Bot Got Your Tongue? Social Learning with Endogenous Action Visibility

John W.E. Cremin*†

September 23, 2026

Latest Version

Abstract

Models of social learning conventionally assume that all actions are visible, whereas in reality, we can often choose whether or not to advertise our choices. In this paper, I study a model of sequential social learning in which *social* agents choose whether or not to let successors see their action, only wanting to do so if they are sufficiently confident in their choice (they have *reputation concerns*), and *noise* agents act randomly. I find that in *sparse* networks, this produces a form of unravelling to the effect that noise agents are overrepresented. This can damage learning to an arbitrary extent if social agents' reputation concerns are sufficiently strong. In *dense* networks, however, no such unravelling occurs, and the combination of noise and reputation concerns can facilitate complete learning even with *bounded* beliefs.

Keywords— Sequential Social Learning, Endogenous Social Networks, Network Theory, Information Economics

1 Introduction

To brag or not to brag? That is an often pertinent question. Having undertaken a particular purchase, made one's mind up on the political controversy du jour, or chosen to invest in a certain stock, we are often able to *choose* whether or not our peers or colleagues are aware of our choice. Models of social learning, however, conventionally neglect this fact and assume that agents make a decision, and will then be observed by their neighbours whether they like it or not. Whenever agents benefit or suffer upon being seen to act wisely or unwisely, perhaps through reputation effects or mockery from peers, and have some

*Aix-Marseille School of Economics, Université d'Aix-Marseille, CNRS, john-walter-edward.cremin@univ-amu.fr.

†I would like to thank Evan Sadler for his guidance with this project, as well as Navin Kartik, Jacopo Perego, Sebastian Bervoets and Romain Ferrali for many helpful comments and suggestions. I would also like to thank seminar participants at Columbia and Aix-Marseille Universities for their feedback. Finally, thanks to the Organisers of the ASSET 2025 Conference for awarding this paper the Louis-André Gerard-Varet Prize.

ability to vary the observability of their action, this assumption is not anodyne. In a world where the visibility of one's action is a choice, any observation network becomes endogenous, and the information a given action carries is altered by the fact that the agent in question *chose* to act visibly.

In this paper, I model agents who only make their action visible to others when sufficiently confident that it is correct, in a classical sequential social learning model à la Acemoglu et al. (2011). Having introduced this *reputation concern* into the model, I consider its interaction with the presence of non-strategic noise agents who could for example represent partisans in political debate, or bots on social media platforms. Their combined effects on first learning and second the overrepresentation of noise agents are then shown to vary as a function of whether the network is sparse or dense, where I define a sparse network as one in which we can find a uniform finite upper bound on the size of all agents' neighbourhoods.

In dense networks, it is possible that reputation concerns interact with noise to help matters by removing the *cascade beliefs*, which are social beliefs at which an agent will choose the same action regardless of their private signal, and facilitating learning. Taking the complete network, I show in Theorem 1 that reputation concerns and noise can deliver learning with bounded beliefs. In this case, we can always find reputation-concern distributions (those with support containing both the very candid *and* the very reputation-sensitive) in which agents converge to certainty on the true state, though this does not occur in their absence. This is due to the fact that without noise agents, at a given cascade belief all agents will choose the corresponding action with probability one regardless of the state, but at least half of noise agents, when they are present, will choose the opposite action. The probability of observing each action will then depend on the likelihood of social agents dropping out, which will be less likely to occur if the cascade belief supports the true state than otherwise since private signals will be less likely to produce moderate beliefs. If the reputation-concern distribution satisfies the condition mentioned above, action one will be strictly more likely if the state is one than if it is zero for any interior social belief. This guarantees learning, and removing either reputation concerns or noise is thus enough to prevent it in this setting.

Sparse networks, conversely, turn out to be vulnerable to the presence of reputation concerns and noise, for a number of reasons. Firstly, I establish that in these networks we will observe a form of *unravelling* in which reputation-concerned *social* agents (i.e. those agents that are not noise agents) will 'drop out' of the game by acting invisibly, and in so doing provoke yet more social agents to do the same, creating a vicious cycle. I characterise a lower bound to the severity of this effect in Theorem 2, and an upper bound on learning that follows from it in Section 3.2. After establishing some comparative statics to this first bound in Lemma 8, I can then demonstrate that introducing *strong enough* reputation concerns can cause an arbitrarily high proportion of social agents to drop out, and smother social learning completely as in Proposition 2. Examples 1 and 2 display the effect of changing the reputation-concern distribution in sparse networks, and help illustrate the breakdown of the *improvement principles* on which much of the literature depends to establish learning in such networks.[1] They also highlight the trade-off between observing more noise agents and more confident social agents that reputation concerns introduce; in the

[1] *Improvement principles* prove learning by noting that in equilibrium, a Bayesian agent must do at least as well as an agent who simply copies their most accurate predecessor, only disagreeing upon receipt of a sufficiently extreme opposing private signal.

immediate predecessor network, Proposition 4 shows that the latter dominates when reputation concerns are weak, whilst the former gains the upper hand when they are strong.

Literature Review: There is an extensive literature on social learning with a binary state, much of this considers only the complete network (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000). Various articles develop this literature by exploring learning on more general network topologies, and my approach follows the workhorse model of Acemoglu et al. (2011), in which it resembles Lobel and Sadler (2015, 2016); Cremin (2025); Lomys (2020). Lobel and Sadler (2016) and Cremin (2025) both consider models in which the classic proof technique of the *improvement principle* breaks down, due in the former to agents having different tastes, and in the latter their rejection of social information via motivated reasoning. The presence of noise agents here has a similar impact to that of motivated reasoners in Cremin (2025), though the fashion in which social information is lost is much more blunt with the former. In Lobel and Sadler (2016), instead the problem is that when unaware of the tastes of a given neighbour it is no longer possible to straightforwardly improve on their action. In copying an agent with different tastes, one does not achieve the same utility, and thus 'improving' on their action with a new private signal may nonetheless yield a lower utility.

A distinct literature extends the original complete network models by varying the visibility of earlier actions.[2] Herrera and Hörner (2013) and Guarino et al. (2011) both model learning when only one of the two actions is visible, therefore making the observation network endogenous. In the first case arrival time is random, and in the second agents do not know at what point in the sequence they arrive. My paper is quite different to these in that the visibility of actions is a choice. Beyond these Song (2016) studies a model in which agents are able to choose which predecessors they observe at some cost, and in which the network topology is also therefore endogenous, but agents cannot choose whether or not their action is observable by successors. To the best of my knowledge, no other papers yet allow agents to themselves choose the visibility of their action, and make it a function of their degree of confidence that they are correct. This seems an important feature of social learning in the real world, in many contexts, and thus forms part of the motivation of this paper.

The paper is organised as follows. In Section 2, I set out the model agents' decision rules. I then move on to discussing my results in Section 3, which includes learning in dense and sparse networks, unravelling, and benchmark results. In Section 3.1, I show how reputation concerns and noise can help learning in dense networks when agents have bounded beliefs with Theorem 1, which also shows that in the complete network they make no difference when signals are unbounded. Section 3.2 then demonstrates the unravelling effect that occurs in sparse networks, providing a lower bound on the fraction of visible agents who are of noise type in Theorem 2. Proposition 1 gives sufficient conditions for such an unravelling to occur (where the fraction of visible agents who are of noise type will almost surely be strictly greater than in the general population). I then discuss learning in sparse networks further with Examples 1 and 2. Section 4 considers to what extent my results depend on the precise model specification, or are robust to extensions.

[2]Models in which agents observe some summary statistics of their predecessors' actions have also been studied, e.g. Guarino and Jehiel (2013), Callander and Hörner (2009), but the connection between those and the model of this paper is much weaker.

2 The Model

We have a binary state $\theta \in \{0,1\}$, which nature draws uniformly at the beginning of the game. The common prior of all agents reflects this, and assigns 0.5 to each state (where this assumption serves only to simplify notation, the patterns explored in this paper do not need an even prior). An infinite sequence of agents arrive at exogenously given times, their indices representing their position in this sequence $n \in \mathbb{N}$, with agent 1 arriving after Nature draws the value of the binary state. Upon arrival, each agent observes some information about the true state of the world $\mathcal{I}_n$, and decides both which action to take $x_n \in \{0,1\}$, and whether or not to make this decision visible $v_n \in \{0,1\}$. Agents come in two unobservable types, *noise* and *social* $\tau_n \in \{N,S\}$. Noise types do not attempt to match their action to the state, but simply randomise uniformly between the two actions, and act visibly $v_n = 1$. Social types use their information to try and match their action to the state, and prefer to have acted visibly if their action is correct, but invisibly if not. The precise details of this can be seen in the utility functions I present momentarily, and each agent is a noise type independently with probability η, where $\eta \in (0,1)$.

Before choosing their action, each agent receives a conditionally independent private signal s_n (this could represent simply intuition, or the result of individual research). The private signal is described by the information structure $(\mathbb{F}_0, \mathbb{F}_1)$ (where $\mathbb{F}_\theta$ is the distribution of the private signal in state of the world θ). I assume these belief distributions do not contain a perfectly revealing signal (are mutually absolutely continuous with respect to each other), but are informative ($\frac{d\mathbb{F}_1}{d\mathbb{F}_0} \neq 1$ on a non-null set). These private signal distributions imply private belief distributions $p_n \sim \mathbb{G}_\theta$ (of the private belief $p_n := \mathbb{P}(\theta = 1|s_n)$, that formed upon observing only the private signal).

In addition to this, I assume each agent observes the actions of some subset of those agents who have visibly-acted before them (chosen $v_n = 1$). Each agent's order in the sequence is of course given by their index, but we can also define an *adjusted index* for each to reflect the number of visible actions that have been taken at the time that agent arrives: $\tilde{n} := 1 + \sum_{i=1}^{n-1} v_i$. Agents do not observe either their own index or those of their neighbours, but they do observe their adjusted index and those of their neighbourhood: each agent with adjusted index $\tilde{n}$ observes the actions $\{x_{\tilde{k}} : \tilde{k} \in B(\tilde{n})\}$ of their neighbourhood $B(\tilde{n}) \subseteq \{1, \ldots, \tilde{n}-1\}$, where $x_{\tilde{k}}$ denotes the action of the visibly-acting agent with adjusted index $\tilde{k}$. Neighbourhoods are therefore sets of adjusted indices rather than of indices, since only visibly-acting predecessors can be observed.[3] These neighbourhoods are drawn according to the commonly-known distributions $\{\mathbb{Q}_{\tilde{n}}\}_{\tilde{n}\in\mathbb{N}} = \mathbb{Q}$ at the beginning of the game, and are independent across $\tilde{n}$. Note that multiple agents will, in general, have the same adjusted index, though only one visibly-acting agent has each. The network is based upon these adjusted indices $\tilde{n}$, and not one's position in the overall sequence of arrivals, n. I shall define network topologies for which there is some $M \in \mathbb{N}$ such that $|B(\tilde{n})| < M$ almost surely as *sparse*, and networks in which this is not the case as *dense*. Upon observing this social information, agents form their social beliefs $sb_n := \mathbb{P}_\sigma\big(\theta = 1|\{x_{\tilde{k}} : \tilde{k} \in B(\tilde{n})\}\big)$,

[3]For applications regarding social media (where the binary visibility choice represents the decision to comment or not) this assumption is particularly defensible, since it amounts to assuming agents can observe how many predecessors have commented on a given hashtag or topic already, without knowing the number of people who have actually logged on and considered a matter without commenting.

which condition on social information only: in each state of the world, a profile of equilibrium strategies induces a probability distribution over histories, and thus the actions of agents in $B(\tilde{n})$, so forming these beliefs involves simply an application of Bayes' rule. One can partition the set of histories (infinite or after a finite amount of time) into sets such that the jth visible action is the same for each history in a given element of this partition for every j, and the number of visible actions is also the same. I shall call each element of this partition a *visible history* h^v, and it will be convenient at points to discuss agents acting after a given visible history, meaning after any history within that element of the partition. Hence I shall use $n(h^v)$ and $\tilde{n}(h^v)$ to refer respectively to the first agent and visible agent who act at history h^v in a given equilibrium (they may be the same agent if the first to arrive chooses to act visibly). Since this game features *self-locating uncertainty*, where one play of the game visits the same information set multiple times, it exhibits a dynamic inconsistency called a *Newcomb tension* that I study in Cremin (2026). In the main model of this paper, I assume that upon observing *only* their adjusted index agents would not update their beliefs about θ. As per Theorem 2.2 and Proposition 2.1(i) of that paper, this is tantamount to assuming that agents have commitment power. I discuss this further, along with the alternative assumption and the fact that it does not substantively change the results, in Section 4.

Given their posterior belief $\mathbb{P}_\sigma\big(\theta = 1 | s_n, \{x_{\tilde{k}} : \tilde{k} \in B(\tilde{n})\}\big)$ (formed with both private and social information) a social agent n chooses (x_n, v_n) to maximise:

$$u_n(x_n, v_n, \theta) = \begin{cases} 1 + v_n & \text{if } x_n = \theta, \\ -\big(1 + r_n \cdot v_n\big) & \text{if } x_n \neq \theta, \end{cases} \tag{2.1}$$

$r_n \in [1, \infty)$ is a parameter representing the agent's reputation concerns, drawn i.i.d. for each agent from some distribution $\Delta(r)$. Social agents are always trying to match the true state of the world, but their payoffs are more extreme for visible actions, though asymmetrically so. They dislike more to be caught publicly saying something incorrect than vice versa, and the extent of this asymmetry is increasing in their reputation concerns. I further define a social agent's 'candour' as the inverse of their r parameter $c_n := \frac{1}{r_n} \sim \Delta(r)$. If $c_n = 1$ a social agent will never choose an invisible action $v_n = 0$, but if we let $c_n \to 0$ they will certainly do so. A social agent choosing $x_n = 1$ will choose for this action to be visible if their posterior is that the true state is 1 with probability strictly greater than $\frac{1}{1+c}$. I break indifference here in favour of invisibility for notational convenience, though with absolutely continuous signal or reputation-concern distributions it does not matter. I shall also assume they randomise uniformly between $x_n \in \{0, 1\}$ if they form belief 0.5 similarly. Figure 1 represents this. The exact form of this utility function is not essential; there are two properties for which it has been chosen: (1) agents would like to match their action to the true state, come what may; (2) agents only want to act visibly if they have strong beliefs. Even then, this second property is not essential. The unravelling argument of Section 3.2 requires only that there is a non-zero probability that any agent observing a unanimous neighbourhood (M neighbours all choosing $x = 1$, or $x = 0$) forms an overall belief that does not provoke a visible action. Hence so long as some open interval around 0.5 induces $x = 0$, the unravelling reasoning is unaffected. For Theorem 1, we need only that the probability of an agent choosing $x = \theta$ visibly is strictly higher in state θ than $\neg\theta$; any

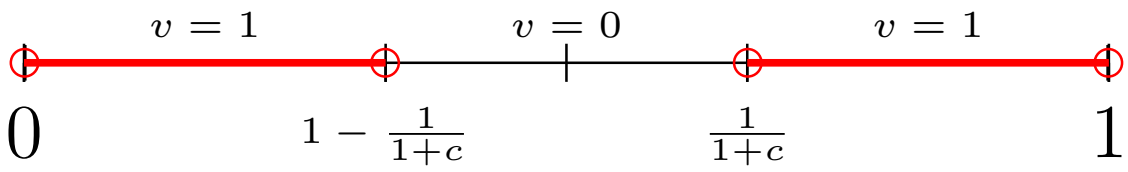


Figure 1: A Social Agent's Action Choices: The position on the number line represents the posterior belief of the agent. If their belief is greater than $\frac{1}{2}$ (that the true state is 1) they will choose $x_n = 1$, and $x_n = 0$ otherwise. In the red regions, their belief is confident enough to choose a visible action, otherwise they will choose an invisible action.

utility function that achieves this will suffice. Any value of r_n below 1 will produce the same behaviour as if the agent had no reputation concerns at all, so there is no point allowing this parameter to take lower values; one can simply assign all probability mass on such values to $r_n = 1$.

A social agent's decision rule for choosing x_n can be represented in terms of the sum of the private and social beliefs, exactly as in Acemoglu et al. (2011, Proposition 2) (reproduced in this paper as Lemma 2 in Appendix B). This lemma is derived only assuming that the private and social signals are conditionally independent of each other, and does not depend at all on the network topology or the assumptions Acemoglu et al. (2011) make about it in the rest of their paper. We can similarly derive a rule governing the agent's choice of action visibility, which of course reflects that stronger beliefs are necessary to choose visible actions:

Lemma 1. *After any history h_n, the agent chooses $v_n = 0$ if:*

$$1 - (r-1)\mathbb{P}(\theta = 1|B(\tilde{n}))\mathbb{P}(\theta = 1|s_n) < S_n < 1 + \frac{r-1}{r}\mathbb{P}(\theta = 1|B(\tilde{n}))\mathbb{P}(\theta = 1|s_n)$$

for $S_n = \mathbb{P}(\theta = 1|B(\tilde{n})) + \mathbb{P}(\theta = 1|s_n)$.

We can clearly see that the threshold for choosing $v_n = 1$ when $S_n > 1$ is higher, since added to the original threshold is a positive cross-product of the social and private beliefs of the agent. The solution concept employed in this paper shall be that of Perfect Bayesian Equilibrium, and writing $\alpha_n := \mathbb{P}(x_n = \theta)$ for an agent's *accuracy*, *complete learning* obtains if the accuracy of social agents, $\mathbb{P}(x_n = \theta|\tau_n = S)$, converges to 1 in every equilibrium. Noise agents match the state with probability $\frac{1}{2}$ whatever the history, so α_n itself can converge at most to $1 - \frac{\eta}{2}$. All omitted proofs can be found in the appendix.

3 Results

As shall become clear, the phenomena of interest in this model result from the interaction between the reputation concerns of social agents and the presence of noise agents. Whereas they can together be expected to limit learning and give noise types an exaggerated presence amongst visibly-acting agents in sparse network topologies, they conversely work together to make it possible in very dense networks. With

respect to learning, this contrast results from the fact that whereas in sparse networks agents depend on a few neighbours to learn about the history of visible actions up to that point, in very dense networks this is less the case (and not at all the case in the complete network) as they will be able to observe much of this history themselves. The combination of reputation concerns and noise agents interferes with the improvement-based learning that achieves complete learning in the first case, but can remove the interior cascade beliefs that prevent it in the second. When we consider the overrepresentation of noise agents, I shall analyse an unravelling effect that occurs in sparse networks (where there is a commonly-known bound M on each agent's number of neighbours), and becomes more extreme the more sparse the network is (the lower this M), but disappears if there is no commonly known integer bound on neighbour numbers ($M = \infty$). These results together show how the impact of noise and reputation concerns depends on the density of the network: in *dense* networks they can help by removing cascade beliefs, in *sparse* networks they produce unravelling, which undermines learning and produces an overrepresentation of noise.

3.1 Learning in Dense Networks

In dense networks, relative to sparse ones, agents are generally not so dependent on recent neighbours to learn about the history of the game, since they can directly observe much of it. Conversely, belief cascades become more concerning, and produce incomplete learning in the canonical model (Smith and Sørensen, 2000) when we have bounded signal structures. As I show in the following, using the complete network as a tractable example of a dense network topology, the combined presence of reputation concerns and noise agents can actually help learning by reducing (or completely removing) the set of cascade beliefs (beliefs at which agents simply ignore their private signals (Smith and Sørensen, 2000)).

Theorem 1 (Learning in the Complete Network)**.** *The following two statements characterise learning in the complete network:*

1. *With unbounded beliefs, there is complete learning. The social belief converges to certainty on the true state almost surely.*
2. *With bounded beliefs supported on $[\underline{p}, \overline{p}] \subset (0,1)$, and a candour distribution $\Delta(c)$ such that for some $\epsilon > 0$, $(0, c^* + \epsilon) \cup (1-\epsilon, 1) \subseteq \text{supp}(\Delta(c))$, where*

$$c^* := \frac{\underline{p}(1-\overline{p})}{\overline{p}(1-\underline{p})},$$

the social belief converges to certainty on the true state almost surely.

