Universitat Pompeu Fabra Barcelona

Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1927

Keeping in the dark with hard evidence

Daniel Bird and Alexander Frug

October 2025

Keeping in the Dark with Hard Evidence

Daniel Bird[*] Alexander Frug[†]

October 16, 2025

Abstract

We present a dynamic learning setting in which the periodic data observed by the decision-maker is mediated by an agent. We study when, and to what extent, this mediation can distort the decision-maker's long-run learning, even though the agent's reports are restricted to consist of verifiable hard evidence and must adhere to certain standards. We introduce the manipulation-proof law of large numbers – that delivers a sharp dichotomy: when it holds, the decision-maker's learning is guaranteed in the long-run; when it fails, the scope for manipulation is essentially unrestricted.

1 Introduction

Classical statistics assures us that increasingly large samples of independent observations lead to arbitrarily accurate inference and correct decisions. But what if the data are filtered by someone with an agenda? In many practical environments – from employee evaluation to policy choices by government agencies – the decision-maker never sees the raw data. Instead, data is processed, summarized, and presented by self-interested agents. When agents have even limited discretion over

[*] Eitan Berglas School of Economics, Tel Aviv University (e-mail: dbird@tauex.tau.ac.il).

[†] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Economics (e-mail: alexander.frug@upf.edu).

Acknowledgments to be added.

how to present the data – e.g., what pieces of evidence to disclose, how to summarize large data into concise reports, which statistical method to apply – the statistical learning problem becomes entangled with strategic behavior. This paper studies, in a dynamic setting, when such strategic reporting undermines the Law of Large Numbers (LLN), and just how far agency frictions can distort long-run inference.

We first illustrate the strength and discontinuous nature of the effect that limited flexibility in periodic reporting can have in a "reputation management" setting where an agent who knows his type—Good or Bad—seeks to convince a decision-maker to "accept" him as often as possible. Periodic information and reports work as follows: in each period, the agent observes two conditionally independent realizations of an informative binary signal and discloses exactly one. The decision-maker then updates her belief and accepts the agent for that period if and only if she believes he is Good with probability at least ν^*. This setting captures, for example, a corporate entity that provides periodic financial reports to obtain or retain an investment-grade rating from a credit rating agency.

We characterize when the decision-maker's long-run beliefs—and, hence, the long-term acceptance rate—are susceptible to manipulation. It turns out that this occurs if and only if a simple condition holds. We term this condition *Keeping in the Dark* (KID): there exists a pair of state-dependent reporting strategies that generate the same periodic distribution of reports across states.

When KID fails, the logic of LLN extends to the mediated learning setting and the DM's learning is (probabilistically) guaranteed. Interestingly, if KID holds, then the long-run manipulation of the DM's beliefs can be remarkably effective. Not only can the long-term acceptance rate be increased for any interior DM's prior belief, we show that, unless the DM's prior belief belongs to a specific interval, she can be induced to accept the agent, *in equilibrium*, with probability arbitrarily close to the theoretical upper bound where any Bayes-plausible belief manipulation is allowed. We refer to this bound as Kamenica-Gentzkow Bound, (or KGB). The above mentioned interval of beliefs can be understood, roughly, as all DM's beliefs "within one positive evidence from acceptance"; this is an open interval (ν^*_{-1}, ν^*), just below the acceptance threshold, bounded away from zero. Figure (1) provides a visual summary of the above discussion.

Recall that, in Kamenica and Gentzkow (2011), achieving maximal persuasion required the agent to have access to all possible (direct) signals and commit to an

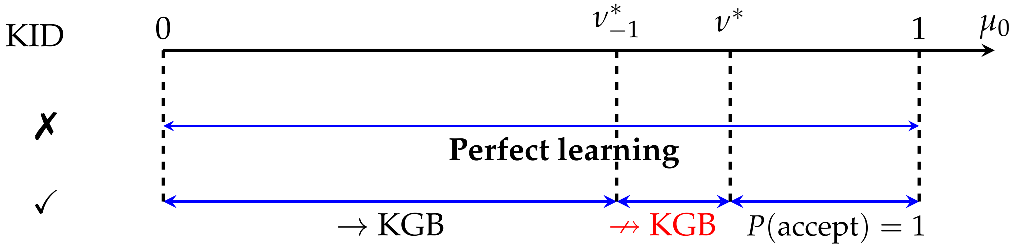


Figure 1: Optimal persuasion and the prior.

experiment before learning his type. In our setting, by contrast, the same long-term acceptance rate can be approximately achieved in equilibrium, using a simple, exogenously given, binary signal.

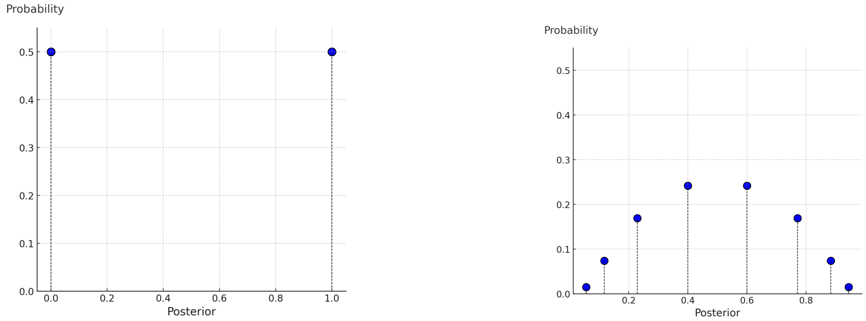
To attain the belief manipulation described above, the agent's reporting strategy must rely on private information. In the reputation-management setting, the agent was fully informed about his type. To study the precise role of this informational advantage in shaping the DM's long-run beliefs, we turn to a setting with two-sided gradual learning: the initially uninformed agent learns from unmediated signals provided by nature, while the DM continues to learn solely through the agent's periodic reports.

Moreover, to isolate the role of informational advantage in the cleanest possible way, we consider a setting in which the agent's goal is not to steer the DM's belief in a particular direction, but simply to prevent her from learning. Beyond clarifying the role of information asymmetry and characterizing the level of informational advantage required to block the DM's learning, we believe this setting is both novel and economically relevant, making it a natural object of study in its own right.

Consider the following "deep state" setting. A politician (DM) wants to learn whether the state of the world is *Green* or *Blue* to enact an appropriate reform (she may also maintain the status-quo). Implementing a reform is costly for a bureaucrat (agent), so he prefers to preserve the status quo. Like before, every period the bureaucrat observes two conditionally independent realizations of a binary signal from which he must disclose exactly one. However, suppose now that both the politician and the bureaucrat are initially uninformed. We ask whether, despite beginning with no informational advantage, the bureaucrat can use strategic disclosure of hard evidence to block the politician's learning and thereby prevent reform.

The main step in the analysis fully characterizes what (partial) informational advantage is sufficient to fully block the DM's learning in a given period. It turns out that whether this informational advantage can be attained in finite time collapses to essentially the same (KID) condition with a mild modification: instead of merely requiring that the state-dependent report distributions intersect, (S-KID) requires their intersection to have a non-empty interior.

We illustrate two qualitatively distinct but representative cases in Figure (2), employing a symmetric version of the setting where the prior is uniform and each realization of the binary signal – of which the bureaucrat observes two realizations – matches the state with probability $\theta > \frac{1}{2}$.



(a) $\theta = \frac{4}{5}$

(b) $\theta = \frac{3}{5}$

Figure 2: Distribution of DM's long-run beliefs.

Suppose first that $\theta = \frac{4}{5}$. In this case, (S-KID) fails and indeed, regardless of the bureaucrat's strategy, the politician will learn the state in the long-run. On the other hand, if $\theta = \frac{3}{5}$, the bureaucrat can induce the long-run belief distribution depicted in Figure (2b). This can be done using a simple strategy: First, reveal one of the two realizations at random in the first 7 periods. During that time, the agent's informational advantage is insufficient to block DM's learning. It turns out that, *regardless of the disclosed signal realizations in the first 7 periods*, the required informational advantage is accumulated and from that period onwards, the DM's learning can be blocked forever (explaining the 8 distinct beliefs in Figure 2b).

To interpret the belief distribution in Figure (2b) within the deep-state scenario, note that if the politician requires a 95% confidence level to initiate a reform, then in this example she will maintain the status-quo with probability 1. That is, the bureaucrat can prevent the reform despite the two-sided learning and the requirement to provide an infinite stream of hard-evidence about the state of nature.

While the reputation-management and the deep-state scenarios differ in the initial information, the players' objectives, and even the frequency of DM's actions, they share important common features. The simplicity of the periodic sampling and reporting (a structure we term *selective forced disclosure*), combined with concrete players' objectives allow us to analyze equilibrium long-run beliefs and, in certain cases, to solve for the induced belief distributions explicitly.

Next, we turn to a broader class of mediated learning problems that abstracts away from particular players' objectives or reporting structure. We consider a general setting in which, each period, the agent observes a random sample drawn from a state-dependent distribution and submits a report subject to sample-specific constraints. These constraints can represent a wide range of real-world partial manipulations, including selective forced disclosure, informal summaries, data cleaning, data fabrication, and more. Again, the agent has some, but limited, freedom in translating samples into reports, and the DM relies on these mediated reports to learn the underlying state.

In this general mediated learning environment, we shift focus away from characterizing equilibrium strategies or optimal behavior under specific objectives. Instead, we ask what are the feasible boundaries of long-run manipulation. We characterize the entire set of long-run belief distributions that can be induced.

We derive a structural property, the *Manipulation-Proof Law of Large Numbers* (MP-LLN), which determines whether long-run learning is guaranteed regardless of the agent's preferences, strategies, or private information. When MP-LLN holds,[1] the decision-maker's beliefs converge (in probability) to the truth over time, extending the classical logic of the LLN to strategically filtered data.

The main message of this part of the paper concerns the limits of manipulation when the MP-LLN fails. Since the decision-maker is Bayesian, any manipulation must satisfy Bayes plausibility: the resulting distribution of beliefs must average to her prior. It turns out that Bayes plausibility is essentially the only constraint. Specifically, we show that if the MP-LLN does not hold on a given subset of the state space, then, within that subset, any Bayes-plausible long-run belief distribution with finite support can be induced. In this sense, the MP-LLN serves as a sharp boundary between mediated learning environments where learning is guaranteed and those where extreme manipulation is possible.

The paper proceeds as follows. In Section 2 we formally define the selective-

[1] This is equivalent to the failure of (KID) in the selective forced-disclosure setting considered above.

forced disclosure model. In Sections 3 and 4 we analyze, respectively, the applications of reputation management and reform avoidance. In Section 5 we study the general mediated-learning framework, and discuss the MP-LLN and its failure. In Section 6 we review the related literature. All proofs are relegated to the Appendix.

2 Selective Forced Disclosure

We now develop a simple model in which the agent has some, but limited, ability to manipulate periodic reports. Within this learning and communication structure, we consider two scenarios to study the scope of long-run manipulation in equilibrium and examine how the agent's informational advantage shapes the DM's beliefs. In Section 5, we extend the analysis to a general mediated-learning framework, where we establish general feasibility bounds on long-run belief manipulation.

Suppose there are two states B and G and let the prior probability that the state is G be $\mu_0 \in (0,1)$. A baseline signal

$$\langle \{b,g\}, \{\theta_\omega\}_{\omega \in \{B,G\}} \rangle,$$

consists of a signal realization space $\{b,g\}$ and a conditional distribution represented by the probability $Pr(g|\omega) = \theta_\omega$ of the signal realization being g at state $\omega \in \{B,G\}$. We assume that $\theta_B < \theta_G$, meaning that a signal realization of g is indicative of state G, while a realization of b is indicative of state B.

