Can Society Function Without Ethical Agents? An Informational Perspective*

Bruno Strulovici
Northwestern University

April 14, 2021
Click here for the most recent version.

Abstract

Many facts must be learned through intermediaries with special expertise or access to information, such as law enforcers, scientists, journalists, and public officials. This paper considers whether society can learn about such facts when intermediaries are devoid of ethical motives and make sequential public announcements. The answer depends on the degree of *information attrition* affecting the amount of discoverable evidence about each fact. Information attrition is nonexistent in fields based on reproducible scientific evidence but can affect the evidence in criminal and corruption investigations. Applications to institution enforcement, social cohesion, scientific progress, and historical revisionism are discussed.

*This project has benefited from numerous conversations and comments, particularly from Alessandro Lizzeri, Alex Frankel, Boli Xu, Doron Ravid, Francisco Poggi, Larry Samuelson, Laura Doval, Ludvig Sinander, Meg Meyer, Piotr Dworczak, Richard Zeckhauser, Ron Siegel, Simone Galperti, and Xavier Duran. Several versions of this project were presented at Microsoft Research New England (Summer 2015), Boston College, Washington University, Universitat Pompeu Fabra, Universitat Autònoma de Barcelona, Johns Hopkins University, University of Chicago, Cambridge University INET Economic Theory Conference (2016), Penn State University, Cornell University, Arizona State University, North American Meeting of the Econometric Society (2017), Stanford University, Duke University, Northwestern University, Bocconi's workshop on Advances in Information Economics, Kyoto University, University of Tokyo (summer school), Stanford SITE 2017, the Transatlantic Theory Workshop 2017, Universidad Carlos III Madrid, Columbia's Economic Theory Conference (2017), Yale University, the Toulouse School of Economics, LACEA/LAMES 2018, the Summer school of the Econometric Society (Sapporo, 2019), the RI/PE conference at UC San Diego, the VSET seminar, Emory University, University of Sussex, the University of Cambridge, Princeton University (Political Economy), and AMETS. Early versions of this project were developed while I was visiting Microsoft Research New England (2015) and Harvard University (2016) whose hospitalities are gratefully acknowledged.

1 Introduction

1.1 Mediated learning

Many facts must be learned through agents with specific expertise or access to information. For example, the net benefits of a vaccine, the validity of a mathematical theorem, the correct resolution of a crime, the historiography of an event, and the anthropogenic nature of climate change cannot be directly verified or disproved by the average citizen, and there is no "public epiphany" at which the truth is exogenously revealed to all. In these and many others instances, citizens have no choice but to rely on intermediaries to learn anything about the fact of interest, a situation that we will call *mediated learning*.

To succeed, mediated learning relies on the investigative efforts and truthfulness of agents who can make false, misleading, or uninformed statements and are subject to biases, pressure, ambition, and other considerations that may distort their behavior. In philosophy and various social sciences, it is not uncommon to consider agents whose rule of behavior is to act truthfully.[1] By contrast, economists usually conceptualize truthful behavior as the result of properly calibrated incentives given to agents who are devoid of ethical motives.

This paper examines the feasibility of mediated learning through agents who are *non-ethical* in the sense that their preferences do not directly depend on the fact to be learned. For example, a prosecutor who wishes to obtain the conviction of a defendant regardless of the defendant's actual guilt is non-ethical.[2]

[1] One branch of epistemic philosophy concerns the vulnerability of testimony, i.e., the fact that a speaker can lie. This vulnerability is resolved through behavioral assumptions driven by norms of truthfulness and principles of ethical behavior, such as Grice's "cooperative principle" (H.P Grice (1975)). Computer scientists consider the related problem of communication through potentially adversarial intermediaries, known in the literature as the *Byzantine generals* problem (Lamport, Shostak, and Pease (1982)). This literature assumes the existence of "loyal" generals, who obediently follow the communication protocol set out by the planner, like Grice's cooperative speaker. In sociology, it is common to assume the existence of pro-social norms that coexist with and sometimes subsume individual incentives, and facilitate truthful behavior. See Granovetter (2017) for a recent comparison of the paradigms in economics and sociology.

[2] Truth-dependent preferences do not imply ethical behavior. For instance, a prosecutor who wishes to convict innocent defendants and acquit guilty ones is unethical. This observation has no bearing on the main analysis of this paper, which focuses on truth-independent preferences. Mediated learning with ethical agents is discussed in Sections 2.2, 2.4, and 4.

1.2 Information Attrition

The cornerstone of the analysis is the concept of *information attrition*, which captures the idea that information about a given fact may be in limited, fragile, or diminishing supply. Consider the problem of determining whether a given athlete used a banned substance during a given competition. Controls typically consist of two blood (or urine) samples, "A" and "B". If the first blood sample, "A", is positive (suggestive of doping), the agency storing the second sample, "B", is then asked to test the second sample. At this point, the second sample is the only direct source of evidence left about whether the athlete ingested a banned substance at the event: the amount of evidence concerning this fact has shrunk, and each blood test reduces the amount of information available in the case.

Information attrition may be caused by various factors. The first one, outlined above, is that the process of learning about a fact sometimes requires transforming the evidence in a way that prevents its use by subsequent investigators: the signals are *disposable*.[3] Second, information attrition may be due to the exogenous degradation of evidence: as time goes by, evidence may deteriorate and become uninformative, for example due to poor storage conditions. Third, information attrition may be caused purposefully by individuals who tamper with or destroy some evidence. Finally, information attrition may stem from social considerations. For example, if some researchers claim to have found that a particular drug or treatment is harmful to human subjects or animals, it may become politically infeasible to run more experiments.

1.3 Core Argument

To appreciate how information attrition affects mediated learning, let us reconsider the blood-testing example and, specifically, the agency in charge of testing the second blood sample. Since this agency possesses the only sample left, it can lie at no risk of being contradicted. If, for instance, the agency stands to gain publicity from incriminating the athlete, it can do so with impunity. If instead the agency benefits from exculpating the athlete (perhaps due to the pressure exerted by some sport governing instance), it can also

[3]Other examples of disposable signals include human subjects in experimental psychology: once a person has been exposed to a particular experiment, this person is irremediably affected and no longer a good subject for the same experiment. In quantum mechanics and other branches of physics, measurements are subject to the *observer effect*, whereby the observation of a physical system affects the state of the system.

do this without risking contradiction. And if the agency is indifferent between incriminating the athlete and absolving him, it has no incentive to incur the cost of running the test to begin with. In all these cases, the second agency's report is untethered to the truth.