The first part of this theorem simply establishes that in the complete network with unbounded beliefs the presence of reputation concerns and noise agents makes no difference. The second part, however, gives the aforementioned setting in which reputation concerns and noise help: for any bounded belief setting we can find some candour distribution that produces complete learning. Precisely, reputation concerns help learning here by providing a signal of the confidence with which agents believe the state of the world

matches their action. When the distribution of this candour parameter has support sufficiently close to 0, it is no longer the case that neighbour actions can be completely uninformative for social beliefs beyond a certain point. One could write a simpler, though weaker, theorem statement by simply assuming the candour distribution is full support; in the following I explain why the chosen assumption is the minimal necessary one. The threshold c^* is determined by the endpoints of the private belief support; requiring candour values below $c^* + \epsilon$ to be supported ensures that for every social belief, some agents' reputation concerns are strong enough that their visibility thresholds (c.f. Lemma 1) fall strictly inside $(\underline{p}, \overline{p})$. This guarantees that there is either strictly positive probability that a social type varies either x_n, v_n, or both upon observing different private signals.[4] However, it is not just *reputation concerns* helping here, as if we removed the noise agents but left reputation concerns we would once again have incorrect learning. In a model in which agents could observe that an agent had decided to act invisibly, reputation concerns would suffice; this is because the problem with cascade beliefs is that an agent with such a belief will always choose the same course of action irrespective of their private signal. This ensures that this action reveals no information about said private signal, and so information ceases to accumulate. If we had no noise agents but could observe when an agent had just chosen to act invisibly, this choice to act invisibly (for a social belief that is strong enough in favour of $\theta = 1$ that choosing $x = 0$ is a probability 0 event) would serve as a noisy signal of that agent's private belief. One way of seeing this is to observe that some social beliefs would be cascade beliefs vis-à-vis the action itself, but not cascade beliefs in terms of an agent's choice of visibility: a social belief would need to be a cascade belief in both senses to stop learning. Of course, in this model agents cannot observe when others have chosen to act invisibly. If there were no noise agents, at cascade beliefs we would simply observe an infinite sequence of all visible agents taking the action consistent with that social belief. No information would be revealed about the proportion of agents choosing to act invisibly. Since there are noise agents, however, there is a minimum probability of $\eta/2$ that each action will be chosen, since it is with this probability that noise agents arrive and choose each action. The proof proceeds by showing that the probability $\phi(sb, \theta)$ with which the next visible agent chooses action 1 is strictly higher in $\theta = 1$ than $\theta = 0$ for every interior social belief. This is established by decomposing ϕ into contributions from agents who are visible and choose 1 (at rate R_θ) and agents who are visible and choose 0 (at rate L_θ), and showing that the first order stochastic dominance of private beliefs in state 1 over state 0 ensures $R_1 > R_0$ or $L_1 < L_0$ (or both) for every interior social belief. The condition on the candour distribution ensures that for every social belief, a positive measure of social agents (or of their reputation-concern values speaking precisely) have visibility thresholds inside the private belief support, where this dominance is strict. As the social belief converges to certainty on the true state, the probability with which the next visible agent is a noise agent will converge to η, as all social agents become confident enough to act visibly. In addition to allowing complete learning with the assumptions of Theorem 1, the reputation-concern distribution can also serve to simply reduce the size

[4]This almost sure convergence to the truth with bounded beliefs and sufficiently strong reputation concerns is reminiscent of Goeree et al. (2006, Theorem 2), where my assumption that the support of the candour parameter contains $(0, c^* + \epsilon)$ for some $\epsilon > 0$ performs the same function as their assumption 3, that the joint distribution of private values over the A available actions has full support: $\text{supp}(f^t) \supseteq [0,1]^A$. Both ensure that the martingale social belief process can only settle on degenerate beliefs.

of the set of cascade beliefs.

3.2 Learning in Sparse Networks

Whereas introducing reputation concerns can be helpful in dense networks, however, in sparse networks the impact is largely negative. As I illustrate with examples 1 and 2 in this section, though they can help learning, more commonly they do not, and in Proposition 2 I show that if reputation concerns are strong *enough* they must harm learning. The major phenomenon that drives this negative impact is an unravelling effect that leads to noise agents being over-represented amongst visibly-commenting agents. Theorem 2 shows the immediate consequences of this unravelling, and much of this section discusses the major intuitive content of the proof.

Theorem 2. *Suppose the network is sparse, with commonly-known bound $M < \infty$ on neighbourhood sizes. Then $\underline{\eta}_\infty = \frac{\eta}{\eta+(1-\lambda_\infty)(1-\eta)}$ is a lower bound on the probability that the $\tilde{n}$th visibly-acting agent is a noise type for any $\tilde{n} \in \mathbb{N}$, and the asymptotic fraction of agents that are noise types is almost surely greater than $\underline{\eta}_\infty$. λ_∞ is the smallest solution in $[0,1]$ of:*

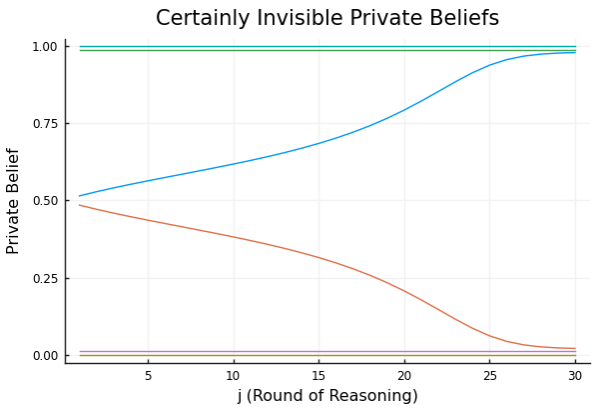
$$\lambda_\infty = \min\left\{ \int_{1/\omega_\infty}^{\infty} \mathbb{G}_0\Big(\Big(\frac{r\omega_\infty}{1+r\omega_\infty}\Big)^-\Big) - \mathbb{G}_0\Big(\frac{1}{1+r\omega_\infty}\Big) d\Delta(r), \right.$$
$$\left. \int_{1/\omega_\infty}^{\infty} \mathbb{G}_1\Big(\Big(\frac{r\omega_\infty}{1+r\omega_\infty}\Big)^-\Big) - \mathbb{G}_1\Big(\frac{1}{1+r\omega_\infty}\Big) d\Delta(r) \right\}$$

where $\omega_\infty := \big(\eta/\big(2-\eta-2\lambda_\infty(1-\eta)\big)\big)^M$.

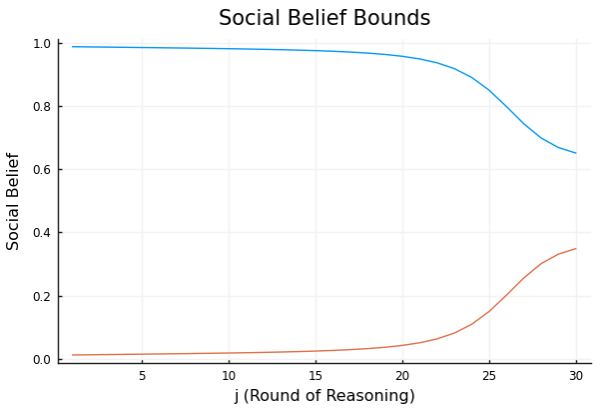
The full proof can be found in the appendix, but the intuition of unravelling (the proof of the first part of this statement) is worth considering here. Even in the absence of reputation concerns, the bound on neighbourhood size itself limits the maximum possible strength of social beliefs. We shall now see that this can produce much tighter constraints on learning and a much greater presence for noise agents than it may at first seem. An implication of these bounds on social beliefs is that if agents' reputation concerns are sufficiently strong, there will be private signals for which they certainly (whatever their social signals) choose to act invisibly. Specifically, using Lemma 1 we can see that private signals within the following interval certainly produce invisible actions, if r_n is large enough for it to be nonempty:

$$\mathbb{P}(\theta = 1|s_n) \in [\underbrace{1 - \frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{L_0}, \underbrace{\frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{U_0}]$$

Suppose at first that $r_n = r$ for all n. If we denote $\lambda_0 := \min\{\mathbb{G}_0(U_0^-) - \mathbb{G}_0(L_0), \mathbb{G}_1(U_0^-) - \mathbb{G}_1(L_0)\}$ (where $\mathbb{G}_\theta(U^-) := \lim_{x\uparrow U} \mathbb{G}_\theta(x)$, so that $\mathbb{G}_\theta(U^-) - \mathbb{G}_\theta(L)$ is the probability of an agent forming a private belief in the open interval (L,U)), we can see that whatever the state of the world, different agents will form private beliefs within this region with at least probability λ_0, where of course agents' private beliefs are



(a)



(b)

Figure 2: In panel (a), the curved blue and orange lines give the bounds of the interval of private beliefs that must produce invisible actions. This interval clearly expands with each round of reasoning. The green and purple lines bounding them give the private signals that are sufficient for a visible action even with a social belief of 0.5. In panel (b), the corresponding interval of possible social beliefs collapses. The parameters in this example are $f_1(x) = 2x$, $f_0(x) = 2 - 2x$, $M = 2$, $\eta = 0.2$, $r = 86$, $\underline{\eta}_\infty = 0.864 (3 d.p.)$.

conditionally independent. This formula ignores, however, the fact that one's neighbours are necessarily drawn from those who chose to act visibly. For any given neighbour, the probability that they area noise agent is therefore higher than the η used in this interval: at most $(1 - \lambda_0)$ of the social agents choose visible actions. The relevant distribution has noise-probability greater than $\underline{\eta}_1 := \eta / \big(\eta + (1 - \lambda_0)(1 - \eta)\big)$.

We can thus run the same argument a second item, iterating the line of reasoning. The upper and lower social belief bounds become tighter, and we can define a private belief interval $[L_1, U_1]$, a new λ_1, and thus a $\underline{\eta}_2$, all by substituting $\underline{\eta}_1$ in place of η. After this second iteration however, we can as before iterate again and again, at each stage getting an interval of social beliefs that is at least weakly tighter, and an at least weakly higher $\underline{\eta}_j$, where j indexes the iteration. Figure 2 shows an example; the region of private beliefs that must produce an invisible action expands with each iteration, and the range of possible social beliefs shrinks. $\{\underline{\eta}_j\}_{j=0}^{j=\infty}$ is a weakly increasing and bounded sequence, so it must converge.

Fundamentally, what this line of reasoning provides is a fixed-point interval of social beliefs. If you observe neighbours with social beliefs in $[\underline{SB}_k, \overline{SB}_k]$, you must have a social belief within $[\underline{SB}_{k+1}, \overline{SB}_{k+1}]$, and thus $[\underline{SB}_\infty, \overline{SB}_\infty]$ is that social belief interval (which notably contains the starting social belief $\frac{1}{2}$) that cannot be 'escaped' by a society. Thus the presence of reputation concerns can produce a tighter range of possible social beliefs.

Without imposing that all agents have a single r, we can apply similar reasoning for any distribution $\Delta(r)$, defining the functions $L_j(r)$ and $U_j(r)$, that give the certainly-invisible private belief regions for each r. Letting r_j^* be the value of r such that $L_j(r) = U_j(r)$, we can similarly define λ_j as the analogous lower-bound probability of an invisible action, precisely with the integral:

$$\lambda_j = \min \left\{ \int_{r_j^*}^{\infty} \mathbb{G}_0(U_j(r)^-) - \mathbb{G}_0(L_j(r)) d\Delta(r), \int_{r_j^*}^{\infty} \mathbb{G}_1(U_j(r)^-) - \mathbb{G}_1(L_j(r)) d\Delta(r) \right\}$$

Once again this produces a weakly increasing and bounded sequence $\{\underline{\eta}_j\}_{j\in\mathbb{N}}$ that converges to a bound $\underline{\eta}_\infty$. The iteration process with a distribution of reputation-concern levels is illustrated by the example in Figure 3; the black lines give $L_0(r)$ and $U_0(r)$, and thus contain the (p_n, r_n) pairs that certainly engender invisible actions for agents with a social belief in $[\underline{SB}_0, \overline{SB}_0]$. The purple line gives the 100th round of this reasoning, and the red the 150th round, illustrating the fact that this region grows and converges to a larger region than the original. The brown line shows the probability density function of the assumed reputation-concern distribution.

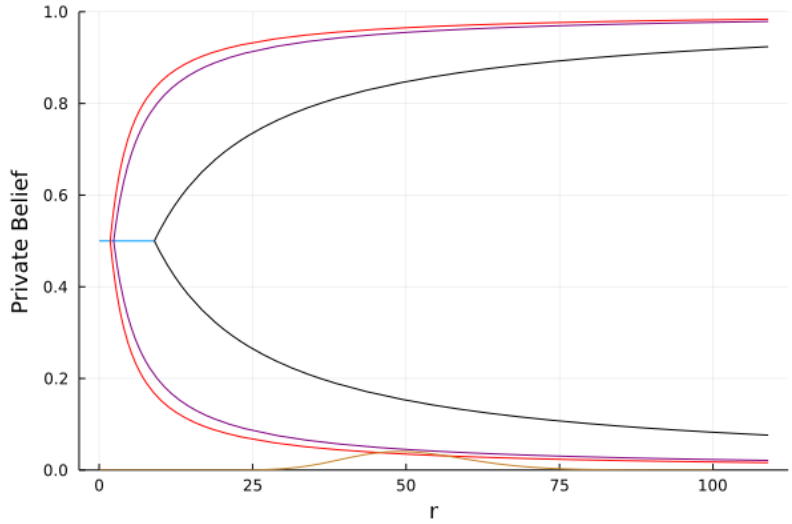


Figure 3: The parameters for this example are $f_1(x) = 2x$, $f_0(x) = 2 - 2x$, $M = 1$, $\eta = 0.2$; the distribution of $r_n - 1$ is a Chi-squared with 50 degrees of freedom.

Note also that this reasoning applies if agents observe their own adjusted index and those of the agents they observe (as mimics Acemoglu et al. (2011)), only their own adjusted index (Smith and Sørensen, 2008), or neither (Monzón and Rapp, 2014). Hence this unravelling will occur in a larger class of social learning problems than that explicitly studied in this paper. Additionally, although the inequalities discussed thus far have been predominantly weak, we do not need strong conditions to produce strict ones (and thus an actual unravelling effect):

Proposition 1 ($\underline{\eta}_\infty$ Properties)**.** *$\underline{\eta}_\infty > \eta$ if and only if $\lambda_0 > 0$, which is guaranteed by assuming that $\Delta(r)$ assigns mass $\epsilon > 0$ to r values strictly greater than r_0^*, and that the interval $(0.5-\delta, 0.5+\delta)$ is contained within the support of private beliefs for some $\delta > 0$.*

We can arrive at the stronger statement that $\underline{\eta}_j > \underline{\eta}_{j-1}$ for any $j \in \mathbb{N}$ by assuming that the r distribution, $\Delta(r)$, is full support; that the private beliefs distribution is full support, and some positive mass $\epsilon > 0$ is assigned to reputation-concern values greater than r_0^*; or finally that private beliefs are unbounded, and the reputation-concern distribution assigns some mass $\epsilon > 0$ to $r_n > \min\{L_0^{-1}(\underline{p}), U_0^{-1}(\overline{p})\}$ where $\overline{p}, \underline{p}$ are respectively the highest and lowest private beliefs with density 0. Beyond noting characteristics of this

fixed point, one can observe the immediate implication that it provides a tighter bound on accuracy than already established in the absence of reputation concerns as the first step of the proof of Theorem 2:

$$\mathbb{P}(x_n = \theta | \tau_n = S) \leq \frac{1}{2}\mathbb{G}_0\left(\frac{(1-\frac{\eta_\infty}{2})^M}{(1-\frac{\eta_\infty}{2})^M + (\frac{\eta_\infty}{2})^M}\right) + \frac{1}{2}\left(1 - \mathbb{G}_1\left(1 - \frac{(1-\frac{\eta_\infty}{2})^M}{(1-\frac{\eta_\infty}{2})^M + (\frac{\eta_\infty}{2})^M}\right)\right)$$

In addition to simply bounding accuracy in a given game however, this result, allowing with comparative statics on η_∞ (collected in Lemma 8 in Appendix B) allow us to see the consequences of increasing reputation concerns without bound. Proposition 2 states that in sparse networks, increasing reputation concerns will always eventually harm learning, and achieve complete domination by noise agents. Similarly increasing the spread of the private belief distributions, in turn leads us to similar negative outcomes as shown in Proposition 3.

Proposition 2. *Suppose we have a learning game with $M < \infty$, and consider a sequence of reputation-concern distributions $\Delta^k(r)$ for $k \in \mathbb{N}$, in which each successive distribution is shifted to the right by some fixed value $\delta > 0$. $\underline{\eta}^k_\infty \to 1$, and from this the limiting visible accuracy $\tilde{\alpha}^* \to 0.5$ and accuracy $\alpha^* \to \frac{1}{2}\mathbb{G}_0(0.5) + \frac{1}{2}\big(1 - \mathbb{G}_1(0.5)\big)$.*

This formalises the fact that whilst some reputation concerns can actually help learning, increasing reputation concerns too much is always harmful. Figure 5a below provides an example that captures this intuition. A series of FOSD increases in the (shifted) Poisson reputation-concern distribution successively reduces the level of asymptotic visible accuracy, though the first two increases instead help learning for a range of values of η. It is also worth noting, as I do in the proof of Proposition 2, that this convergence is faster the lower we set M. Hence it is not only true that sparse networks are vulnerable to reputation concerns in the limit, but that the *more* sparse they are, the *sooner* increasing reputation concerns will do this damage. We could also write this result in terms of a sequence of FOSD shifts, so long as we insist that each term has at least $\delta_1 > 0$ mass shifted at least $\delta_2 > 0$ distance to the right (simply shifting the distribution to the right is of course an example of an FOSD shift).

Turning our attention from shifting the reputation-concern distribution to reducing the spread of the private belief distributions we have Proposition 3. Since we have a single reputation-concern distribution in this result, we can define $\underline{r}$ as the lowest r value supported by this reputation-concern distribution. Finally, for a sequence of private belief distributions $\{\mathbb{G}^k_0, \mathbb{G}^k_1\}_{k\in\mathbb{N}}$, define $U^k_\infty(\underline{r})$ and $L^k_\infty(\underline{r})$ as the corresponding series of functions evaluated at $\underline{r}$.

Proposition 3. *Consider a learning game with some $M < \infty$ and some set of private belief distributions $\{\mathbb{G}_0, \mathbb{G}_1\}$ such that $U^1_\infty(\underline{r}) > L^1_\infty(\underline{r})$. If we take a sequence of new unbounded belief distributions $\{\mathbb{G}^k_0, \mathbb{G}^k_1\}_{k\in\mathbb{N}}$ such that each is a mean-preserving (or Birnbaum-) spread of its successor, and the mass assigned to $[L^1_\infty(\underline{r}), U^1_\infty(\underline{r})]$ converges to 1, then $\underline{\eta}^k_\infty \to 1$.*

As with Proposition 2, this convergence will happen faster when M is lower. The fact that this last proposition places a constraint on the lowest supported r value does of course make it less widely applicable

than Proposition 2, but it should be noted that similar convergence will occur without it, except that $\underline{\eta}^k_\infty$ will not converge to 1, but to some limit strictly below 1. So long as $[L_\infty(r), U_\infty(r)]$ is non-empty for some r and the mass of the $\mathbb{G}$ distributions converges to within the interval defined by some supported r, thinner tails will ensure that this lower bound $\underline{\eta}^k_\infty$ climbs upward.