In each period, the agent privately observes two conditionally independent realizations of the baseline signal and must disclose exactly one of them.

This model imposes simple bounds on the distribution of the agent's periodic reports in each state; an object that we will show below is inherently tied to the agent's ability to manipulate the DM's long-term beliefs. The lower bound on the probability of disclosing a signal realization of g in state ω corresponds to the agent's strategy where he discloses a realization of b whenever possible. Under this strategy g is disclosed only if both of the signal realizations are g; an event that occurs with probability θ_ω^2. The upper bound on this probability is attained by the opposite strategy where the agent discloses a realization of g whenever possible. Under this strategy a signal realization of g is disclosed unless both the signal realizations are b; an event that occurs with probability $1-(1-\theta_\omega)^2$. By randomizing between disclosing g or b when both are available, the agent can implement any disclosure probability between the two bounds above.

If the minimal probability with which g must be disclosed in state G (i.e., θ_G^2) is greater than the maximal probability with which it can be disclosed in state B (i.e., $1 - (1 - \theta_B)^2$) then, regardless of the agent's disclosure strategy, the DM will eventually learn the state of nature since the long-run frequencies of disclosing g in the two states are bounded away from one another.

Therefore, a necessary condition for manipulation to be feasible is,

$$\theta_G^2 \leq 2\theta_B - \theta_B^2. \tag{KID}$$

When this condition is satisfied, it opens the possibility for different agent-types to "collude" on a periodic uninformative disclosure strategy. Specifically, since the sets of distributions over disclosed evidence across different states intersect, there exists a strategy profile under which the probability of disclosing a signal realization of g is the same in both states. Therefore, we refer to this condition as Keeping In the Dark.

3 The story of reputation management

Consider an agent who knows his type – Good (G) or Bad (B) – and seeks repeated acceptance from a DM who updates her beliefs based on the hard evidence disclosed over time. The DM accepts the agent in a given period if her belief that the agent is Good, in that period, meets or exceeds a given threshold $\nu^* \in (0, 1)$. The agent discounts future payoffs at rate $\beta < 1$. We conduct an equilibrium analysis where, at every period, each agent-type selects a disclosure strategy that maximizes his expected discounted continuation payoff.[2]

We find that, if condition KID holds, long-run persuasion can be remarkably effective. Specifically, in many cases, the agent can approximate the strongest long-run belief manipulation consistent with Bayes plausibility, despite the exogenous constraints on the feasible reporting strategies and his inability to commit to a dis-

[2] Since any finite sequence of disclosed realizations occurs with positive probability, the agent's strategy profile uniquely pins down the DM's belief via Bayes' law. Hence, we do not specify the DM's belief or consequent actions. We assume that the DM accepts the agent if she holds a belief of exactly ν^*.

closure strategy. That is, the equilibrium long-run acceptance rate can be made arbitrarily close to the theoretical upper bound characterized in Kamenica and Gentzkow (2011). Recall, that we refer to this upper bound as the Kamenica-Gentzkow Bound (KGB).

However, as we illustrated in Figure (1) in the Introduction, this is not always true. Whether or not this can be done depends on a specific relation between the DM's prior belief about the agent's type and the baseline signal. To describe this relation, consider the following intuitive persuasion strategy: Every period, if at least one of the signal realizations is g, the agent discloses g. We term this strategy *within-period sanitization*.[3]

Now, let $\nu^*_{-1} \in (0, \nu^*)$ denote the DM's belief at which, if the agent plays within-period sanitization and discloses g, the DM will update her belief to exactly ν^* (the acceptance threshold). The main result of this section (Propositions 1 and 2) is that the theoretical upper bound on persuasion (KGB) can be reached (or approximated) in equilibrium, unless the DM's prior belief μ_0 belongs to the interval (ν^*_{-1}, ν^*).

3.1 Mechanisms of reputation management: freeze, mix & wait

A natural idea for boosting the long-run acceptance rate above what can be attained via the naive strategy of within-period sanitization is to use the *early sanitization* strategy: so long as the DM's period t belief is outside the acceptance region (i.e., is less than ν^*) the agent plays within-period sanitization, and once the belief enters the acceptance region the agent switches to playing an uninformative disclosure strategy profile (the existence of which follows from condition KID) thereby, freezes the DM's beliefs.

This strategy is simple and effective: it replaces full learning in the limit (which would occur if the agent used, for example, within-period sanitization) with a distribution of posterior beliefs that includes 0 (reflecting the DM's certainty that the agent is of type B) and beliefs within the interval $[\nu^*, 1)$. Moreover, this strategy and the corresponding DM's beliefs constitute an equilibrium.[4]

Clearly, if the DM's prior belief lies within the acceptance region, $\mu_0 \geq \nu^*$, the

[3] This terminology follows the one introduced by Song Shin (2003).

[4] To see that early sanitization is an equilibrium, note that the DM's beliefs are not altered by periodic disclosures once the agent enters the acceptance region, and disclosing g before this occurs increases the DM's belief. Therefore, disclosing g leads to a quicker entrance into the acceptance region.

early-sanitization strategy trivially achieves maximal acceptance, as the agent is accepted in every period. If $\mu_0 < \nu^*$, early sanitization still increases the long-run acceptance rate relative to full DM's learning, but it typically leads to overshooting: beliefs cross ν^* at points strictly above the threshold, and so the resulting distribution of long-term beliefs is bounded away from KGB.[5] To further increase the long-term acceptance rate, the agent must exert finer control over the belief path.

Consider first the case where $\mu_0 \in (0, \nu^*_{-1})$. We now illustrate a modification of the early sanitization strategy that will serve as a key building block in constructing an equilibrium strategy profile that approximates KGB. To do so, we need to introduce another threshold belief: let $\nu^*_{-2} \in (0, \nu^*_{-1})$ denote the DM's belief at which, if the agent plays within-period sanitization and discloses g, the DM will update her belief to exactly ν^*_{-1}. Consider the following strategy.

> **1-mixing early-sanitization strategy:** In each period, the agent plays within-period sanitization until the DM's belief enters (ν^*_{-2}, ν^*_{-1}) for the first time. When this occurs, the agent's strategy depends on his type. Type B continues to play within-period sanitization. In contrast, type G – if he samples evidence $\{bg\}$ – mixes over which signal realization to disclose. Specifically, he discloses g with probability α, chosen so that upon observing g, the DM's updated belief is exactly ν^*_{-1}.
>
> For this mixing to be part of an equilibrium, type G must be indifferent between disclosing g or b. Since the choice of α ensures that the disclosure of g raises the DM's belief, it follows that a disclosure of b must lower it. Hence, to obtain such indifference, if g is disclosed, the agent enters a waiting phase of τ (possibly stochastic) periods, during which he plays an uninformative periodic disclosure strategy profile.[6] After the waiting phase ends, the agent resumes early sanitization (if b is disclosed in the mixing period, early sanitization resumes immediately).

The 1-mixing early sanitization strategy uses periodic uninformative disclosure

[5] In the KGB, the prior belief $\mu_0 \in (0, \nu^*)$ is replaced with a Bayes-plausible distribution over two posterior beliefs: 0, and ν^*.

[6] Implementing a stochastic waiting time requires a public randomization device or the ability to engage in a two-sided cheap-talk communication to create an appropriate jointly controlled lottery á la Aumann and Hart (2003) and Krishna and Morgan (2004). As we show below we only require simple lotteries with two potential outcomes but, nevertheless, to facilitate exposition we avoid the formal construction of such lotteries.

strategy profiles as an incentivization device. Specifically, the temporary waiting phase introduces a cost of delay as, during this phase, the agent is outside the acceptance region and receives his lowest periodic payoff. Crucially, as the DM perceives any disclosure as uninformative during this phase, the agent cannot bypass the waiting phase to accelerate acceptance.[7] The duration of this phase is calibrated so that the agent's expected discounting until entry to the acceptance region does not depend on which signal is disclosed during the mixing period.

To approximate the maximal Bayes-plausible acceptance rate in the long run, we extend the idea described above to allow for more than one mixing period. Intuitively, suppose that after the first mixing period the belief falls below ν^*_{-1}. Instead of continuing with the early-sanitization strategy (like in the 1-mixing early-sanitization strategy), we can apply the logic of the 1-mixing strategy again. From the ex-ante perspective, this would then allow for two mixing periods, and the number of mixing periods can be extended further. By allowing for multiple potential mixing periods, we show that the probability of entering the acceptance region at ν^* can be made arbitrarily close to the overall probability of eventual acceptance.

Of course, the proof of Proposition 1 below requires additional work and several refinements. For example, anticipating the possibility of future mixing periods affects the expected time to acceptance following earlier mixing periods, which in turn alters the duration of the waiting phases throughout. Nonetheless, the high-level intuition for this result is captured by the structure described above.

Proposition 1. *Assume that Condition KID holds. The maximal Bayes-plausible acceptance rate can be achieved/approximated, whenever* $\mu_0 \notin (\nu^*_{-1}, \nu^*)$.

An interesting insight from Proposition 1 is that to attain maximal persuasion in equilibrium, it may be necessary to employ a periodic uninformative disclosure strategy profile not only after, but also before, the DM's belief enters the acceptance region. Once the belief is already in that region, making the periodic disclosure uninformative allows the agent to protect himself from ever being rejected. In contrast, applying this strategy profile prior to entering the acceptance region helps shape equilibrium incentives and increase the overall acceptance rate in the long run.

[7] The idea of using costly waiting times to influence behavior has already appeared in a number of other settings (see, e.g., Escobar and Zhang, 2021, Antler, Bird and Oliveros, 2023, and Eliaz, Fershtman and Frug, 2024).

The proof of Proposition 1 is constructive: we derive a specific strategy and analyze its properties. This approach is of course not valid for arguing that if the DM's prior belief belongs to the interval (ν^*_{-1}, ν^*), the maximal Bayes-plausible acceptance rate cannot be approximated. Nevertheless, we can derive the following result.

Proposition 2. *If $\mu_0 \in (\nu^*_{-1}, \nu^*)$, then KGB cannot be approximated in equilibrium.*

In the subsequent analysis (Proposition 5) we show that if the agent could commit to a strategy, he could induce any Bayes-plausible long-term belief distribution. Thus, we can conclude that it is the combination of the agent's dynamic incentive and the intermediate priors that jointly preclude efficient long-run persuasion. Roughly speaking, when $\mu_0 \in (\nu^*_{-1}, \nu^*)$, a type G agent who has never disclosed b in the past, must use a mixed strategy to prevent overshooting. However, we show that to support such mixing in equilibrium the type G agent must be accepted with a probability that is bounded away from one.

4 The stroy of reform avoidance: deep state

To achieve maximal long-run persuasion in the scenario analyzed in Section 3, the different agent-types engaged in a delicate and coordinated manipulation of the DM's beliefs. Importantly, the prescribed agent's strategy relied on him *knowing* the state of the world. In many situations, however, the agent may not have such information. He might begin the interaction with only partial knowledge—perhaps as uninformed as the DM, slightly more informed, or even less. Is informational advantage necessary for the agent to manipulate the DM's long-run beliefs? And if so, what degree of informational advantage is required?

To answer these questions, we now analyze a setting with two-sided learning. The initially uninformed agent gradually learns from periodic samples provided by nature, while the DM learns from the agent's periodic reports.

For tractability and clean parametric analysis, we employ the same selective forced disclosure evidence structure that we defined in Section 2 and used in Section 3. Moreover, building on the observation from the previous section – that a necessary condition for belief manipulation is the ability to generate indistinguishable report distributions across states – we now consider a setting in which preventing the DM's learning is the agent's *objective*, rather than a means to some other end. In particular, we develop and analyze a simple "deep state" model.