Consider now the agency in charge of the first blood sample. Whatever report this agency produces, the second agency's report will not be based on the truth, as shown in the previous paragraph. The first agency's expected payoff is therefore independent of the content of its blood sample. Like the second agency, it has no incentive to report the true content of its blood sample, and incentives unravel. In this example, mediated learning is infeasible if agencies are not ethical.[4] When agencies are non-ethical, mediated learning is infeasible regardless of the agencies' material incentives.[5]

1.4 Reproducible Evidence and Incentive Design

Some investigations are immune to information attrition. Consider the question of determining whether a mathematical proof is correct. The proof remains available for anyone to read no matter how many times it has been checked in the past. The unraveling argument of the previous section no longer applies and this paper shows (Theorem 2) that mediated learning by non-ethical agents is feasible provided that agents' material incentives are designed appropriately. The incentives that deliver successful mediated learning have an intuitive structure: they reward an agent whose report is vindicated by subsequent findings and punish him otherwise, and the magnitude of the rewards and punishments of each agent depends on how surprising the agent's report is relative to earlier beliefs about the state.[6]

Information attrition does not arise, either, in scientific inquiries that rely on reproducible experiments. For instance, no matter how many times physicists measure the weight of an electron, the experiment can be replicated more times, ad infinitum. Facts based on reproducible evidence can, under the appropriate incentive structure, be learned through

[4]In practice, the same laboratory is often in charge of storing and testing *both* samples, which may lead to even more severe incentives problems, as in the case of nationally organized doping schemes.

[5]If there are three or more agencies, the results are the same when the agencies proceed sequentially, regardless of the rule (e.g., majority rule) used to convict the athlete or incentives given to the agencies. Section 4.2 discusses the case in which agencies move simultaneously as well as other learning structures.

[6]Some of these features are similar to the socially optimal policy analyzed by Smith, Sørensen, and Tian (2020) in the context of herding models. In herding models, all equilibria are at least somewhat informative: cascades may occur, but only after some information has been publicly revealed.

non-ethical agents.

As a result, the theory can explain why truthfulness oaths are used in some applications, for which information attrition is a concern, and not others. For instance, it may explain why oaths are used in courts of law but not in physical science. See Section 4.3.

1.5 Analytical Challenge: Uncertain and Endogenous Attrition

To explore the consequences of information attrition, this paper considers an analytical framework in which investigators act sequentially and communicate through public reports.[7] This framework generalizes the core argument in three ways. First, the supply of evidence available in a case may be unknown a priori, even to investigators. Second, the number of potential investigators may be large and a priori unknown, even to current investigators. Third, investigators may not know how past investigators have affected the evidence available. For example, suppose that a criminal has left several pieces of evidence on a crime scene, which may be discovered through some investigative effort. Investigators do not know a priori the exact number of pieces left behind. Moreover, if an investigator inherits the case from a previous investigator, she may not know how diligent the previous investigator has been and, hence, what fraction of the evidence has already been discovered, fabricated, or destroyed. An investigator thus faces exogenous and endogenous uncertainty about the amount of discoverable evidence.

This paper provides conditions on the probability distribution of the supply evidence, taking into account the impact of each agent on the evidence, under which mediated learning is feasible, and under which it is not (Theorems 1 and 2). When the necessary conditions are violated, mediated learning by non-ethical agents is impossible in a strong sense: even when (i) there is an unbounded sequence of potential investigators, and (ii) investigators' incentives may be arbitrarily designed and administered without any commitment or agency problem, there does not exist *any* equilibrium in which at least one investigator provides an informative report with positive probability.

It is easy to design incentives for which mediated learning always fails or, given some incentive structure, to construct an equilibrium for which mediated learning fails. The

[7] Alternative approaches to mediated learning are discussed in Section 4.2. In particular, it is argued that any investigatory activity realistically begets another one, which introduces a sequential component to any mediated learning system.

main analytical challenge is to show that, in the presence of information attrition, mediated learning fails *for all incentive structures and all equilibria*, in the strong sense that *all agents surely fail to reveal any information about the fact.*

Formally, this paper studies a sequential game of incomplete information in which the state variable is a probability distribution over the set of evidence that remains to discover. This probability distribution, defined on an infinite-dimensional space, represents agents' belief about the supply of evidence and does not have simple monotonicity properties. For instance, while the discovery of a piece of evidence may suggest that the remaining supply of evidence is now smaller by one piece, this discovery could also reveal the tip of an iceberg of evidence, pointing to many more pieces to discover. A discovery could also, if it contradicts past reports, indicate that previous investigators have lied and that whatever evidence they purported to have found (and thus "removed" from the supply of evidence) was in fact fabricated, resulting in a more optimistic belief about the amount of evidence that actually remains to discover.

An agent's belief about the evidence supply affects his incentives directly, through the probability that he discovers new evidence, as well as indirectly, through the probability that subsequent agents also look for evidence. An agent cares about subsequent agents' beliefs about the supply of evidence, their beliefs about subsequent agents' beliefs, and so on. Moreover, an agent can manipulate other agents' beliefs by lying in his report.

Information attrition does not preclude that (a) a long sequence of agents work, or that (b) many pieces evidence remain to be discovered, but it imposes a negative correlation between these two events. To rule out the existence of an informative equilibrium, one has to control the sequence of agents' beliefs. This is achieved by showing that the probability that an agent discovers evidence is probabilistically linked to the impact that this agent's report has on subsequent agents' beliefs. This key link is established by Proposition 2.

2 Implications of the Theory

2.1 Political Consequences of Information Attrition

Consider the perspective of a citizen who is *cynical* in the sense that he does not trust information intermediaries to behave ethically. In this citizen's mind, all agents involved in

the learning process have non-ethical motives and are collectively aware of this.

To see how information attrition affects the views of cynical citizens, let us first revisit the blood-testing example. Given that the blood samples are subject to information attrition, it is reasonable—indeed, rational—for a cynical individual to treat as uninformative any finding reported by the blood-testing agencies, for the reasons explained above.[8] As a result, two citizens with cynical beliefs about agencies' behavior and otherwise different views of the world may rationally entertain completely different beliefs about whether a particular athlete used banned substance, even after agencies have made their reports public: mediated learning fails to convince citizens and bring their views closer.

Mediated-learning failures can have severe consequences for social cohesion and political stability. When a politician is accused of corruption, for instance, the average citizen must rely on declarations made by intermediaries, such as officials and journalists, to learn anything about whether the corruption charge is true or, instead, an attempt to smear and neutralize some politician. There is no *deus ex machina* available to lift all confusion and finally reveal the truth to the public. Moreover, information attrition may be a concern in these environments. In a corruption case, for instance, incriminating documents can be destroyed and witnesses intimidated or eliminated.[9]

Now let us consider citizens holding very different priors, perhaps based on their party affiliation or ideology, about whether some politician was guilty of corruption or some governmental agency abused its power. If these citizens are cynical in the above sense, they may dismiss official statements and journalistic reports about the case and maintain strong disagreements.

[8]Blood tests could in principle be certified by a third party, which would check wether the agency did its job properly and thus increase the trust in its report. One would then have to understand the third party's incentives. This "monitoring the monitor" structure is discussed in Section 4.2.