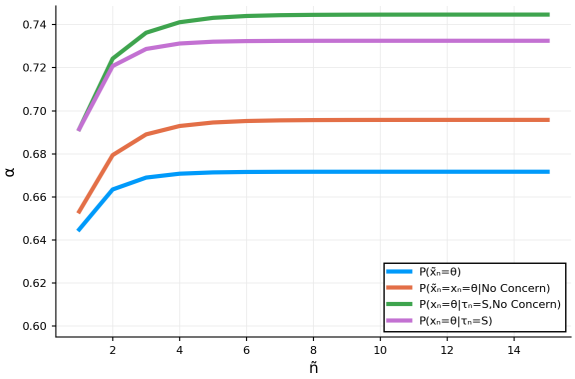
Though after a point they must harm learning, more generally reputation concerns can have two countervailing effects in the presence of noise agents. On the one hand, if social agents choose not to act visibly out of reputation concern and the proportion of noise agents observed increases, a smaller fraction of each agent's neighbours are informative in expectation; they observe more noise. On the other hand, those agents who are most likely to act invisibly are those with the weakest beliefs, i.e. those who are closest to being noise anyway. If the first agent to act at a given visible history is social, and chooses to act invisibly, the next could be either a noise type who acts visibly without information, or another social agent who receives a more informative signal and whose action provides a better indication of the state. Proposition 2 shows how with strong enough reputation concerns this first effect completely dominates (and Proposition 3 shows that with thin enough tails on our private belief distributions and a reputation-concern distribution that does not support low r values, this is also the case), but in general one can find examples where either the first or second effect dominates asymptotically i.e. one can find parameters such that agents match the state with a higher or lower probability in the limit when they have reputation concerns than if we modelled them all as having $r = c = 1$, as in the standard model (Acemoglu et al., 2011). Examples 1 and 2 illustrate this fact, with Figure 5a showing a case in which the second effect dominates when introducing mild reputation concerns, before eventually being overwhelmed. I show that this phenomenon, displayed in these examples, can be shown to hold generally in immediate predecessor networks with antisymmetric private signals in Proposition 4. If we write $\tilde{\alpha}^*(r)$ and $\alpha^*(r)$ for the limits as $\tilde{n} \to \infty$, when they exist, of visible accuracy $\tilde{\alpha}_{\tilde{n}} := \mathbb{P}(\tilde{x}_{\tilde{n}} = \theta)$ and of the accuracy of the first agent to arrive after the $\tilde{n}$th visible action, conditional on their being social, $\mathbb{P}(x_{n(h^v)} = \theta | \tau_{n(h^v)} = S)$ (the blue and purple lines of Figure 4b).

Proposition 4 (Countervailing Effects)**.** *Suppose that agents form an immediate predecessor network, all have the same $r \geq 1$, and that the private belief distributions are antisymmetric, with a density that is bounded above, continuous and strictly positive on $(0, 1)$ (so that private beliefs are unbounded). Then low-level reputation concerns improve learning, and strong ones harm it:*

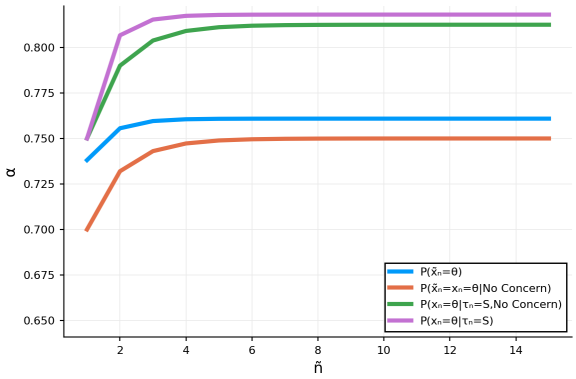
$$\alpha^*(r) > \alpha^*(1) \text{ for all } r > 1 \text{ close enough to } 1, \qquad \alpha^*(r) < \alpha^*(1) \text{ for all } r \text{ large enough.}$$

Here both limits exist, in every equilibrium, and the same inequalities hold for $\tilde{\alpha}^*$.

The proof also shows that $\tilde{\alpha}^*$ and α^* are strictly increasing in r for r close enough to 1, and that $\tilde{\alpha}^*(r) \to \frac{1}{2}$ as $r \to \infty$, so that $\alpha^*(r)$ falls to $\frac{1}{2}\mathbb{G}_0(0.5) + \frac{1}{2}\big(1 - \mathbb{G}_1(0.5)\big)$, the accuracy of an agent with no social information, as in Proposition 2. For intermediate r $\tilde{\alpha}_{\tilde{n}}$ can have several fixed points, and the limiting accuracy could then jump up or fall discontinuously as r grows.



(a) Normal Signals



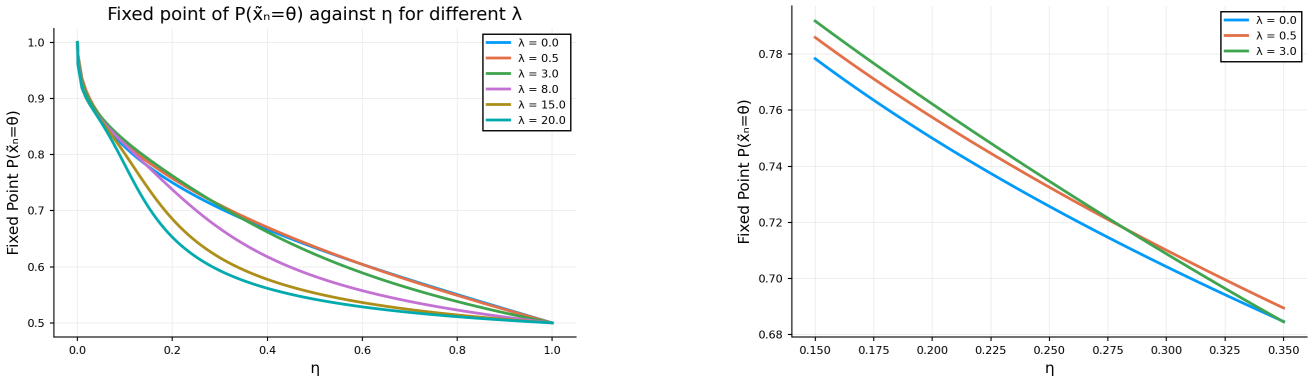
(b) Signals with densities $\{2(1-s), 2s\}$

Figure 4: Graphs for Example 1

Example 1. *In this example, simply varying the private signal distribution and leaving all other parameters unchanged is enough to move from a setting in which (some) reputation concerns harm asymptotic accuracy to one in which it helps. In both instances, assume that agents form an immediate predecessor network (each agent who arrives in the game sees only the agent who most recently acted visibly), all have* $r = 5$*, and* $\eta = 0.2$*. In the first case, plotted in Figure 4a, the private signal distribution is normal, with variance* 1 *and mean* $\mu = \theta$*. In the second, plotted in Figure 4b the probability density functions are* $\{2(1-s), 2s\}$*. As is visible in the graphs, in the first case agents achieve higher asymptotic accuracy without reputation concerns. In the second however, removing reputation concerns reduces their asymptotic accuracy, even though nothing has changed but the signal distributions, and in each case beliefs are unbounded. Whether or not reputation concerns help or hinder depends on the precise parameters assumed.*

Example 2. *In this example we have signals distributed according to densities* $\{2(1-s), 2s\}$*, but* $r - 1$ *follows a Poisson distribution with mean* λ*. The network topology is an immediate predecessor network, and I plot in Figure 5a the fixed point probability such that each visibly-acting agent performs exactly as well as his immediate predecessor. This shows that generally increasing the strength of reputation concerns with a first-order stochastic shift in the reputation-concern distribution reduces asymptotic accuracy for any value of* η*, and that accuracy in turn falls for any reputation-concern distribution as we increase the proportion of noise agents. There is, however, a slight exception to the first of these points in that moving from* $\lambda = 0$ *(where all agents have* $r = 1$*, and thus have no reputation concerns) to* $\lambda = 0.5$ *increases accuracy for a large range of values of* η *(of which a portion is displayed in Figure 5b), and moving on to* $\lambda = 3$ *increases it further for a smaller range.*

It would be nice to establish general solutions that characterise the level of asymptotic accuracy that should result from general classes of parameter assumptions here, but unfortunately the presence of noise agents, reputation concerns and bounded neighbourhoods produces a number of analytical challenges that foil the standard tools of this literature. First of all, since we are interested here in sparse networks, martingale (for nested network topologies) and large sample techniques are of no use. Conventionally,



(a) (b)

Figure 5: Graphs for Example 2

in more general network topologies, one then resorts to improvement principles. These often allow the analyst to establish complete learning in sufficiently connected networks (those that satisfy Expanding Observations, $\lim_{\tilde{n}\to\infty} \mathbb{Q}_{\tilde{n}}\big(\max_{\tilde{b}\in B(\tilde{n})} \tilde{b} < K\big) = 0$ for all $K \in \mathbb{N}$, so that agents' most recent neighbours are eventually arbitrarily late in the sequence of visible agents; see Definition 1 in the appendix), as is still the case when $\eta = 0$ (c.f. Theorem 3). However, the presence of noise agents and reputation concerns together impede this technique here.

Figures 4b and 6 illustrate the problem, and a number of points can be gleaned from them. First of all, notice that improvement reasoning does enable us to observe that social agents acting after a given visible history h^v (whether or not they act visibly) must match the state with higher probability than do the visible agents in their neighbourhood (c.f. Lemma 7 in Appendix B). In Figure 4b, the fact that the purple line (the accuracy of social agents at that history, visible or not) is always strictly above the value of the blue line one step before (this shows the accuracy of visibly-acting agents) follows from this fact. We can also observe that the gradient of the purple line is always the same sign as the gradient of said blue counterpart one step before; this follows from the fact that if one visibly-acting agent matches the state with a strictly higher probability than the one before, his successors will observe a signal (social and private combined) that Blackwell dominates the one he and other agents with his adjusted index received. Hence, they will perform better and the purple line will climb. Having observed that the behaviour of the purple line is very intuitive, however, we can note that the curve in Figure 6(which tracks the accuracy of visible agents conditional upon their being social) is not at all so. The fact that it decreases when moving between $p = \frac{1}{2}$ and $p = 0.74$ (the first and second visible agents in Figure 4b) shows that when moving from a visible history at which agents are observing one signal, to another in which they are observing a *strictly Blackwell better* signal, the probability with which they match the state actually goes down. At first glance this may seem to contradict Blackwell's theorem (alarmingly!), but it does of course not in fact do so. Blackwell's theorem (Blackwell, 1953) asserts that when one signal Blackwell dominates another, this is equivalent to saying that an agent observing the former is better off in their expected utility faced with any decision problem. Here, however, we are not discussing the welfare of a single agent, but the probability with which the first agent who happens to act visibly matches the state (conditional on their

being social in this case). Selecting any specific agent, we can assert from Blackwell's theorem that they will obtain higher expected utility and match the state with higher probability. Additionally, since we are conditioning on the event that the first agent to act visibly is of social type, we are making a statement about the tails of a belief distribution, rather than the welfare that results from acting on the basis of it in expectation. Figure 6 illustrates that providing a Blackwell-better signal does not necessarily increase the probability with which the next agent matches the state, even conditional on their being of social type.

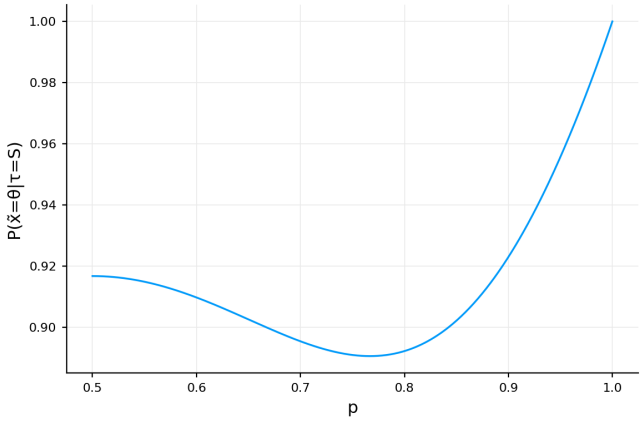


Figure 6: If at visible history h^v agents observe a Bernoulli trial with success probability p in place of a social signal, conditional on the first visible-acting agent being Bayesian, they match the state with probability given by this graph. Parameters as in Figure 4b.

In addition to this, a more informative social signal can increase the probability with which the next visible agent is a noise type. An example of this is shown by Figure 7b in Appendix B, where the probability with which the next visible agent is a noise type increases with the accuracy p of their social signal for low values of p (with antisymmetric private beliefs this cannot happen, c.f. Lemma 5). These two channels can both therefore lead to a given visible agent matching the state with lower probability than their neighbours. Improvement principles are thus no longer effective in this model, which reflects why reputation concerns and noise are together harmful in sparse networks, where improvement principles reflect more closely the nature of the actual reasoning by which Bayesian agents choose their action. In line networks, observing a predecessor and copying them unless one observes a sufficiently strong opposing private signal provides an exact description of what agents do. The more agents one observes, the worse an approximation this becomes.

4 Discussion

Finally, a few extensions merit discussion, to establish to what degree the results of this paper depend on the exact model specification I work with. Before this, however, I discuss my assumption that agents do not update their beliefs in response to observing $\tilde{n}$.

In such a model with endogenous neighbourhoods, social agents will form beliefs over their true indices given the adjusted indices they observe, and could in principle update their beliefs about θ upon observing their adjusted index (I call these 'indexical beliefs'). As I mention in Section 2, this model formally features a dynamic inconsistency I call a *Newcomb tension* that I study further in Cremin (2026): were each agent able to commit themself to a strategy in some planning stage before the game they would choose to act *as if* they held beliefs other than the *Bayesian* beliefs they actually form at their information set themselves.[5] In Cremin (2026) I call the beliefs that induce the action to make the planning-optimal decision *one-boxer beliefs* (Theorem 2.2 of that paper shows under what conditions such beliefs exist, they do here, and Proposition 2.1 (i) establishes that in social learning settings such as this one-boxer beliefs on the state are simply the prior belief), and have written *Bot Got Your Tongue?* assuming that agents form exactly these beliefs. I note another fact that suggests this choice is sensible: Bayesian beliefs are not well defined at the information set with infinitely many agents and a uniform prior over indices unless one is willing to adopt the convention that $\frac{0}{0} = 1$. One alternative response to this point is to instead imagine agents adopting an improper uniform prior over their indices, and forming Bayesian beliefs over the state with it. I consider this in an online appendix (Appendix D), and show that all of the results still hold due to the presence of noise agents, with the exception of Theorem 3 (where by assumption there are no noise agents).

Turning to the robustness of these results to the model specification, I note first of all that though they are convenient for concreteness, the exact assumptions I have made on the nature of the social signal are not necessary for the unravelling results of this paper. Instead it suffices to assume that every agent observes at most $M \in \mathbb{N}$ agents with common knowledge of M, and that the private signals and subset of previous agents seen are independent of θ and n. Instead of working with a model à la Acemoglu et al. (2011), we could follow Smith and Sørensen (2008), where agents observe an unordered, anonymous sample of predecessors, or Monzón and Rapp (2014) where they do not even observe their own adjusted index. Beyond this one could consider other specifications, for example without binary x and v action sets. For my results to hold, it is important that agents cannot tell whether or not their neighbours are noise or social types. If we introduce another level of visibility, for example where agents can make their action observable to only some subset of predecessors and diminish the reward and punishment when seen to act correctly or incorrectly respectively, we would need to specify that noise agents also make this choice with some probability. The same is true if we allow agents to choose from a large set of feasible actions $x_n \in X \supset \{0, 1\}$: we must in turn assume noise agents randomise over that same set of actions X. It is of course true that the failures of learning that reputation concerns can prevent, as per Theorem 1, become less severe the finer the set of available actions, disappearing entirely when this set is continuous. The binary setting is a convenient and simple one in which to illustrate the possible impact of visibility in any case, and extending it to finitely many actions will not similarly render Theorem 1 redundant.

In a similar vein, we do not need to assume that noise agents all act visibly. If they instead act visibly with some probability, this is equivalent to nature selecting noise agents with lower probability. Adjusting

[5] I discuss the Sleeping Beauty literature in further detail in Cremin (2026), but note here that I am following the '*thirder*' consensus of said literature. Some would argue that the planning-optimal beliefs are the correct Bayesian beliefs in any case. Winkler (2017) provides a good review of this literature for interested readers.

noise agent behaviour in a different way, one may also be interested in a model where noise agents all choose the same action $x_n = 1$, as in a disinformation campaign to influence an election for example. I consider this possibility in Online Appendix E and find that most of my results extend. Counterintuitively, in sparse networks such bots would actually do more damage to the state they are trying to promote than to the opposing state.

A more general robustness question on noise agents, however, is to ask what happens if we remove them from the model entirely. I have already discussed the importance of noise for dense networks in Theorem 1, but can establish that the conditions that suffice for complete learning in the standard Bayesian model of Acemoglu et al. (2011) (that the network satisfies expanding observations, defined in Section 3.2, and the signal structure is unbounded) actually continue to guarantee it here in the absence of noise. Hence, the presence of *at least some* noise is important for the results of this paper both in sparse and dense networks:

Theorem 3. *With $\eta = 0$, unbounded signals and expanding observations are sufficient for complete learning in every equilibrium.*

Recalling the logic of the improvement principle behind this result: when expanding observations is satisfied, each agent will be at the end of an arbitrarily long *improvement path* (a chain of agents in which each observes his immediate predecessor in said chain) in the limit, and at each step in any such improvement path the agent must be able to improve upon his predecessor. Acemoglu et al. (2011) then prove that, with unbounded signals, the size of these improvements must be large enough that accuracy converges to 1 along these paths. Here, the presence of reputation concerns simply ensures that each agent must asymptotically be at the end of an arbitrarily long improvement path of visible agents, where the improvement between each successive visible agent is greater than in the model without reputation concerns. To clarify, if in the standard model (without reputation concerns or noise agents) an agent observing a predecessor with accuracy α_b would have an accuracy at least Δ greater than α_b, in the model with reputation concerns (but still no noise agents) we can find a lower bound on the improvement strictly greater than Δ. After any visible history h_n^v, it follows simply from the nature of Bayesian belief formation that the probability with which the next agent (visible or not) matches the state must be lower than the probability that the next agent with a strong enough belief to act visibly will do so; formally this last intuition is proved in Lemma 6.

Naturally this result does not suffice to establish that reputation concerns are completely anodyne. Though expanding observations is a relatively minimal connectivity condition, without which complete learning is impossible with any signal structure and reputation-concern distribution, the literature on social learning does not exclusively consider learning with unbounded signal structures. In the bounded setting, it is always possible to find a reputation-concern distribution that completely prevents learning by assuming that all agents' reputation concerns are at least strong enough that they never act visibly with a social belief of 0.5. In this case, no social agent will ever be the first social type to act, and so knowing this all agents will forever have the social belief 0.5. This reasoning holds with or without noise agents; in their absence no agent ever visibly acts, in their presence exclusively noise agents do.

However, I consider this quite a strong assumption on the reputation-concern distribution (that there is exactly *zero* probability of an agent's candour exceeding a certain point), particularly if the signal structure is not 'very' bounded, i.e. such that the convex hull of the support is some small interval around 0.5. Furthermore, the case of unbounded signals and a network topology satisfying expanding observations is a very important one in such models, since they are jointly sufficient for complete learning in the purely Bayesian case; that reputation concerns alone do not prevent learning also provides a contrast to what we shall see as we introduce noise agents into the model. In at least this prominent instance therefore reputation concerns alone are anodyne.

5 Conclusion

That individuals have a choice whether or not to reveal their actions is of clear relevance in many situations in life involving social learning, yet despite this such models have thus far neglected to model said choice. Particularly in debate, and even more so on social media platforms, that agents may choose to keep their views to themselves is likely to play an important role in the outcome, as would the presence of noise.

Studying a model in which agents only reveal their actions when they are sufficiently confident in them, I discover that the combination of reputation concerns and noise agents will affect learning in different ways in sparse and dense networks. In sparse networks, it can create an unravelling effect that leads to an exaggerated presence for noise agents amongst those who act visibly. With strong enough reputation concerns, social learning can be completely shut down in sparse networks, and with moderate levels it will nonetheless be damaged. As a counterpoint to this, weak reputation concerns can actually help here by silencing very uncertain agents.

In contrast, with dense networks the picture is generally positive. The unravelling that threatens sparse networks does not occur, and for bounded beliefs reputation concerns and noise agents can remedy incomplete learning by eliminating the cascade beliefs that imply it in their absence. This factor depends on both the presence of reputation concerns *and* noise agents, and provides a setting in which, counterintuitively, malicious agents creating bots to harm society will inadvertently help it. More generally, we can observe that the conclusions of the sequential social learning literature should not be taken for granted with endogenous action visibility, which can interact with other features of a learning game in surprising ways.

