose the same action. Moreover, when the network topology does produce learning, increasing the extent of motivated reasoning directly reduces the amount of consensus, since as 'prior shifting' becomes more extreme, stronger

[3]This is to reflect experimental evidence, particularly Oprea and Yuksel (2022). I discuss this evidence further in Section 3.1

and stronger private signals are necessary to overwhelm an agent's bias. In the limit, this produces complete tribalism, in which motivated agents always side with their ideology irrespective of the information available to them. The absence of even complete-network[4] consensus with unbounded beliefs, despite common knowledge of the true model, reflects the comparatively extreme nature of this bias, 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 (Bohren and Hauser, 2021, Theorem 6). Finally, Propositions 3 and 4 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 observes, Proposition 3 establishes that with uniformly informative beliefs, any network topology in which agents observe an ever-growing number of nested neigbors will achieve complete Bayesian learning, and the pattern of tribal polarization this implies. Proposition 4 then establishes that even if agents do not asymptotically observe infinitely many nested neighbors, the probability an agent rejects their social signal can be made arbitrarily close to 1 by supposing that they observe sufficiently many. As I argue in section 5, 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, Dmitriev, and Galperti, 2023), and echo chambers (Levy and Razin, 2019) 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. Beyond this, they also suggest a certain fragility to results on learning with Bayesians: introducing any non-zero fraction of motivated reasoners will prevent complete Bayesian learning on any network that fails to satisfy expanding sample sizes, where expanding observations suffices without them. What's more, we can evaluate the robustness of the Bayesian model by asking what happens when all Bayesian agents are replaced with agents engaging in a very small amount of motivated reasoning. Section 5.1 shows that even in such cases, not only can learning fail, but the probability with which agents match the correct state may not converge at all.

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. My main result on learning is Theorem 1, in Section 5, 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

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

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

uniformly informative beliefs in Proposition 3. Consensus is discussed in Section 6, with Theorem 2 setting out the implications of motivated reasoning for this. Discussion of extensions is contained in section 7, and section 8 concludes.

2 Literature Review

There is a large literature on sequential social learning, go back to classic articles by Bikhchandani, Hirshleifer, and Welch (1992) and Banerjee (1992). These and much of the early literature are nested 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 was the first to discover the distinction between bounded and unbounded beliefs. Following on from this initial benchmark, there are two clear sub-literatures extending 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' neighbourhoods 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 sub-literature following Acemoglu et al. (2011) has developped, 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 5.

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, Babichenko, Müller, Pourbabaee, and Tamuz, 2023). This fact ensures that the analyst can study 'the' social belief at period n, and study 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 5.1, this possible loss of information is a major problem

[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 (2004b) 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.

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, Tahbaz-Salehi, and Jadbabaie (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 here.

A third line of extension from Smith and Sorensen (2000) is the sub-literature exploring learning with more than two states (Goeree, Palfrey, and Rogers, 2006; Mueller-Frank and Neri, 2021; Kartik, Lee, Liu, and Rappoport, 2022). Both Goeree et al. (2006) and Mueller-Frank and Neri (2021) consider only the complete network, but Kartik et al. (2022) studying a model with countably many states and arbitrary network topologies. Unlike in my setting, they find minimally sufficient conditions on preferences and private signals (with multiple states, their interplay is important) for learning even with expanding observations. In my model, the results are substantially negative, in that learning occurs only in very highly connected networks: expanding observations does not suffice. Since increasing the number of states intuitively adds to the difficulty of the inference problem agents are required to solve, my conjecture is that extending this model to countably many states would yield qualitatively similar results. Nonetheless, I discuss the possibility of this extension in Section 7.

3 Model

Following the standard sequential social learning model, an infinite sequence of agents $n \in \mathbb{N}$ 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 begin with 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 a private signal $s_n \sim \mathbb{F}_\theta$, 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, which is equivalent to insisting that there are no perfectly informative signals, and be informative (their Radon-Nikodym derivative is not almost surely 1). The analysis in this paper shall mostly use the private *belief* distributions $(\mathbb{G}_0, \mathbb{G}_1)$.

In addition to their private signals, each agent observes some ordered subset of the actions of those agents before them, and the indexes of those agents: $\{x_k : k \in B(n)\}$. 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 of course form their social belief.

Agents can be of type 1,0 or B, 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 B agents shall be Bayesian (showing no bias either way). I assume types are independent and identically distributed, and independent of the network topology and state of the world. Let the probability a given agent is Bayesian be β, for notational simplicity I assume there is equal chance of an agent being type 1 or 2: $\frac{1}{2}(1-\beta)$.[8] The specific reasoning procedure I assume for the social belief is based on Little (2021), which provides a tractable and elegant representation of an agent trading-off a desire for accuracy against a desire to belief the state of the world takes particular values.[9]

I represent motivated reasoning with two parameters (R, s). 'R' denotes a threshold for '*signal rejection*', and 's' is a parameter governing '*prior shifting*'. If an agent n has type $\tau_n = 0$ and reasons according to parameters $(R, s) \in (\frac{1}{2}, 1) \times (0, 1)$, they shall reject any social signal generating a belief (that the state is $\theta = 1$) strictly greater than R, and then they shall behave as a Bayesian updating their belief with the prior $s \times \frac{1}{2} + (1-s) \times (0) = \frac{s}{2}$ (which compared to the true prior, $\frac{1}{2}$, is biased towards their type). Conversely, an agent of type $\tau_n = 1$ shall reject social signals generating belief

[8] 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.

[9] 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.

below $1-R$, and interprets signals according to the prior $s \times \frac{1}{2} + (1-s) \times (1) = 1 - \frac{s}{2}$. [10,11]

3.1 Evidence on Motivated Reasoning

There is some experimental evidence suggesting that people are distinctly irrational when asked to process social information, whilst managing to update beliefs correctly in response to private information. For example, Conlon, Mani, Rao, Ridley, and Schilbach (2022) run an experiment in which participants must guess the proportion of balls in a jar of a given colour, and find that whilst agents update their beliefs correctly upon drawing a ball at random and observing a colour, they do not update correctly upon receiving information from another participant. Astonishingly, this holds even when 'receiving information from another participant' means directly observing the other participant drawing a ball, and seeing its colour! Oprea and Yuksel (2022) similarly find that in an experiment concerning the formation of beliefs about one's own IQ (a subject chosen because it will trigger motivated reasoning), agents there too respond correctly to private signals they receive[12], but do not respond correctly to social information.[13] Hence I suppose that agents form their private beliefs in a Bayesian fashion and their social beliefs behaviorally.[14]

Not only do Oprea and Yuksel (2022) find that agents do not update optimally in response to social information, but that some agents adjust in the incorrect direction upon observing a social signal that is 'bad news' (Oprea and Yuksel, 2022, Result 4). This fits with agents replacing rejected social signals with the biased prior. To more fully reflect their results, we could adopt a distribution of s values for different agents. This is discussed in Section 7.