Consider an interaction between a newly elected politician (DM), who seeks to enact a reform—provided she becomes sufficiently convinced about which reform is appropriate—and a bureaucrat (agent), who must provide periodic reports and implement whatever reform the DM decides to pursue. Implementing any reform is costly for the bureaucrat, so he prefers to preserve the status quo; and therefore has an incentive to prevent the DM's learning.

Formally, the state is either *Blue* or *Green*, and at the start of the interaction, the players share a common prior belief μ_0 that the state is *Green*. The politician can make an irreversible decision – ending the interaction – to enact one of two potential reforms, $\mathcal{B}$ or $\mathcal{G}$. She enacts reform $\mathcal{B}$ (respectively, $\mathcal{G}$) the moment her belief that the state is *Blue* (*Green*) exceeds the threshold $1 - \pi$, for some $\pi < 1/2$. As long as her confidence in the realized state remains below $1 - \pi$, she maintains the status quo and the interaction continues.

The impact of the agent's private information – or lack thereof – is most noticeable in the first period of the interaction. If the agent's disclosable evidence is $\{bb\}$ or $\{gg\}$, he has no choice over what to disclose, and the reported signal is informative about the state. To make the overall signal disclosure uninformative, the agent would need to counterbalance the informativeness of these "no-choice" cases by strategically choosing what to disclose when his available evidence is $\{bg\}$. However, upon sampling $\{bg\}$, the agent's belief is the same in both states, and so cannot dilute the informational content of disclosure in both of the no-choice cases.

Intuitively, for the agent to make the disclosed signal uninformative he must possess *private* information – not emanating from the disclosable evidence – that allows him to apply different rules for choosing what to disclose in a manner that the DM cannot perfectly invert. Since it takes time for the agent to accumulate private information, and information cannot be taken back from a rational DM, the main focus of this section is to determine when is it possible for the agent to stop the DM's learning, at least from some period onwards. In Proposition 3 we show that the agent can do so if and only if

$$\theta_G^2 < 2\theta_B - \theta_B^2, \tag{S-KID}$$

that is, if Condition KID holds with a strict inequality.

We begin by characterizing the agent's (private) information structures that enable him to make the periodic disclosure uninformative. We then apply this characterization to determine whether the bureaucrat can prevent reform in the long run, and what strategies he might use to do so.

4.1 One-shot forced-disclosure with hard and soft information

Consider a single-period forced-disclosure game where prior to collecting and disclosing evidence, the agent privately observes an undisclosable signal (soft information),

$$\langle \mathcal{S}, \{Pr(\cdot|\omega)\}_{\omega \in \{B,G\}} \rangle,$$

where $\mathcal{S}$ is a finite realization space and $Pr(\cdot|\omega)$ is a probability distribution over $\mathcal{S}$ given the state ω. Next, as before, the agent samples two realizations from the baseline signal, of which he must disclose exactly one. We assume that the soft information and baseline signal realizations are conditionally independent.

The agent's type-space at the time of selecting what evidence to disclose is

$$\overline{Y} = \{\{gg\}, \{bg\}, \{bb\}\} \times \mathcal{S},$$

where the first component is the sampled evidence from which he can disclose and the second is soft information. For each $y \in \overline{Y}$, we denote by $\mu(y)$ the agent's updated belief based on both components of his information. To state our characterization we first define a specific disclosure strategy and an intuitive notion over the inferences made from disclosure.

Definition (Belief-contrarian strategy)**.** *We say that the agent uses a belief-contrarian strategy if an agent-type with evidence $\{bg\}$ discloses g if $\mu(y) < \mu_0$ and discloses b if $\mu(y) > \mu_0$.*

Roughly speaking, an agent that uses such a strategy attempts to mislead the DM by disclosing the evidence that goes against the overall information that he has received.

Definition (Meaning reversal strategy)**.** *We say that the agent's strategy leads to (weak) meaning reversal if disclosing g (weakly) decreases the DM's belief that the state is G.*

The next result characterizes the agent's information structures under which an uninformative disclosure strategy profile exists.

Lemma 1. *Suppose the agent's type space is $\overline{Y}$. The belief-contrarian strategy leads to weak meaning reversal if and only if there exists an uninformative disclosure strategy profile.*

Lemma 1 gives rise to a simple condition that determines whether the agent has sufficient private information to render the periodic disclosure uninformative. To present this condition formally, let

$$D \equiv \{s \in S : \mu(\langle bg, s\rangle) \leq \mu_0\}.$$

That is, D is the set of agent-types who can disclose either b or g and have updated the belief that the state is G (weakly) downward after they receive their information.

Lemma 2. *The belief-contrarian strategy results in meaning reversal if:*

$$Pr(gg|G) - Pr(gg|B) \leq Pr(bg, D|B) - Pr(bg, D|G). \tag{1}$$

Moreover, if (1) *holds for some $\mu_0 \in (0,1)$, then it holds for all such μ_0.*

To understand why our characterization is independent of the prior, note that $s \in D$ if the likelihood of receiving information $\langle gb, s\rangle$ in state B is greater than in state G. Since this assessment is independent of the prior, our characterization is also prior-independent.

4.2 From one shot to dynamics: reform avoidance revisited

Building on the condition given in (1), we now return to the dynamic setting. Suppose that the agent's disclosure strategy in a given period prescribes revealing each of the realizations with probability $\frac{1}{2}$. This has the following implications. First, the disclosed realization updates the DM's belief as if she had observed a single draw of the baseline signal. Second, due to the prior independence property of Condition 1, the agent's informational advantage can be represented by the number of privately observed draws from the baseline signal (and is therefore independent on which realizations were disclosed). This is the case because the disclosed and concealed signal realizations are conditionally independent – a property that is preserved not only since the draws are conditionally independent but also through the agent's disclosure strategy.

Therefore, if there exists a time period such that, following random disclosure up to that period, the agent can block the DM's future learning for some history of disclosed realizations, then he can do so for any history of disclosed realizations. In particular, it reduces the task of checking whether the required informational advantage can eventually be attained to checking whether privately observing sufficiently many draws of the baseline signal – or equivalently, whether being arbitrarily close to being fully informed – provides the required informational advantage. The following result shows that this requirement is equivariant to Condition S-KID.

Proposition 3. *There exists $T^* \in \mathbb{N}$ such that the agent can prevent the DM from learning any information from time T^* onward if and only if Condition S-KID holds. Moreover, when such T^* exists it is independent of the prior.*

Although Proposition 3 may appear asymptotic, this is not the case. To illustrate this concretely, consider a symmetric version of the selective forced disclosure model in which both signals have the same precision. That is, assume that $\theta \equiv \theta_G = 1 - \theta_B$. Note that in this symmetric model we require that $\theta > 1/2$ to make signal realization g indicative of state G, and that Condition S-KID becomes $\theta < \frac{1}{\sqrt{2}}$. Figure (3) depicts T^* as a function of θ and shows that, for most of the relevant range of θ, the DM's learning can be stopped after a relatively short amount of time.

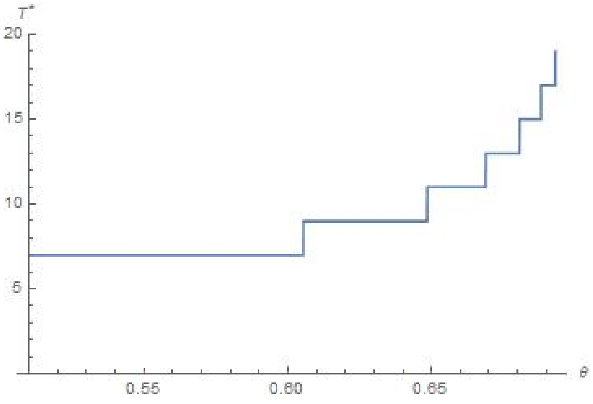


Figure 3: T^* as a function of θ.

Recall the example from the Introduction in which $\theta = \frac{3}{5}$. As can be seen in Figure (3), given the agent's random disclosure strategy, the informational advantage required to block the DM's learning is attained after 7 periods. Hence, the DM will observe 7 informative signals before her learning is blocked (under the random disclosure strategy the order in which of these signals are disclosed is immaterial). This leads to the 8 possible beliefs presented in Figure (2b).

Consider the disclosure strategy where, in the first 7 periods the agent discloses randomly and then switches to uninformative disclosure. Whether this strategy is optimal depends on the DM's policies. In our example, the induced DM's belief after 7 periods belong to the interval $[0.055, 0.945]$. Thus, if, for example, the DM enacts the reform only when she is certain about the state with at least 95%, this simple strategy prevents the reform with probability 1. This shows that our simple strategy is optimal for a range of DM's policies.

5 Manipulation in general mediated learning

Up to this point we have focused on the selective forced disclosure framework, applied to concrete environments, to identify forms of persistent manipulation that arise in equilibrium. The potential for long-run manipulation in mediated learning is clearly not confined to these cases; it exists in any setting where an agent has limited, but nontrivial, discretion over how to present data to a DM.

In this section, we consider a general mediated learning setting. We introduce the Manipulation-Proof Law of Large Numbers (MP-LLN) that specifies when the DM's long-run learning is unaffected by strategic reporting.[8] We then study the limits of manipulation when the MP-LLN fails. Rather than analyzing equilibrium outcomes in a particular setting, we investigate the extent to which the DM's learning can be manipulated, as determined directly by the structure of the periodic data available to the agent and by the reporting protocol that ties his reports to the privately observed samples. We show that the scope of long-run manipulation is essentially bounded only by Bayes plausibility, restricted to subsets of states that satisfy a simple joint manipulability condition introduced below.

Framework

Consider a dynamic mediated learning problem in which the state of nature $\omega \in \Omega$ is distributed according to a prior distribution μ_0. Information about the state is generated by a sequence of periodic samples $\{x_t\}_{t=1}^{T}$ that are drawn i.i.d. according

[8] When the MP-LLN holds, even though the DM will learn the true state, the agent may be able to impact the rate at which the DM learns. In this paper we set aside this aspect of dynamic mediated learning.

to a family of distinct state-dependent distributions $\{\chi_\omega\}_{\omega\in\Omega}$. Let X be the set of possible periodic sample realizations. We assume that both Ω and X are finite.

Every period the agent privately observes the realized sample $x \in X$, and submits a report to the DM. The key notion that captures the agent's flexibility in manipulating the raw data is the feasible report correspondence $R(x)$ that specifies the set of feasible reports for each periodic sample $x \in X$.[9] Let $R = \bigcup_x R(x)$ be the set of all feasible reports. We assume that this set is finite. The agent's periodic strategy is a mapping $\sigma : X \to \Delta(R)$ such that $\sigma(x) \in \Delta(R(x))$, which, for every realization of the periodic sample, specifies a probability distribution over feasible reports.

This framework captures a wide array of environments in which the agent's reports are grounded in real data, but typically do not reveal the full truth. We illustrate how this framework can be used to capture phenomena such as cherry picking, date fabrication, and selective cleaning of the data in Appendix B.1.

5.1 The Manipulation-Proof Law of Large Numbers

In this class of mediated learning problems the DM attempts to infer the state of nature from the sequence of the agent's reports $\{\sigma_t(x_t)\}_{t=1}^T$, where $\sigma_t(\cdot)$ is the reporting strategy used in period t, rather than from the sequence of periodic samples $\{x_t\}_{t=1}^T$. We now establish a condition under which the DM's long-term learning is asymptotically equivalent in both cases.