[9]The same observations apply to government agencies suspected of abusing their power or violating some rules. Examples of evidence destruction by governmental agencies abound, even in the world's most prominent democracies. In the Unites States, for instance, CIA director Richard Helms ordered in 1973 that all documents pertaining to the CIA's infamous MK Ultra program on mind-control experiments be destroyed ("An interview with Richard Helms", `https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol44no4/html/v44i4a07p_0021`
In 2005, the CIA's Director of Operations at the time ordered the destruction of all interrogation tapes of Abu Zudaydah and Abd al-Rahim al-Nashiri that featured "enhanced" interrogations ("Tapes by C.I.A. Lived and Died to Save Image," `https://www.nytimes.com/2007/12/30/washington/30intel.html`.)

The theory suggests that reasonable citizens may remain divided on questions that are subject to information attrition. For such questions, the lack of trust in investigative institutions can rationally perpetuate polarization. The theory thus also provides a specific mechanism to explain why *eroding citizens' trust in institutions* harms social cohesion.

2.2 Ethical Necessity

The arguments presented so far bring us to one of the paper's key motivating questions: is ethical behavior *necessary* for society to function? Ethical necessity is not salient in economic analysis, which typically focuses on "selfish" agents evolving within the boundary of well-defined institutions, such as markets or democratic institutions, whose enforcement is taken for granted.[10] In reality, institutions cannot be taken for granted: if agents are unethical, they may attempt to violate these institutions in various ways, and it is unclear why agents should be presumed to act selfishly within the behavioral boundaries imposed by these institutions but ethically with respect to these boundaries. This artificial separation amounts to a "heroic dichotomy" that, at the very least, deserves closer inspection.

By focusing on mediated learning, this paper provides a tractable framework to study ethical necessity.[11] Since mediated learning is required to investigate criminal cases, political corruption, and other possible violations of institutions, successful mediated learning is a necessary condition for society to function: whenever mediated learning requires ethical behavior, so does society.

While ethical necessity has not been a salient issue in economics, it did receive some attention from economists.[12] Hurwicz (2007) explicitly discusses the existence of "intervenors,"

[10]Economists have studied various forms of non-selfish behaviors, such as pro-social behavior arising in dictator and ultimatum games and various forms of altruism. Unlike earlier works, this paper does not study ethical behavior per se, but whether ethical behavior is necessary for society to function. Self-interest remains the default assumption in economic models and is deeply rooted in the discipline. For example, Edgeworth (1881) observed that "self-interest is the first principle of pure economics."

[11]The present framework relies on a sequential learning structure. One could consider alternative structures. For example, a central agency may ask several intermediaries to seek and report the truth simultaneously and independently of one another. Under this parallel structure, the incentives of such a centralized agency must however also be scrutinized, which reintroduces a sequential element. This and other designs are discussed in Section 4.

[12]For instance, Myerson (2006) shows that in a federalist democracy, the existence of some virtuous politicians can guarantee that democracy succeeds either at a national or a provincial level, whereas democracy

which he defines as ethical monitors in a monitoring-the-monitor problem, and expresses his personal belief in the existence of intervenors. Unlike the present paper, however, Hurwicz argues that intervenors are not needed for successful monitoring hierarchies.[13] Hurwicz describes an environment with three agents A, B, C, in which B monitors A's actions, C monitors B's monitoring of A, A monitors C's monitoring of B's monitoring A, and so on. This "Hurwicz triangle," in which the monitoring hierarchy is folded in a loop going through the three agents, omits the possibility of corruption across monitors, studied in Strulovici (2020).[14]

Holmström (1982) shares a similar concern, noting in his conclusion that "another important issue relates to monitoring hierarchies. (...) The question is what determines the choice of monitors; and how should output be shared so as to provide all members of the organization (including monitors) with the best incentives to perform?"

The existence of a reliable monitor is implicitly assumed in Becker and Stigler's (1974) study of wrongdoing and malfeasance by enforcement officers. The authors assume that any wrongdoing by enforcement officers is exogenously detected with some probability. When ethical monitors are unavailable, the question of how this detection is generated becomes important.[15]

---

can fail at both levels when such politicians are absent. By contrast, Glazer and Rubinstein (1998) describe an information aggregation environment in which if all agents are purely concerned with achieving the social optimum, there always exist equilibria in which the optimum is not achieved, whereas if all agents are *also* (but not only) concerned with their individual recommendation being followed, the social optimum is uniquely selected. Matsushima (2008) studies the possibility of full implementation when agents prefer truthtelling whenever their material payoffs are unaffected by their message.

[13]Rahman (2012) proposes an approach that is well suited for repeated monitoring tasks, such as controls used for airport security or in sting operations: a principal can ask agents to violate the rules on purpose to check whether these violations are caught by the monitor. Such violations are detected "for free" by the principal since he instigates them. The approach is well suited when violations can be faked at little social cost, the principal has commitment power, and collusion between the principal and agents is impossible.

[14]Levine and Modica (2016) consider a similar structure, in which agents in a group take some initial action, then verify the action taken by their neighbor, then verify the earlier verification task of their neighbors, and so on.

[15]Milgrom, North, and Weingast (1990), who study the enforcement of trade institutions by law merchants in medieval Europe, consider some of the law merchants' incentives to lie and take bribes.

2.3 Historical Revisionism

Historical revisionism is another instance of mediated learning, in which information attrition plays an important role. Understanding historical events is of obvious importance, not only to learn lessons from the past but also to assess claims based on such events, such as territoriality or reparation claims. Mediated learning is necessary because citizens cannot directly verify historical events.[16] They must rely on experts and officials to access and correctly interpret archives, artifacts, and other sources of information. Information attrition is both exogenous (e.g., witnesses die) and endogenous (e.g., documents may be destroyed on purpose).

To give a concrete example,[17] consider the fire of the German parliamentary building (Reichstag) on February 27, 1933. The importance of this event can hardly be overstated. The Nazis, who had lost seats in the previous parliamentary election, claimed that the fire had been caused by communists and used this event to pressure president Hindenburg into imposing martial law on Germany (the "Reichstag fire decree") and arrest and weaken communists. This allowed the Nazis to form a majority coalition after the March 5, 1933 parliamentary elections, consolidated Hitler's power, and led to the Enabling Act.

While communist involvement in the fire has long been rejected, there was until recently a consensus among mainstream historians that the Reichstag fire had *not* been caused by the Nazis. Fritz Tobias, one of the most respected historians on this subject in the postwar period, published a series of articles purporting to show that van der Lubbe, the person who was convicted for the arson, had acted alone.[18] Historians accepted Tobias's version of the event until 2001, when two historians studying Gestapo archives raised the possibility that it was a group of SA officers who had set the fire (Bahar and Kugel (2001)). Hett (2014) used recent scientific advances to convincingly argue that it would have been impossible for a single individual to set the fire. In 2019, an affidavit written in 1955 by former SA Hans-Martin

[16]This point is obvious with regard to events that took place before citizens' lifetime. Even with regard to contemporaneous events, aggregating information and forming a global picture of an event is a highly complex task that requires expertise, time, and a special access to information. The pitfalls of such information aggregation have been famously illustrated by Stendhal's *La Chartreuse de Parme* whose protagonist, Fabrice del Dongo, takes part to the Battle of Waterloo with the Napoleonic army and construes a completely erroneous version of the battle.