4 Equilibrium Strategies and Outcomes

As established in Proposition 2 of Acemolgu et al. 2011, Bayesian agents with a prior of 1/2 choose between the two actions by comparing their sum to 1: if the sum of private and social beliefs is greater than 1 they choose $x_n = 1$, otherwise they choose $x_n = 0$. Since our behavioral agents

[10] It may seem odd that upon signal rejection the agents then shift the common prior according to prior shifting, and treat that as the signal, rather than just taking the common prior. I do this to follow the specification used by Little, but qualitatively not much changes if instead rejection leads to the common prior. See Appendix E.1 for more discussion on this point.

[11] 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.

[12] In the terminology of their paper, these are called *public signals*, but they correspond to private signals in this context.

[13] These results make motivated reasoning of particular interest in social learning, where the distinction between private and social information is of such importance. Yet more such evidence can be seen in Weizsäcker (2010) and Conlon, Mani, Rao, Ridley, and Schilbach (2021).

[14] Appendix E.1 considers an alternative model in which agents process all information in a motivated fashion, finding substantially similar results.

behave as if both their social and private beliefs were formed in a Bayesian fashion, they do the same, as outlined in Proposition 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, each agent's motivated decision rule can be represented as in the following proposition:

Proposition 1. *Consider an agent of type $\tau_n = 1$, who forms social belief 'MSB'. They will choose $x_n = 1$ if the following condition is satisfied:*

$$MSB + \mathbb{P}_\sigma(\theta = 1|s_n) \geq 1$$

where the motivated social belief is formed according to:

$$MSB = \begin{cases} \frac{(1+s)\mathbb{P}_\sigma(\theta=1|B(n))}{\left(2s\mathbb{P}_\sigma(\theta=1|B(n))+(1-s)\right)} & \text{if } \mathbb{P}_\sigma(\theta = 1|B(n)) \geq (1-R) \\ \frac{1}{2}(1+s) & o.w. \end{cases}$$

Otherwise they will choose $x_n = 0$.

Proof. See Appendix C. □

Having now defined the belief formation process, I note that the equilibrium concept I shall use is simply 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:*

1. *Sequential Rationality: Every agent's strategy is an optimal response to their belief given the strategies of other agents σ_{-n}.*
2. *(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 this definition, equilibrium existence is immediate, as is standard in models of sequential social learning:

Proposition 2. *A Motivated Equilibrium exists.*

Proof. The Motivated Reasoning procedure always uniquely defines a belief for every history and private signal (s_n, h_n). The set of optimal actions is non-empty for every belief (and is not a function

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

of the strategies of other agents except through this belief), thus recursively applying these facts clearly gives equilibrium existence. □

Naturally, 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 (beyond its intrinsic interest as a question) 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*:

Definition 2. *(Complete Bayesian Learning)* ***Complete Bayesian learning*** *obtains if Ξ_n (i.e. x_n if $\tau_n = B$) converges to θ in probability (according to measure $\mathbb{P}_\sigma$), i.e.*

$$\lim_{n\to\infty} \alpha_n := \lim_{n\to\infty} \mathbb{P}_\sigma(\Xi_n = \theta) = 1$$

Consenus *obtains if x_n converges to ω for any $\omega \in \{0, 1\}$ in probability (according to measure $\mathbb{P}_\sigma$), i.e.*

$$\lim_{n\to\infty} \mathbb{P}_\sigma(x_n = \omega) = 1$$

Either of these outcomes occurs within infinite subset $S \subseteq \mathbb{N}$ if the limit probability of agents within S converges to 1.

I shall refer to α_n as the *Bayesian accuracy* of agent n. It will also be convenient, when agent n is a motivated reasoner, to refer to a hypothetical Bayesian agent in their place (observing the same information) as their *Bayesian equivalent*. Whereas the actual agent's action is x_n, the action their Bayesian Equivalent would choose is Ξ_n; of course, if $\tau_n = B$, these two are the same: $x_n = \Xi_n$.

Finally, the standard partition of information structures into *bounded* and *unbounded* is not the most useful one in this setting. In this partition, 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 $\overline{B} < 1$. This is usually an important distinction, and only with unbounded beliefs are there no social beliefs for which, should they hold that belief, an agent's action carries no information at all. Here however, motivated reasoning reduces the pertinence of this partition. To replace it, I use a different partition of information structures between those that are *uniformly informative* and those that are *severely bounded*. Fundamentally, severely bounded information structures shall be those for which there exist some social beliefs that are decisive, i.e. an agent with this social belief will take the corresponding action, whatever their type and private signal. Uniformly informative signals, in contrast, are those for which there are *no* such decisive social beliefs.[16]

The presence of prior shifting first expands the region of social beliefs that can be overturned (the orange region in Figure 1), since type 0 agents will form social beliefs more in favor of $\theta = 0$, and type 1 agents in favor of $\theta = 1$. In Figure 1 the region of over-turnable social beliefs grows from $[1 - b_B, b_B]$

[16] In the standard Bayesian model, this definition of uniformly informative information structures is equivalent to the definition of unbounded signals.

to $[1-b_1, b_0]$. Secondly, signal rejection entails that Bayesian social beliefs in the regions $[0, 1-R)$ and $(R, 1]$ also imply the agent will choose each action with strictly positive probability, as if they are of non-congenial type they will simply reject this belief entirely.

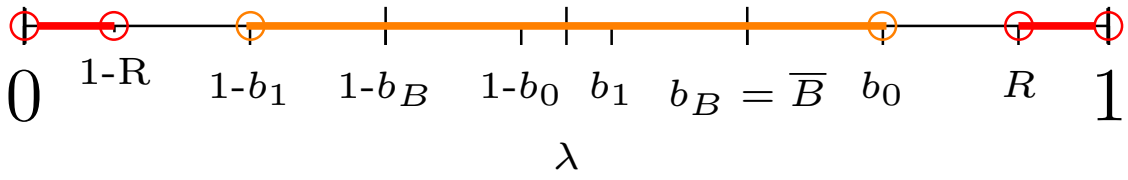


Figure 1: *Uniformly Informative vs Severely Bounded Signal Structures:* Any social belief in the red region may be rejected, and any in the orange region can be overturned. Social beliefs in neither are definitive, and it is the absence of these that defines uniformly informative information structures.