Given a reporting strategy σ and a state ω, let

$$f(r \mid \omega, \sigma) \;=\; \sum_{x\in X} \sigma(r \mid x)\, \chi_\omega(x),$$

denote the induced distribution of reports in state ω under reporting strategy σ. Note that if the agent uses the strategy σ in every period, then, by the law of large numbers, the long-run frequency of report r in state ω is $f(r \mid \omega, \sigma)$. Denote the set of feasible report distributions in state $\omega \in \Omega$ by

$$\mathcal{F}(\omega) \;=\; \Big\{ f(\cdot \mid \omega, \sigma) : \sigma \text{ is a reporting strategy} \Big\}.$$

Note that these sets are closed and convex.

[9] This object is akin to the feasible reporting set in Dye (1988) or the set of statements in Glazer and Rubinstein (2006).

The key question in determining if the DM will be able to learn the state regardless of the agent's strategy, is whether the sets $\{\mathcal{F}(\omega)\}_{\omega\in\Omega}$ overlap. Let $\mu(\cdot)$ denote the DM's updated belief, and δ_ω the degenerate belief assigning probability one to state ω.

Proposition 4 (Manipulation-proof law of large numbers). *If the sets $\{\mathcal{F}(\omega)\}_{\omega\in\Omega}$ are pairwise disjoint, then for any agent's periodic strategy $\{\sigma_t\}_{t=1}^{\infty}$ that may depend arbitrarily on the past or the agent's private information,*

$$\mu(\{\sigma_t(x_t)\}_{i=1}^{T}) \xrightarrow[T\to\infty]{D} \delta_{\omega'},$$

where ω' is the realized state of the world.

Intuitively, even if the agent distorts which reports are sent in each period arbitrarily, the law of large numbers ensures that the long-run reporting frequencies in state ω converge to an element of the convex set $\mathcal{F}(\omega)$. Since these sets are also closed and disjoint, the long run-reporting frequencies in different states can be separated, and so the DM can eventually identify the true state.

5.2 The scope of long-run manipulation

To study the scope of manipulation when the sets $\{\mathcal{F}(\omega)\}_{\omega\in\Omega}$ are not pairwise disjoint, we begin with the following definition.

We say that a subset $S \subseteq \Omega$ with $|S| > 1$ is *strictly manipulable* if

$$|\bigcap_{\omega\in S} \mathcal{F}(\omega)| > 1.$$

Since the sets $\mathcal{F}(\omega)$ are convex, if S is strictly manipulable, there is a continuum of report distributions that can be induced in every state in S.[10] Note that in the (binary) selective forced disclosure model introduced in Section 2, Condition S-KID is equivariant to the set Ω being strictly manipulable. Thus, the condition for manipulability in that model is a special case of our general condition.

To describe the scope of manipulation on strictly manipulable sets, we need to formalize the notion of conditional Bayes plausibility. Given a prior μ_0 on Ω and a

[10] We discuss the nongeneric case where $|\bigcap_{\omega\in S} \mathcal{F}(\omega)| = 1$ in Appendix C.

subset $S \subseteq \Omega$, let

$$\mu_0^S(\omega) \;=\; \frac{\mu_0(\omega)}{\mu_0(S)} \quad \text{for } \omega \in S$$

be the prior distribution restricted to S. A finite-support distribution $\{(\nu_k, p_k)\}_{k=1}^K$ with $\nu_k \in \Delta(S)$ is said to be *Bayes-plausible conditional on S* if

$$\sum_{k=1}^{K} p_k \, \nu_k(\omega) \;=\; \mu_0^S(\omega) \quad \text{for all } \omega \in S.$$

In words, Bayes-plausibility conditional on S requires that the proposed distribution of posteriors averages back to the prior beliefs conditional on S.

Proposition 5. *Let $S \subseteq \Omega$ be a strictly manipulable set and let $\{(\nu_k, p_k)\}_{k=1}^K$ be a finite-support Bayes-plausible distribution of posteriors conditional on S. There exists a reporting strategy σ such that:*

1. *The DM will identify almost surely whether or not $\theta \in S$.*
2. *If the true state lies in S, then the DM's long-run belief converges to ν_k with probability p_k, for each $k = 1, \ldots, K$.*

Proposition 5 shows that within a strictly manipulable set the only restriction on the agent's ability to manipulate the DM is Bayes-plausibility. That is, the fact that the agent uses periodic samples – which arrive according to a known exogenous process – to craft reports – based on a known reporting rule – does not inherently limit his ability to manipulate the DM's long-term beliefs.

The intuition behind this result is straightforward. For states in S, the fact that $\bigcap_{\omega \in S} \mathcal{F}(\omega)$ contains a continuum of report distributions provides enough flexibility to vary the induced long-run frequencies without interfering with those used in states outside of S. By randomizing across different points in this set in a specific state-dependent manner, the agent can generate in the limit any finite-support distribution over posteriors that is Bayes-plausible conditional on S.

A direct implication of Proposition 5 is that if Ω itself is strictly manipulable, then the scope of manipulation is essentially unrestricted.

Corollary 1 (Global manipulability)**.** *Suppose that Ω is a strictly manipulable set. Then in the long run the agent can induce any finite-support Bayes-plausible distribution of the DM's beliefs.*

We now return to the selective forced disclosure setting to illustrate in a parametrized example how manipulability may be restricted to certain subsets and how the agent may have to choose on which subsets to manipulate the DM's beliefs.

Example 1 (Selective forced disclosure with three states)**.** *Consider a selective forced disclosure setting in which the agent observes two i.i.d. binary signal realizations each period (interpreted now as good or bad evidence) and is required to disclose exactly one of them. Recall that for a state with probability θ of drawing good evidence, the set of inducible long-run frequencies of reporting good evidence is*

$$F(\theta) \;=\; \left[\theta^2,\, 1-(1-\theta)^2\right].$$

Suppose that there are three states, $\Omega = \{Low, Medium, High\}$, with probabilities of good evidence $\theta_L = 0.2$, $\theta_M \in (0.2, 0.8)$, and $\theta_H = 0.8$, respectively. In this case,

$$\mathcal{F}(\theta_L) = [0.04, 0.36], \quad \mathcal{F}(\theta_M) = [\theta_M^2, 1-(1-\theta_M)^2], \quad \mathcal{F}(\theta_H) = [0.64, 0.96].$$

Since $\mathcal{F}(\theta_L) \cap \mathcal{F}(\theta_H) = \varnothing$, neither the pair $\{Low, High\}$ nor the full set Ω is strictly manipulable. By contrast, $\{Low, Medium\}$ is a strictly manipulable set whenever θ_M satisfies, $0.2 < \theta_M < 0.6$, because in this range the intervals $\mathcal{F}(\theta_L)$ and $\mathcal{F}(\theta_M)$ overlap with non-empty interior. Similarly, $\{Medium, High\}$ is strictly manipulable whenever $0.4 < \theta_M < 0.8$, as $\mathcal{F}(\theta_M)$ and $\mathcal{F}(\theta_H)$ then have overlapping interiors.

Thus, for intermediate values of θ_M both pairs $\{Low, Medium\}$ and $\{Medium, High\}$ are strictly manipulable, while for smaller or larger θ_M only one of these pairs is. In other words, depending on the position of the Medium state, the agent can sometimes pool it with Low, sometimes with High, and sometimes with either. However, for any value of θ_M the agent will be forced to reveal at least one of the states.

6 Related Literature

Research on strategic disclosure of hard evidence has largely focused on static environments, where an agent seeks to prove he is of high quality. The seminal works of Grossman and Hart (1980); Grossman (1981); Milgrom (1981) establish an "unraveling" result: if disclosure is unconstrained, all private information is revealed in equilibrium, driven by the incentive of the best type to separate from lower types.

Later contributions impose various constraints on disclosure. For instance, Verrecchia (1983) studies costly disclosure, while Dye (1985) and Jung and Kwon (1988) incorporate uncertainty about the sender's information endowment.

Closer to our approach is Fishman and Hagerty (1990), who consider a static game in which an agent, observing multiple signals, must disclose exactly one from a restricted set. They examine how varying the agent's discretion over which signals can be reported (i.e., changing the size of the set of the signals that can be disclosed) affects social efficiency. By contrast, our focus is on a dynamic environment, where the agent can select which evidence to present over time, in order to circumvent the DM's learning via the logic of the Law of Large Numbers.[11]

A related literature explores constrained persuasion and selective disclosure. Glazer and Rubinstein (2004, 2006) focus on characterizing the mechanism that minimizes the probability that the listener accepts a sender she should reject in a one-shot game. Our notion of feasible report correspondence is related to their modeling approach. More recently, Antic and Chakraborty (2024) study a static persuasion model in which a sender must choose a subset of verifiable facts to disclose and characterize the optimal disclosure rule via a maximum-weight matching representation. Antic, Chakraborty and Harbaugh (2024) examine how communication between informed parties can be structured to share information while concealing it from an external monitor. Finally, Di Tillio, Ottaviani and Sørensen (2021) analyze how sub-sampling affects the informativeness of the produced evidence, contrasting random and selective sampling. These papers form the closest antecedents to our framework: they study how verifiable evidence can be strategically filtered. Our contribution differs in that we move beyond a one-shot perspective and show how repeated selective provision of evidence can manipulate learning in the long-run.

Our work also contributes to the literature on Bayesian Persuasion started with Kamenica and Gentzkow (2011), where the sender is limited in the amount of information he can provide (see e.g., Bird and Neeman, 2022; Eilat, Eliaz and Mu, 2021; Galperti and Perego, 2024). We add to this body of work by showing that a sender that is subject to a baseline forced-reporting rule can replicate, in the long

[11] A related strand in accounting and finance, often referred to as Earnings Management (See Healy and Wahlen, 1999; Beyer et al., 2010 for extensive reviews of this literature), explores how managers selectively report information under various rules. While that literature shares the theme of selective disclosure, its typical assumptions differ from ours in three key respects: (i) it rarely incorporates long-run learning about a fixed underlying state, (ii) managers typically aim to maximize beliefs about firm value, and (iii) report manipulation is generally costly.

run, the outcomes of an unconstrained sender and, in certain cases, can even attain the optimal persuasion outcome in equilibrium.

Finally, long-run manipulation of beliefs is also the focus on extensive literature that studies reputation in dynamic games (see e.g., Kreps and Wilson 1982; Milgrom and Roberts 1982; Fudenberg and Levine 2009; Pei 2020; Ekmekci et al. 2022). In contrast to our paper, in that literature it is typically assumed that it is costly for the agent to take the action that generates the desirable reputation and it is clear which reputation agent would like to acquire.

References

Antic, Nemanja, and Archishman Chakraborty. 2024. "Selected Facts." CMS-EMS Disscsuion paper #1608.

Antic, Nemanja, Archishman Chakraborty, and Rick Harbaugh. 2024. "Subversive Conversations." *Journal of Polotical Economy*, Forthcoming.

Antler, Yair, Daniel Bird, and Santiago Oliveros. 2023. "Sequential learning." *American Economic Journal: Microeconomics*, 15(1): 399–433.

Aumann, Robert J, and Sergiu Hart. 2003. "Long cheap talk." *Econometrica*, 71(6): 1619–1660.

Beyer, Anne, Daniel A. Cohen, Thomas Z. Lys, and Beverly R. Walther. 2010. "The financial reporting environment: Review of the recent literature." *Journal of Accounting and Economics*, 50(2): 296–343.

Bird, Daniel, and Zvika Neeman. 2022. "What should a firm know? Protecting consumers' privacy rents." *American Economic Journal: Microeconomics*, 14(4): 257–295.

Di Tillio, Alfredo, Marco Ottaviani, and Peter Norman Sørensen. 2021. "Strategic sample selection." *Econometrica*, 89(2): 911–953.

Dye, Ronald A. 1985. "Disclosure of nonproprietary information." *Journal of Accounting Research*, 123–145.