A Useful Definitions

Definition 1 (Expanding Observations)**.** *A network topology has expanding observations if for all* $K \in \mathbb{N}$*, we have*

$$\lim_{\tilde{n}\to\infty} \mathbb{Q}_{\tilde{n}}\Big(\max_{\tilde{b}\in B(\tilde{n})} \tilde{b} < K\Big) = 0$$

Definition 2 (Antisymmetric Distributions)**.** *A pair of private belief distributions* $\{\mathbb{G}_0, \mathbb{G}_1\}$ *are antisymmetric if for all* $p \in [0,1]$ $\mathbb{G}_0(p) = 1 - \mathbb{G}_1(1-p)$*.*

Definition 3 (Complete Learning)**.** *Complete learning obtains if, for social agents, x_n converges to θ in probability (according to measure $\mathbb{P}_\sigma$) in all equilibria $\varsigma \in \Sigma$, i.e.*

$$\lim_{n\to\infty} \mathbb{P}(x_n = \theta | \tau_n = S) = 1$$

Definition 4 (Birnbaum Spread Birnbaum (1948))**.** *A distribution F' on $[0,1]$ is less peaked about $\frac{1}{2}$ than F if $F'\big(\{p : |p - \frac{1}{2}| \geq z\}\big) \geq F\big(\{p : |p - \frac{1}{2}| \geq z\}\big)$ for all $z \geq 0$; equivalently, F' assigns weakly less mass than F to every interval centred on $\frac{1}{2}$. A pair $\{\mathbb{G}'_0, \mathbb{G}'_1\}$ is a Birnbaum-spread of $\{\mathbb{G}_0, \mathbb{G}_1\}$ if $\mathbb{G}'_\theta$ is less peaked about $\frac{1}{2}$ than $\mathbb{G}_\theta$ for both $\theta \in \{0,1\}$.*

B Useful Lemmas

In this section I state and prove lemmas that will be useful throughout the paper.

Lemma 2 (Acemoglu et al. (2011) Proposition 2)**.** *After any history, h_n, agent n will choose $x_n = 1$ if*

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

and $x_n = 0$ if this sum is strictly less than 1. If they exactly equal 1, the agent is indifferent and assumed to randomise uniformly.

Lemma 3. *If the private belief distributions are antisymmetric, every social belief distribution is also antisymmetric. That is to say, the probability agent n forms belief $sb > 0.5$ if the state is 1 is equal to the probability they form social belief $1 - sb$ if the state is 0.*

Proof. In either state of the world θ, a realisation for the first n private signals directly implies the action choices of these first n agents, and a distribution over all possible choices of visibility (implied by the reputation-concern distribution). If we take these private beliefs, and replace each with the difference between that belief and 1, we will exactly reverse the action choices, but change in no way the probability with which any selection of these agents acts visibly. Observe that with antisymmetric private signals, the probability with which the private beliefs take one profile of realisations in state $\theta = 1$ is identical to the probability they take the difference between those beliefs and 1 if $\theta = 0$. Thus the probability of any sequence of actions and visibility choices in $\theta = 1$ is exactly the same as the probability of the opposite sequence of actions (swapping every 0 with a 1 and vice versa) and the same visibility choices in $\theta = 0$. Hence a social signal which produces social belief sb is exactly as likely in state 1 as one that produces $1 - sb$ in state 0. This means the social belief distributions are antisymmetric whenever private belief distributions are. □

The following lemma justifies my speaking of social signals with a higher 'p' being more Blackwell informative in graphs such as Figure 6.

Lemma 4. *Games with antisymmetric private signal structures are games in which observing only $\tilde{x}_{\tilde{n}}$ is more Blackwell informative if and only if the unconditional probability with which it matches the state is higher.*

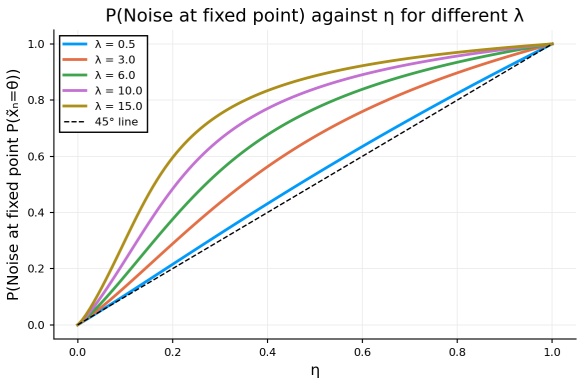
Proof. According to Lemma 3 the social belief distributions are antisymmetric in this case. If one is observing only one action there are only two beliefs that can be generated. By antisymmetry, whichever belief sb is induced by the observation of $\tilde{x}_{\tilde{n}} = 1$, it must be equally likely to form social belief $1 - sb$ when $\theta = 0$ as sb when $\theta = 1$. Hence in each state of the world these are the two possible beliefs, and that belief which supports the true state is necessarily most likely, occurring with the same probability in each state. The unconditional probability of matching the state is then simply the conditional probability of doing so in each state, they are the same. □

Lemma 5 (Monotonic Noise with Antisymmetric Private Beliefs)**.** *Suppose the private signals induce antisymmetric private belief distributions. If we take two social signals, SS_1 and SS_2, in which SS_1 Blackwell dominates SS_2, then the probability with which the visible agent who comments upon observing the signal is a noise agent is lower for SS_1 than SS_2*

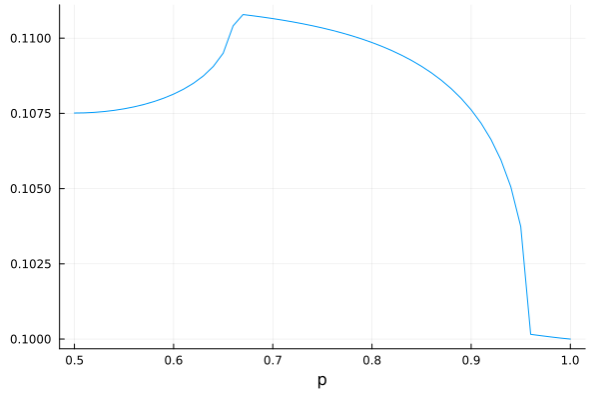
Proof. If SS_1 dominates SS_2, then the information set of SS_1 and a private signal Blackwell dominates the information set of SS_2 and a private signal by Blackwell (1951, Theorem 12). This means the unconditional (on the state) distribution of posteriors induced by SS_1 and the private signal is a mean-preserving spread of the set induced by SS_2 (Blackwell, 1953, Theorem 2).[6] Therefore there will be a higher unconditional probability of those beliefs being in the extremes demarcated by the thresholds in Lemma 1 for SS_1 than SS_2. Using Lemma 3, we know that the fact that the private belief distributions are antisymmetric means that every possible social belief $sb > 0.5$ occurs with the same conditional probability when $\theta = 1$ as $1 - sb$ does when $\theta = 0$. The probability of a private signal that, in $\theta = 1$ and combined with sb or $1 - sb$, induces an overall belief within the invisible action threshold, is the same as the probability of one that in $\theta = 0$ induces an invisible belief when combined with $1 - sb$ and sb respectively. Therefore, the conditional probabilities of an invisible signal induced by SS_1 in $\theta = 0$ and $\theta = 1$ are equal (and the same is true for SS_2). For a given agent, these facts imply probabilities $\mathbb{P}(Silent|SS_1)$ and $\mathbb{P}(Silent|SS_2)$ where the former is larger than the latter. The probability the next visible agent will be a noise agent is:

$$\begin{aligned}
&\eta + (1-\eta) \times \mathbb{P}(Silent|SS) \times \eta + \Big((1-\eta) \times \mathbb{P}(Silent|SS)\Big)^2 \times \eta + ... \\
&= \eta \sum_{j=0}^{\infty} \Big((1-\eta) \times \mathbb{P}(Silent|SS)\Big)^j \\
&= \frac{\eta}{1 - \Big((1-\eta) \times \mathbb{P}(Silent|SS)\Big)}
\end{aligned} \tag{B.1}$$

[6] This theorem states that Blackwell dominance implies a mean-preserving spread on the experimental outcome space. To apply this to posterior beliefs, simply define a new set of Blackwell experiments that map directly to the posterior belief space (applying the original probability mapping, and then Bayesian belief updating).



(a) Antisymmetric private signal pdf $\{2(1-s), 2s\}$ and shifted Poisson reputation concerns ($r-1$ distributed according to a Poisson distribution with mean λ).



(b) This plots the probability with which the next visible-acting agent is a noise type, if their social signal is a Bernoulli trial with success probability p, and the parameters are as set out below Lemma 5. It is clearly non-monotonic.

Figure 7: The right panel shows that without antisymmetric private belief distributions the probability with which an agent is a noise agent is not necessarily monotonic. The left panel shows a case with antisymmetric private beliefs, in which the presence of this monotonicity gives that for every η, asymptotic noise is increasing in reputation concerns, and for all reputation-concern distributions it is also increasing in η.

Since $\mathbb{P}(Silent|SS_1) \geq \mathbb{P}(Silent|SS_2)$, the above expression is smaller for SS_1 than SS_2. This completes the proof. □

Note that this is not true for more general private signal distributions. For example, if we have a line network with $\eta = 0.1$, a r-distribution that assigns 50% probability to 1 and 3.5 each, and normally distributed signals $N(-0.15, 0.1^2)$ and $N(0.15, 0.6^2)$ in states 0 and 1 respectively, then the probability with which an agent is a noise agent is non-monotonic as seen in Figure 7b.

A consequence of Lemma 5 is that if the private belief distributions are antisymmetric, we can compute an upper bound for the probability with which a visibly acting agent is a noise agent (as opposed to $\underline{\eta}_\infty$ in the text, which is a lower bound) for any network topology, but for a given reputation-concern distribution and private signal structure, by computing this probability for $p = 0.5$ (when the social signal is just noise).

Corollary 3.1 (Corollary to Lemma 5)**.** *If the private belief structure is antisymmetric, the probability with which any visibly acting agent $\tilde{n}$ is a noise agent can be bounded above by:*

$$\mathbb{P}(\tilde{\tau}_{\tilde{n}} = N) \leq \frac{\eta}{1-(1-\eta)\int_1^\infty \mathbb{G}_1(\frac{r}{r+1}) - \mathbb{G}_1(\frac{1}{r+1})d\Delta(r)}$$

Proof. As explained above, Lemma 5 gives us that this probability is decreasing in the informativeness of the social signal, and thus at its maximum when the social signal is pure noise. When this is so, the social belief it produces is 0.5. Given this, the overall belief of the agent is simply their private belief

and we can directly apply Lemma 1 to the private belief. This gives us the visible-action thresholds, and agents will choose to act invisibly if they observe a private signal within this range for a given r. Then we need only integrate across r, and substitute this probability into Equation B.1. That the distributions are antisymmetric means that it does not matter which distribution we use (that for $\theta = 0$ or that for $\theta = 1$). □

For example, if all agents have the same reputation-concern parameter and the density functions are $\{2(1-s), 2s\}$, the upper bound probability with which an agent is a noise agent can be computed as follows:

$$\left(\frac{r}{r+1}\right)^2 - \left(\frac{1}{r+1}\right)^2 = \frac{r-1}{r+1}$$

Substituting this into our expression from Corollary 3.1

$$\mathbb{P}(\tilde{\tau}_{\tilde{n}} = N) \leq \frac{\eta}{1-(1-\eta)\frac{r-1}{r+1}}$$

If $r = 1$ this is simply η, as one would hope. If $r = 2$ it becomes $\frac{3\eta}{2+\eta}$, and as $r \to \infty$ it converges to 1 for any η.

The following lemma will be useful in the proof of Theorem 3.

Lemma 6 (Strong beliefs imply high accuracy)**.** *Suppose that we refer to the event that an agent n acting at visible history h^v forms belief weakly stronger than $\frac{r}{r+1}$ that the true state is θ for either $\theta \in \{0, 1\}$ as 'Strong'. Then it follows that $\mathbb{P}(x_n = \theta | Strong) \geq r/r + 1 > \mathbb{P}(x_n = \theta | \neg Strong)$.*

Proof. This statement is effectively an elaborate restatement of the definition of a belief, where in this paper all beliefs are defined as probabilities the agent assigns to the event $\theta = 1$. An agent's *'Belief that $\theta = 1$'* $= \mathbb{P}(\theta = 1 | Belief)$. (A similar tautology is also used to prove Lemma A1 in Acemoglu et al. (2011).) Similarly their *'Belief that $\theta = 0$'* $= \mathbb{P}(\theta = 0 | Belief)$. Let us define *Belief* as the agent's belief that the true state is $\theta = 1$. Since an agent will choose x_n to be whichever state they consider most likely, we have that $\mathbb{P}(x_n = \theta | Belief) = \max\{\mathbb{P}(\theta = 0 | Belief), \mathbb{P}(\theta = 1 | Belief)\} \geq 0.5$.

Using the definition of '*Strong*', we can observe that

$$\max\{\mathbb{P}(\theta = 0 | Belief, Strong), \mathbb{P}(\theta = 1 | Belief, Strong)\} \geq \frac{r}{r+1}.$$

Thus it follows that $\mathbb{P}(x_n = \theta | Strong) \geq \frac{r}{r+1}$. Similarly conditioning on the event $\neg Strong$ gives us that $\max\{\mathbb{P}(\theta = 0 | Belief, \neg Strong), \mathbb{P}(\theta = 1 | Belief, \neg Strong)\} < \frac{r}{r+1}$, which gives the second half of our inequality. □

Lemma 7. *Suppose we are at visible history h^v, where the action of agent $\tilde{n}$ can be observed; the following four statements concerning the success probabilities of agents $\tilde{n}(h^v)$ and $n(h^v)$ are true:*

1. $\mathbb{P}(x_{n(h^v)} = \theta | \tau_{n(h^v)} = S) \geq \mathbb{P}(\tilde{x}_{\tilde{n}} = \theta)$

2. $\mathbb{P}(\tilde{x}_{\tilde{n}(h^v)} = \theta | \tilde{\tau}_{\tilde{n}(h^v)} = S) \geq \mathbb{P}(x_{n(h^v)} = \theta | \tau_{n(h^v)} = S)$

3. $\mathbb{P}(\tilde{x}_{\tilde{n}(h^v)} = \theta | \tilde{\tau}_{\tilde{n}(h^v)} = S) \geq \mathbb{P}(x_{n(h^v)} = \theta)$

4. $\mathbb{P}(\tilde{x}_{\tilde{n}(h^v)} = \theta | \tilde{\tau}_{\tilde{n}(h^v)} = S) \geq \mathbb{P}(\tilde{x}_{\tilde{n}} = \theta)$

Proof. Point 1 is information monotonicity and follows from basic improvement reasoning (Acemoglu et al., 2011, Lemma A2), it says that a social observer of the $\tilde{n}$th visible agent must do better than them (this says and assumes nothing about whether or not the observer then chooses to act visibly). Point 2 follows from Lemma 6. It states that conditional on choosing to act (or conditional on the agent that chooses to act being social, which is equivalent) a social agent must have a higher accuracy than one (another social agent that is) who does not necessarily choose to act. Point 3 observes that since a social agent must do better than a noise agent at a given visible history (which is just a specific information set before observing private signals), and since point 2 implies that a social visibly-acting agent following $\tilde{n}$ must do better than social agents at the same information set not necessarily acting visibly, then the accuracy of the agent acting at that visible history conditional on their being social must be higher than agents of arbitrary type and visibility. Point 4 then follows from combining points 1 and 2, and tells us that the curve in diagrams such as Figure 6 (that plot the success probability of a visible agent in an immediate predecessor conditional on their being social) must dominate the 45-degree line i.e. that information monotonicity holds conditional on no noise agents being drawn (which makes sense given Theorem 3, although they are not identical situations since in Theorem 3 all agents know that no agent is a noise agent). Hence whether or not this curve is monotonic, it will converge to 1 as the accuracy of the social signal increases to 1.[7] □

Lemma 8 ($\underline{\eta}_\infty$ Comparative Statics)**.** *$\underline{\eta}_\infty$ is increasing in η, increasing with First order stochastic shifts in $\Delta(r)$, decreasing in M, and decreasing with any 'Birnbaum-spread' about 0.5 (c.f. Definition 4) in $\mathbb{G}_0$ and $\mathbb{G}_1$.*

Now of course, since these are the comparative statics of a bound, they do not necessarily reflect the comparative statics of the actual asymptotic fractions of agents who are noise agents. The use of this lemma is rather in deriving the Propositions 2 and 3.

[7] It is easy to speak of the accuracy of a given agent converging to one as the accuracy of the social signal increases to 1 in a line network with an antisymmetric private belief structure, as then the accuracy of the social signal is well defined as the probability with which it (a symmetric Bernoulli trial) matches the state. Beyond this, however, the statement intuitively holds for any signal structure and social information set. To formalise this, one could consider the maximally and minimally informative symmetric Bernoulli trials that an agent confronted with the decision problem of this paper would disprefer and prefer respectively to a given social signal A, with success probabilities $\underline{p}$ and $\overline{p}$. If then we pick a series of social signals such that both of these probabilities are converging to 1, the probability a social agent acting at this visible history will correctly match the state will then also converge to 1.

C Omitted Proofs

C.1 Decision Rule Results

Proof of Lemma 1. To maximise their expected utility, an agent will choose $v_n = 1$ if they believe the true state is $\theta = 1$ either with probability strictly greater than $\frac{r}{1+r}$ or strictly less than $\frac{1}{1+r}$. Suppose without loss that $\mathbb{P}(\theta = 1) \geq \frac{1}{2}$, a symmetric argument will apply otherwise. We can derive the desired right-hand side inequality as follows.

An agent will choose $v_n = 1$ upon observing information set I_n iff.:

$$\mathbb{P}(\theta = 1|I_n) > \frac{r}{1+r}$$

The first steps proving this statement are the same as for Lemma 2, which is simply Acemoglu et al. (2011, Proposition 2), except that we have the conditions that $\mathbb{P}(\theta = 1|I_n) > \frac{r}{1+r}$ rather than $\mathbb{P}(\theta = 1|I_n) > \frac{1}{2}$. Following these steps one arrives at:

$$d\mathbb{P}_\sigma(I_n|\theta = 1) > d\mathbb{P}_\sigma(I_n|\theta = 0)r$$

Using the fact that $d\mathbb{P}_\sigma(I_n|\theta = j) = d\mathbb{P}_\sigma(p_n|\theta = j)\mathbb{P}_\sigma(B(\tilde{n})|\theta = j)$, this becomes:

$$\begin{aligned}&\Big[d\mathbb{P}_\sigma(p_n|\theta = 1) + d\mathbb{P}_\sigma(p_n|\theta = 0)\Big]\mathbb{P}_\sigma(B(\tilde{n})|\theta = 1)\\ &> d\mathbb{P}_\sigma(p_n|\theta = 0)\mathbb{P}(B(\tilde{n})|\theta = 1) + d\mathbb{P}_\sigma(p_n|\theta = 0)\mathbb{P}_\sigma(B(\tilde{n})|\theta = 0)r\end{aligned}$$

Now we divide by $\Big[d\mathbb{P}_\sigma(p_n|\theta = 1) + d\mathbb{P}_\sigma(p_n|\theta = 0)\Big]$, apply Bayes' Rule (using again that the common prior is $\frac{1}{2}$), divide by $\mathbb{P}_\sigma(B(\tilde{n})|\theta = 1) + \mathbb{P}_\sigma(B(\tilde{n})|\theta = 0)$ before using Bayes' rule again to get:

$$\mathbb{P}_\sigma(\theta = 1|B(\tilde{n})) > d\mathbb{P}_\sigma(\theta = 0|p_n)\mathbb{P}_\sigma(\theta = 1|B(\tilde{n})) + d\mathbb{P}_\sigma(\theta = 0|p_n)\mathbb{P}_\sigma(\theta = 0|B(\tilde{n}))r$$

With further algebra this yields:

$$S_n > 1 + \frac{r-1}{r} d\mathbb{P}_\sigma(\theta = 1|p_n)d\mathbb{P}_\sigma(\theta = 1|B(\tilde{n}))$$

□

C.2 Proofs for Section 3.1

Proof of Theorem 1. This proof uses standard tools from Smith and Sørensen (2000). Without loss, suppose the true state is $\theta = 1$, and define the likelihood ratio

$$\ell_{\tilde{n}} = \frac{1 - sb_{\tilde{n}}}{sb_{\tilde{n}}}.$$

Let $\phi(sb, \theta)$ denote, in state θ, the probability that the first visibly-acting agent at a history with social belief sb chooses action 1. For each part of the statement, three steps provide the proof:

1. By the Martingale Convergence Theorem (Breiman, 1968, Theorem 5.14), ℓ almost surely does not converge to infinity (*Incorrect Learning* a.s. does not obtain).