As we increase the maximum signal strength of our private signal $\overline{B}$, the amount of prior shifting s, and/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. Hence a more useful partition divides them into *severely bounded* structures such that, given the parameters (R, s), there are some social beliefs at which an observed agent is certainly ignoring their private signal, and *uniformly informative* structures where this is never the case. Using the notation $\psi(\lambda, \theta)$ to denote the probability an agent n of unknown type chooses $x_n = 1$ if they have social belief λ and the state of the world is θ, we can formally define the two elements of this partition as follows:

Definition 3 (Uniformly Informative & Severely Bounded Signal Structures)**.** *A signal structure* $(\mathbb{F}_0, \mathbb{F}_1)$ *is uniformly informative for a given parameter tuple* (R, s, β)*, if the set* $\{\lambda : \psi(\lambda, 1) = \psi(\lambda, 0)\} \cap (0, 1)$ *is non-empty. Otherwise, it is severely bounded for* (R, s, β)*.*

5 Learning with Expanding Sample Sizes

Under what precise conditions does learning occur with motivated reasoners, if any? The standard martingale techniques of Smith and Sorensen 2000 are sufficient to establish that, with a uniformly informative information structure, we will get learning in the complete network (in fact it gives us a stronger statement, since complete Bayesian learning is defined in terms of convergence in probability and martingale techniques give almost sure convergence). Lemma 3, in appendix A, states this result, which effectively amounts to proving part (b) of their Theorem 1 in this context. Note that, since some bounded signals are nonetheless uniformly informative, this demonstrates that the presence of motivated reasoners can actually help learning. This is a consequence of motivated reasoners performing the function of *sacrifical lambs*.[17]

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 1) might be enough.

[17]This terminology is from Golub & Sadler 2017, who review the literature on learning in social networks.

Assumption 1 (Expanding Observations)**.** *A network topology $\mathbb{Q}$ satisfies expanding observations if for all $K \in \mathbb{N}$,*

$$\lim_{n\to\infty} \mathbb{Q}_n\left(\max_{b\in B(n)} b < K\right) = 0$$

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 privates signals, but is no longer sufficient in this setting. This insufficiency follows from Theorem 1, which establishes that a much stronger conditions '*Expanding Sample Sizes*' is necessary for complete Bayesian learning to obtain.

Assumption 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.[18] *If for some subset $S \subseteq \mathbb{N}$ we have:*

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

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.

Theorem 1. *With any signal structure $(\mathbb{F}_0, \mathbb{F}_1)$, complete Bayesian learning obtains only if the network topology satisfies expanding sample sizes.*

Proof. See Appendix C. □

I offer a proof by contradiction of this result, but the rough intuition behind it is based on the fact that any neighbor is either acting on the basis of a weak social belief, and is thus not very informative to observe, or a very strong social belief, and will be rejecting this belief if of non-congenial type. For any finite neighbourhood, the probability that all neighbors happen to be of non-congenial type is bounded away from zero. The ex-ante probability with which an agent's Bayesian Equivalent matches the true state thus partly reflects the possibility that they will be acting on the basis of an entire neighbourhood of agents rejecting their social signals. They may not be acting on the basis of information accumulated over the course of the history of the game, but only the private signals of neighbours.

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

[18] Note that this condition implies expanding observations, which is still of course a necessary condition.

5.1 A Line Network Example & Information Loss

In addition to illustrating the insufficiency of expanding observations, this example also serves to illustrate the unique problems of learning with a society of motivated reasoners. Signal rejection, specifically when applied to the social belief, engenders the repeated loss of all accumulated information. Learning results derived using improvement principles are thus inapplicable, as the neighbor upon which a given agent is improving may be acting on the basis of only his own private information.

The line network is a particularly convenient one to study, especially with some specific parameter assumptions. 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. If, for example, we impose that (a) each agent is almost surely a motivated reasoner $\beta = 0$; (b) no agents engage in prior shifting $s = 0$; (c) the private signal distributions are anti-symmetric (Definition 4); then the accuracy of an agent whose predecessor has Bayesian accuracy α must himself have accuracy $H(\alpha)$ as in Equation 5.1.

$$H(\alpha) := \frac{\alpha}{2}\big[\mathbb{G}_0(\alpha) + 1 - \mathbb{G}_1(1-\alpha)\big] + \frac{(1-\alpha)}{2}\big[\mathbb{G}_0(1-\alpha) + 1 - \mathbb{G}_1(\alpha)\big] \tag{5.1}$$

Specifically taking the parameters $R = 0.7$ and $(f_0(s), f_1(s)) = (2(1-s), 2s)$ this H function boils down to a simple quadratic. Thorough readers can find the full working for this example in appendix B, and these are the parameters used in figure 2. 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 $1/2 < R < 1$.

The first panel of figure 2, subfigure 2a, 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 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.[19]

Figure 2b 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 2c gives the

[19]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.

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.