Dye, Ronald A. 1988. "Earnings management in an overlapping generations model." *Journal of Accounting research*, 195–235.

Eilat, Ran, Kfir Eliaz, and Xiaosheng Mu. 2021. "Bayesian privacy." *Theoretical Economics*, 16(4): 1557–1603.

Ekmekci, Mehmet, Leandro Gorno, Lucas Maestri, Jian Sun, and Dong Wei. 2022. "Learning from Manipulable Signals." *American Economic Review*, 112(12): 3995–4040.

Eliaz, Kfir, Daniel Fershtman, and Alexander Frug. 2024. "Clerks." Working paper.

Escobar, Juan F, and Qiaoxi Zhang. 2021. "Delegating learning." *Theoretical Economics*, 16(2): 571–603.

Fishman, Michael J, and Kathleen M Hagerty. 1990. "The optimal amount of discretion to allow in disclosure." *The Quarterly Journal of Economics*, 105(2): 427–444.

Fudenberg, Drew, and David K Levine. 2009. "Reputation and equilibrium selection in games with a patient player." In *A Long-Run Collaboration On Long-Run Games*. 123–142. World Scientific.

Galperti, Simone, and Jacopo Perego. 2024. "Games with information constraints: seeds and spillovers." *Theortical Economics*.

Glazer, Jacob, and Ariel Rubinstein. 2004. "On optimal rules of persuasion." *Econometrica*, 72(6): 1715–1736.

Glazer, Jacob, and Ariel Rubinstein. 2006. "A study in the pragmatics of persuasion: a game theoretical approach." *Theoretical Economics*, 1: 395–410.

Goodstein, David. 2010. *On fact and fraud: Cautionary tales from the front lines of science.* Princeton University Press.

Grossman, Sanford J. 1981. "The informational role of warranties and private disclosure about product quality." *Journal of Law and Economics*, 24(3): 461–483.

Grossman, Sanford J, and Oliver D Hart. 1980. "Disclosure laws and takeover bids." *Journal of Finance*, 35(2): 323–334.

Healy, Paul M, and James M Wahlen. 1999. "A review of the earnings management literature and its implications for standard setting." *Accounting Horizons*, 13(4): 365–383.

Jung, Woon-Oh, and Young K Kwon. 1988. "Disclosure when the market is unsure of information endowment of managers." *Journal of Accounting Research*, 146–153.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review*, 101(6): 2590–2615.

Kreps, David M, and Robert Wilson. 1982. "Reputation and imperfect information." *Journal of Economic Theory*, 27(2): 253–279.

Krishna, Vijay, and John Morgan. 2004. "The art of conversation: eliciting information from experts through multi-stage communication." *Journal of Economic Theory*, 117(2): 147–179.

Milgrom, Paul, and John Roberts. 1982. "Predation, reputation, and entry deterrence." *Journal of Economic Theory*, 27(2): 280–312.

Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *Bell Journal of Economics*, 380–391.

Pei, Harry. 2020. "Reputation effects under interdependent values." *Econometrica*, 88(5): 2175–2202.

Song Shin, Hyun. 2003. "Disclosures and asset returns." *Econometrica*, 71(1): 105–133.

Verrecchia, Robert E. 1983. "Discretionary disclosure." *Journal of Accounting and Economics*, 5: 179–194.

A Proofs

Proof of Proposition 1. If $\mu_0 \geq v^*$, then by using a strategy profile under which the periodic disclosure is uninformative, all agent-types are accepted with probability one in all periods. That is, all agent-types attain their maximal payoff in the interaction, and so this strategy profile must be an equilibrium.

Next, we consider the case where $\mu_0 \leq \nu^*_{-1}$. We begin, by formally constructing the equilibrium candidate.

Definition of K-Mixing Early-Sanitization Strategy

If the DM's belief that the agent is good is at least ν^*, then the agent use an uninformative disclosure strategy profile. For lower beliefs, we define the equilibrium candidate via a three dimensional state space $\langle \mu, \tau, k \rangle$, where $\mu \in [0, \nu^*)$ is the DM's belief that the agent is good, $\tau \in \mathbb{N} \cup \{0\}$ is the duration of the current waiting phase, and $k \in \mathbb{N} \cup \{0\}$ counts the maximal number of waiting phases that can occur in the future (whiles still outside the acceptance region). The initial state is given by $(\mu_0, 0, K)$.

We now specify the agent's disclosure strategy as a function of the state, and the evolution of the state over time. For each state $\langle \mu, \tau, k \rangle$ and agent's strategy described below, denote by μ'_i the belief derived via Bayes updating following the disclosure of signal i. Furthermore, let $\Xi \equiv (\nu^*_{-2}, \nu^*_{-1})$, where ν^*_{-1} is formally defined by

$$\nu^* = \frac{\nu^*_{-1}(1-(1-\theta_G)^2)}{\nu^*_{-1}(1-(1-\theta_G)^2) + (1-\nu^*_{-1})(1-(1-\theta_B)^2)},$$

and ν^*_{-2} is defined in analogous manner. The state space can be partitioned into three classes.

1. *Waiting states* are states of the form $\langle \mu, \tau, k \rangle$, where $\tau > 0$. In such states the agent uses an uninformative disclosure strategy profile. Following each waiting period, the state transitions to state $\langle \mu, \tau - 1, k \rangle$.

2. *Sanitization states* are states of the form $\langle \mu, 0, k \rangle$, where $\mu \notin \Xi$. In such states the agent uses within-period sanitization, regardless of his type, and the state transitions to state $\langle \mu'_i, 0, k \rangle$.

3. *Mixing states* are states of the form $\langle \mu, 0, k \rangle$, where $\mu \in \Xi$. In such states, a good agent that observes $\{bg\}$ discloses g with probability $\alpha^*(\mu)$ (defined below), and a bad agent discloses g if possible. Following the disclosure of b, the state transitions to state $\langle \mu'_b, \sigma^*_b(\mu, k), k-1 \rangle$, whereas following the disclosure of g the state transitions to state $\langle \nu^*_{-1}, \tau^*(\mu, k) + \sigma^*_g(\mu, k), k-1 \rangle$, where $\sigma^*_i(\cdot, \cdot)$ and $\tau^*(\cdot, \cdot)$ are lotteries with a support of at most two consecutive integers (defined below).

Note that, by construction, since $\mu_0 < \nu^*_{-1}$ the DM's beliefs cannot reach the interval (ν^*_{-1}, ν^*) while $k > 0$.

Definition of the Mixing Probabilities. In this step we define the function

$$\alpha^* : \Xi \to (0,1)$$

that specifics the probability that a good agent that observes $\{bg\}$ discloses g in mixing states. Consider a mixing state with belief $\mu \in \Xi$. If a good agent discloses g whenever possible (i.e., $\alpha = 1$), then, according to the definition of Ξ, the DM's belief following the disclosure of g will exceed ν^*_{-1}. Conversely, if a good agent discloses b whenever possible (i.e., $\alpha = 0$), then, due to assumption (KID) the DM's beliefs will (weakly) decrease following the disclosure of g. Since the DM's updated belief is continuous in α, by the Intermediate Value Theorem there exits $\alpha^*(\mu) \in (0,1)$ such that the DM's belief following the disclosure of g is exactly ν^*_{-1}.

Duration of the Waiting Phases. In this step we define the functions

$$\tau^* : \Xi \times \mathbb{N} \to \tilde{\Delta} \quad ; \quad \sigma^*_i : \Xi \times \mathbb{N} \to \tilde{\Delta}$$

where $\tilde{\Delta}$ is the space of lotteries with a support of at most two consecutive integers. We construct these functions in K iterative steps.

Step 1: Suppose we are in the last mixing phase with belief $\mu \in \Xi$, i.e., let $k = 1$. We begin by specifying the waiting time $\tau^*(\mu, 1)$ that is designed to make a good agent indifferent between disclosing either signal at mixing state $\langle \mu, 0, 1 \rangle$, under the assumption that $\sigma^*_i(\cdot, 1) = 0$. By Bayes plausibility the disclosure of b must reduce the DM's belief if a good agent uses the strategy $\alpha^*(\mu)$. If $\tau^*(\mu, 1) = 0$, then a good agent prefers to disclose g whenever possible. Conversely, if $\tau^*(\mu, 1) \to \infty$, then a good agent prefers to disclose b whenever possible. Let $\overline{\tau}$ denote the maximal integer such that a good agent weakly prefers to disclose g for all waiting phases of duration $\tau(\mu, 1) \leq \overline{\tau}$. If a good agent is indifferent between disclosing b and g for a waiting time of $\overline{\tau}$, set $\tau^*(\mu, 1) = \overline{\tau}$. Otherwise, there exists a lottery with support $\{\overline{\tau}, \overline{\tau} + 1\}$ that makes a good agent indifferent. In this case, set $\tau^*(\mu, 1)$ to be this lottery.

Next, we specify the additional waiting time following the disclosure of g, $\sigma^*_g(\mu, 1)$. This lottery is chosen so that the distribution of time it takes a good agent that discloses g in a mixing state to reach the acceptance region is the same across all mixing states with $k = 1$. As we show below, this will be useful to incentivize the disclosure of g in sanitization states prior to reaching this mixing state.

Formally, define

$$\hat{\tau}(1) = inf_{\tilde{\mu} \in \Xi}\{\mathbb{E}(\beta^{\tau^*(\tilde{\mu},1)})\},$$

and define $\sigma_g^*(\mu, 1) \in \tilde{\Delta}$ implicitly by

$$\hat{\tau}(1) = \mathbb{E}(\beta^{\tau^*(\mu,1)+\sigma_g^*(\mu,1)}).$$

To maintain a good agent's indifference at mixing state $\langle \mu, 0, 1 \rangle$ we must also add a waiting phase following the disclosure of b. Let $\sigma_b^*(1, \mu) \in \tilde{\Delta}$ be the lottery over waiting times that attains this indifference.[12] This completes step 1.

After $n < K$ steps, we have defined $\tau^*(\mu, k)$ and $\sigma_i^*(\mu, k)$ for all $\mu \in \Xi$ and $k \leq n$. We now consider step $n + 1$. Using $\{\tau^*(\mu, k), \sigma_i^*(\mu, k) | k \leq n, \mu \in \Xi, i \in \{b, g\}\}$, by the same arguments as before there exists a lottery $\tau^*(\mu, n + 1) \in \tilde{\Delta}$ that makes a good agent indifferent between disclosing b or g at mixing state $\langle \mu, 0, n + 1 \rangle$, provided that $\sigma_i^*(\mu, n + 1) = 0$. After defining $\tau^*(\mu, n + 1)$, we define $\sigma_i^*(\mu, n + 1)$ analogously to the way we did in step 1.

Equilibrium

Having defined the equilibrium candidate, we now show that, for every $K \in \mathbb{N}$, the K-mixing early-sanitization strategy is an equilibrium.

In waiting states, the periodic disclosure is meaningless, and so any periodic disclosure strategy is optimal. Similarly, in sanitization phases of the form $\langle \mu, 0, 0 \rangle$ the continuation strategy profile is early-sanitization, which we already established to be a continuation equilibrium.

We proceed by backward induction over k.

Step 1: At mixing state $\langle \mu, 0, 1 \rangle$ a good agent is, by the construction of the lotteries $\tau^*(\mu, 1)$ and $\sigma_i^*(\mu, 1)$, indifferent between disclosing either signal. A bad agent, on the other hand, strictly prefers to disclose g. To see this, note that a bad agent is likely to have less g signal realizations than a good agent in the future, and thus his opportunity cost from forgoing the chance to increase the DM's belief is greater

[12] The existence of $\sigma_i^*(\mu, 1) \in \tilde{\Delta}$ follows from the same reasons that there exists $\tau^*(\mu, 1) \in \tilde{\Delta}$.

than the opportunity cost of a good agent. This implies that the proposed strategy is optimal at mixing states of the form $\langle \mu, 0, 1 \rangle$.