2. By Smith and Sørensen (2000, Theorem B.1), since this is a *Markov-Martingale System* (which Smith and Sørensen describe immediately before they present this theorem), ℓ almost surely converges to a stationary social belief; i.e., one such that $\phi(sb, 0) = \phi(sb, 1)$, and almost surely *not* to any other point.

3. Under the condition of each part, this condition is only satisfied at $\ell = 0$ ($sb = 1$), therefore ℓ almost surely converges to this value. This establishes that complete learning obtains.

It remains only to verify $\phi(sb, 1) > \phi(sb, 0)$ for all interior social beliefs in each part.

For a social agent with reputation concern r, the private belief space divides into three regions. By Lemma 1, the agent acts invisibly when $p \in (p^{**}(r), p^{*}(r))$, and visibly otherwise. By Lemma 2, the agent chooses $x_n = 1$ if $p > 1 - sb$ and $x_n = 0$ if $p < 1 - sb$. Since $p^{**}(r) \leq 1 - sb \leq p^{*}(r)$, these two rules combine as follows: if $p > p^{*}(r)$ the agent is *visible and chooses* $x_n = 1$; if $p \in (p^{**}(r), p^{*}(r))$ the agent is *invisible*; and if $p < p^{**}(r)$ the agent is *visible and chooses* $x_n = 0$. A noise agent always acts visibly and chooses $x_n = 1$ with probability $\frac{1}{2}$. By Lemma 1:

$$p^{*}(r) = \frac{r\big(1 - sb\big)}{sb + r\big(1 - sb\big)}, \qquad p^{**}(r) = \frac{1 - sb}{1 + (r - 1)sb}.$$

Note that p^{*} is continuous and strictly increasing in r, from $p^{*}(1) = 1 - sb$ to $p^{*}(r) \to 1$ as $r \to \infty$; while p^{**} is continuous and strictly decreasing, from $p^{**}(1) = 1 - sb$ to $p^{**}(r) \to 0$. Define the average visible-right and visible-left probabilities, integrated over the type distribution:

$$R_\theta := \int \big[1 - \mathbb{G}_\theta\big(p^{*}(r)\big)\big] d\Delta(r), \qquad L_\theta := \int \mathbb{G}_\theta\big(p^{**}(r)\big) d\Delta(r).$$

Then

$$\phi\big(sb, \theta\big) = \frac{\eta/2 + (1 - \eta)R_\theta}{\eta + (1 - \eta)(R_\theta + L_\theta)}. \tag{C.1}$$

Write $\phi = h(R, L)$ where $h(R, L) := \frac{\eta/2 + (1-\eta)R}{\eta + (1-\eta)(R+L)}$. Since $\eta > 0$, h is strictly increasing in R and strictly decreasing in L. By Acemoglu et al. (2011, Lemma A1(c)), $\mathbb{G}_1$ first-order stochastically dominates $\mathbb{G}_0$ ($\mathbb{G}_1(x) \leq \mathbb{G}_0(x)$ for all x, with strict inequality on the interior of the support of $\mathbb{G}_\theta$), giving pointwise in r:

$$1 - \mathbb{G}_1(p^{*}(r)) \geq 1 - \mathbb{G}_0(p^{*}(r)), \qquad \mathbb{G}_1(p^{**}(r)) \leq \mathbb{G}_0(p^{**}(r)),$$

with strict inequality whenever $p^{*}(r)$ or $p^{**}(r)$ lies in the interior of the support where FOSD is strict. Integrating over r: $R_1 \geq R_0$ and $L_1 \leq L_0$, and by the monotonicity of h, $\phi(sb, 1) > \phi(sb, 0)$ whenever at

least one of these inequalities is strict. It suffices to verify this strictness under the hypotheses of each part.

Part 1. With unbounded private beliefs, Acemoglu et al. (2011, Lemma A1(c)) gives $\mathbb{G}_1(x) < \mathbb{G}_0(x)$ for all $x \in (0,1)$. Since $p^*(r), p^{**}(r) \in (0,1)$ for every $r \geq 1$ and every interior sb, the inequalities are strict for every r: $R_1 > R_0$ and $L_1 < L_0$, giving $\phi(sb,1) > \phi(sb,0)$. No interior sb is stationary.

Part 2. With bounded private beliefs supported on $[\underline{p}, \overline{p}] \subset (0,1)$, Acemoglu et al. (2011, Lemma A1(c)) gives $\mathbb{G}_1(x) < \mathbb{G}_0(x)$ for $x \in (\underline{p}, \overline{p})$, with $\mathbb{G}_1 = \mathbb{G}_0$ outside this interval. I show that for every interior sb, at least one threshold $p^*(r)$ or $p^{**}(r)$ lies in $(\underline{p}, \overline{p})$ for a positive-Δ-measure set of reputation-concern values. Define

$$c^* := \frac{\underline{p}(1-\overline{p})}{\overline{p}(1-\underline{p})}.$$

The hypothesis $(0, c^* + \epsilon) \cup (1-\epsilon, 1) \subseteq \text{supp}(\Delta(c))$ for some $\epsilon > 0$ is used as follows.

- $sb \in (1-\overline{p}, 1-\underline{p})$*:* $p^*(1) = 1 - sb \in (\underline{p}, \overline{p})$. By continuity, $p^*(r) \in (\underline{p}, \overline{p})$ for all r in a neighbourhood of 1, i.e. for $c = 1/r$ in a neighbourhood of 1. Since $(1-\epsilon, 1) \subseteq \text{supp}(\Delta)$, these reputation-concern values carry positive mass.

- $sb \geq 1 - \underline{p}$*:* $p^*(1) = 1 - sb \leq \underline{p}$, but $p^*(r) \to 1$, so as r increases $p^*(r)$ enters $(\underline{p}, \overline{p})$. The set of reputation-concern values with $p^*(r) \in (\underline{p}, \overline{p})$ corresponds to an open interval of c-values with left endpoint

  $$c_{\min}(sb) = \frac{(1-sb)(1-\overline{p})}{\overline{p} \cdot sb},$$

  which is at most c^* (attained at $sb = 1 - \underline{p}$) and tends to 0 as $sb \to 1$. Since $(0, c^* + \epsilon) \subseteq \text{supp}(\Delta)$ and $c_{\min} \leq c^*$, the interval of reputation-concern values which produce visible and invisible action with strictly positive probability always intersects the support.

- $sb \leq 1 - \overline{p}$*:* Here we can make a symmetric argument applied to $p^{**}(r)$, which enters $(\underline{p}, \overline{p})$ from above as r increases. The left endpoint of the corresponding c-interval is again at most c^* (attained at $sb = 1 - \overline{p}$), so the same condition $(0, c^* + \epsilon) \subseteq \text{supp}(\Delta)$ provides the needed positive-measure interval of reputation-concern values such that agents will act visibly and invisibly with strictly positive probability.

In each case, strict FOSD at the relevant threshold for a positive-measure set of reputation-concern values gives $R_1 > R_0$ or $L_1 < L_0$, and hence $\phi(sb,1) > \phi(sb,0)$. No interior sb is stationary, so the social belief converges to certainty on the true state almost surely. □

C.3 Proofs for Section 3.2

Proof of Theorem 2. Suppose agent n observes M others, and suppose each of these chooses action 1. Further suppose that at this point in the game social agents are informed that their actions are perfectly informative. This is the strongest possible signal an agent can observe in favour of $\theta = 1$, and thus

establishes the bound, though one could go through the basic reasoning of the following argument for any social signal, since for each possible social signal there is some probability it was produced by observing only noise agents. Using Bayes' rule, the probability that upon observing all 1s one's entire neighbourhood is comprised of 1-noise agents, we can compute that:

$$\mathbb{P}(\theta = 1 | M \text{ 1's observed}) = \frac{(1-\frac{\eta}{2})^M}{(1-\frac{\eta}{2})^M + (\frac{\eta}{2})^M}$$

This provides an upper bound on the social beliefs agent n can form, and symmetrically we can compute a bound on how strong they can be in favour of $\theta = 0$. Given this, for any private signal, we can compute a range of possible overall beliefs, and if r_n is high enough, it may be that for some private beliefs close to 0.5 the entire range of possible overall beliefs would lead to an agent choosing to act invisibly $v_n = 0$. Specifically, using Lemma 1 we can see that private signals within the following interval certainly produce invisible actions, if r_n is large enough for it to be nonempty:

$$\mathbb{P}(\theta = 1 | s_n) \in [\underbrace{1 - \frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{L_0}, \underbrace{\frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{U_0}]$$

In each state of the world, therefore, the probability that an agent receives a private belief within this region gives an i.i.d lower bound on the probability the will choose to act invisibly. Let us define r_0^* as the value of r such that $U_0 = L_0$, noting that U_0 is increasing in r_n and L_0 is decreasing in it, so for all lower r values the interval of interest is empty, and for all higher r values it is non-empty. The lower bound probability can then be expressed:

$$\lambda_0 = \min \left\{ \int_{r_0^*}^{\infty} \mathbb{G}_0(U_0(r)^-) - \mathbb{G}_0(L_0(r)) d\Delta(r), \int_{r_0^*}^{\infty} \mathbb{G}_1(U_0(r)^-) - \mathbb{G}_1(L_0(r)) d\Delta(r) \right\}$$

Since neighbours are drawn only from the $v_n = 1$ pool, if one's neighbourhood is drawn from agents with at most social beliefs in the region $[\underline{SB}, \overline{SB}]$ the type distribution of one's neighbourhood is no longer $\eta, 1-\eta$, since at most $(1-\lambda_0)$ of the social agents choose visible actions. Instead, the relevant distribution is one with lower bound noise-probability $\underline{\eta}_1 := \eta / \big(\eta + (1-\lambda_0)(1-\eta)\big)$.

This reasoning, however, can be iterated. The upper and lower bounds on the social beliefs of an agent with such a neighbourhood are tighter since the neighbours they observe are more likely to be noise agents. Thus if an agent's neighbourhood is drawn from a set of agents whose social beliefs are necessarily within $[\underline{SB}_1, \overline{SB}_1]$, we can define a new private belief interval $[L_1, U_1]$, a new λ_1, and thus a $\underline{\eta}_2$, all by substituting $\underline{\eta}_1$ in place of η. Iterating this reasoning infinitely many times, and indexing the rounds by j, at each stage we get an at least weakly tighter range of social beliefs, and a weakly higher $\underline{\eta}_j$. Figure 2 shows a homogenous-r example; the region of private beliefs that must produce an invisible action can be seen to expand, and correspondingly the range of possible social beliefs shrinks. Figure 3 shows an example displaying this reasoning for general r distributions.

Taking $\underline{\eta}_j$, this is clearly a weakly increasing and bounded sequence (since η is a probability) and thus

converges. The limit of this process, η_∞, then equals $\underline{\eta}_\infty = \frac{\eta}{\eta+(1-\lambda_\infty)(1-\eta)}$.

The reasoning in the preceding paragraphs clearly establishes that for any set of agents whose social beliefs are within $[\underline{SB}_\infty, \overline{SB}_\infty]$, the probability that the first to visibly act is a noise agent is at least $\underline{\eta}_\infty$. The Strong Law of Large Numbers then gives that: (1) the asymptotic fraction of social agents choosing invisible actions is almost surely at least λ_∞, since private signals are independent and identically distributed (the SLLN is applied not to the visibility decisions, which are endogenous, but to the indicators that each agent's private belief lies in $(L_\infty(r_n), U_\infty(r_n))$: these depend only on that agent's private signal and reputation concern, so are i.i.d. across agents conditional on θ, and each implies $v_n = 0$ whatever the history), and (2) the asymptotic fraction of agents who are of noise type is almost surely η, since types are also independent and identically distributed. Thus the asymptotic fraction of agents who are social and act visibly is at most $(1-\lambda_\infty)(1-\eta)$, and the asymptotic fraction of agents who are noise agents is η. The remainder are social agents who act invisibly, thus the fraction of visibly-acting agents who are noise types is at least $\eta/\big((1-\lambda_\infty)(1-\eta)+\eta\big) = \underline{\eta}_\infty$.

That completes the proof of the first part of our theorem, what remains is the characterisation of λ_∞ in terms of the parameters of our model. For this, let us substitute $\underline{\eta}_j$ in place of η in the interval derived above, and write $\omega_j := \big(\underline{\eta}_j/(2-\underline{\eta}_j)\big)^M$, then we can express the bounds of our interval thus:

$$L_j(r) = \frac{1}{1+r\omega_j}, \qquad U_j(r) = \frac{r\omega_j}{1+r\omega_j}, \qquad r_j^* = \frac{1}{\omega_j}$$

Everything in round j depends only on ω_j, and since $\underline{\eta}_j = \eta/\big(\eta+(1-\lambda_{j-1})(1-\eta)\big)$, we can write ω_j directly in terms of λ_{j-1}:

$$\begin{aligned}\omega_j &= \left(\frac{\underline{\eta}_j}{2-\underline{\eta}_j}\right)^M = \left(\frac{\eta}{2\big(\eta+(1-\lambda_{j-1})(1-\eta)\big)-\eta}\right)^M \\ &= \left(\frac{\eta}{2-\eta-2\lambda_{j-1}(1-\eta)}\right)^M =: \Omega(\lambda_{j-1})\end{aligned}$$

Then combining this with the iterated-reasoning above we can express λ_j as the following recursion in terms of $\Psi(\lambda_{j-1})$:

$$\begin{aligned}\lambda_j =& \Psi(\lambda_{j-1}) := \\ &\min\left\{\int_{1/\Omega(\lambda_{j-1})}^{\infty} \mathbb{G}_0\left(\left(\frac{r\Omega(\lambda_{j-1})}{1+r\Omega(\lambda_{j-1})}\right)^-\right) - \mathbb{G}_0\left(\frac{1}{1+r\Omega(\lambda_{j-1})}\right) d\Delta(r),\right. \\ &\left.\int_{1/\Omega(\lambda_{j-1})}^{\infty} \mathbb{G}_1\left(\left(\frac{r\Omega(\lambda_{j-1})}{1+r\Omega(\lambda_{j-1})}\right)^-\right) - \mathbb{G}_1\left(\frac{1}{1+r\Omega(\lambda_{j-1})}\right) d\Delta(r)\right\}\end{aligned}$$

Ψ is non-decreasing; Ω is increasing in λ, since $2-\eta-2\lambda(1-\eta)$ is decreasing in λ and bounded below by $\eta > 0$; a higher ω then increases $U_j(r)$, decreases $L_j(r)$, and decreases r_j^*, i.e. both integrals in the statement are non-decreasing in λ; and the minimum of two non-decreasing functions is itself non-decreasing.

Since Ψ also maps $[0,1]$ into itself, and $\lambda_0 = \Psi(0) \geq 0$, the sequence $\{\lambda_j\}_{j\in\mathbb{N}}$ is non-decreasing and bounded above by 1, which is the convergence already noted for η. Its limit must be a solution to $\lambda_\infty = \Psi(\lambda_\infty)$: as $\lambda_j \uparrow \lambda_\infty$ we have $U_j(r) \uparrow U_\infty(r)$ and $L_j(r) \downarrow L_\infty(r)$ for every r, so monotone convergence applies to each integral.

Finally, λ_∞ must be the smallest solution (there could in principle be more than one): since Ψ is non-decreasing, an iteration starting from 0 can never pass a fixed point λ^*, as $\lambda_j \leq \lambda^*$ implies $\lambda_{j+1} = \Psi(\lambda_j) \leq \Psi(\lambda^*) = \lambda^*$. If some $\lambda^* \in [0,1]$ satisfies $\lambda^* = \Psi(\lambda^*)$, then $\lambda^* \geq 0$, and applying Ψ to both sides j times we have $\lambda^* = \Psi^j(\lambda^*) \geq \Psi^j(0) = \lambda_{j-1}$ for every $j \in \mathbb{N}$. Letting $j \to \infty$ gives $\lambda^* \geq \lambda_\infty$.

□

Proof of Proposition 1. If there are values of r strictly greater than r_0^* in the support of the r-distribution, there is some interval $[L_0(r), U_0(r)]$ associated with each of them around 0.5. Since there is some interval $(0.5-\delta, 0.5+\delta)$ included within the support of the belief distribution, then for each r $[L_0(r), U_0(r)] \cap (0.5-\delta, 0.5+\delta)$ is also within the support of the private signal distribution.

The region $\bigcup_{r>r_0^*} \left([L_0(r), U_0(r)] \cap (0.5-\delta, 0.5+\delta)\right)$ is then a subset of the region of pairs of reputation-concern parameters and private beliefs that produce invisible actions. Integrating across this region then produces a strictly positive probability that is strictly below λ_0. Hence $\lambda_0 > 0$ and the result follows. □

After this proposition, I note that additional conditions can guarantee that $\underline{\eta}_j > \underline{\eta}_{j-1}$ for any j. The following reasoning establishes these claims.

With a full support r-distribution, for any p within the support of the private belief distribution, there is an interval of r-values $[L_0^{-1}(p), \infty)$ or $[U_0^{-1}(p), \infty)$ (the former if p is below 0.5, the latter if above) such that the agent will choose $v_n = 0$ if they observe private signal p and have a r_n within this interval. The integral of these probabilities across the support of p is at least $\lambda_1 > 0$: (1) This is *at least* λ_1 since we have not specified the true state of the world, as λ_1 is the minimum value of the probability of a (p_n, r_n) pair that gives $v_n = 0$ in either of the two states of the world (2) This is strictly greater than zero as for any value in the support of the private signal there is a strictly positive probability the agent has a r value that implies $v_n = 0$, since we have assumed our r distribution is full support.

The fact that $\lambda_1 > 0$ implies that for any p, $U_1^{-1}(p) < U_0^{-1}(p)$ and $L_1^{-1}(p) > L_0^{-1}(p)$. This in turn implies we are integrating our private signal and r-distributions over a strictly larger region to obtain λ_2, thus obtaining a strictly larger value for λ_2: $\lambda_2 > \lambda_1$. This implies in turn that $\underline{\eta}_2 > \underline{\eta}_1$. The same reasoning holds for any $j \in \mathbb{N}$, thus $\underline{\eta}_j > \underline{\eta}_{j-1}$ for any $j \in \mathbb{N}$.

If the private beliefs distribution is full-support, then whatever r-values above r_0^* are contained within the support of $\Delta(r)$, each iteration of the reasoning must increase the interval of invisible-action private beliefs. Since the distribution of private beliefs is full support, this must imply that invisible actions are chosen with a strictly higher probability each round $\lambda_j > \lambda_{j-1}$ for all $j \in \mathbb{N}$. Thus $\underline{\eta}_j > \underline{\eta}_{j-1}$.

If the private beliefs distribution is unbounded, and the r-distribution satisfies the condition specified in the third sufficient condition, then even though mild private beliefs are not in the support there are r values in the support that can bring an agent with social belief 0.5 back into the invisible-action belief region (the r condition is defined to guarantee this). For agents with stronger social beliefs in either

direction, since the $\underline{p}$ and $\overline{p}$ values are the values beyond which all private beliefs have positive support, there is necessarily also a positive probability that one of the private beliefs of sufficient strength to push the agent's belief back into the invisible action region will be selected. The conclusion follows from this as in the previous two cases.