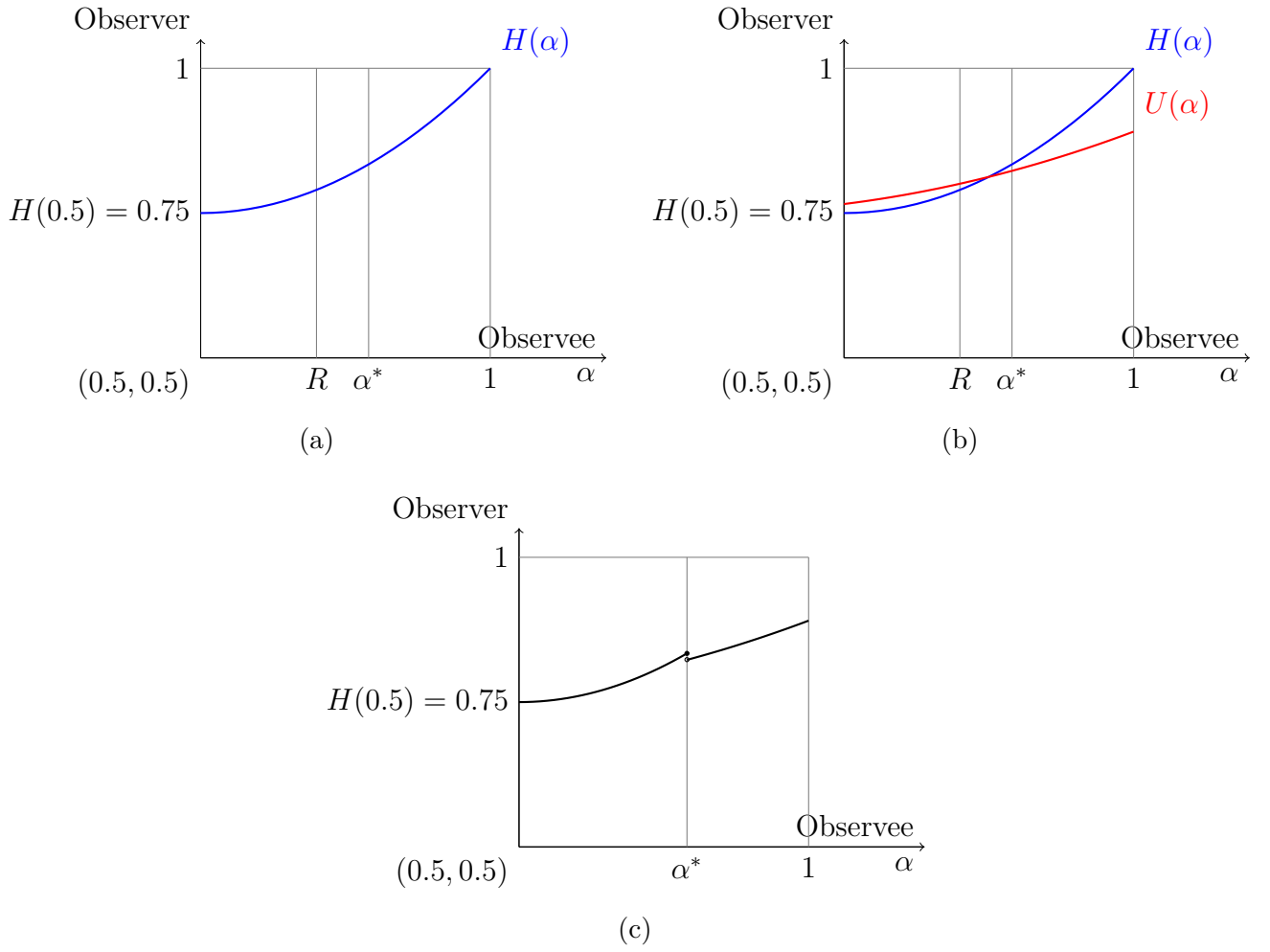


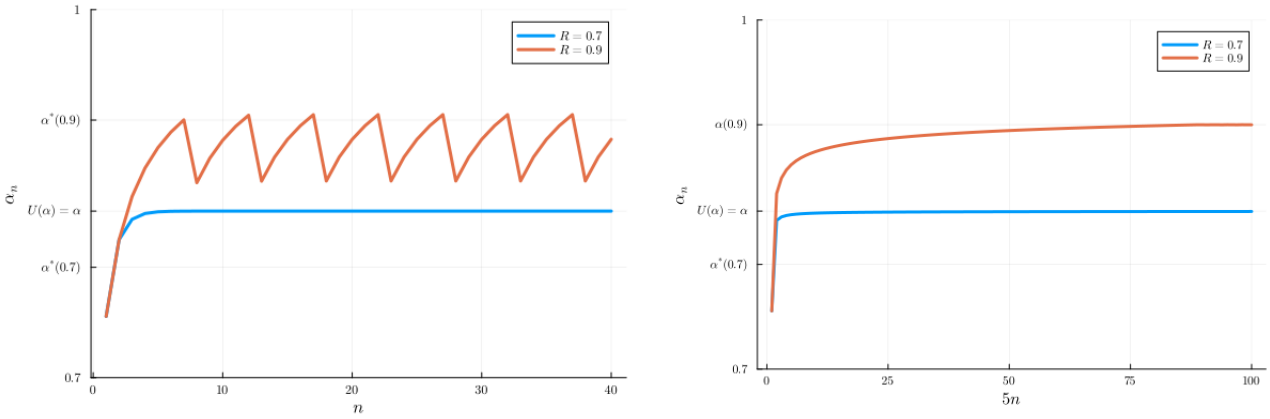
Figure 2

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 3a, 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 3b 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 3: The path of α_n against n, with $R \in \{0.7, 0.9\}$, $s = 0$, $\beta = 0$, $(f_0, f_1) = (2(1-s), 2s)$.

Another intuition explaining the failure of complete Bayesian learning here, and contrasting it to the fully Bayesian case, concerns the periodic 'deletion' of social information in this model. In the standard Bayesian model, the assumptions 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. Once again, even this intuition applies to both the current model with unobservable types and one in which we allowed neighbors to observe neighbor types. In this model, the added uncertainty over whether or not rejections have in fact occurred, and how recently, will exacerbate learning yet further.

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 Ξ_n were converging to 1 as $\alpha_n \to 1$, and the probability $\mathbb{P}(x_n \neq \Xi_n)$ 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 are receiving a rejection-region social belief. Since the probability with which a given agent is of the non-congenial type is $\frac{1}{2}(1-\beta)$, this implies that $\mathbb{P}(x_n \neq \Xi_n) \to \frac{1}{2}(1-\beta)$ as $\alpha_n \to 1$.

This breakdown in information monotonicity is reminiscent of that seen in Lobel & Sadler 2016, where

agents have heterogeneous preferences over the two actions. However, it is even worse in this 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 a sufficiently large amount of homophily in their Proposition 3). In this model, however, knowing the type of the agent observed agent does not fully fix the issue. If you observe the type of your most accurate neighbor, and also observe (through observing their neighborhood) that they are of congenial type, this tells you that x_n is highly likely to be equal to Ξ_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.

5.2 Sufficient Conditions for Learning

Theorem 1 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 condition for learning within S is that it satisfies *expanding nested neighbor samples*:

Assumption 3 (Expanding Nested Neighborhood Samples)**.** *For agent n, let $B^n(n) \subseteq B(n)$ be the set of agents within $m \in B(n)$ such that $B(m) \subset B(n)$. A network topology has expanding nested neighbor samples for S if for all $K \in \mathbb{N}$, we have*

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

A *nested neighbors* of n is an agent n observes, whose entire neighborhood, $B(n)$, n also observes. Expanding nested neighbor samples then requires than agents observe ever greater numbers of such agents. Though a very demanding condition on the network topology, the requirement that agents observe a large number of nested neighbors is not so unreasonable when studying social networks. Modern social networks 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 neigbors.

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

Proof. See Appendix C □

One implication of this is that any network topology that satisfies expanding nested neighbor 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 $B(n) = \{m \in \mathbb{N} : m < n\} \cap \{1, 100\} \cap \{i \times 100 + 50, ..., i \times 100 + 99, \forall i \in \mathbb{N}\}$ then the network topology satisfies expanding nested neighborhood samples. When establishing complete Bayesian learning within proper subsets of $\mathbb{N}$, we can demonstrate that even network topologies that don't 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 neighborhood 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 that take 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 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 sample sizes.

Though expanding nested neighborhood samples for S is a sufficient condition for complete Bayesian learning within S, it is not a necessary condition. To see this, consider the following counter-example, which also illustrates that motivated reasoning can in a sense help with learning by reintroducing a certain amount of independence between agent's actions, roughly speaking.