[17]Huq (2018) discusses similar, very recent examples, in which the possibly false threats of terrorism or coups were used to weaken democratic institutions and shift power to more authoritarian regimes.

[18]The articles were published in *Der Spiegel* under the title „Stehen Sie auf, van der Lubbe!," in 1959 and 1960, in issues 43 to 52.

Lennings was discovered in Tobias's personal files and published by RedaktionsNetzwerk Deutschland, which stated that Lennings and other SAs had driven van der Lubbe from an infirmary to the Reichstag when the fire had already started, effectively setting up van der Lubbe.

Information attrition took several forms: all but one of the SA officers who were allegedly involved in the Reichstag fire were killed (and, hence, silenced) during the Night of the Long Knives; van der Lubbe was beheaded in 1934 for his alleged role in the arson; and any forensic evidence about the Reichstag fire has been long gone. Furthermore, Fritz Tobias hid Lennings' affidavit, which contradicted Tobias' single perpetrator theory, until his death.[19] To this day, the strongest case for Nazi involvement thus seems to come from Lennings' affidavit, and hence boils down to one man's statement. It is unclear what Lennings' motivation for incriminating the SA may have been, except perhaps for setting the record straight. For observers who doubt Lennings' statement, the question of who set the Reichstag on fire may reasonably represent a failure of mediated learning. Lutjens (2016) believes that "the continuous reshaping of the Reichstag fire by those with a stake in the matter has fragmented the truth beyond recovery."

2.4 Existence and Observability of Ethical Agents

The arguments developed so far assume that agents are known to be non-ethical. Without this knowledge, mediated learning can be achieved if agents put sufficiently high weight on the probability that other agents are ethical. To see this, let us consider once more the blood-testing example. If the first agency and citizens all believe that the second agency behaves ethically, the first agency can be incentivized to test and truthfully report the content of the first blood sample because its finding can be compared to the second agency's finding, which all believe to be truthful, and the first agency can be held accountable for any discrepancy.

More generally, any agent can be incentivized to behave truthfully as long as the agent believes that subsequent agents are likely to provide informative reports.

Conversely, if some agent is ethical but other agents believe that he is not, mediated learning can fail just as when all agents are non-ethical, for two reasons: First, the findings of the ethical agent are (wrongly) believed to be uninformative. Second, precisely because this

[19] Tobias is now suspected of protecting former Nazi officers after the war and having a private interest in dissimulating the Nazis' role in the Reichstag fire.

agent's findings are believed to be uninformative, the findings cannot be used to incentivize other agents to behave truthfully. Therefore, mediated learning requires that agents believe that other agents behave ethically with high enough probability.[20]

The theory thus provides a specific mechanism for why eroding trust in institutions is damaging: society may need everyone to believe in the existence of ethical agents in order to sustain ethical behavior. Events that erode the strength of this belief can have severe consequences for the feasibility of mediated learning and the functioning of society.[21]

Furthermore, it may be empirically difficult to distinguish between agents who have ethical preferences and agents who merely behave ethically because they believe in the existence of ethical agents. Put differently, the belief in ethical behavior can be self-fulfilling.[22]

2.5 Organization of Remaining Sections

Section 3 describes the formal model and the main results of the paper. In the baseline model, all agents must incur a cost (which may be arbitrarily small, but strictly positive) in order to acquire information. The model is then extended to allow for the existence of witnesses, who receive information for free and are subject to idiosyncratic, private shocks that affect their reporting preferences. Section 4.1 describes the relation between the concept of hard evidence, information attrition, and intermediation. Section 4.2 discusses alternative investigation designs and settings. Section 4.3 explores how to foster truth-dependent preferences among investigators. All results are proved in the Appendix.

[20]It would be interesting to study, beyond the analysis of the present paper, just how high and concentrated the probability that agents are ethical needs to be in order to sustain mediated learning.

[21]Belief in ethical behavior gets rid of virtual attrition, but not of real attrition. For example, if the second blood-testing agency is, in fact, non-ethical, it cannot be incentivized to tell the truth: real attrition interrupts mediated learning. But if everyone erroneously believes that the second agency is ethical, then the first agency may be incentivized to behave truthfully.

[22]Bénabou, Falk, and Tirole (2020) study theoretically and empirically the extent to which moral preferences can be elicited.

3 Formal Analysis

3.1 Baseline Model

The fact of interest, $\omega \in \Omega$, must be inferred from a sequence $S = (s^1, \ldots, s^{\tilde{K}})$ of signals, each of which takes values in some finite signal space Σ. The sequence S and its length $\tilde{K} \leq \infty$ are stochastic. The joint distribution of (ω, S) is arbitrary.[23]

In each round $i \geq 1$, a new agent arrives and makes two decisions: First, the agent privately chooses between seeking a signal ("working") at cost $c > 0$ and doing nothing ("shirking"). Second, the agent publicly sends a message m_i from some finite message space M. The agent can randomize his decisions.

Let S_i denote the sequence of signals that remain to discover at the beginning of round i (in particular, $S_1 = S$). If the agent in round i (hereafter, "agent i" or simply "i") works and $S_i \neq \emptyset$, then i discovers some element of S_i with probability $\lambda \in (0, 1]$. The discovered signal is denoted s_i.[24,25] With probability $1 - \lambda$, i discovers no signal. If S_i is empty, i surely discovers no signal.

For simplicity, we will not model the possibility that agents destroy signals without discovering them, or that signals disintegrate exogenously. These additional forms of information attrition would only strengthen the paper's impossibility results (Theorem 1 and Theorem 3).[26] The extension is more complex when some agents, such as witnesses, can discover signals for free. This extension is analyzed explicitly in Section 3.4.

After the information-seeking stage, i sends a report m_i whose distribution in $\Delta(M)$ can arbitrarily depend on what (if anything) i has observed in the information-seeking stage and on the reports $m_1^{i-1} = (m_1, \ldots, m_{i-1})$ made by previous agents.

[23] One could impose more structure guaranteeing that S reveals ω perfectly or with a certain level of precision. Such additional structure is orthogonal to the analysis.

[24] To index signals, note that s_i is the signal discovered by agent i (when applicable) and s^j is the j^{th} signal in the sequence S.

[25] No constraint is imposed on how likely each element of S_i is of being discovered by agent i. This likelihood could depend arbitrarily on i's identity and on S_i, and the discovery of some specific signal may contain information about other signals in S_i that depends on i's identity. Moreover, Theorem 1 can be generalized to the case in which the probability λ of discovering a signal is nondecreasing in the length of S_i.