Proof of Lemma 8. We have the relation:

$$\underline{\eta}_{\infty}\Big(1-(1-\eta)\lambda_{\infty}(\eta^{*}_{\infty})\Big)=\eta$$

λ_{∞} is the minimum of two probabilities (one for each value of θ) that we have a private belief–reputation-concern pair in the invisibility region:

$$\lambda_{\infty}=\min\left\{\int_{r^{*}_{\infty}}^{\infty}\mathbb{G}_0(U_{\infty}(r)^{-})-\mathbb{G}_0(L_{\infty}(r))d\Delta(r),\int_{r^{*}_{\infty}}^{\infty}\mathbb{G}_1(U_{\infty}(r)^{-})-\mathbb{G}_1(L_{\infty}(r))d\Delta(r)\right\}$$

In general the left-hand side of our first relation is increasing in $\underline{\eta}$ so there is clearly a unique solution for $\underline{\eta}$, which is increasing in the right-hand side, η. Secondly, the left-hand side is decreasing in λ_{∞}, so a higher λ_{∞} implies a higher $\underline{\eta}$ for the equation to balance. The comparative statics are then implied by the impact of each variable on λ_{∞}.

- For higher values of r, $\mathbb{G}_{\theta}(U_{\infty}(r)^{-})-\mathbb{G}_{\theta}(L_{\infty}(r))$ necessarily takes a higher value for either $\theta\in\{0,1\}$. It follows that shifting probability mass to higher r values necessarily increases λ_{∞}, and thus $\underline{\eta}$. Therefore a FOSD shift in $\Delta(r)$ exacerbates the problem and increases the representation of noise agents.

- Increasing M necessarily decreases $U_{\infty}(r)$ and increases $L_{\infty}(r)$ for all levels of r, so it reduces λ_{∞}, and then reduces $\underline{\eta}$.

- Also, the more mass the belief distributions assign to central values (value within $U_{\infty}(r)$ and $L_{\infty}(r)$), the higher λ_{∞}. A Birnbaum-spread in $\mathbb{G}_0$ or $\mathbb{G}_1$ achieves this as stated in Definition 4, and thus this reduces the representation of noise agents. Since mean-preserving spreads do not have this property, such a spread is not sufficient (hence my use of Birnbaum here).

□

Proof of Proposition 2. For bounded private belief signals, let $\overline{p}$ and $\underline{p}$ be the most extreme private beliefs in favour of $\theta=1$ and $\theta=0$ respectively. Once the r-distribution has been shifted far enough that every supported r exceeds $\max\{\frac{\overline{p}}{1-\overline{p}},\frac{1-\underline{p}}{\underline{p}}\}$, the first agent, with social belief 0.5, acts invisibly with probability 1. The first visible agent is therefore a noise type, their action carries no information, and by the same reasoning at every adjusted index no social agent ever acts visibly. For unbounded signals, we have from the above working that any r implies a (possibly empty) interval of private signals that imply the agent

receiving them will certainly act invisibly i.e. choose $v_n = 0$:

$$\mathbb{P}(\theta = 1|s_n) \in [\underbrace{1 - \frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{L_0}, \underbrace{\frac{r_n(\frac{\eta}{2})^M}{r_n(\frac{\eta}{2})^M + (1-\frac{\eta}{2})^M}}_{U_0}]$$

This lower limit is decreasing in r and converging to 0, and the upper limit is increasing in r and converging to 1, hence the shape of the curves in Figure 3. This process of shifting the r distribution to the right actually implies that λ_0 converges to 1, before we even worry about our rounds of iterated reasoning as above ($\lambda_0 \to 1$ itself implies that $\lambda_\infty \to 1$). That this in turn implies that $\underline{\eta}_\infty \to 1$ follows immediately from the definition of $\underline{\eta}_\infty$ in Theorem 2. It remains therefore only to argue that λ_0 does in fact converge to 1 as we shift our r-distribution as described.

In computing λ_0, we integrate (for both $\theta = 0$ and $\theta = 1$, before taking the minimum) over the region within the first black curve depicted in Figure 3, and shifting the r-distribution to the right simply implies that each region of this distribution is now paired to an at least weakly wider interval of private signals, and therefore an at least weakly higher probability mass. Whatever the private belief distributions, the amount of mass assigned to $[\epsilon, 1-\epsilon]$ must be converging to 0 and ϵ converges to 0, so as the r-distribution allocates all of its mass to wider and wider intervals (converging to the complete $[0,1]$ interval), the integral must converge to 1 (as we are integrating over the entire support of both r and p_n).

Notice that we could make exactly the same argument for any λ_j, and that whilst this would not change the limit, it would make convergence faster. Integrating between the curves $[L_\infty(r), U_\infty(r)]$ would produce the fastest convergence. Since we are only establishing convergence here anyway, it is easiest to just use λ_0. Notice also that the width of intervals are decreasing in M for all values of r, so the lower M the faster the convergence. □

Proof of Proposition 3. The reasoning in this proof is very similar to that of Proposition 2. The sequence of distributions is defined such that a greater and greater proportion of the region supported by the product distribution over p_n and r is within the region in which agents certainly comment invisibly. As before, since the mass assigned outside this region converges to 0, λ_∞ must converge to 1, which gives us that $\underline{\eta}_\infty \to 1$, given the definition of $\underline{\eta}_\infty$ in Theorem 2.

Note also that we could have used $[L_0^1(\underline{r}), U_0^1(\underline{r})]$ instead of $[L_\infty^1, U_\infty^1]$ if we do not want to have to compute $[L_\infty^1, U_\infty^1]$, though this is a stronger condition on the sequence of new private belief distribution pairs.

□

Proof of Proposition 4. Recall the R_θ and L_θ from the proof of Theorem 1, and the $p^*(r)$ and $p^{**}(r)$ in terms of which they are defined:

$$R_\theta := \int \big[1 - \mathbb{G}_\theta\big(p^*(r)\big)\big] d\Delta(r), \qquad L_\theta := \int \mathbb{G}_\theta\big(p^{**}(r)\big) d\Delta(r),$$

$$p^*(r) = \frac{r\big(1-sb\big)}{sb + r\big(1-sb\big)}, \qquad p^{**}(r) = \frac{1-sb}{1+(r-1)sb}.$$

With all agents given the same r-parameter, these can be expressed as $R_\theta = 1 - \mathbb{G}_\theta\big(p^*(r)\big)$ and $L_\theta = \mathbb{G}_\theta\big(p^{**}(r)\big)$, and $\phi(sb,\theta)$ is given by (C.1). For $\alpha \in (0,1)$ let

$$\Gamma(\alpha; r) := \alpha\phi(\alpha, 1) + (1-\alpha)\phi(1-\alpha, 1).$$

If the visible action at a given history is correct with probability α, the next visible action is right with probability $\Gamma(\alpha; r)$ (this is what Step 1 proves). Since $p_n = \mathbb{P}(\theta = 1|s_n)$, $(1-p)\mathbb{G}_1'(p) = p\,\mathbb{G}_0'(p)$ by Bayes' rule, and by antisymmetry $\mathbb{G}_0'(p) = \mathbb{G}_1'(1-p)$, so that $(1-p)\mathbb{G}_1'(p) = p\,\mathbb{G}_1'(1-p)$.

Step 1: $\tilde{\alpha}_{\tilde{n}} = \Gamma(\tilde{\alpha}_{\tilde{n}-1}; r)$, and the first agent with adjusted index $\tilde{n}+1$ matches the state, if social, with probability $\big(\Gamma(\tilde{\alpha}_{\tilde{n}}; 1) - \frac{\eta}{2}\big)/(1-\eta)$. Suppose $\mathbb{P}(\tilde{x}_{\tilde{n}} = \theta|\theta) = \tilde{\alpha}_{\tilde{n}}$ in both states, as is true for $\tilde{n} = 0$ with $\tilde{\alpha}_0 := \frac{1}{2}$ (agents with adjusted index 1 observe nothing). An agent who observes $\tilde{x}_{\tilde{n}} = 1$ then has social belief $\tilde{\alpha}_{\tilde{n}}$, and one who observes $\tilde{x}_{\tilde{n}} = 0$ has $1 - \tilde{\alpha}_{\tilde{n}}$ (agents do not update on their adjusted index). As all agents with adjusted index $\tilde{n}+1$ share this social belief, (C.1) gives $\mathbb{P}(\tilde{x}_{\tilde{n}+1} = 1|\theta = 1) = \Gamma(\tilde{\alpha}_{\tilde{n}}; r)$. As $p^{**}(r)$ at social belief $1 - sb$ equals $1 - p^*(r)$ at sb, antisymmetry makes R_0 and L_0 at $1-sb$ equal to L_1 and R_1 at sb, so that $\phi(sb, 0) = 1 - \phi(1-sb, 1)$ for every sb, and $\mathbb{P}(\tilde{x}_{\tilde{n}+1} = 0|\theta = 0) = \Gamma(\tilde{\alpha}_{\tilde{n}}; r)$ too, which closes the induction. The first of these agents, if social, chooses $x_n = 1$ at sb when $\theta = 1$, and $x_n = 0$ at $1 - sb$ when $\theta = 0$, with the same probability $1 - \mathbb{G}_1(1-sb)$, by Lemma 2 and whatever r. At $r = 1$, $p^*(r) = p^{**}(r) = 1 - sb$, so $R_1 + L_1 = 1$ and $\phi(sb, 1) = \frac{\eta}{2} + (1-\eta)\big(1 - \mathbb{G}_1(1-sb)\big)$; averaging over the two social beliefs gives the second claim.

Step 2: $\Gamma(\cdot; r)$ maps $[\frac{1}{2}, 1 - \frac{\eta}{2}]$ into $(\frac{1}{2}, 1 - \frac{\eta}{2})$. Private beliefs are unbounded, so by the proof of Part 1 of Theorem 1, $\phi(sb, 1) > \phi(sb, 0)$ for every $sb \in (0,1)$, and $\phi(sb, 0) = 1 - \phi(1-sb, 1)$ by Step 1, so that $\phi(sb, 1) + \phi(1-sb, 1) > 1$. As $p^*(r)$ and $p^{**}(r)$ are decreasing in sb, R_1 is increasing and L_1 decreasing in sb, so $\phi(\cdot, 1)$ is strictly increasing, and

$$\Gamma(\alpha; r) = \tfrac{1}{2}\big(\phi(\alpha,1) + \phi(1-\alpha,1)\big) + \big(\alpha - \tfrac{1}{2}\big)\big(\phi(\alpha,1) - \phi(1-\alpha,1)\big) > \tfrac{1}{2} \quad \text{for } \alpha \geq \tfrac{1}{2},$$

while $\Gamma < 1 - \frac{\eta}{2}$ because $L_1 > 0$ and $R_1 + L_1 \leq 1$ give $\phi(sb, 1) < 1 - \frac{\eta}{2}$. In particular $\tilde{\alpha}_{\tilde{n}} \in [\frac{1}{2}, 1 - \frac{\eta}{2}]$ for all $\tilde{n}$.

Step 3: contraction. Γ is continuously differentiable on $[\frac{1}{2}, 1 - \frac{\eta}{2}] \times [1, \infty)$, with

$$\partial_\alpha \Gamma = \phi(\alpha, 1) - \phi(1-\alpha, 1) + \alpha\,\partial_{sb}\phi(\alpha, 1) - (1-\alpha)\,\partial_{sb}\phi(1-\alpha, 1). \tag{C.2}$$

Whenever $\sup|\partial_\alpha \Gamma(\cdot; r)| < 1$ (suprema are over $[\frac{1}{2}, 1 - \frac{\eta}{2}]$ throughout), $\Gamma(\cdot; r)$ is a contraction of $[\frac{1}{2}, 1 - \frac{\eta}{2}]$ into itself, so it has a unique fixed point $\tilde{\alpha}^*(r)$ there, and $\tilde{\alpha}_{\tilde{n}} \to \tilde{\alpha}^*(r)$; by Step 1, $\alpha^*(r)$ then exists and equals $\big(\Gamma(\tilde{\alpha}^*(r); 1) - \frac{\eta}{2}\big)/(1-\eta)$. If this holds on an interval of r, the implicit function theorem makes $\tilde{\alpha}^*$ continuously differentiable there (including at $r = 1$, as the same formulae define Γ as a continuously differentiable function for $r < 1$), with $\frac{d\tilde{\alpha}^*}{dr} = \partial_r \Gamma / (1 - \partial_\alpha \Gamma)$ at $(\tilde{\alpha}^*(r), r)$, whose denominator is positive.

Step 4: r close to 1. By Step 1, $\Gamma(\alpha;1) = \frac{\eta}{2} + (1-\eta)\big[\alpha\big(1-\mathbb{G}_1(1-\alpha)\big) + (1-\alpha)\big(1-\mathbb{G}_1(\alpha)\big)\big]$, so

$$\begin{aligned}\partial_\alpha\Gamma(\alpha;1) &= (1-\eta)\big[\mathbb{G}_1(\alpha) - \mathbb{G}_1(1-\alpha) + \alpha\mathbb{G}_1'(1-\alpha) - (1-\alpha)\mathbb{G}_1'(\alpha)\big]\\ &= (1-\eta)\big(\mathbb{G}_1(\alpha) - \mathbb{G}_1(1-\alpha)\big),\end{aligned}$$

which lies in $[0, 1-\eta]$ for $\alpha \geq \frac{1}{2}$ and is positive for $\alpha > \frac{1}{2}$, so $\Gamma(\cdot;1)$ is strictly increasing on $[\frac{1}{2}, 1)$. By uniform continuity of $\partial_\alpha\Gamma$ on $[\frac{1}{2}, 1-\frac{\eta}{2}] \times [1,2]$, $\sup|\partial_\alpha\Gamma(\cdot;r)| < 1$ for all r close enough to 1. At $r=1$, $\partial_r p^* = -\partial_r p^{**} = sb(1-sb)$, so $\partial_r R_1 = \partial_r L_1 = -sb(1-sb)\mathbb{G}_1'(1-sb)$, and differentiating (C.1), whose denominator is then 1, $\partial_r\phi(sb,1) = (1-\eta)^2 sb(1-sb)\mathbb{G}_1'(1-sb)\big(1-2\mathbb{G}_1(1-sb)\big)$. Using $(1-\alpha)\mathbb{G}_1'(\alpha) = \alpha\mathbb{G}_1'(1-\alpha)$ again,

$$\partial_r\Gamma(\alpha;1) = 2(1-\eta)^2\alpha^2(1-\alpha)\mathbb{G}_1'(1-\alpha)\big[1-\mathbb{G}_1(\alpha)-\mathbb{G}_1(1-\alpha)\big] > 0,$$

as by antisymmetry the bracket is $\mathbb{G}_0(\alpha) - \mathbb{G}_1(\alpha)$, which is positive by Acemoglu et al. (2011, Lemma A1(c)). Hence $\frac{d\tilde{\alpha}^*}{dr} > 0$ at $r=1$, and by continuity for all r close enough to 1; as $\Gamma(\cdot;1)$ is strictly increasing, so is α^*.

Step 5: large r. For $sb \in [\frac{\eta}{2}, 1-\frac{\eta}{2}]$, $1-p^*(r)$ and $p^{**}(r)$ are at most $\frac{2-\eta}{2-\eta+r\eta}$, so R_1 and L_1 vanish uniformly as $r\to\infty$; as $\partial_{sb}p^* = -\frac{p^*(1-p^*)}{sb(1-sb)}$, $\partial_{sb}p^{**} = -\frac{p^{**}(1-p^{**})}{sb(1-sb)}$ and $\mathbb{G}_1'$ is bounded, so do $\partial_{sb}R_1$ and $\partial_{sb}L_1$. The denominator of (C.1) being at least η, it follows that, uniformly on $[\frac{\eta}{2}, 1-\frac{\eta}{2}]$, $\phi(\cdot,1) \to \frac{1}{2}$ and $\partial_{sb}\phi(\cdot,1)\to 0$, so by (C.2) $\sup|\partial_\alpha\Gamma(\cdot;r)| \to 0$ and $\sup|\Gamma(\cdot;r) - \frac{1}{2}| \to 0$. So Step 3 applies for all r large enough, and $|\tilde{\alpha}^*(r) - \frac{1}{2}| = |\Gamma(\tilde{\alpha}^*(r);r) - \frac{1}{2}| \to 0$. Since $\tilde{\alpha}^*(1) > \frac{1}{2}$ by Step 2 and $\Gamma(\cdot;1)$ is strictly increasing, $\alpha^*(r) \to \big(\Gamma(\frac{1}{2};1) - \frac{\eta}{2}\big)/(1-\eta) < \alpha^*(1)$, so $\alpha^*(r) < \alpha^*(1)$ for all r large enough. As $\Gamma(\cdot;1)$ is strictly increasing, these inequalities for α^* are equivalent to the same ones for $\tilde{\alpha}^*$, and by Step 4 and antisymmetry $\big(\Gamma(\frac{1}{2};1) - \frac{\eta}{2}\big)/(1-\eta) = 1 - \mathbb{G}_1(0.5) = \frac{1}{2}\mathbb{G}_0(0.5) + \frac{1}{2}\big(1-\mathbb{G}_1(0.5)\big)$. □

C.4 Proofs for Section 4

Proof of Theorem 3. As noted in Section 2, an agent with reputation concern r will choose to act visibly if they form belief stronger than $\frac{r}{r+1}$ that the state is θ, for any $\theta \in \{0,1\}$. In the following argument, I refer to the event of the agent having such a strong belief as *Strong*.

- The first visibly acting agent after $\tilde{n}$, $\tilde{n}+1$, will be correct with probability $\mathbb{P}(x_{\tilde{n}+1} = \theta) = \mathbb{P}(x_{n(\tilde{n})+1} = \theta | Strong)$. In words, the probability the first visibly-acting agent after $\tilde{n}$ makes the correct choice is simply the probability that the first agent (visible or not) makes the correct choice conditional on their having a belief stronger than $\frac{r}{r+1}$ in favour of one state or the other.

- This probability must be higher than the unconditional probability that $\tilde{n}$'s immediate successor (visible or not) will be successful, since $\mathbb{P}(x_{n(\tilde{n})+1} = \theta | Strong) \geq \mathbb{P}(x_{n(\tilde{n})+1} = \theta | \neg Strong)$ by Lemma 6 and:

$$\begin{aligned}\mathbb{P}(x_{n(\tilde{n})+1} = \theta) &= \mathbb{P}(x_{n(\tilde{n})+1} = \theta | Strong)\mathbb{P}(Strong) + \mathbb{P}(x_{n(\tilde{n})+1} = \theta | \neg Strong)\mathbb{P}(\neg Strong)\\ &= \mathbb{P}(x_{n(\tilde{n})+1} = \theta | Strong)\mathbb{P}(Strong) + \mathbb{P}(x_{n(\tilde{n})+1} = \theta | \neg Strong)\big(1-\mathbb{P}(Strong)\big)\end{aligned}$$

- That $\mathbb{P}(x_{n(\tilde{n})+1} = \theta)$ must be higher than $\mathbb{P}(x_{n(\tilde{n})} = \theta)$ however, follows from the improvement principle reasoning of Acemoglu et al. (2011). It thus follows that an improvement path between any visibly-acting agents must produce a greater lower bound on the final agent's accuracy than in an analogous path in the model of Acemoglu et al. (2011).

- Thus unless after a point no agent will ever act visibly again, the argument of Acemoglu et al. (2011, Theorem 2) implies that the accuracy of visibly-acting agents must converge in probability to 1. Since Borel-Cantelli establishes that after any given history there must almost surely be another visibly-acting agent (in fact, infinitely many of them), there cannot be such a point. Hence the accuracy of visibly-acting agents converges to 1 in probability.

- Any invisibly-acting agent is also improving on these agents, and so their accuracy converges in probability to 1 as well. The probability with which an agent acts visibly must also converge to 1, since the strength of social beliefs is converging to 1.

The above argument assumed a given r, but works for an arbitrary value of it. Integrating across r values establishes that it holds for any distribution of r values.

□

References

Acemoglu, D., M. A. Dahleh, I. Lobel, and A. Ozdaglar (2011): "Bayesian learning in social networks," *The Review of Economic Studies*, 78, 1201–1236.

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

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.

Birnbaum, Z. W. (1948): "On random variables with comparable peakedness," *The Annals of Mathematical Statistics*, 19, 76–81.

Blackwell, D. (1951): "Comparisons of experiments," *Berkeley Symp. on Math. Statist. and Prob.*, 93–102.

——— (1953): "Equivalent comparisons of experiments," *The annals of mathematical statistics*, 265–272.