Example 1 (ENNS is not Necessary for Complete Bayesian learning within S)**.** *Consider again a set $S' = \{10^m : m \in \mathbb{N} \cup \{0\}\}$ in which all agents bar agent 0 observe there immediate predecessor, and take the parameter values of the example represented by the upper curve (orange) in Figure 3a ($\beta = 0$,$s = 0$, $R = 0.9$, $(f_0, f_1) = (2(1-s), 2s)$). 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 is in a similar position $\{8, 13, 18, 23, ...\}$, let us call this subset of S', S''. Suppose also that agents in the set $\mathbb{N} \setminus S'$ have neighborhoods $B(n) = S'' \cap \cap \{1, ..., n-1\}$. Expanding nested sample sizes does not hold for $N \setminus S'$, but there is complete Bayesian learning within this set.*

Proof. See Appendix C □

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 $\theta = 0$. The agents 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).

Interestingly, 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 arguably 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.

These sufficient conditions are more powerful when considered in the context of the extensions considered in section 7, since the assumption of uniformly informative 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 uniformly informative 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 the 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.

None of this, it should also be noted, establishes that severely bounded beliefs preclude learning. Much as with Acemoglu et al. 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 uniformly informative information structures are most relevant to investigating modern polarization, but I include some discussion of severely bounded beliefs in Appendix D for completeness.

6 Consensus with Motivated Reasoning

Having considered in which networks learning will obtain, we can immediately observe a number of implications for the possibility of consensus; I present these as Theorem 2. First and foremost, it can be easily seen that complete Bayesian learning and consensus are incompatible with this model of motivated reasoning (for any network topology), as the former implies extreme social beliefs, which are necessarily

rejected by agents of non-congenial type. This in turn gives that uniformly informative beliefs preclude learning, as either complete Bayesian learning obtains, or 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 severely bounded beliefs there are some network topologies and parameter values that do produce it (the complete network can provide such a case).

Theorem 2.

1. *Complete Bayesian learning implies consensus does not obtain.* $(\forall \mathbb{Q})$
2. *Thus consensus cannot obtain with uniformly informative beliefs.* $(\forall \mathbb{Q})$
3. *A consensus can occur with severely bounded beliefs.* $(\exists \mathbb{Q})$

Proof. Let $\theta = 0$,

1. 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 $\mathbb{P}(\theta = 1|s_n) > \frac{1}{2}(1-s)$.
2. In a model of uniformly informative beliefs therefore, either the Bayesian belief does not converge to certainty, and any social belief can asymptotically be overturned (and we do not have consensus), or it does converge to certainty (and we do not have consensus).
3. This can be seen by taking the example of the complete network (which is particularly convenient, since it allows the almost direct application of SS2000), with bounded beliefs: figure 4 illustrates the set of all possible social beliefs. A complete proof with details is relegated to appendix C, but the intuition is the following. The assumption of bounded beliefs implies that there is a minimum social belief that can possibly be overturned by a private signal for each type $(1-b_1 < 1-b_B < 1-b_0)$, and similarly maximum social beliefs that can be overturned by a weak signal $(b_1 < b_B, b_0)$. The exact levels of these are given by the type-parameters, and implied by proposition 1. Thus if the 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 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 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.

□

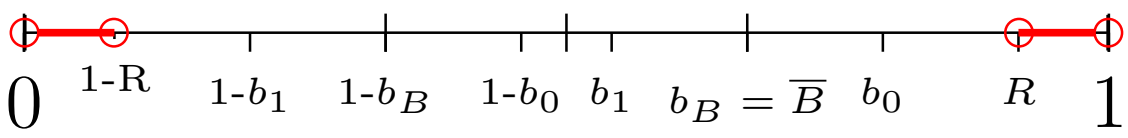


Figure 4

Whereas severely bounded beliefs maintain the possibility of consensus, the level of polarization in networks exhibiting expanding sample sizes can be pinned-down exactly, and is increasing in the level of motivated reasoning. The exact degree of polarization this produces will depend on the level of prior-shifting, what fraction of agents are Bayesian, and the belief distributions. If the true state is $\theta = 1$ agents will choose $x_n = 1$ with asymptotic probability $\beta + \frac{1}{2}(1-\beta)(2 - \mathbb{G}_1(\frac{1}{2}(1-s)))$. If s converges to 1, the proportion of type zero agents choosing $x = 0$, despite the strong Bayesian social belief in favor of $\theta = 1$, also converges to 1, and we have complete polarization. Though the fact of belief rejection is essential in achieving this, the level of polarization is not affected by the value of R in such network topologies, except in so far as this R needs to be small enough to make the signal structure uniformly informative.

If we are willing to suppose that real observation networks are sufficiently large and connected for expanding sample sizes to reasonably model them, a number of interesting points follow. Firstly, in the presence of a fixed pair of parameters (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 severely bounded to a uniformly informative information structure. 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 'severely bounded' 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. Thirdly, if instead we accept that reversing this change is simply a lost cause- all that can be done to *partially* recover consensus is to either reduce the s parameter or increase the fraction of Bayesians β. Since converting a partisan agent into one who cares about nothing but the truth would be intuitively difficult, reducing the extent and relevance of value polarization (since s reflects and is increasing in the magnitude of the 'directional motives' of our agents) is the only realistic target for remedial policy. 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.[20]

This particular finding also presents the most easily testable prediction of this paper. With a complete network of experimental subjects, running treatments with first very weak private signals, and then increasing the strength of signals provided to subjects, should lead to such a jump in the level of observed disagreement. What's more, increasing the polarization parameters by priming subjects to be partisan should make this disagreement more extreme, and lead to a high proportion of subjects siding with the position associated with their party affiliation.

Even if we are not quite willing to suppose that a network topology satisfies expanding nested neighborhood samples, we can nonetheless show that non-congenial type agents will reject their social signals

[20]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 Ξ_n.

with probability $1-\epsilon$ for arbitrary $\epsilon > 0$ if they have neighborhoods with at least M nested neighbors[21] 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 5, high clustering and connectivity of online social networks gives cause to hope that this may hold.

Proposition 4. *For any $\epsilon > 0$, there is an $M \in \mathbb{N}$ such that, if agents within set S asymptotically have at least M nested neighbors and the information structure is uniformly informative, the asymptotic ex-ante probability that any given non-congenial type agent rejects their social signal is at least $1-\epsilon$.*

Proof. See Appendix C. □

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 polarisation. If for a given network topology the asymptotic distribution of 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.