[26] The possibility result (Theorem 2), which allows a geometrically distributed number of signals, may be interpreted as allowing an exogenous decaying rate of evidence.

Entering round $i+1$, we have $S_{i+1} = S_i$ if i did not discover any signal. If i discovered a signal, then S_{i+1} is a subsequence of S_i with length $|S_{i+1}| = |S_i| - 1$.

Let $m = (m_1, m_2, \ldots)$ denote the sequence of reports made by all agents. The realized utility of agent i is given by

$$U_i = V_i(m, \omega) - c\mathbb{1}_{i \text{ works}} \tag{1}$$

where V_i takes values in some compact interval $[-P, R]$.

We will say that agent i is *non-ethical* if the function V_i is independent of ω, which captures the idea that i does not care directly about the truth.[27] In this case, V_i may be defined on the restricted domain $M^{\mathbb{N}}$. V_i can depend arbitrarily on the entire sequence $m = (m_1, m_2, \ldots)$. In particular, the impossibility results presented in this paper hold regardless of whether V_i is an exogenously given utility function or one that is specifically designed (or, at least, influenced) by a regulator or social planner.

To illustrate the various forms that V_i may take, i could be punished if his report is contradicted by subsequent investigators or rewarded if his report differs from past investigators' (as in a journalistic scoop). V_i may aggregate a discounted stream of rewards and punishments.[28] The formulation captures situations in which i's utility is affected by reports indirectly through the actions that these reports trigger. For example, i could be a prosecutor at a trial, whose outcome $a(m) \in \{$'guilty', 'not guilty'$\}$ depends on the statements of all agents involved in the case. If i is non-ethical, his utility may be modeled by $V_i(m) = R \times \mathbb{1}\{a(m) =$ 'guilty'$\}$, as in Landes (1971). Agent i's realized utility could depend stochastically on other agents' reports. For example, suppose that the number of investigators is stochastic. We can model this by interrupting mediated learning at some stopping time τ, in which case i's utility depends only on $(m_1, \ldots, m_\tau)$. Similar variations, such as randomizing the order of investigators, how investigator j's report affects i's utility, or including a private type that affects i's utility, are also easily encompassed by the model.

Agents have a common prior about the distribution of S. The equilibrium concept is

[27] This concept is somewhat similar to the notion of being purely "extrinsically motivated" in Bénabou and Tirole (2003), and one could say that i is intrinsically motivated *by the truth* if V_i depends on ω.

[28] For example, if i receives utility $v_{i,j}(m_1, \ldots, m_j)$ in round $j \geq i$ and discounts future utility with some factor $\delta < 1$, then

$$V_i(m) = \sum_{j \geq i} \delta^{j-i} v_{i,j}(m_1, \ldots m_j).$$

(weak) Perfect Bayesian Equilibrium.[29]

3.2 Main Results

For each $k \geq 1$, let $F^k = \Pr(|S| \geq k)$ denote the prior probability that there are at least k signals to discover at the beginning of the investigation process.

DEFINITION 1 *An equilibrium is* **informative** *if at least one agent works with positive probability, and* **uninformative** *otherwise.*

Uncovering any modicum of information about S (and, hence, ω) with positive probability, no matter how small, suffices to qualify an equilibrium as "informative." The following theorem shows, however, that informative equilibria fail to exist when information is subject to attrition in a specific sense.

THEOREM 1 *For any parameters (R, P, c, λ), there exist strictly positive thresholds $\{\underline{F}^k\}_{k\geq 1}$ with the following property:*

$$\textit{All equilibria are uninformative unless } F^k \geq \underline{F}^k \textit{ for all } k \geq 1.$$

For an informative equilibrium to exist, the support of $|S|$ must therefore be unbounded: if not, $F^K = 0$ for some K, which violates the threshold condition regardless of the parameters (R, P, c, λ) and utility functions $\{V_i\}_{i\geq 1}$. Even if $|S|$ has unbounded support, Theorem 1 implies that mediated learning is feasible only if the survival function $F^k = \Pr(|S| \geq k)$ does not decrease too fast in k.[30]

Intuition: To understand mediated learning failures, suppose for simplicity that there is a fixed number of discoverable signals. Successful learning requires that at least one agent discovers one of the signals, which can be incentivized only if a subsequent agent discovers (with positive probability) a second signal, which can be incentivized only if a third agent discovers a third signal with positive probability, and so on. Therefore, agents

[29]The impossibility results in this paper hold for stronger concepts of equilibrium, such as sequential equilibrium, and would hold even for Bayes-Nash equilibria since off-path beliefs play no role in the analysis. Reciprocally, the equilibrium constructed to prove the positive result, Theorem 2, is a sequential equilibrium.

[30]It would be interesting to characterize the decrease rate of the cutoffs. This question does not seem to have a simple answer. For example while Theorem 2 shows that there are settings in which geometric decay is compatible for mediated learning, there is no indication that this condition is necessary.

must reach with positive probability a round in which (i) the probability that a signal remains is arbitrarily low and (ii) the probability that some agent makes an informative report is positive. These conditions are incompatible: no agent wants to work when the expected benefit is arbitrarily close to zero. This shows that mediated learning must fail down this path, which causes incentives to unravel all the way back to the first agent and leads to a complete and global failure of mediated learning.[31] This intuition ignores important challenges, which are presented in Appendix A together with the rigorous proof.

When the survival function F^k decreases at most at a geometric rate, there are instances of the model and utility functions $\{V_i\}_{i\in\mathbb{N}}$ for which mediated learning is feasible, as indicated the next theorem.

THEOREM 2 *For any $\rho \in (0,1]$ and $\lambda \in (0,1]$, there exists an instance of the model and utility functions $\{V_i\}_{i\geq 1}$ for which the distribution of S satisfies $F^k = \rho^{k-1}F^1$ for all $k \geq 1$ and an informative equilibrium exists.*

To illustrates this positive result, the next section focuses for expositional simplicity on the case of reproducible evidence, which corresponds to $\rho = 1$ (the supply of signals is unlimited), and shows that mediated learning can be made arbitrarily precise as long as the rewards and punishments are high enough.

3.3 Reproducible Evidence

Suppose that S consists of infinitely many signals taking binary value, "H" or "L".[32] The signals are conditionally i.i.d.: there is an unknown parameter $\omega \in \{H, L\}$—the underlying fact of interest—such that each signal $\tilde{s}$ satisfies $\Pr(\tilde{s} = \text{“}H\text{”}|\omega = H) = \Pr(\tilde{s} = \text{“}L\text{”}|\omega = L) = \pi \in (1/2, 1)$, and the signals are independently distributed conditional on ω. Agents' message space is chosen to be binary: $m_i \in \{\text{“}H\text{”}, \text{“}L\text{”}\}$ for all i.