Next we consider sanitization phases of the form $\langle \mu, 0, 1 \rangle$. By the construction of the waiting time lotteries, a good agent's continuation utility is identical at all mixing states of the form $\langle \mu, 0, 1 \rangle$. Moreover, his continuation utility at each such state is positive, whereas his periodic payoff at sanitization phases is zero. It follows that at any sanitization state $\langle \mu, 0, 1 \rangle$, the objective of a good agent is to reach a mixing state as quickly as possible. Since posterior beliefs are increasing in the prior, within period sanitization minimizes the time – in a first order stochastic sense – it takes the agent to reach the next mixing state. As a bad agent's cost of forgoing the opportunity to disclose g is greater than a good agent's cost of doing so, within-period sanitization is optimal for both agent types at any sanitization phase with $k = 1$.

Step n: Given $\tau^*(\cdot, \cdot)$ and $\sigma_i^*(\cdot, \cdot)$, for any given $k < n$ the expected time it takes a good agent to get from a mixing state $\langle \cdot, 0, k \rangle$ to the acceptance region is independent of $\mu \in \Xi$. Thus, at any sanitization state $\langle \cdot, 0, k \rangle$, the agent's objective is to reach the next mixing state as quickly as possible. By the definition of $\tau^*(\mu, n)$ and $\sigma_i^*(\mu, n)$, from the perspective of a good agent, at every mixing state of the form $\langle \mu, 0, n \rangle$, disclosing either g or b leads to the same distribution of time until the next mixing state is reached (i.e., a state of the form $\langle \mu, 0, n-1 \rangle$ where $\mu \in \Xi$), and thus, until a good agent enters the acceptance region. Hence, as before, his objective is to arrive at a mixing state $\langle \mu, 0, n-1 \rangle$ as soon as possible (the belief $\mu \in \Xi$ at which he enters is not important). Therefore, he will disclose g whenever possible, and by the same single-crossing type of argument we used in step 1, it is in the best interest of a bad agent type to disclose g (whenever possible) during the sanitization states as well.

Maximal Persuasion

Finally, we show that the acceptance rate of a bad agent under these equilibria approaches the KGB as $k \to \infty$.

Under any K-mixing early-sanitization equilibrium, a good agent is (eventually) accepted with probability one. To see this note that, by construction, it is optimal for a good agent to disclose g whenever possible. By the Borrel-Cantalli Lemma, such an agent will eventuality be accepted. Either by disclosing g following a waiting phase or once the continuation strategy reverts to early sanitization (i.e., once $k = 0$).

Moreover, the probability that a good agent will pass through K mixing phases converges to zero as K increases to infinity. To see this, note that with a probability of $\theta_G^4 > 0$, a good agent will have only g signals in two given periods. It follows that with strictly positive probability a good agent will be forced to disclose g both in a mixing state and in the next sanitization state (that occurs after $\tau^* + \sigma_g^*$ periods). Hence, the probability that a good agent will pass through K mixing phases is bounded from above by $(1 - \theta_G^4)^K$; an expression that converges to zero as $K \to \infty$. By constriction, if the agent is accepted in a state with $k > 0$, he enters the acceptance region at belief ν^*. Hence, the distribution of the DM's belief upon entry to the acceptance region converges to a degenerate belief on ν^* as $K \to \infty$. ■

Proof of Proposition 2. Any equilibrium disclosure strategy σ defines a probability space over binary sequences of disclosed evidence $\{b, g\}$, and a corresponding martingale of end-of-period beliefs $\mu_\sigma(h_t)$ following the disclosure in the first t periods. Let $\hat{H}(\sigma)$ be the set of (infinite) histories along which beliefs never fall below $\frac{\mu_0+\nu_{-1}^*}{2}$ under σ. Since beliefs are a martingale with a support that is bounded in $[0,1]$, Doob's Martingale Inequality implies that

$$Prob[\hat{H}(\sigma)] \geq \frac{\mu_0 + \nu_{-1}^*}{2 - \mu_0 + \nu_{-1}^*} \equiv M_1.$$

Fix and equilibrium disclosure strategy. One of the following must hold:

i): With a probability of at least $\frac{1}{3}$ a history in $\hat{H}$ never enters the AR (acceptance region: $\mu(\cdot) \geq \nu^*$). In this case, the outcome is bounded away from KGB since a good agent is never accepted with a probability of at least $\frac{M_1}{3}\frac{\mu_0+\nu_{-1}^*}{2}$.

ii): With a probability of at least $\frac{1}{3}$, under a history in $\hat{H}$ the first entry to AR occurs with a report of g at a period where a good agent uses within-period sanitization. In this case, the outcome is bounded away from KGB since the belief following entry to AR is at least $\bar{\nu} > \nu^*$ (a belief that results following within-period sanitization that begins at belief $\frac{\mu_0+\nu_{-1}^*}{2}$). Note that, at such histories, the report g is made with a probability of at least θ_B^2, so that the DM's beliefs upon entry AR will be at least $\bar{\nu}$ with a portability of at least $\theta_B^2 \frac{M1}{3}$. Following such beliefs at entry, Doob's Martingale inequality implies that the DM's long-run belief will be above $\frac{\bar{\nu}+\nu^*}{2}$ with probability of at least $M_2 \equiv \frac{\bar{\nu}+\nu^*}{2-\bar{\nu}+\nu^*}$. Thus, the long run beliefs are bounded away from KG.

iii): With a probability of at least $\frac{1}{3}$, under a history in $\hat{H}$ the first entry to AR occurs when a good agent that observes $\{b, g\}$ discloses b with positive probability. Denote all the prefixes at which such entrance to AR occurs by $\mathcal{H}^{AR}_{b \succeq g}$.

For any $V < \frac{1}{1-\beta}$, let $\overline{T}(V)$ denote the minimal integer such that

$$\sum_{t=0}^{\overline{T}(V)} \beta^t \geq V.$$

Moreover, let $\kappa \equiv \min\{(1-\theta_G)^2, \theta_B^2\}$. Intuitively, κ is a uniform lower bound on the probability that a given signal is disclosed in a specific period. Next, we derive the following lemma about such histories (proof below).

Lemma A.1. *The continuation payoff of both types at any history at $\mathcal{H}^{AR}_{b \succeq g}$ is below $\frac{1}{1-\beta} - \Delta^*$, where $\Delta^* \equiv (\kappa\beta)^{\overline{T}(\frac{\beta}{1-\beta})}$.*

For each $h_t \in \mathcal{H}^{AR}_{b \succeq g}$, we can define a sequence $u(h_t) = (u_j(h_t))_{j=1}^{\infty}$ such that the DM's belief is non-decreasing along the history $(h_t, u(h_t))$. We now partition the set

$$\mathcal{H}^{AR}_{b \succeq g} = I_1 \cup I_2 \cup ... \cup I_{T^\dagger},$$

where $T^\dagger = \overline{T}(\frac{1}{1-\beta} - \Delta^*)$, as follows.

For each $h_t \in \mathcal{H}^{AR}_{b \succeq g}$, if in period $t+1$, following h_t, i) both agent-types disclose evidence according to pure strategy and ii) disclosure is informative, let $h_t \in I_1$. Otherwise, if in period $t+2$, i) both agent-types uses a pure strategy at $(h_t, u_1(h_t))$ and ii) disclosure is informative, let $h_t \in I_2$. Continuing in this manner, $h_t \in I_{j+1}$ if at at every period along $(h_t, (u_i(h_t))_{i=1}^{j-1})$ either disclosure was noninformative of at least one agent-type uses a mixed strategy, whereas at $(h_t, (u_i(h_t))_{i=1}^{j})$ disclosure is informative and both agent types use a pure strategy.

By the definition of $T^\dagger$,

$$h_t \in I_1 \cup_2 \cup \ldots \cup I_{T^\dagger}.$$

Note that there exists $S \in \{0, ..., T^\dagger\}$ for which the measure of I_S is at least $\frac{M_1 \kappa^{T\dagger}}{3T^\dagger} > 0$.

If i) the DM's prior belief is μ, ii) both agent types use a pure strategy, and iii) disclosure is informative, then the support of the belief in the following period is disjoint from $[\mu - \epsilon(\mu), \mu + \epsilon(\mu)]$ for some $\epsilon(\mu) > 0$. Furthermore, note that $\epsilon(\mu)$ is continuous in μ. Hence, there exists $\epsilon^* > 0$, such that $\epsilon^* \leq \epsilon(\mu)$ for every $\mu \in [\nu^*, \frac{\nu^*+1}{2}]$.

Note that if $\mu \in [\nu^*, \frac{\nu^*+1}{2}]$, then the DM's belief in the following period is at least $\nu^* + \epsilon^*$ with a probability of at least κ. Intuitively, exactly one of the signals will lead to an increase in the DM's belief, and κ is the a lower bound on the probability that the agent must disclose that signal. It follows that the DM's belief following I_S is at least $\nu^* + \epsilon^*$ with a probability of at least κ. Hence, the outcome is bounded away from the KGB outcome. ■

Proof of Lemma A.1. Fix a history in $h_t \in \mathcal{H}^{AR}_{b \succeq g}$. Note that at h_t entry to AR occurs following exactly one of the disclosed signal realizations, either b or g.

First, consider the case where entry to AR occurs via the disclosure of g. At period t at a prefix of h_t, a good agent that has evidence $\{b, g\}$ weakly prefers disclosing b. As disclosing b will lead to a payoff of zero in period t, it must be the case that the continuation payoff from disclosing g, from the perspective of that agent, is at most $\frac{\beta}{1-\beta}$.

This means that, upon entry to AR, from the perspective of a good agent, there is a non-zero vector $(r_1, r_2, ...)$, where r_j represents the probability of leaving AR (for the first time) at period $t + j$, from the perspective of period t. Moreover, since this continuation payoff is less than $\frac{\beta}{1-\beta}$, the first $\overline{T}(\frac{\beta}{1-\beta})$ components of this vector cannot all equal zero. Note that any nonzero component from the perspective of a good agent also corresponds to a non-zero component from the perspective of a bad agent, and moreover, each such component is bounded from below by $\kappa^{\overline{T}(\frac{\beta}{1-\beta})}$ for both agent types.

Next, consider the case where entry to AR occurs via the disclosure of b. In this case, at period t (the prefix of h_t), a bad agent that has evidence $\{b, g\}$ must disclose g with positive probability. Otherwise, disclosing b leads to a reduction of the DM's belief, contradicting the assumption that AR is entered via the disclosure of b. Thus, reporting g (and staying outside of AR in period t) is optimal for a bad agent, and so his continuation payoff is at most $\frac{\beta}{1-\beta}$. An analogous argument to the one used in the previous case, establishes the lemma. ■

Proof of Lemma 1.

Part one (1 ⇒ 2)*:* Let σ_U denote the strategy under which a agent-type with evidence $\{bg\}$ discloses each signal realization with probability $\frac{1}{2}$, and let σ_{BC} denote the belief-contrarian strategy.

Next, define the strategy σ_λ to be the mixture between σ_{BC} with probability λ, and σ_U with probability $1-\lambda$. Finally, let $\psi : [0,1] \to \mathbb{R}$ be the function

$$\psi(\lambda) = q_G(\lambda) - q_B(\lambda),$$

where $q_\omega(\lambda)$ is the probability that signal g is disclosed under the strategy σ_λ when the state is ω.