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

Breiman, L. (1968): *Probability*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

Callander, S. and J. Hörner (2009): "The wisdom of the minority," *Journal of Economic Theory*, 144, 1421–1439.e2.

Cremin, J. W. (2025): "Too Much Information & The Death of Consensus," Working paper or preprint.

Cremin, J. W. E. (2026): "Sleeping Newcomb," Working paper, Aix-Marseille University.

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

Guarino, A., H. Harmgart, and S. Huck (2011): "Aggregate information cascades," *Games and Economic Behavior*, 73, 167–185.

Guarino, A. and P. Jehiel (2013): "Social learning with coarse inference," *American Economic Journal: Microeconomics*, 5, 147–174.

Herrera, H. and J. Hörner (2013): "Biased social learning," *Games and Economic Behavior*, 80, 131–146.

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*.

Monzón, I. and M. Rapp (2014): "Observational learning with position uncertainty," *Journal of Economic Theory*, 154, 375–402.

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

——— (2008): "Rational social learning with random sampling," Tech. rep., working paper.

Song, Y. (2016): "Social learning with endogenous observation," *Journal of Economic Theory*, 166, 324–333.

Winkler, P. (2017): "The sleeping beauty controversy," *The American Mathematical Monthly*, 124, 579–587.

D Online Appendix: Indexical Beliefs

Upon observing their adjusted index, ignoring other information for the moment, a Bayesian agent should update their belief about the state θ. Whilst the main text assumes agents act according to the *one-boxer beliefs* of Cremin (2026), the representation results of which show that agents with commitment power would behave as if their state-beliefs ignore indexical information, this appendix considers the alternative: agents who act without commitment power, forming Bayesian beliefs (beliefs the *Sleeping Beauty* literature refers to as 'thirder').

For such agents, since the common prior is $\frac{1}{2}$, the correct posterior belief is:

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

Given that the prior over n is uniform by the principle of insufficient reason however, this is equal to $0/(0+0)$ (the probability of being 'assigned' any $n \in \mathbb{N}$ is zero, and thus the conditional probability of having any $\tilde{n} \in \mathbb{N}$ is also zero). Assuming that agents do not update their belief upon observing $\tilde{n}$ can be defended as the sensible response to this, in which case Bayesians adopt one-boxer beliefs anyway, but an alternative approach is to endow them with an improper prior.

In this appendix I show that proceeding thus, we can nonetheless bound agents' 'indexical' beliefs away from 0 and 1, thanks to the presence of noise agents. This fact ensures that the unravelling results all still hold. Furthermore, I shall explain that the remaining material in Section 3.2 is unaffected, and that Theorem 1 also stands with this alternative assumption. The only result whose proof does not carry forward is Theorem 3, since this benchmark result specifically considers the case in which we have no noise agents ($\eta = 0$).

Imagine the agent has an improper prior, assigning probability $\epsilon > 0$ to each possible $n \in \mathbb{N}$. If their adjusted index is $\tilde{n} = 1$, they consider the hypotheses: $\{n = 1\} \cap \{\theta = 0\}$, $\{n = 2\} \cap \{\theta = 0\}$, ..., and $\{n = 1\} \cap \{\theta = 1\}$, $\{n = 2\} \cap \{\theta = 1\}$, ... This is because if you observe that no agent has visibly acted before you, this could simply be because you are the first agent to arrive, but it could also be that you are the 100th and that all 99 of your predecessors chose to act invisibly. The more agents are required to have acted invisibly, the more unlikely a hypothesis, but a Bayesian with an improper prior will consider all such possible explanations.

The probability of having $\tilde{n} = 1$ conditional on $\theta = 0$ is then, for example:

$$\begin{aligned}\mathbb{P}(\tilde{n} = 1|\theta = 0) = &\frac{1}{2} \times \epsilon + \frac{1}{2} \times \epsilon \times \mathbb{P}(v_1 = 0|\theta = 0) \\ &+ ... + \frac{1}{2} \times \epsilon \times \mathbb{P}(v_1 = 0|\theta = 0) \times ... \times \mathbb{P}(v_{k-1} = 0|\theta = 0) \\ &+ ...\end{aligned}$$

Suppose we have upper and lower bounds for the probabilities $\{\mathbb{P}(v_k = 0|\theta = 0) : k \in \mathbb{N}\}$. Call these

$\underline{p}$ and $\overline{p}$. It then follows that:

$$\mathbb{P}(\tilde{n}=1|\theta=0) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2}\overline{p} + ... + \frac{\epsilon}{2}\overline{p}^{k-1} + ... = \frac{\epsilon}{2}\frac{1}{1-\overline{p}}$$
$$\text{and } \mathbb{P}(\tilde{n}=1|\theta=0) \geq \frac{\epsilon}{2}\frac{1}{1-\underline{p}}$$

This line of reasoning applies equally for probabilities conditional on $\theta = 1$ as $\theta = 0$, and indeed with infinitely many agents (where there will almost surely be infinitely many visibly-acting agents, thanks to the presence of noise) identical bounds can be derived for any $\tilde{n}$ (not just $\tilde{n} = 1$). Thus we can set out the following upper bound:

$$\mathbb{P}(\theta=1|\tilde{n}) = \frac{1}{1+\frac{\mathbb{P}(\tilde{n}|\theta=0)}{\mathbb{P}(\tilde{n}|\theta=1)}} \leq \frac{1}{1+\frac{1-\overline{p}}{1-\underline{p}}}$$

Since noise agents always act visibly, a candidate for $\overline{p}$ is $1 - \eta$. Simply taking $\underline{p} = 0$ we can then see that:

$$\mathbb{P}(\theta=1|\tilde{n}) = \frac{1}{1+\frac{\mathbb{P}(\tilde{n}|\theta=0)}{\mathbb{P}(\tilde{n}|\theta=1)}} \leq \frac{1}{1+\eta}$$

Alternatively, in any given equilibrium we can define λ_∞ as in the main text to provide a lower bound on the probability with which any social agent acts invisibly. $(1-\eta)\lambda_\infty$ then serves as a candidate for $\underline{p}$. This gives a tighter bound:

$$\mathbb{P}(\theta=1|\tilde{n}) \leq \frac{1-(1-\eta)\lambda_\infty}{1-(1-\eta)\lambda_\infty+\eta}$$

The tighter the bound, the lower r-values will be necessary to establish that unravelling takes place. However, for the sake of this appendix we need simply establish that we can bound indexical beliefs away from 0 and 1, so let us use the simpler one $\frac{1}{1+\eta}$. We could have run through the same steps with $\mathbb{P}(\theta = 0|\tilde{n})$, so since $\mathbb{P}(\theta = 1|\tilde{n}) = 1 - \mathbb{P}(\theta = 0|\tilde{n})$ we have also a lower bound for $\mathbb{P}(\theta = 1|\tilde{n}) = \frac{1+\eta}{1+\eta} - \frac{1}{1+\eta} = \frac{\eta}{1+\eta}$. Hence $\mathbb{P}(\theta = 1|\tilde{n}) \in [\frac{\eta}{1+\eta}, \frac{1}{1+\eta}]$.

D.1 Which results still hold?

As I have mentioned, the only result that does not carry over is Theorem 3, since it assumes $\eta = 0$ and a positive probability of noise is necessary for all information sets to be reached with probability 1.

All the results in Section 3.2 on unravelling follow from the fact that we can bound social beliefs away from one and zero, generating a central interval of private beliefs that certainly induce invisible actions for high enough r. This line of reasoning remains valid with agents who have an improper indexical prior, since the prior beliefs are bounded away from zero and one. Hence Theorem 2 and Propositions 1 and 3 still hold, and there is no need to even adjust the expressions contained within. For the upper bound on accuracy in Section 3.2, the exact expression no longer applies, though we could provide another straightforwardly by asking what accuracy agents would obtain if in state $\theta = 0$ they happened to have the most extreme indexical belief in favour of 0, and vice versa in $\theta = 1$. The qualitative point that one can find a tighter bound that reflects unravelling however, is unchanged. Similarly Proposition 2 only changes

in that the accuracy value to which agents converge will depend on their adjusted index (as different adjusted indices imply different indexical beliefs), instead of converging to $\frac{1}{2}\mathbb{G}_0(0.5) + \frac{1}{2}\big(1 - \mathbb{G}_1(0.5)\big)$ for all agents.

In Section 3.1, I discuss learning in dense networks. Theorem 1 depends on the martingale convergence theorem, and the required application of this does not change here. It will still be the case that the social belief of agents will be a martingale that will converge to 1 on the true state, though agents will now be combining this social belief with both a private signal *and* their indexical beliefs. Much as private beliefs are eventually dominated by a social belief that is converging to 1, indexical beliefs become insignificant also (since again they are crucially bounded away from 1 and 0). Point 2 of this theorem leverages the fact that with bounded beliefs, we can remove any cascade beliefs with strong enough reputation concerns; this is also unchanged by the improper prior assumption.

Therefore, the results of the paper do not crucially depend on the fashion in which agents treat indexical information. Assuming they operate with an improper prior may increase the strength of reputation concerns needed to produce unravelling, but the qualitative argument of the paper remains substantially unaffected.

E One-Sided Noise

In this paper I have assumed noise agents randomise uniformly between the two actions, but what, one might ask, if they were to all choose $x = 1$?[8] Perhaps they represent a targeted disinformation campaign by a hostile state during an election?

In this model, such agents would successfully prevent learning, but would have an exaggeratedly negative impact on the very state they were trying to promote ($\theta = 1$) in sparse networks. In dense networks they help learning in a smaller set of circumstances than in the main model (c.f. Proposition 6 below). This sparse network result obtains since upon observing a neighbourhood in which all choose $x = 1$, an agent will account for the possibility that they are all noise agents, and form a social belief no stronger than $\frac{(1-\eta)^M}{(1-\eta)^M+\eta^M}$. (In the main paper the equivalent expression replaces the η here with $\eta/2$.) However, no such bound holds for beliefs that $\theta = 0$. We could then ask, what private beliefs will certainly **not** induce $\{x_n = 1, v_n = 1\}$? We can then rehearse much the same unravelling argument as in the main article, except applying it uniquely to the agents visibly choosing action 1. More agents with beliefs above $\frac{1}{2}$ will choose to act invisibly, which will reduce the maximum strength of social beliefs in favour of $\theta = 1$, which will cause more such agents to act invisibly, which will lower the bound, and so on and so forth.

[8] Thanks to Sebastian Bervoets and Franz Ostrizek for posing this question to me.

E.0.1 Unravelling with One-Sided Noise

Let us observe that for any private belief in the following interval, agents will not choose $\{x_n = 1, v_n = 1\}$:

$$\mathbb{P}(\theta = 1|s_n) \in [0, \underbrace{\frac{r_n\eta^M}{r_n\eta^M + (1-\eta)^M}}_{U_0^1}]$$

Following the argument of the main text, we can then define $\lambda_0^1 := \min\{\mathbb{G}_0(U_0), \mathbb{G}_1(U_0)\}$. Whatever the value of θ, agents will form private beliefs within this region with at least probability λ_0, where of course agents' private beliefs are conditionally independent. Hence the probability with which agent $\tilde{n}$ is a noise agent is in fact:

$$\underline{\eta}_1^1 := \frac{\eta}{\big(\eta + (1-\lambda_0^1)(1-\eta)\big)}$$

The major difference here is of course that upon observing a predecessor has chosen $x = 0$, agents immediately learn that this agent cannot be a noise agent. Hence, this probability is only relevant when interpreting the actions of neighbours who have chosen $x = 1$. Nonetheless, we can feed this back into the interval above to find U_0^2, generate a new λ_0^2, and a new η_2^1 just as in the main article. We can define $\underline{\eta}_\infty^1$ and λ_∞^1 through this process exactly as before, keeping in mind that $\underline{\eta}_\infty^1$ is now a lower bound on the probability with which any visible agent $\tilde{n}$ will be a noise agent, and relevant when forming beliefs about the type of a neighbour who chose $x = 1$, but irrelevant when observing $x = 0$.

Proposition 5 (Extending Unravelling Results)**.** *We can find equivalents of the following unravelling results:*

1. *Theorem 2: $\underline{\eta}_\infty^1$ is a lower bound on the probability that the $\tilde{n}^{th}$ visibly-acting agent is a noise type for any $\tilde{n} \in \mathbb{N}$, and the asymptotic fraction of agents that are noise types is almost surely greater than $\underline{\eta}_\infty^1$.*

2. *Bound on accuracy with η: Conditional on θ, accuracy can be bounded by*

$$\mathbb{P}(x_n = \theta|\theta = 1) \leq \left(1 - \mathbb{G}_1\left(1 - \frac{(1-\eta)^M}{(1-\eta)^M + \eta^M}\right)\right) \tag{E.1}$$

$$\mathbb{P}(x_n = \theta|\theta = 0) \leq (1 - \eta^M) + \eta^M\mathbb{G}_0(0.5) \tag{E.2}$$

$$\mathbb{P}(x_n = \theta) = 0.5\mathbb{P}(x_n = \theta|\theta = 0) + 0.5\mathbb{P}(x_n = \theta|\theta = 1) \tag{E.3}$$

3. *Bound on accuracy with $\underline{\eta}_\infty^1$: as in Section 3.2, we can tighten the bound of Point 2 by replacing η with $\underline{\eta}_\infty^1$.*

4. *Proposition 1 can be exactly restated, substituting $\underline{\eta}_\infty^1$ for $\underline{\eta}_\infty$ and λ_0^1 for λ_0.*

5. *Proposition 2: This result holds without adjustment.*

6. Proposition 3: This result also holds without adjustment.

Proof. To see the above points observe that:

1. Point 1 follows from the proof of Theorem 2 and the reasoning above, bearing in mind the changed definitions of η and λ.

2. Point 2: The portion of the upper bound on α_n with $\mathbb{G}_1$ can be derived exactly as in the first step of the proof of Theorem 2, except that every $\eta/2$ can be replaced with a η, to bound accuracy conditional on $\theta = 1$. This gives Inequality E.1. A similar bound cannot be produced for accuracy conditional on $\theta = 0$, though one can observe that with probability η^M, they cannot perform better than $\mathbb{G}_0(0.5)$ (since with this probability their network is entirely noise, and they must underperform an agent given no social information at all). This observation produces the second bound, Inequality E.2. One can substitute each of these (E.1 and E.2) into Equation E.3 for an overall bound on accuracy.

3. Point 4, the proof as written in Appendix C.3 stands, with only the two substitutions mentioned in the statement.

4. Point 5: As with the proof of Proposition 2, repeatedly shifting the r distribution to the right implies that λ_0^1 converges to 1 before even considering unravelling. This implies that $\underline{\eta}_\infty^1 \to 1$, as can be seen in our above expression for $\underline{\eta}_1^1 < \underline{\eta}_\infty^1$. The rest of the proof of Proposition 2 can be applied without adjustment.

□

Therefore, this change to one-sided noise will still damage learning in sparse networks when $\theta = 0$, as can be seen in Points 2, 3 and 5 of Proposition 5, relative to the model with no noise agents. Without noise, as established in Theorem 3, we would achieve complete learning with unbounded beliefs. With these noise agents however, this is no longer the case, as any agent will, with some probability, still observe a neighbourhood of pure noise. Increasing the strength of reputation concerns will also eventually completely smother learning (Point 5).

In dense networks, too, these noise agents can still help learning. The combined presence of noise agents and *strong enough* reputation concerns, will ensure that the probability of observing a predecessor choosing $x = 1$ visibly will be different in the two states of the world. Here, however, the benefits to learning set out in Part 2 of Theorem 1 are lost. Since there are no noise agents who take action 0, it is now possible that there is some region of social beliefs in favour of $\theta = 1$ ($\lambda > 0.5$) such that no agent can receive a private signal strong enough to choose $x_n = 0$ *and* $v_n = 1$. With noise agents who uniformly randomise, the fact that any $\tilde{n}$ will always choose $x_{\tilde{n}} = 0$ with some probability ensures that the ratio of agents choosing $\{x_n = 1, v_n = 1\}$ against $\{x_n = 1, v_n = 0\}$ is observable; this means learning will continue. With one-sided noise however, this is no longer the case. We can resurrect learning here under a similar condition to that in the main text, only we now need one-sided unbounded beliefs (where they are unbounded in favour of $\theta = 0$, the state that is not being supported by bots).

Proposition 6 (Theorem 1 Part 2: Version with One-sided Noise)**.** *Suppose noise agents now always choose $x_n = 1$. If private beliefs have support $(0, \overline{b})$, i.e. they are bounded on the $\theta = 1$ side only, and the candour distribution $\Delta(c)$ satisfies $[0, \epsilon) \cup (1 - \epsilon, 1) \subseteq \text{supp}(\Delta(c))$ for some $\epsilon > 0$, the social belief converges to certainty on the true state almost surely.*

Proof. Since private beliefs are unbounded in favour of $\theta = 0$, there are no cascade beliefs λ greater than 0.5. The upper bound on private beliefs $\overline{b}$ implies that social beliefs below $1 - \overline{b}$ are cascade beliefs in favour of $x_n = 0$. Agents acting at social beliefs below this point will choose $\{x_n = 1, v_n = 1\}$ with at least probability η (since all noise types choose $x_n = 1$). For any social belief below $1 - \overline{b}$, the probability with which agents at that social belief will choose $v_n = 1$ will be determined by the reputation-concern distribution, and the distribution of private beliefs. Since the private beliefs are unbounded in favour of 0, $\mathbb{G}_0(\lambda) > \mathbb{G}_1(\lambda)$ for any $\lambda < 0.5$. Thus the probability with which the next visible action is 1 must be strictly lower if $\theta = 0$ than if $\theta = 1$ for any $\lambda < 0.5$. As we observed at the beginning of this proof, there are no cascade beliefs $\lambda \geq 0.5$. Hence this is true for all $\lambda \in [0, 1]$. As with the proof of Theorem 1, the result then follows from Smith and Sørensen (2000, Theorem B.1). □

That, in sparse networks, bots would actually do more damage to the cause they are promoting than to the other seems strange. A crucial feature of the model that drives this is the assumption that η and the behaviour of bots is common knowledge. If a foreign power sets up a series of bots to spout disinformation, in reality the effectiveness of such a move is clearly based on the fact that unsuspecting voters in the target country will omit to account for this fact, and simply be duped into believing that some of the bot-content they observe is legitimate. In my model, when the adversary chooses to programme enough bots to ensure 5% of all commenters are said bots, agents in the model are assumed automatically to know this. Introducing misspecification into the model would be necessary to capture the true impact of such a scheme, and would no doubt produce more failures of learning, as in Bohren and Hauser (2021). Nonetheless, even with common knowledge of η and bot-behaviour some damage is still done.

F Online Appendix: Line Example Working

Some of the graphs in this paper concern learning on immediate predecessor (or line) networks where the private signal structure induces antisymmetric private beliefs, and for these networks we can compute the probability with which an agent correctly guesses the state when their predecessor has accuracy α analytically. I include working for some examples of this here that I used to plot some of the figures above, so that interested readers can reproduce them as easily as possible.

F.1 Beta Private Signals and Pareto Reputation Concerns

- The signal densities are $\{2(1 - s), 2s\}$.

- Reputation concern r follows a Pareto distribution with $\alpha = 1$ and $r_{min} = 2$:

$$f_T(r) = \frac{2}{r^2}, \quad r \geq 2$$

This implies $(r-1)/r$ is distributed according to a uniform distribution on $[0.5, 1]$. Using $\tilde{z}_n$ as notation for the probability with which the n^{th} visible agent matches the state, we can write the following:

$$\tilde{z}_n := \mathbb{P}(\tilde{x}_{\tilde{n}} = \theta) = \mathbb{P}(\tilde{\tau}_n = N|\tilde{z}_{n-1})\big(\frac{1}{2}\big) + \mathbb{P}(\tilde{\tau}_{\tilde{n}} = S)(\tilde{z}_{n-1}) \times \mathbb{P}(\tilde{x}_{\tilde{n}} = \theta|\tilde{\tau}_{\tilde{n}} = S)(\tilde{z}_{n-1})$$

The social signal matches the state with probability $\tilde{z}_{n-1}$, suppose the true state is 1 and the social signal is 1, what are the relevant probabilities in this case?