7 Discussion

There are a number of simple extensions to which we can easily extend the results in this paper. The details of the more involved extensions are relegated to appendix E. 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 regions in which non-congenial types will or will not necessarily reject the 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 $[\underline{R}, \overline{R}]$, we must have $\underline{R} < 1$. Specifically, this condition is sufficient to reproduce the proof of Theorem 1, the reasoning within the line network example, and the first two parts of Theorem 2. 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 uniformly informative and severely bounded information structures (Definition 3) still hold in this case, except that the tuple of parameters that determine if an information structure is one or the other are now $(\underline{R}, s, \beta)$ 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 social beliefs, could be defined to continuously increase the rejection probability for non-congenial social

[21] neighbors whose neighborhoods they also observe

beliefs so long as it assigns strictly positive probability to the rejection of some social beliefs. To be precise, we can simply endoww each agent with two parameters $R_{min} > \frac{1}{2}$, $R_{max} < 1$ and a rejection function $\mathcal{R} : [\frac{1}{2}, 1] \to \{0, 1\}$ such that:

$$\mathcal{R}(z) = \begin{cases} 0 & \text{if } z < R_{min} \\ f(z) & \text{if } R_{min} \leq z \leq R_{max} \\ 1 & \text{if } z > R_{max} \end{cases}$$

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 1, the line network example and the first two parts of Theorem 2 carry through so long as agents reject some sufficiently extreme non-degenerate social beliefs with some strictly positive probability. There are no severely bounded information structures without the condition $R_{min} > \frac{1}{2}$, much as $\underline{R} > \frac{1}{2}$ was necessary for this in the previous extension.

An additional alteration that makes only a minimal difference, and in fact even less of one that the two preceding extensions, would be to have agents adopt either the common prior or a less extreme social belief in the event of a social belief rejection. For example, were a type-0 agent to reject a social belief of 0.9, whereas I have assumed the adopt the social belief $\frac{1}{2}(1 + s)$, one could instead assume that they adopt the belief $\frac{1}{2}$, or perhaps some convex combination of this and the rejected belief. In each case, this does not in anyway impact any of the main results. Its only impact is the vary the exact proportions of agents who choose each action in the event of complete Bayesian learning; the extreme polarization achieved when $s = 1$ is no longer achievable if we replace this assumption. The present assumption reflects observed behavior in Oprea & Yuksel 2022, but if it seems an extreme response to a strong non-congenial signal, it is not crucial.

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 its sufficiency 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 E.1, but in short the fundamental obstacles that motivated reasoning poses to learning remain.

In addition to allowing us to study learning in societies with motivated reasoners, this model also allows

us to study the robustness of the traditional Bayesian model to motivated reasoning. When thinking about robustness, we need to consider two different sense in which the Bayesian model might be robust, and two different outcomes (either learning or consensus- identical in the standard model) with which we might be concerned. The two different sense in which the Bayesian model might be robust to motivated reasoning for a given outcome are: (1) the outcome varies little when we consider a model with $\beta = 1$ and $\beta = 1 - \epsilon$ for small ϵ; and (2) the outcome various little between the model with $\beta = 1$ and the model with $\beta = 0$ and $(s, R) = (\epsilon_1, 1-\epsilon_2)$ with small $(\epsilon_1, \epsilon_2) >> 0$. With consensus, theorem 2 makes clear that the Bayesian model is not robust in the first sense, since for any uniformly informative information structure consensus becomes impossible. This is quite a contrast to the Bayesian model, in which expanding observations and unbounded beliefs are enough to ensure it. In the second sense, the line network example demonstrates that we do not have robustness with respect to learning or consensus, since despite satisfying expanding observations, even the $(s = 0, R = 0.9)$ case produces an eternal cycle in α_n. Learning is possible in this model, but requires very connected network topologies, and consensus can only be achieved in those networks that do not provide learning, and have very limited access to information.

Finally, it is worth commenting on the possibility of extending this model to a non-binary framework. As I observed in Section 2, the major results of this paper are negative: we need very connected networks to achieve learning, and consensus is fragile if not impossible. Intuitively, increasing the number of states increases the difficulty of the inference problem agents face, so it seems likely that results in a non-binary model will be qualitatively similar. This view, and the tractability of binary state models, justifies the choice to study only the binary setting. However, were one to want to extend this model to more than 2 states, what would this entail? Firstly, the model of motivated reasoning in this paper is an application of Little (2021) to this setting, and involves agents who assign weights to each state, reflecting to what degree the state in questions is congenial to them. Hence, v_j parametrises to what degree the agent 'likes' state j. Outside the binary framework, one would need to think carefully about this distribution of weights. In this sense, the interplay between preferences and signal structures that Kartik et al. (2022) discuss is complicated further, as now there are two sets of preferences: the explicit preferences of the agent, and the psychological 'preferences' (weights) that underpin motivated reasoning. There will no longer be a simple (s, R) parameter pair to represent motivated reasoning.

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 this 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, that consensus can always be broken by increasing value-polarisation or access to information.

Arguably, however, this shift in network topology is descriptive of the change that the internet has brought about. When paired with the shift from severely bounded to uniformly informative 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 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 to reducing polarization with motivated agents in our ever more connected and informationally-overwhelmed societies.

References

Acemoglu, D., M. A. Dahleh, I. Lobel, and A. Ozdaglar (2009): "Rate of Convergence of Learning in Social Networks," .

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

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

Arieli, I., Y. Babichenko, S. Müller, F. Pourbabaee, and O. Tamuz (2023): "The Hazards and Benefits of Condescension in Social Learning," *arXiv preprint arXiv:2301.11237*.

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

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

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

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

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

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

Blackwell, D. (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. A. (2016): "Informational herding with model misspecification," *Journal of Economic Theory*, 163, 222–247.

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

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

Brady, D. W., J. A. Ferejohn, and B. Parker (2022): "Cognitive Political Economy: A Growing Partisan Divide in Economic Perceptions," *American Politics Research*, 50, 3–16.

Campbell, A., P. E. Converse, W. E. Miller, and D. E. Stokes (1980): *The american voter*, University of Chicago Press, originally published in 1960.

Chamley, C. (2007): *Economics and psychology : a promising new cross-disciplinary field / edited by Bruno S. Frey and Alois Stutzer Rational herds : economic models of social learning / Christophe P. Chamley*, Cambridge, Mass. : Cambridge :: MIT Press Cambridge University Press.

Conlon, J. J., M. Mani, G. Rao, M. W. Ridley, and F. 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. (2023): "Learning through Anonymous Speech," .

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

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

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

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

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

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

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

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

Hamilton, J. D. (2020): *Time series analysis*, Princeton university press.

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

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

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

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

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

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

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

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

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

Mueller-Frank, M. and C. Neri (2021): "A general analysis of boundedly rational learning in social networks," *Theoretical Economics*, 16, 317–357.

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

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

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

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

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

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

Westen, D., P. S. Blagov, K. Harenski, C. Kilts, and S. 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, 1947–1958.