[31] This intuition differs from herding models (Bikhchandani, Hirshleifer, and Welch (1992) and Banerjee (1992)), in which learning failures ("cascades") occur only after so much public information has been revealed that agents prefer to forgo their own signals. In the present model, the learning failure occurs from the first very agent. One way to visualize this feature is that cascades occurring in the future "ripple back" to early rounds because agents' payoffs in early rounds depend on the messages produced during the cascades.

[32] This amounts to assuming that $\rho = 1$ in Theorem 2. The argument is almost identical if $\rho < 1$ as explained in the proof.

For expositional simplicity, we assume in this section that $\lambda = 1$ which means that every agent who works surely discovers a signal.

In the equilibrium that we consider, it is always in an agent's interest to follow his signal if he acquired one. The relevant pure strategies are: (i) work at cost $c > 0$ and report one's signal ($m_i = s_i$), and (ii) shirk and send a message $m_i \in \{\text{“}H\text{”}, \text{“}L\text{”}\}$ at no cost.

Let $p_1 = P(\omega = H)$ denote the prior about ω before the investigation process. For any given equilibrium, let γ_i denote the probability that i works and p_i denote the probability that $\omega = H$, both conditional on the past reports m_1^{i-1}.

PROPOSITION 1 *For any thresholds p_- and p_+ such that $0 < p_- < p_1 < p_+ < 1$, there exist $P, R > 0$, utility functions $\{V_i\}_{i\geq 1}$ taking values in $[-P, R]$, and thresholds $\underline{p}, \bar{p}$ such that $0 < \underline{p} < p_-$ and $1 > \bar{p} > p_+$ for which the following strategy profile constitutes an equilibrium: $\gamma_i = 1$ if $p_i \in (\underline{p}, \bar{p})$ and $\gamma_i = 0$ otherwise.*

The state ω can thus be learned with arbitrary precision if the rewards and punishments used to incentivize agents are high enough. In equilibrium, agents work with probability 1 until the posterior belief becomes extreme enough, at which point learning stops.

Since p_i is a martingale and each signal has the same level of informativeness, p_i must exit $[\underline{p}, \bar{p}]$ with probability 1 in the candidate equilibrium. Incentives are provided as follows: if i reported “H”, he gets a reward if $\bar{p}$ is reached and a punishment if $\underline{p}$ is reached, and vice versa. These rewards and punishment depends on the belief p_i before i's report. If p_i was very close to one of the boundaries and i's report takes the posterior away from this boundary, i gets a high reward if the belief process ends up exiting through the other boundary (a low probability event) and a very mild punishment if the belief process ends up crossing the nearby boundary.

3.4 Witnesses

We now introduce witnesses, who differ from the previous agents along two dimensions:

- Witnesses discover a signal for free.
- They are subject to (possibly, arbitrarily small) preference shocks that affect which report they prefer to send.

In each round $i \geq 1$, agent i can be an investigator, identical to the agents of the baseline model, or a witness. Whether i is an investigator or a witness is public information.

If i is an investigator, the structure of i's round, information, and utility is as in the baseline model. If i is a witness, he receives at no cost a signal $s_i \in S_i$ at the beginning of the round, where S_i is the set of signals remaining at the end of round $i-1$. In particular, i can be a witness only if S_i is nonempty. If i is a witness, the sequence S_{i+1} of available signals at the end of round i satisfies $|S_{i+1}| = |S_i| - 1$.

At the beginning of round i, the probability φ_i that i is a witness is zero if $S_i = \emptyset$ and can take any value in $[0,1]$ otherwise, and depend on past reports m_1^{i-1}.[33]

In principle, a witness' signal could be informative about the number of signals that remain to discover and, in particular, about whether information attrition is an issue for subsequent agents.[34] This would significantly weaken the relevance of the initial belief about the supply of information, on which previous theorems are based.

We rule out this possibility and focus on the case in which a witness' signal never increases expectations about the total number of signals. This is achieved as follows: the sequence S of signals is obtained by, first, generating an infinite sequence S^∞ of signals, which may exhibit any arbitrary correlation between one another and, second, by truncating this sequence at some integer-valued random variable $\tilde{K}$ that is independently distributed from S^∞. Agents observe signals in the order of the sequence. Thus, writing $S = (s^1, \ldots, s^{\tilde{K}})$ (using superscripts to avoid confusion with the signals discovered in round i, which are denoted with subscripts), suppose that q_i signals have been uncovered by the beginning of round i, so that $S_i = (s^{q_i+1}, s^{q_i+2}, \ldots, s^{\tilde{K}})$. If i discovers a signal s_i, then necessarily $s_i = s^{q_i+1}$ and $S_{i+1} = (s^{q_i+2}, \ldots s^{\tilde{K}})$.

Moreover, we assume that $\tilde{K}$ has an increasing hazard rate, i.e., $\Pr(\tilde{K} = k)/\Pr(\tilde{K} \geq k)$ is increasing in k.

Assumption 1 *The total number of signals $\tilde{K}$ has an increasing hazard rate.*

[33]For example, there could be a fixed subset $\mathcal{N} \subset \mathbb{N}$ of rounds such that $\varphi_i = \mathbb{1}_{i \in N}$ (which puts a lower bound on the support of $|S|$). Alternatively, there could be a fixed time T beyond which all witnesses have appeared, so that $\varphi_i \in (0,1)$ for $i \leq T$ and $\varphi_i = 0$ for $i > T$. In general, the number of witnesses and the timing of their appearance may be stochastic and depend on past reports. This captured by allowing φ_i to depend arbitrarily on past reports and on calendar time.

[34]For example, $|S|$ could be a finite or infinite with equal probability, and a witness' signal could reveal whether the latter is true, in which case mediated learning may be feasible, as described by Theorem 2.

Intuitively, this assumption guarantees that the more signals have been discovered, the more likely it is that there are no signals left to discover.

After observing his signal s_i, witness i sends a report $m_i \in M$. His realized utility has two parts:

$$U_i(m) = V_i(m) + \epsilon_i(m_i)$$

where $V_i(m)$ plays the same role as investigators' utility function, and $\epsilon_i(m_i)$ is a shock affecting i's preferences.

ASSUMPTION 2 *The random variables $\{\epsilon_i(m_i)\}_{m_i \in M_i}$ are privately observed by i. Conditional on m_1^{i-1}, they are independently distributed from one another and from all other variables in the model. Their density functions $\{f_{i,m_i}\}_{m_i \in M_i}$ are uniformly bounded above by some arbitrary constant $\bar{f}$.*

The bound $\bar{f}$ plays for witnesses the same role as the cost inverse $1/c$ does for investigators. Intuitively, i's disutility of sending a less preferred message m_i instead of a more preferred message m_i' is of order $1/\bar{f}$ in expectation, as explained in Lemma 8.