The probability the next agent is a noise agent is:

$$\begin{aligned}
&\eta + (1-\eta) \times \mathbb{P}(p_n \in \{\text{Silent region for 1,1}\}) \times \eta \\
&+ \Big((1-\eta) \times \mathbb{P}(p_n \in \{\text{Silent...}\})\Big)^2 \times \eta + ... \\
&= \eta \sum_{j=0}^{\infty} \Big((1-\eta) \times \mathbb{P}(p_n \in \{\text{Silent ...}\})\Big)^j = \frac{\eta}{1 - \Big((1-\eta) \times \mathbb{P}(p_n \in \{\text{Silent...}\})\Big)}
\end{aligned}$$

Decision Rule Inequalities:

In $1,1$ (when $\theta = 1$ and $\tilde{x}_n = 1$), according to Lemma 1 an S agent gets the right answer visibly if $p_n\big(1 - \frac{(r-1)}{r}(1 - z_{n-1})\big) > 1 - z_{n-1}$. In $1,0$, an S agent gets the right answer visibly if the same left-hand expression is strictly greater than z_{n-1}. In $1,1$ they get the wrong answer visibly if $p_n(1 + (r-1)z_{n-1}) < (1 - z_{n-1})$ and in $1,0$ they get it wrong visibly if this left-hand expression is strictly less than z_{n-1}.

Since the private beliefs are antisymmetric, we can just consider the case in which $\theta = 1$. As can be understood from Lemma 3, the relevant probabilities are symmetric. Next, we must compute the distributions of the expressions on the left hand side of each of the above inequalities, of which two of these expressions are distinct. In each case, Jacobian transformation and convolution formulae suffice to compute the probability density functions.

One can compute that the probability density function of a variable distributed as $p_n \times (1 - \frac{r-1}{r} z_{n-1})$ is:

$$f_{VR}(\zeta) = \begin{cases} \dfrac{4\zeta}{c}\left(\dfrac{1}{1-c} - \dfrac{1}{1-0.5c}\right), & 0 \leq \zeta \leq 1-c, \\ \dfrac{4}{c}\left(1 - \dfrac{\zeta}{1-0.5c}\right), & 1-c \leq \zeta \leq 1-0.5c, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly one distributed as $p_n\Big(1+(r-1)(1-z_{n-1})\Big)$ has density function:

$$f_{VW}(\zeta)=\begin{cases}0, & \zeta\leq 0,\\ \dfrac{4c\zeta}{(1-c)^3}\left[\dfrac{1+c}{2c}-\dfrac{2c}{1+c}-2\ln\left(\dfrac{1+c}{2c}\right)\right], & 0<\zeta<1+c,\\ \dfrac{4c}{(1-c)^3}\left\{\dfrac{\zeta^2}{\zeta-(1-c)}+2\zeta\ln\left(\dfrac{\zeta-(1-c)}{\zeta}\right)-\Big[\zeta-(1-c)\Big]\right\}, & \zeta\geq 1+c.\end{cases}$$

To compute the probabilities that the above probabilities are satisfied we need the CDFs, which are unwieldy to write out and therefore omitted, but serve to calculate the probabilities that our above inequalities hold:

- In $1,1$, c takes value z_{n-1} in the above, and we want probabilities that $\zeta\geq 1-z_{n-1}$.
- In $1,0$, c takes value $(1-z_{n-1})$ in both of these, and we want probabilities that $\zeta\geq z_{n-1}$.

$$F_{VR11}(1-z_{n-1})=\frac{(1-z_{n-1})^2}{(1-z_{n-1})(1-0.5z_{n-1})}=\frac{(1-z_{n-1})}{(1-0.5z_{n-1})}$$

$$\begin{aligned}F_{VW11}(1-z_{n-1})&=\frac{(1-z_{n-1})^2}{2}\Big(\frac{4z_{n-1}}{(1-z_{n-1})^3}\left[\frac{1+z_{n-1}}{2z_{n-1}}-\frac{2z_{n-1}}{1+z_{n-1}}-2\ln\left(\frac{1+z_{n-1}}{2z_{n-1}}\right)\right]\Big)\\&=\frac{2z_{n-1}}{(1-z_{n-1})}\left[\frac{1+z_{n-1}}{2z_{n-1}}-\frac{2z_{n-1}}{1+z_{n-1}}-2\ln\left(\frac{1+z_{n-1}}{2z_{n-1}}\right)\right]\end{aligned}$$

$$F_{VR10}(z_{n-1})=\frac{z_{n-1}^2}{(1-1+z_{n-1})(1-0.5(1-z_{n-1}))}=\frac{2z_{n-1}}{1+z_{n-1}}$$

$$\begin{aligned}F_{VW10}(z_{n-1})&=\frac{z_{n-1}^2}{2}\frac{4(1-z_{n-1})}{(z_{n-1})^3}\left[\frac{2-z_{n-1}}{2(1-z_{n-1})}-\frac{2(1-z_{n-1})}{2-z_{n-1}}-2\ln\left(\frac{2-z_{n-1}}{2(1-z_{n-1})}\right)\right]\\&=\frac{2(1-z_{n-1})}{(z_{n-1})}\left[\frac{2-z_{n-1}}{2(1-z_{n-1})}-\frac{2(1-z_{n-1})}{2-z_{n-1}}-2\ln\left(\frac{2-z_{n-1}}{2(1-z_{n-1})}\right)\right]\end{aligned}$$

And we can finally begin plugging-in these probabilities to compute the conditional probabilities (conditional on the social signal), which will then allow us to compute the unconditional probabilities required in our original equation:

Probability of being Noise agent in $1,1$:

$$\begin{aligned}&\mathbb{P}(\tilde{\tau}_n=N|\theta=1,\tilde{x}_{n-1}=1,z_{n-1})=\\&=\frac{\eta}{1-(1-\eta)\big(1-(1-F_{VR11}(1-z_{n-1})+F_{VW11}(1-z_{n-1}))\big)}\\&=\frac{\eta}{1-(1-\eta)\big(F_{VR11}(1-z_{n-1})-F_{VW11}(1-z_{n-1})\big)}\\&=\frac{\eta}{1-(1-\eta)\Big(\frac{1-z_{n-1}}{1-0.5z_{n-1}}-\frac{2z_{n-1}}{1-z_{n-1}}\Big[\frac{1+z_{n-1}}{2z_{n-1}}-\frac{2z_{n-1}}{1+z_{n-1}}-2\ln\Big(\frac{1+z_{n-1}}{2z_{n-1}}\Big)\Big]\Big)}\end{aligned}$$

Probability of being Noise agent in $1,0$:

$$\begin{aligned}
&\mathbb{P}(\tilde{\tau}_n = N|\theta = 1, \tilde{x}_{n-1} = 0, z_{n-1}) = \\
&= \frac{\eta}{1-(1-\eta)\big(F_{VR10}(z_{n-1}) - F_{VW10}(z_{n-1})\big)} \\
&= \frac{\eta}{1-(1-\eta)\Big(\frac{2z_{n-1}}{1+z_{n-1}} - \frac{2(1-z_{n-1})}{z_{n-1}}\Big[\frac{2-z_{n-1}}{2(1-z_{n-1})} - \frac{2(1-z_{n-1})}{2-z_{n-1}} - 2\ln\Big(\frac{2-z_{n-1}}{2(1-z_{n-1})}\Big)\Big]\Big)}
\end{aligned}$$

Probability of $\tilde{n}$ being visibly correct conditional on $\tilde{\tau}_n = S$ agent in $1,1$:

$$\begin{aligned}
&\mathbb{P}(\tilde{x}_n = \theta \mid \tilde{\tau}_n = S, \theta = 1, \tilde{x}_{n-1} = 1) = \frac{1 - F_{VR11}(1-z_{n-1})}{1 - F_{VR11}(1-z_{n-1}) + F_{VW11}(1-z_{n-1})} \\
&= \frac{1 - \frac{1-z_{n-1}}{1-0.5z_{n-1}}}{1 - \frac{1-z_{n-1}}{1-0.5z_{n-1}} + \frac{2z_{n-1}}{1-z_{n-1}}\Big[\frac{1+z_{n-1}}{2z_{n-1}} - \frac{2z_{n-1}}{1+z_{n-1}} - 2\ln\Big(\frac{1+z_{n-1}}{2z_{n-1}}\Big)\Big]}
\end{aligned}$$

Probability of $\tilde{n}$ being visibly correct conditional on $\tilde{\tau}_n = S$ agent in $1,0$:

$$\begin{aligned}
&\mathbb{P}(\tilde{x}_n = \theta \mid \tilde{\tau}_n = S, \theta = 1, \tilde{x}_{n-1} = 0) = \frac{1 - F_{VR10}(z_{n-1})}{1 - F_{VR10}(z_{n-1}) + F_{VW10}(z_{n-1})} \\
&= \frac{1 - \frac{2z_{n-1}}{1+z_{n-1}}}{1 - \frac{2z_{n-1}}{1+z_{n-1}} + \frac{2(1-z_{n-1})}{z_{n-1}}\Big[\frac{2-z_{n-1}}{2(1-z_{n-1})} - \frac{2(1-z_{n-1})}{2-z_{n-1}} - 2\ln\Big(\frac{2-z_{n-1}}{2(1-z_{n-1})}\Big)\Big]}
\end{aligned}$$

The Unconditional Quantities of Interest:

$$\begin{aligned}
\mathbb{P}(\tilde{\tau}_n = N|z_{n-1}) &= \mathbb{P}(\tilde{\tau}_n = N|\theta = 1, z_{n-1}) && \text{By symmetry} \\
&= z_{n-1}\mathbb{P}(\tilde{\tau}_n = N|\theta = 1, \tilde{x}_{n-1} = 1, z_{n-1}) \\
&\quad + (1-z_{n-1})\mathbb{P}(\tilde{\tau}_n = N|\theta = 1, \tilde{x}_{n-1} = 0, z_{n-1})
\end{aligned}$$

The unconditional probability of success assuming an agent is social is:

$$\begin{aligned}
\mathbb{P}(\tilde{x}_n = \theta|\tilde{\tau}_n = S) = z_{n-1}&\frac{\mathbb{P}(\tilde{\tau}_n = S|\theta = 1, \tilde{x}_{n-1} = 1, z_{n-1})}{\mathbb{P}(\tilde{\tau}_n = S|z_{n-1})}\mathbb{P}(\tilde{x}_n = \theta \mid \tilde{\tau}_n = S, \theta = 1, \tilde{x}_{n-1} = 1) \\
&+ (1-z_{n-1})\frac{\mathbb{P}(\tilde{\tau}_n = S|\theta = 1, \tilde{x}_{n-1} = 0, z_{n-1})}{\mathbb{P}(\tilde{\tau}_n = S|z_{n-1})}\mathbb{P}(\tilde{x}_n = \theta \mid \tilde{\tau}_n = S, \theta = 1, \tilde{x}_{n-1} = 0)
\end{aligned}$$

Substituting these into our earlier expressions in here we get closed form expressions for $\mathbb{P}(\tilde{\tau}_n = N|z_{n-1})$ and $\mathbb{P}(\tilde{x}_n = \theta|\tilde{\tau}_n = S)$, which can in turn be substituted into our original expression:

$$\tilde{z}_n := \mathbb{P}(\tilde{x}_n = \theta) = \mathbb{P}(\tilde{\tau}_n = N|\tilde{z}_{n-1})\big(\frac{1}{2}\big) + \mathbb{P}(\tilde{\tau}_n = S|\tilde{z}_{n-1}) \times \mathbb{P}(\tilde{x}_n = \theta|\tilde{\tau}_n = S, \tilde{z}_{n-1})$$

Since these are very messy expressions, I shall omit to actually write them out here. Nonetheless, plotting the relevant curves illustrates their properties. When I include figures from this example in the body of the main article, I am simply plotting these curves.

F.2 Poisson Reputation Concerns and Beta Signals

In this subsection I consider when $r-1$ is distributed according to a Poisson distribution with parameter λ. In this instance, one can compute that the two relevant pdfs are:

$$f_{VR}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \frac{2\zeta}{\left(1-\frac{ck}{k+1}\right)^2} \cdot \mathbf{1}\left(0 \leq \zeta \leq 1 - \frac{ck}{k+1}\right)$$

$$f_{VW}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \frac{2\zeta}{(1+ck)^2} \cdot \mathbf{1}\left(0 \leq \zeta \leq 1 + ck\right)$$

Following the same steps as before this produces the four CDFs we need:

$$F_{VR11}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \begin{cases} \left(\frac{\zeta}{1-\frac{z_{n-1}k}{k+1}}\right)^2 & \text{if } \zeta < 1 - \frac{z_{n-1}k}{k+1} \\ 1 & \text{if } \zeta \geq 1 - \frac{z_{n-1}k}{k+1} \end{cases}$$

$$F_{VR10}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \begin{cases} \left(\frac{\zeta}{1-\frac{(1-z_{n-1})k}{k+1}}\right)^2 & \text{if } \zeta < 1 - \frac{(1-z_{n-1})k}{k+1} \\ 1 & \text{if } \zeta \geq 1 - \frac{(1-z_{n-1})k}{k+1} \end{cases}$$

$$F_{VW11}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \begin{cases} \left(\frac{\zeta}{1+z_{n-1}k}\right)^2 & \text{if } \zeta < 1 + z_{n-1}k \\ 1 & \text{if } \zeta \geq 1 + z_{n-1}k \end{cases}$$

$$F_{VW10}(\zeta) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \cdot \begin{cases} \left(\frac{\zeta}{1+(1-z_{n-1})k}\right)^2 & \text{if } \zeta < 1 + (1-z_{n-1})k \\ 1 & \text{if } \zeta \geq 1 + (1-z_{n-1})k \end{cases}$$

Substituting these values into the expressions we have already derived allows us to plot the graphs with the parameters distributed according to these new distributions. Figures 5a, 5b and 7a are computed using these quantities.

F.3 Normal Signals

Here I record formulae that give the distribution of private beliefs when the private signals are distributed normally. I do not go through the same analytic computation as above as it is too laborious. Instead I have simply written some code to compute the thresholds within which a private signal produces an invisible action for each possible neighbour-action, and work out the required probabilities directly.

F.3.1 Antisymmetric Case ($\sigma_0 = \sigma_1$)

Let μ_0, μ_1 be the signal means in state 0 and 1 respectively, assuming the variance of each is 1. Then, the probability that an agent forms a posterior belief less than or equal to p, given that the true state is

0, is:

$$\mathbb{P}_{\theta=0}\left(\mathbb{P}(\theta=1 \mid s) \leq p\right) = \Phi\left(\frac{1}{2(\mu_0-\mu_1)}\left[2\log\left(\frac{1-p}{p}\right) - (\mu_1^2-\mu_0^2)\right] - \mu_0\right)$$

where Φ is the standard normal CDF. If, to reduce the number of parameters without the loss of any degrees of freedom, we can set the signal means as $\mu_0 = -\mu$, $\mu_1 = \mu$, we have:

$$\mathbb{P}_{\theta=1}\left(\mathbb{P}(\theta=1 \mid s) \leq p\right) = \Phi\left(-\frac{\log\left(\frac{1-p}{p}\right)}{2\mu} - \mu\right)$$

$$\mathbb{P}_{\theta=0}\left(\mathbb{P}(\theta=1 \mid s) \leq p\right) = \Phi\left(-\frac{\log\left(\frac{1-p}{p}\right)}{2\mu} + \mu\right)$$

F.3.2 Non-antisymmetric case

Now suppose the variances of the two signals are not the same (which allows us to generate non-antisymmetric private beliefs). They are now distributed as:

$$s \mid \theta = i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

where I further assume that $\mu_0 = -\mu$, $\mu_1 = \mu$ for some $\mu > 0$.

We wish to compute:

$$\mathbb{P}_{\theta=1}\left(\mathbb{P}(\theta=1 \mid s) \leq p\right)$$

This is equivalent to solving:

$$\log\left(\frac{f_1(s)}{f_0(s)}\right) \leq \log\left(\frac{p}{1-p}\right)$$

where the f functions are the density functions of the signals.

The log-likelihood ratio is:

$$\log\left(\frac{f_1(s)}{f_0(s)}\right) = \log\left(\frac{\sigma_0}{\sigma_1}\right) - \frac{(s-\mu)^2}{2\sigma_1^2} + \frac{(s+\mu)^2}{2\sigma_0^2}$$

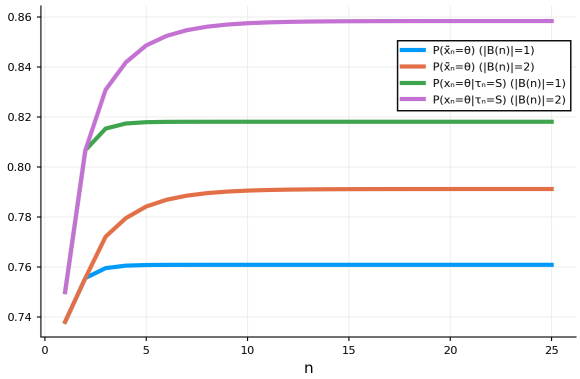
Hence the probability of a private belief lower than p is the probability that the following quadratic inequality holds:

$$as^2 + bs + c \leq 0$$

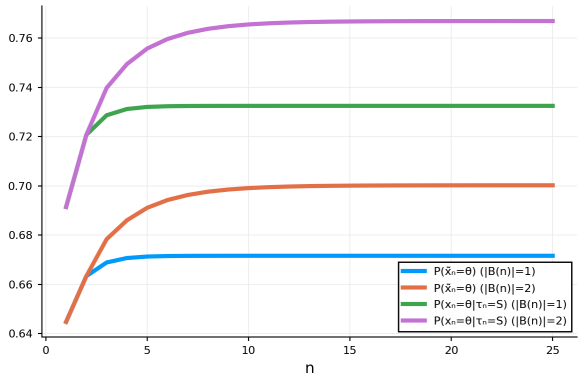
where

$$\begin{aligned} a &= \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2} \\ b &= \frac{\mu}{\sigma_0^2} + \frac{\mu}{\sigma_1^2} \\ c &= \log\left(\frac{\sigma_0}{\sigma_1}\right) + \frac{\mu^2}{2\sigma_0^2} - \frac{\mu^2}{2\sigma_1^2} - \log\left(\frac{p}{1-p}\right) \end{aligned}$$

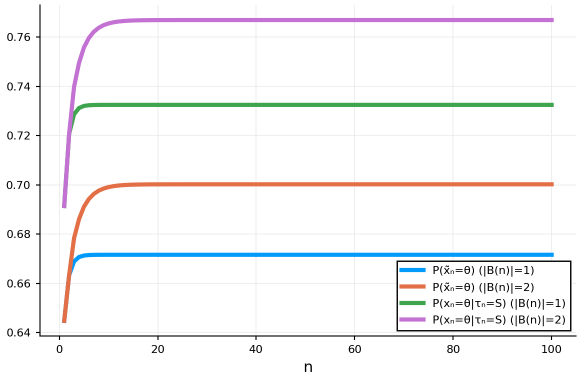
Let s_1, s_2 be the roots of the quadratic. Then:



(a) With signal structure $\{2(1-s), 2s\}$.



(b) With signal structure $\{N(0,1), N(1,1)\}$



(c) Normal but showing up to $n/\tilde{n} = 100$

Figure 8: The parameters for these graphs are the same as in Example 1, except that I plot both the immediate predecessor network and the 2-immediate predecessor network.

$$\mathbb{P}_{\theta=1}\left(\mathbb{P}(\theta = 1 \mid s) \leq p\right) = \Phi\left(\frac{\max(s_1, s_2) - \mu}{\sigma_1}\right) - \Phi\left(\frac{\min(s_1, s_2) - \mu}{\sigma_1}\right)$$

As with the normal-signal antisymmetric case, we can then compute the required thresholds using this. Figure 7b is computed with this procedure (using the parameters specified just above it). Since this computational method does not require the same simplifying assumptions as the analytical case, I also use it to study what happens in a 2-immediate predecessor network. Doing this has produced no important insights so I exclude them from the main article, but here I present some graphs showing how the Example 1 graphs in Figure 4a change. The first two agents are always identical in these graphs since the network topology is identical for them, but in the 2-immediate predecessor agents perform better asymptotically.