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

Çelen, B. and S. Kariv (2004a): "Distinguishing Informational Cascades from Herd Behavior in the Laboratory," *American Economic Review*, 94, 484–498.

——— (2004b): "Observational learning under imperfect information," *Games and Economic Behavior*, 47, 72–86.

A Useful Lemmas

The first two lemmas here serve only to reproduce important results in ADLO2011 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 SS2000, which is the central theorem describing the fixed points of Markov-Martingale processes.

Lemma 1 (ADLO2011 Proposition 2)**.** *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 ADLO2011 Proposition 2. □

This next lemma is Lemma 1 from ADLO2011, and gives several useful properties of the belief distributions.

Lemma 2 (ADLO2011 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 nonincreasing in r and is strictly larger than 1 for all $r \in (\underline{\beta}, \overline{\beta})$

Proof. See ADLO2011 Lemma 1. □

Definition 4 (Anti-Symmetry)**.**
A pair of distributions $(\mathbb{G}_0, \mathbb{G}_1)$ is said to be anti-symmetric if $\mathbb{G}_0(r) = 1 - \mathbb{G}_1(1-r)$ for all $r \in [0,1]$.

Lemma 3 (SS2000 Theorem 1b)**.** *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 uniformly informative suffices) beliefs, the likelihood ratio, $l_n = \frac{(1-q_n)}{q_n}$, of the 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 SS2000 Theorem 1b. □

Lemma 4 (Social Belief Distribution Relationship)**.** *If the 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 Lemma A1 (a) from ADLO2011- adjusted in necessary ways. By then definition of a 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:

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

(*Note this differs from the analogous expression in ADLO2011 since there are only a finite number of possible social beliefs at any point.)

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

Using the notation that h^n_θ is the pmf for the social beliefs of agent n in state θ:

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

□

Lemma 5. *If we take a set A containing N agents, each of whom has an ex-ante success probability of $\alpha_n > 1 - \frac{\delta}{N}(1 - \mathbb{G}_0(R))$, the ex-ante probability that their 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. Lemma 5

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

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

The more mass we place on lower 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 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 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 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.

Thus the probability that at least one agent does not have a rejection-region signal (of N agents observed all of whom have α_n above this threshold) is less than P, where:

$$P = \sum_{i=1}^{N}(1 - \mathbb{P}(i \text{ in rejection region})) \leq \sum_{i=1}^{N}(1-(1-\rho))$$
$$= \sum_{i=1}^{N}\rho$$
$$= N\rho$$

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

Lemma 6. *If a set of $n \in \{1, ..., N\}$ Bernoulli-Blackwell/ 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'[22] of this experiment are:

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

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:

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

And similarly:

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

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

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 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 7. *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:

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

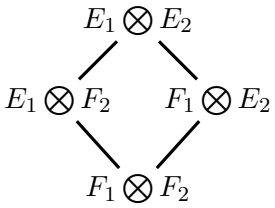
Clearly any experiment with $p^1 \leq \overline{p}^1$ and $p^0 \geq \underline{p}^0$ will have lower $(\kappa, \mathcal{L})$ than this upper bound experiment, and be dominated in the confidence order. As per the equivalent reasoning in the previous lemma, they must therefore give a lower value in the 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 8. *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' 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 Line Network Example Working for Section 5.1

Recall that in this simplified example we assume there is no prior-shifting ($s = 0$), $R = 0.7$, $\beta = 0$, and the private signals have density functions: $(f_0(s), f_1(s)) = (2(1-s), 2s)$. These signal distributions produce identical belief distributions: $f_\theta(\cdot) = g_\theta(\cdot)$.

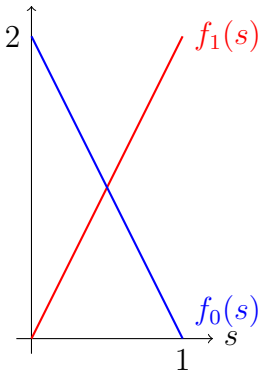


Figure 5: The Density functions for signal (and belief) distributions $f_0(s) = 2(1-s)$ and $f_1(s) = 2s$

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 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 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 4 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:

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

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 4 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 5.1, reproduced here as B.1. Given the particular belief distributions specified above, this simplifies to B.2.

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

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

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 $\frac{1}{2}$. $U(\alpha)$ is thus defined as $H(0.5\alpha_{n-1} + 0.5H(0.5))$.

C Omitted Proofs

Proof of Proposition 1. The fact that summing the private and social beliefs can may added, and compared to 1 in order to choose an action results from the fact that my agents are assumed to combine their private and social beliefs as a Bayesian, as if they had both been formed in a Bayesian fashion. The sum representation is then simply an application of ADLO2011 Proposition 2.

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

The form of MSB when $\mathbb{P}_\sigma(\theta = 1|B(n)) \geq (1 - R)$ 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 $\tilde{\mathbb{P}}$:

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

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

Dividing C.1 by C.2, we get:

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

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

Now we can apply Bayes' rule (using the true prior) twice on the RHS to get:

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

P.S. Following an equivalent line of working, for type 0 agents the equivalent expression is:

$$\tilde{\mathbb{P}} = \frac{(1-s)\mathbb{P}}{(1+s)-2s\mathbb{P}}$$

□

Proof of Theorem 1. 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:

$$\epsilon < \min\{\overbrace{\Big(\frac{1}{2}(1-\beta)\Big)^{M_1-1}(1-\delta)\Big(1-\alpha_{IOR}^{M_1-1}\Big)}^{Reverse\ Engineered}, \overbrace{\frac{\delta}{M_1-1}(1-\mathbb{G}_0(R))}^{For\ Lemma\ 5}\}$$

- This first term is reverse engineered to provide a contradiction, this second allows us to use Lemma 5.
- Consider an agent m such that $m > N_\epsilon + M_2$. WLOG 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 $\left(\frac{1}{2}(1-\beta)\right)^{M_1-1} P_m$.
- Thus we have

$$\alpha_m < \overbrace{\left(1 - \left(\frac{1}{2}(1-\beta)\right)^{M_1-1} P_1\right)}^{\text{Not Rej}} \cdot (1) + \left(\frac{1}{2}(1-\beta)\right)^{M_1-1} P_1 \cdot \alpha_{IOR}^{M_1-1}$$

- Our selection of ϵ then implies

$$1 - \epsilon < \alpha_m < 1 - \left(\frac{1}{2}(1-\beta)\right)^{M_1-1} (1-\delta)\left(1 - \alpha_{IOR}^{M_1-1}\right)$$