If i is a witness, we will say that i's message is uninformative it is statistically independent of i's signal conditional on m_1^{i-1}. Otherwise, i's message is informative. An informative equilibrium is defined as before: there is at least one agent (investigator or witness) who produces an informative message with positive probability. Continuation informative equilibria are defined analogously. The following result is proved in Appendix D.

THEOREM 3 *There exist strictly positive thresholds $\{\underline{F}^k\}_{k \geq 1}$ such that if $F^k < \underline{F}^k$ for some $k \geq 1$, there does not exist any informative equilibrium.*

4 Discussion

Mediated learning fails if the following conditions hold jointly: (i) intermediaries do not care about the truth, (ii) information is subject to attrition, (iii) there is no exogenous, public revelation of the truth at any future time, and (iv) intermediaries proceed sequentially.

This result holds for a general class of utility functions, for all equilibria, in the strong sense that nothing at all is learned about the state of the world, and despite the fact that agent incentives can be administered without any further agency problem: whatever rewards and

punishments are promised to the agents as function of reporting histories can be perfectly enforced.

This impossibility result may be viewed as a reference point: to succeed, mediated learning must break at least one of the four conditions above. In particular, mediated learning can succeed if some intermediaries are motivated by the desire to seek the truth, or by a belief that other intermediaries have such a motivation. Several ways out are discussed below.

4.1 Escaping Attrition with Hard evidence

In some cases, signals are hard to fabricate and not disposable. Consider the video footage of a crime, in which the criminal is clearly identifiable. Such footage can be viewed numerous times with little degradation and is conceptually similar to reproducible evidence.

Even this kind of evidence is need not be perfectly reliable. For example, video footage can fabricated, as exemplified by the emergence of baffling deepfakes.[35]

DNA testing also illustrates this issue. The amount of usable DNA samples on a crime scene is finite. The procedure of DNA testing involves a replication phase, such as PCR amplification or DNA cloning, which creates more "evidence" that can be stored and verified by subsequent investigators. Crucially, however, these replications are only as reliable as the original DNA sample. They do not constitute new, independent evidence.

This is all the more important as DNA samples can be synthesized, i.e., literally "fabricated," to match any desired DNA profile.[36] Moreover, DNA samples can be erroneously or malevolently taken outside of the crime scene and presented as coming from the scene.[37] Ultimately, agents joining an investigation can either test the DNA material that previous laboratories have left them, which may be the result of manipulation or fabrication, or

[35]Deepfakes can be detected by computer scientists who specialize in the field. However, this kind of detection *increases* the reliance on mediated learning. With deepfakes, video evidence no longer constitutes public, hard evidence.

[36]Frumkin, Wasserstrom, Davidson, and Grafit (2010) show the possibility of creating saliva or blood samples with the desired DNA. The authors, as well as subsequent work by other researchers, show that identifying methylation patterns in DNA samples can help distinguish synthetic and natural DNA, although such identification is challenging.

[37]A famous example is the "Phantom of Heilbronn," a presumed serial killer whose DNA was found on 40 crime scenes over a fifteen-year span in Germany, Austria, and France. The DNA turned out to belong to a woman working in the factory that made the cotton swabs used to collect DNA samples.

look for genuinely new DNA samples, which brings us back to the problem of information attrition.

The role of technology on mediated learning is complex and deserves a separate exploration. For example, DNA testing, video footage, and other technological advances have increased the set of reliable evidence. However, technology can also be used to manipulate evidence and do so more anonymously than before, increasing the reliance on experts and, hence, on mediated learning.

4.2 Alternative Designs

Several remedies may be considered to address the unraveling results of Theorems 1 and 3. First, agents could be asked to investigate and report their findings simultaneously, a structure that we will call *parallel monitoring*. Second, agents could investigate past investigators. Third, in some cases, such as criminal cases, agents can in principle be incentivized by the perspective that their findings have an influence of subsequent crimes or events. Finally, the same agents could be called to make statements repeatedly over time. These possibilities are examined in turn.

1. Parallel Monitoring with a Centralized Authority

In applications such as historical revisionism, it is realistic to assume that agents make their statements sequentially. It is nonetheless natural to consider, from a mechanism-design perspective, the case in which several agents simultaneously and independently investigate the fact of interest and report their findings to some central authority. A conceptually similar design is to hide past reports from investigators until they have made their own report. Gershkov and Szentes (2009) show that such a design is optimal in a voting model with costly information acquisition.[38]

For example, blood-testing laboratories could be asked to test their respective samples independently from each other and report their findings simultaneously to some overseeing agency. The laboratories would be rewarded if their findings match and punished otherwise.[39]

[38]In their model, voters' preferences depend on the state of the world and informative equilibria exist even when all reports are public. However, optimal learning requires that past reports be hidden.

[39]The academic refereeing process has reporting features that resemble the parallel-monitoring design, although the incentives for referees are different and arguably more complex than a mere coordination motive.

An immediate concern with this solution stems from the incentives of the central authority. If the authority has a material interest in a specific outcome (which cannot be ruled out, especially in politically charged investigations), it can secretly help agents coordinate on some report or influence the agents' reports in various ways. In order to avoid this, the central agency must be itself monitored, which brings us back to a sequential monitoring problem. In the blood-testing example, if laboratories must report their findings simultaneously, there is no recourse for an athlete accused of doping if laboratories conspire to accuse the athlete.

In some cases, parallel monitoring may be challenging to implement. It is difficult, for instance, to send multiple investigators on a crime scene to independently interrogate witness and collect evidence, without the investigators being able to communicate, either directly or through witnesses and possibly coordinate. Moreover, when the evidence is in limited supply (such as the weapon of a crime), such a limitation creates negative correlation in investigators' reports, since at most one of them can discover the evidence (weapon).[40]

Finally, successful parallel monitoring can coexist only with other equilibria, many of which are uninformative. Indeed, there always exist equilibria in which monitors coordinate on a predetermined sequence of reports. Among these equilibria, successful parallel monitoring, when it is feasible, may be non-robust. Even small amounts of strategic uncertainty about other monitors can suffice to destroy the informative equilibrium.[41]

2. Monitoring the Monitor

Instead of all agents investigating some initial fact of interest, agents investigate one another. For example, agent 1 investigates the initial question of interest, agent 2 investigates agent 1's investigation of the initial question, agent 3 investigates agent 2's investigation of 1, and so on. Such a sequence is called a "monitoring hierarchy." Hurwicz (2007) considers such a hierarchy, except that the number of agents is finite and the monitoring chain cycles repeatedly across the a fixed number of agents. A first conceptual difficulty with this approach concerns the simultaneity and complexity of these monitoring tasks: agents are supposed to conduct an infinite amount of monitoring tasks and are indirectly the subject of the tasks that they are investigating.

Even if we consider an unbounded sequence of distinct agents, each of which is tasked with

[40]The deleterious impact of negative correlation on the informativeness of agents with a coordination motive is a central finding in Pei and Strulovici's (2020) analysis of strategic crime.