By the premise (part 1 of the lemma) we have that $\psi(1) \leq 0$. On the other hand, under σ_U the disclosure of signal realization g is more likely in state G than in state B. That is, $\Psi(0) > 0$.

Note that $q_\omega(\cdot)$ is continuous, and so $\psi(\cdot)$ is also continuous. Hence, by the Intermediate Value Theorem there exists $\lambda^\star \in (0,1]$ such that $\psi(\lambda^\star) = 0$. That is, the strategy $\sigma_{\lambda^\star}$ makes the disclosed signal realization uninformative about the state.

Part two $(2 \Rightarrow 1)$: To establish this part of the lemma, we first establish an auxiliary result. Namely, that for any weak meaning reversal strategy, σ, (other than the belief-contrarian strategy), altering σ so that agent-type $\hat{y}$ with evidence $\{bg\}$ and belief $\mu(\hat{y})$ goes against his belief – discloses $g(b)$ if $\mu(\hat{y}) < (>)\mu$ – maintains the weak meaning reversal property.

To see why this auxiliary results is true, fix σ and an agent-type $\hat{y}$ such that: 1) $\hat{y}$ has evidence $\{bg\}$, 2) WLOG $\mu(\hat{y}) > \mu_0$, and 3) $\hat{y}$ discloses g with probability $\alpha > 0$. Denote the DM's belief under σ by $P_\sigma(\cdot)$. Let σ' be the strategy that is identical to σ for all $y \neq \hat{y}$, and $\hat{y}$ discloses b with probability one (goes against his belief).

$$\begin{aligned}
\mu_0 \geq P_\sigma(G|g) = & \frac{P_\sigma(G \wedge g)}{P_\sigma(g)} \\
= & \frac{Pr(y \neq \hat{y})P_\sigma(G \wedge g|y \neq \hat{y}) + Pr(y = \hat{y})P_\sigma(G \wedge g|y = \hat{y})}{Pr(y \neq \hat{y})P_\sigma(g|y \neq \hat{y}) + Pr(y = \hat{y})P_\sigma(g|y = \hat{y})} \\
= & \frac{Pr(y \neq \hat{y})P_\sigma(G \wedge g|y \neq \hat{y}) + Pr(y = \hat{y})\mu(\hat{y})\alpha}{Pr(y \neq \hat{y})P_\sigma(g|y \neq \hat{y}) + Pr(y = \hat{y})\alpha} \\
> & \frac{Pr(y \neq \hat{y})P_\sigma(G \wedge g|y \neq \hat{y}) + Pr(y = \hat{y})\mu_0\alpha}{Pr(y \neq \hat{y})P_\sigma(g|y \neq \hat{y}) + Pr(y = \hat{y})\alpha}.
\end{aligned}$$

The first inequality follows from the assumption that σ reverses the meaning of the signal, and the second inequality follows from the assumption that $\mu(\hat{y}) > \mu_0$. By simple algebra, the above inequality implies that

$$\mu_0 > \frac{Pr(t \neq \hat{t})P_\sigma(G \wedge g|t \neq \hat{t})}{Pr(t \neq \hat{t})P_\sigma(g|t \neq \hat{t})} \equiv P_{\sigma'}(G|g).$$

This establishes the auxiliary result.

Now, assume that a strategy $\hat{\sigma}$ makes the disclosure uninformative. By definition, $\hat{\sigma}$ satisfies the weak meaning reversal property. The auxiliary result established above implies that if we iteratively alter $\hat{\sigma}$ so that every agent-type goes against her belief, the resulting strategy will not only be the belief-contrarian strategy, but will also satisfy the weak meaning reversal property. ■

Proof of Lemma 2.

Denote by d the disclosed signal. Under the belief-contrarian strategy

$$Pr(d = g|\omega) = Pr(gg|\omega) + Pr(gb, D|\omega)$$

Weak meaning reversal requires that the likelihood ratio $\frac{Pr(d=g|G)}{Pr(d=g|B)}$ be less than one. Plugging the above probability into $\frac{Pr(d=g|G)}{Pr(d=g|B)} \leq 1$ and rearranging yields Condition (1).

Note that the left-hand side of Equation (1) is independent of μ_0. To establish that the right-hand side of this expression is also independent of μ_0 observe that

$$Pr(gb, D|\omega) = Pr(gb, \mu(y) < \mu_0|\omega) = Pr(bg|\omega)Pr(\mu(y) < \mu_0|gb, \omega).$$

By definition, $Pr(bg|\omega)$ is independent of μ_0. To show that $Pr(\mu(y) < \mu_0|gb, \omega)$ is independent of μ_0, note that (conditional on evidence $\{g, b\}$) a private signal realization of s decreases the agent's beliefs only if $\frac{Pr(gb,s|G)}{Pr(gb,s|B)} < 1$. Since the evidence and private signal are conditionally independent, this is equivalent to

$$\frac{Pr(gb|G)Pr(s|G)}{Pr(gb|B)Pr(s|B)} < 1 \Leftrightarrow \frac{Pr(s|G)}{Pr(s|B)} < \frac{Pr(gb|B)}{Pr(gb|G)}.$$

Therefore, D is independent of μ_0. ■

Proof of Proposition 3.

We begin by considering a single period auxiliary setting in which the agent observes K private realizations of the baseline signal at the beginning of the interaction. Denote by μ_K the agent's (stochastic) belief after observing K private signal realizations. By the law of large numbers,

$$\lim_{K\to\infty} Pr(\mu_K < \mu_0|G) \to 0 \text{ and } \lim_{K\to\infty} Pr(\mu_K < \mu_0|B) \to 1.$$

Denoting by μ_{K+2} the agent's (stochastic) belief after observing the K private signal realizations and the 2 disclosable signal realizations, it follows that

$$\lim_{K\to\infty} Pr(gb, \mu_{K+2} < \mu_0|B) = \lim_{K\to\infty} Pr(gb|B)Pr(\mu_{K+2} < \mu_0|gb, B) = 2\theta_B(1-\theta_B),$$

and that

$$\lim_{K\to\infty} Pr(gb, \mu_{K+2} < \mu_0|G) = \lim_{K\to\infty} Pr(gb|G)Pr(\mu_{K+2} < \mu_0|gb, G) = 0.$$

The LHS of (1) is $\theta_G^2 - \theta_B^2$ regardless of K, and the above calculations show that the RHS of (1) converges to $2\theta_B(1-\theta_B)$ as $K \to \infty$. Thus, Condition (S-KID), implies that Equation (1) is satisfied for large enough K. In combination with Lemma 1, this implies that there exits K^*, such that there exists an uninformative disclosure strategy for all $K > K^*$.

The above implies that in the dynamic model, where the agent is initially uniformed, he can stop the DM's learning from period $T^* = K^*$ and onward. In particular, assume that for T^* periods the agent chooses at random which of the two signal realizations to disclose. This implies, that at the beginning of period $T^* + 1$ the agent's informational advantage is exactly T^* draws of the baseline signal. As the baseline signal realizations are independent of one another (conditional on the state), the DM has no information regrading the realization of these signals. Hence, regardless of the information revealed to the DM in the first T^* periods, there exists a strategy that makes the periodic disclosure in period $T^* + 1$ uninformative. If each agent-type at period $T^* + 1$ keeps using the same periodic strategy in all future periods – regardless of the information they receive from period $T^* + 1$ and onward – the disclosed signal realization remains uninformative in every subsequent period.

Next, we show that if (S-KID) does not hold, then the disclosed signal realization is informative in every period. To see this note that at every period the sequences of periodic signal realizations $(gg), (gg), (gg), \ldots$ and $(bb), (bb), (bb), \ldots$ occur with positive probability in both states of nature. Thus, at every period there is a positive probability that the agent updated his belief in the wrong direction (i.e., increases his belief relative to the prior when the state is B and vice-versa). This, in turn, implies that regardless of the agent's strategy, $Pr(gb, \mu(y) < \mu_0|B) - Pr(gb, \mu(y) < \mu_0|G) < 2\theta_B(1-\theta_B)$. Hence, Condition (1) does not hold, and by Lemma 1 an uninformative disclosure strategy does not exist. ■

Proof of Proposition 4. By the Law of Large Number, for any reporting strategy that the agent may choose, the long-run frequency of reports in state ω will be an element of $\mathcal{F}(\omega)$ with arbitrarily high probability. The sets $\{\mathcal{F}(\omega)\}_{\omega\in\Omega}$ are closed and convex subsets of $\Delta(R)$. Since these sets are pairwise-disjoint, then they can be separated. Hence, the DM will eventually be able to distinguish between any two states with arbitrarily high probability. ■

Proof of Proposition 5.

Fix a strictly manipulable set S and a finite support belief distribution $\{p_k, \nu_k\}_{k=1}^{K}$ that is Bayes-plausible conditional on S.

We begin by constructing a specific reporting strategy. Fix an arbitrary state-contingent reporting strategy σ_0. For states $\omega \in \Omega \setminus S$ the agent uses σ_0 in every period. To construct the rest of the agent's strategy, we need the following notation.

For each $\omega \in \Omega \setminus S$ and $\epsilon > 0$, let $g_\epsilon(\omega)$ denote the open ϵ ball around the long-run reporting frequency $f(\cdot|\omega, \sigma_0)$. Let

$$G_\epsilon = \bigcup_{\omega\in\Omega\setminus S} g_\epsilon(\omega).$$

Fix ε for which the set

$$H = \bigcap_{\omega\in S} \mathcal{F}(\omega) \setminus G_\varepsilon$$

contains a line of strictly positive length. Such ε exists since S is strictly manipulable.

To construct the agent's strategy in S, select K distinct reporting distributions $\{r_k\}_{k=1}^{K}$ that belong to H. For each such distribution, denote the corresponding state-contingent reporting strategy in state $\omega \in S$ by $\sigma_k(\omega)$. In state $\omega \in S$ the agent uses $\sigma_k(\omega)$ (in all periods) with probability $\frac{p_k\nu_k(\omega)}{\mu_0^S(\omega)}$.

By the Law of Large Numbers implies that the distribution of the long-term reporting frequencies will converge to the report distribution conditional on the realized state and the chosen reporting strategy. Since G_ε and H can be separated, the DM will be able to infer weather or not the state belongs to S. Moreover, after observing a long-term reporting frequency of $r_k \in H$ the DM's belief of the state being ω is

$$\frac{\mu_0^S(\omega)\frac{p_k \nu_k(\omega)}{\mu_0^S(\omega)}}{\sum_{\omega' \in S}(\mu_0^S(\omega')\frac{p_k \nu_k(\omega')}{\mu_0^S(\omega')}}) = \frac{p_k \nu_k(\omega)}{\sum_{\omega' \in S}(p_k \nu_k(\omega'))} = \frac{\nu_k(\omega)}{\sum_{\omega' \in S} \nu_k(\omega')} = \nu_k(\omega).$$

■

B Examples of mediated learning settings

In this appendix, we first illustrate how our mediated learning framework can capture other forms of strategic reporting. We then characterize when the MP-LLN holds in each of these examples.

B.1 Examples

Cherry picking

An agent who collects data often observes a large periodic sample. However, the DM may not have the capacity to review all the data-points collected by the agent in a given period, and so requires the agent to submit a report about only part of the observed data-points. This opens the door to cherry-picking by the agent. A specific case of this class is described in Section 2. To embed this in our general framework, we would set $X = \{bb, bg, gg\}$, $R = \{b, g\}$ and have

$$R(bb) = \{b\}, R(gg) = \{g\}, R(bg) = \{b, g\}.$$

Fabricating Data

Even though agents that collect data are supposed to report truthfully, there is substantial evidence of such agents fabricating data.[13] To (concisely) embed such a phenomena in our framework, consider a setting with a binary state-space, where in each period the agent observes one binary signal with precision $\phi > 1/2$ that he must report to the DM. However, with probability $p \in (0,1)$ the agent can fabricate the realization of the signal before submitting the report. In this setting, a periodic sample is $\langle s, f \rangle$, where $s \in \{b, g\}$ is the realization of the signal, and $f \in \{0,1\}$ is an indicator for whether or not the agent can fabricate the signal realization. The set of reports is $R = \{b, g\}$, and

$$R(s,0) = \{s\}, R(s,1) = \{b,g\}.$$

Cleaning the data

The role of an agent that collects data is often to summarize the data into a single summary statistic. However, such agents often have the flexibility to "clean the data," i.e., remove a small number of "non-representative" observations from the sample, before calculating the required statistic.