- 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 5 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:

$$\epsilon < \min\{\epsilon^*, \left(\frac{1}{2}(1-\beta)\right)^{M_1-1} (1-\delta)\left(1 - \alpha_{IOR}^{M_1-1}\right), \frac{\delta}{M_1 - 1}(1 - \mathbb{G}_0(R))} \tag{C.3}$$

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:

$$\alpha_m < \left(1 - (\frac{1}{2})^{M_1-1-|B_1(m)|}P_2\right) \cdot 1 + (\frac{1}{2})^{M_1-1-|B_1(m)|}P_2\alpha^*(|B_2(m)|)$$
$$\alpha_m < 1 - (\frac{1}{2})^{M_1-|B_1(m)|}P_2(1 - \alpha^*(|B_2(m)|))$$
$$\alpha_m < 1 - (\frac{1}{2})^{|B_2(m)|}\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)| + |B_2(m)|}\Big)(1 - \alpha^*(|B_2(m)|))$$

Now let us define $\epsilon^* = min_{B_2(m)}\Big((\frac{1}{2})^{|B_2(m)|}\Big(1 - |B_2(m)|\frac{\delta}{|B_1(m)|+|B_2(m)|}\Big)(1 - \alpha^*(|B_2(m)|))\Big)$.
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 assumption 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 Theorem 3. Take an agent n with M nested neighbors.

- Make point that Bayesians must do at least as well as just updating belief in response to Blackwell experiment represented by each nested neighbor's action.

Using Lemma 6, we can observe that each nested neighbor action must be more informative than the lower bound Blackwell experiment.

By Lemma 8, 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 Example 1. A covariance stationary process $\{Y_t\}_{t=1}^{\infty}$ with mean μ is defined by the following three properties (Hamilton, 2020):

$$\begin{aligned} \mathbb{E}(Y_t) &= \mu && \forall t \\ \mathbb{E}(Y_t - \mu)(Y_{t-j} - \mu) &= \gamma_j && \forall t \\ \sum_{j=0}^{\infty} |\gamma_j| &< \infty \end{aligned}$$

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 following conditional covariance formalu to establish that the covariances are absolutely convergent:

$$cov(X, Y) = \mathbb{E}[\mathbb{E}[XY|Z]] - \mathbb{E}[\mathbb{E}[X|Z]]\mathbb{E}[\mathbb{E}[Y|Z]]$$

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

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.

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

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:

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

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 Chapter 7.2, pg. 186) that,:

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

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.

□

Proof of Theorem 2 Part 3. To recall the assumptions already mentioned in the main text, assume:

- A complete network
- Bounded beliefs: $[1 - \underline{B}, \overline{B}]$
- b_1, b_0, b_B defined to represent the highest social beliefs that can be overturned by a private signal. Proposition 1 gives us that $(1 - b_1) = \frac{\overline{B}(1+s)}{1-s+2s\overline{B}}$, and a similar expression b_0.

For any 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 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 social belief will any observation be able to push the updated social belief into a rejection in which some types will reject it.

These definitions establish that the social belief cannot enter a rejection region, so to complete the proof we must simply establish that:

1. The social belief will necessarily enter the region $(1 - b_1, b_0)^c$ at some point.
2. If it does so, agents will all take the same action.

This second point is easy to see; if, for example, the 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 $1 - b_B$ 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 SS2000 Theorem B.1; the Martingale-Markov social belief must eventually settle at some fixed-point.

□

Proof of Proposition 4. 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.[23] 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 6 and 8 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 $m > M$. □

D Learning with Severely Bounded Beliefs

As is established by part 3 of theorem 2, expanding nested neighborhood samples is clearly not a sufficient condition for learning when the signal structure is severely bounded: the complete network

[23] An example of such a problem can be produced by adjusting the decision problem of Cremin (2023) 'Learning through Anonymous Speech'. 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.

exhibits expanding sample sizes, and learning does not occur here with severely bounded beliefs. However, it does not follow that one cannot achieve complete Bayesian learning with severely bounded beliefs; in ADLO2011, one of their most surprising results (since their article was written as an extension to SS2000, 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 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 severely bounded beliefs here.

The specific class of network topologies used in ADLO2011 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 5 (Non-Persuasive neighborhoods)**.** *A finite set $B \subset \mathbb{N}$ is a non-persuasive neighborhood in equilibrium $\sigma \in \Sigma$ if*

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

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

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. Instead they will need expanding sample sizes with respect to this set of learning agents, allowing learning along the same lines as in theorem **??**.

Theorem 3. *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. $C_i \in \mathcal{U}_\sigma$ 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 ADLO2011, and learning outside follows from the same argument used to establish the sufficiency of expanding nested neighborhood samples for learning with uniformly informative beliefs in theorem 3. □

Thus learning is certainly still possible with severely bounded 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 Propositions 3 and 4 in ADLO2011, 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 Extensions

E.1 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 $\lambda \in (0,1)$ there is some non-zero probability that the agent will form a combined belief within $[0, 1-R) \cup (R, 1]$. 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. 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 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, ENNS is no longer a sufficient condition for learning.

In this setting the results of Theorem 2 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 severely bounded beliefs, there will be some region of 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.

E.2 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 social beliefs above 0.5 with probability $\frac{1}{2}(1-SB)^2 + \epsilon$ for some $\epsilon > 0$.

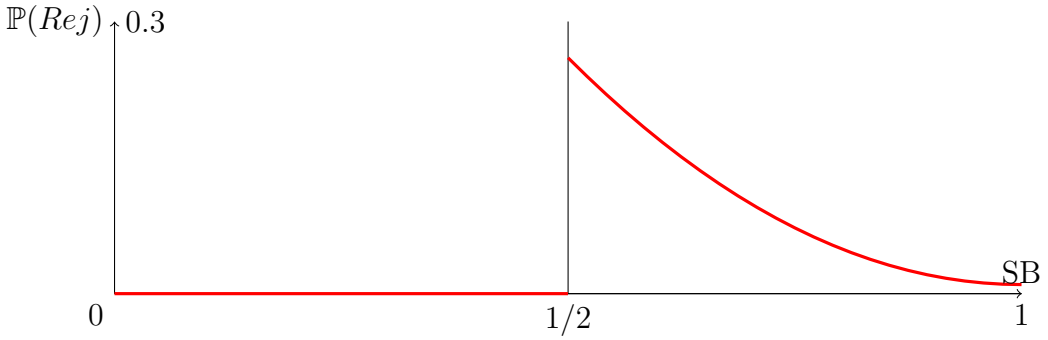


Figure 6: $(1-x)^2 + 0.01$ rejection probability beyond $\frac{1}{2}$

Such a rejection procedure preserves the uniform informativeness 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 severely bounded and uniformly informative becomes moot- all signal structures are uniformly informative). 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 social beliefs converge to 1, the fraction of agents rejecting their 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 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 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 occur). 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.