To illustrate how such a setting can be embedded in our framework, assume that there are two states **Negative** and **Positive**, and that the agent observes a sample of three conditionally independent draws of a random variable with support $\{-2, 0, 2\}$, where higher (lower) realizations are more likely in the positive (negative) state.

The agent must report the realized average of the periodic sample at every period, however, before calculating the average the agent excludes one of the three realizations. In this example, $X = \{-2,0,2\}^3$, $R = \{-2,-1,0,1,2\}$, and

$$r \in R(x) \Leftrightarrow \exists i, j \in \{1,2,3\} \text{ s.t. } i \neq j \text{ and } x_i + x_j = 2r.$$

B.2 Checking whether the condition for the MP-LLN holds

Cleaning the data: Despite the apparent simplicity of this setting, the reporting space is high-dimensional. In particular, this implies that report distributions cannot

[13] See, e.g., Goodstein (2010) for a review of fabrication cases in scientific research.

be easily summarized by a single parameter, and so, it is not trivial to determine whether the condition for the MP-LLN holds.

We demonstrate this using the following (simple) parametric distribution function for the random variable, where $\zeta \in (0, \frac{1}{3})$:

States \ Outcomes	-2	0	2
Negative	$\frac{1}{3}+\zeta$	$\frac{1}{3}$	$\frac{1}{3}-\zeta$
Positive	$\frac{1}{3}-\zeta$	$\frac{1}{3}$	$\frac{1}{3}+\zeta.$

In this setting, regardless of the agent's strategy, the sign of the reported average in a given period corresponds to the state with probability bounded away from zero (in particular, this probability is always greater than $\frac{1}{27}$). Now, take $\zeta \to \frac{1}{3}$ and suppose for a moment that the state is Positive. In this case, the probability of a negative report in a given period approaches zero. Therefore, for sufficiently large ζ, the sign of the average report will match the state with arbitrarily high probability, implying that the condition for the MP-LLN holds.

By always excluding the highest of the three realizations in the Positive state, the agent reduces the periodic report as much as possible. Due to the symmetric structure of the model, excluding the lowest realization in the Negative state pushes the reports upwards at exactly the same rate. Accordingly, there exists a threshold $\zeta_0 \approx 0.141$ at which, under this strategy, the expected periodic reports in both states coincide, and are thus equal to zero.

However, if $\zeta = \zeta_0$, the condition for the MP-LLN still holds. To see this, it is sufficient to note that even though, for $\zeta = \zeta_0$, the strategy described above generates the same expected report in both states (and the same variance of reports), the distribution of reports is different between the two states:

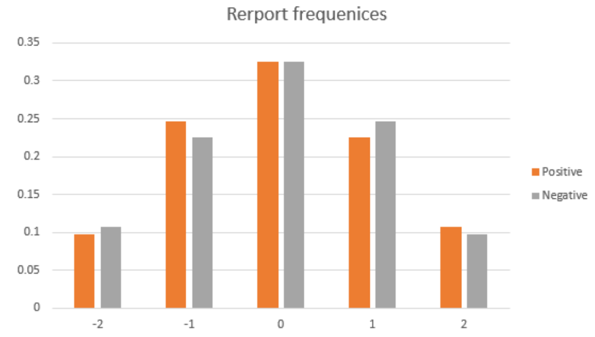


Figure 4: Reporting frequencies for extreme exclusion strategies.

The attentive DM will, therefore, not be misled by the uninformative average of the periodic reports but will infer the state based on the higher moments of the empirical frequencies of reports.

The above discussion implies that for the MP-LLN to fail, it must be the case that $\zeta < \zeta_0$, which, in turn, implies that we must consider reporting strategies that are more nuanced than simply omitting extreme draws. Due to the symmetric nature of this setting, generating a report distribution that is symmetric around $r = 0$ in one state is equivalent to finding a reporting distribution that can be induced in both states. For example, when $\zeta = \frac{1}{9}$ a symmetric distribution of reports is attained in the Positive state by the following reporting strategy:

For any of the following realized samples, $\{-2, -2, 0\}, \{-2, 0, 0\}, \{-2, -2, 2\}$, remove the highest realization with probability $\frac{2}{3}$ (and the lowest realization with probability $\frac{1}{3}$); if the realized sample is $\{-2, 0, 2\}$, remove the highest realization with probability $\frac{5}{6}$ and the lowest one with probability $\frac{1}{6}$; for any other realized sample, remove the highest realization.

Employing such mixed strategies is necessary because achieving a perfectly symmetric distribution of reports requires simultaneous and careful finetuning of different regions of the distribution. Therefore, it is typically more computationally challenging than, for example, reaching a given threshold of a uni-dimensional variable. However, the above example shows that for some values of $\zeta > 0$, crafting the reporting strategy so that the resulting distribution of reports is symmetric about zero is possible and we conjecture that there exists $\zeta^* \in (\frac{1}{9}, \zeta_0)$ such that, the agent can prevent the DM's learning for any $\zeta \leq \zeta^*$.

Fabricating Date: In this example the minimal probability of reporting g in state G is $(1-p)\phi$, whereas the maximal probability of reporting g in state B is $p + (1 - p)(1 - \phi)$. The condition for the MP-LLN holds if the former is greater than the latter, that is, if

$$\frac{1}{2} < \phi(1-p).$$

C Weak manipulability

When $|\bigcap_{\omega \in S} \mathcal{F}(\omega)| = 1$ the construction used in the proof of Proposition 5 cannot be applied. In such cases the overlap is too thin to directly support the richness of frequencies needed both to separate from states outside S and to mix between distinct frequencies within S. Nevertheless, as we illustrate below, the scope of manipulation can still be large. Achieving this, however, typically requires a more complex reporting strategy than the one used in the proof of Proposition 5 .

Consider the selective forced disclosure model, and assume that $|\mathcal{F}(G) \cup \mathcal{F}(B)| = 1$, i.e., that $\theta_G^2 = 2\theta_B - \theta_B^2$. Denote the state contingent strategy that induces the same report distribution in both states by σ_U. Moreover, denote the strategy of disclosing either signal at random by σ_I. Note that if the agent uses σ_I the disclosed signal is informative about the state. For $\gamma \in [0,1]$, denote by $\sigma_M(\gamma)$ the strategy under which the agent uses strategy σ_I with probability γ and σ_U with probability $1-\gamma$.

Fix a target long-term belief distribution with finite support that satisfies Bayes-plausibility. That is select $F^* = \{p_k, \nu_k\}_{k=1}^K$ such that $0 \leq \nu_1 < \nu_2 < \cdots < \nu_K \leq 1$, $\sum_k p_k = 1$, and $\sum_k p_k \nu_k = \mu_0$. To simplify exposition, assume that $\mu_0 \neq \nu_k$.[14] Define k_0 by $\nu_{k_0} < \mu_0 < \nu_{k_0+1}$, and let G_t denote the distribution of beliefs at the beginning of period t.

Targeting: For a belief distribution G_t and two target beliefs $x < y$ such that $supp(G) \subseteq (x,y)$, we say that the agent targets $\langle x,y \rangle$ if, regardless of the disclosure history, the agent uses $\sigma(\gamma(x,y,G_t))$, where $\gamma(x,y,G_t)$ is the maximal probability such that $supp(G_{t+1}) \subseteq [x,y]$. Such a strategy exists since for $\gamma = 0$ there is no updating, the updated beliefs are continuous in λ, and the maximal (minimal) element of G_{t+1} is strictly increasing (decreasing) in γ.

The Algorithm: Now we are ready to construct an algorithm that enables the agent to generate beliefs $\{G_t\}_{t=1}^{\infty}$ that converge (in distribution) to F^* as $t \to \infty$.

Define the initial targets as $x = \nu_{k_0}, y = \nu_{k_0+1}$, the initial capacities as $\kappa_x = p_{k_0}, \kappa_y = p_{k_0+1}$, and the let G_1 denote the degenerate belief distribution that assigns probability one to the prior. Let $t = 1$, and apply the following procedure.

Assume that from G_t the agent targets beliefs $\langle x,y \rangle$, and consider two cases: *Case a: no overfilling:* We say that targeting $\langle x,y \rangle$ from beliefs G_t does not lead to overfilling if the mass on $x(y)$ under G_{t+1} is at most $\kappa_x(\kappa_y)$. In this case, the agent uses this targeting strategy in period t. For any history that leads to one of the targets, the belief is frozen. That is, in any continuation of such a history the agent uses σ_U. Furthermore, if the mass on a $x(y)$ is exactly $\kappa_x(\kappa_y)$, then reset the target and capacity to $x = \nu_{k_0-1}, \kappa_x = p_{k_0-1}(y = \nu_{k_0+2}, \kappa_y = p_{k_0+2})$. For all non-frozen histories in G_{t+1}, recalculate that targeting strategy for $\langle x,y \rangle$, and repeat.

Case b: Overfilling We focus on the case where the mass of beliefs on y in G_{t+1}

[14] It can be shown that this is WLOG.

exceeds κ_y. The cases where there is overfilling of x or overfilling of both targets is analogous. Note that under the targeting strategy the disclosure of signal realization g leads to an increases of the updated beliefs, and that beliefs are updated to y only from the maximal non-frozen belief in G_t, denote this belief by $\overline{g}_t$.

Let q denote the probability of disclosing g under the targeting strategy $\gamma(x, t, G_t)$ from belief $\overline{g}_t$. Since σ_I is interior in both states of nature, the reporting strategy can be changed so that beliefs following the disclosure of g remain y but the probability of disclosing g is any element of $B_\epsilon(q)$, for some $\epsilon > 0$. Moreover, there exists $\delta > 0$ such that for any belief in $B_{\overline{g}_t}(\delta)$ targeting y is possible.

For beliefs in $B_{\overline{g}_t}(\delta)$, the set of probabilities with which signal g can be disclosed so that the updated belief following the disclosure of g is exactly y is continuous in the initial belief. Hence, if the agent first uses the strategy $\sigma(\gamma)$ N times, where γ is sufficiently small and N is sufficiently large, and then, from each resulting belief, either targets $\langle x, y \rangle$ or temporarily freeze the belief, he will be able induce belief y with a probability of exactly κ_y in period $t + N + 1$ round. Assume that the agent does so, and in period $t + N + 2$ freezes all histories that reach belief that reach y and unfreeze the beliefs that were temporarily frozen in these N periods. Finally, reset the target and capacity to $y = \nu_{k_0+2}, \kappa_y = p - k_0 + 1$, and for all unfrozen beliefs in G_{t+N+2} recalculate that targeting strategy, and repeat.

Since the updated beliefs converge in distribution to beliefs with support $\{0, 1\}$ if $\gamma = 1$, this process will, almost surely, fill all beliefs apart for ν_1 and ν_K in finite time. Note that by Bayes-plausibility if all but one of the targets has been achieved, the last target will have been archives as well. Hence, this algorithm generates a sequence of belief distributions $\{G_t\}_{t=1}^{\infty}$ that converge to the target F^*.