[41]This point is explored in ongoing work with Harry Pei.

investigating the previous agent in the sequence, another issue emerges: what if an agent who discovers incriminating evidence about the agent he was monitoring can hide or destroy the evidence, in exchange for a payment from the guilty agent? Such a transfer amounts to a local form of corruption among nearby agents. In a separate paper, I show that this even this very local form of corruption may suffice to destroy the possibility of mediated learning (Strulovici (2020)).

3. Repeated Setting

When mediated learning concerns the identification of a criminal and opportunities to commit crime are repeated over time, it is a priori possible that investigators care about the truth indirectly, through the impact that their findings have on citizens' future behavior.

Suppose that a citizen's decision to commit crime depends on whether his past actions were accurately called by past investigators: for example, a citizen who was wrongfully accused in the past or mistakenly acquitted of crimes that he did commit may be more likely to commit crime in the future. In this setting, investigators could in principle have an endogenous incentive to report accurate findings. Provided that players are sufficiently patient, this kind of strategy profile could a priori be used to incentivize accurate reporting.

However, for this argument to work, a citizen's strategy must depend on his private history, where the public history consists of official findings about citizens' past actions and a citizen's private history records his actual past actions. In order for a citizen's private history to affect his decision of whether to commit crime, the citizen must indifferent between committing crime and abstaining from it, a knife-edge condition that is violated if, for instance, citizens are subject to small private shocks affecting their benefits from committing crime.[42]

4. Alternating Statements

Finally, one could ask a fixed set of agents to take turns investigating and reporting on the question of interest. The key difference with the baseline model is that agents now have private information about what they did in the past, which affects how they interpret the declarations of other agents. If an agent has discovered disposable signals in the past, he knows that other agents have fewer signals to discover. The analysis becomes more complex because agents' decisions now depend on their beliefs about the amount of evidence left, about other agents' beliefs about the amount evidence, their belief about agents' beliefs

[42]The argument is somewhat similar to the section on witnesses in this paper.

about their beliefs about the amount of evidence left and so on. While information attrition is likely to have a similar effect as in this paper's model, confirming this intuition and exploring this question is left for future research.

4.3 Designing Truth-Dependent Preferences: Oaths, Capitalism, and Popular Juries

A more direct approach to improving mediated learning is to increase the salience of truth-dependent preferences.

This may be achieved by fostering agents' ethical sense, from inculcating an ethical education and culture to strengthening trust in institutions and developing effective vetting and selection processes for key learning responsibilities.

Professional oaths, from the Hippocratic oath in medicine to journalistic oaths such as Walter Williams' Journalist's Creed (Farrar (1998)) aim at eliciting ethical behavior. This paper suggests a positive correlation between the need for oaths in various professions and the severity of information attrition in these professions.

Even if a small fraction of agents is swayed by such oaths, this may in principle suffice for incentivizing truthful behavior by other agents. Studying the mechanisms and behavioral features through which ethical agents can incentivize mediated learning is beyond the scope of this paper, but it is easy to conceive of simple examples:[43] suppose that an agent, who is known to truthfully seek and report the truth is commonly know to appear in round $N > 1$. This agent's report provides reliable information, akin to an exogenous public signal about the state of the world, which can be used to incentivize all agents coming in rounds $i < N$. Even if agent N has only a small probability $p < 1$ of behaving ethically, his report may still be used to incentivize agents in earlier around as long as these agents' rewards and punishments are of order $1/p$.

Another approach is to organize society in a way that increases information mediators' material dependence on the truth, i.e., gives them "skin in the game." Eliciting information from agents about a scientific fact or the social value of a new product or process is easier

[43]There are various approaches to modeling ethical behavior. For instance, Ellingsen and Mohlin (2020) distinguish three dimensions: decency, integrity, and punitivity. Harsanyi (1980) and Feddersen and Sandroni (2006) study rule-utilitarian agents and Roemer (2019) consider Kantian agents.

when the agents stand to gain financially from this information, which may broadly interpreted as capitalistic incentives. Thus interpreted, the theory offers a new perspective on the "virtue" of capitalism relative to systems in which agents have low-powered incentives.[44] The theory also emphasizes that violations of the rules of capitalism (or any other system, for that matter) may be difficult to detect and reveal truthfully, and thus suggests a potential tradeoff between the incentives provided within a given economic or political system and the incentives required to guarantee that the system is respected by its participants.

A final angle to attack mediated learning failures is to democratize the learning process by enlarging the pool of potential information intermediaries. Large pools can increase the alignment—real or perceived—between the intermediaries and society as a whole, in contrast to the baseline model of the paper, in which society's objective is dissociated from the intermediaries'. Large pools of intermediaries are conceivable when the expertise required to learn the fact of interest is limited. The institution of popular juries may be viewed as one such application, which trades off intermediaries' expertise with their representativeness of a more global and diffuse body of stakeholders.

References

Bahar, A. and Kugel, W. (2001) *Der Reichstagsbrand. Wie Geschichte gemacht wird,* Edition Q Verlag, Berlin.

Banerjee, A. (1992) "A simple model of herd behavior," *Quarterly Journal of Economics*, Vol. 107, pp. 797–817.

Barlow, R., Marshall, A., and Proschan, F. (1963) "Properties of probability distributions with monotone hazard rate," *Annals of Mathematical Statistics*, Vol. 34, pp. 375–389.

Becker, G. and Stigler, G. (1974) "Law enforcement, malfeasance, and compensation of enforcers," *The Journal of Legal Studies*, Vol. 3, pp. 1–18.

Bénabou, R. and Tirole, J. (2003) "Intrinsic and extrinsic motivation," *Review of Economic Studies*, Vol. 70, pp. 489–520.

[44] A large literature emphasizes capitalism's ability to reduce moral hazard problems relative to socialistic systems. See Myerson (2007) and Tirole (2006) for a review of relevant papers and corporate-finance models capturing this idea. The question of incentive compatibility and its relation to various economic systems is at the heart of Leonid Hurwicz's development of mechanism design (Hurwicz (1973)).

Bénabou, R., Falk, A., and Tirole, J. (2020) "Eliciting moral preferences: Theory and experiment ," *Working Paper*, Princeton University.

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992) "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of Political Economy*, Vol. 100, pp. 992–1026.

Edgeworth, F. (1881) *Mathematical psychics: An essay on the application of mathematics to the moral sciences,* Kegan Paul.

Ellingsen, T. and Mohlin, E. (2020) "Dutiful behavior: A model of moral sentiments," *Working Paper*, Stockholm School of Economics.

Farrar R. (1998) *A Creed for My Profession: Walter Williams, Journalist to the World.* University of Missouri Press.

Feddersen, T., Sandroni, A. (2006) "A theory of participation in elections," *American Economic Review*, Vol. 96, pp. 1271–1282.

Frumkin, D., Wasserstrom, A., Davidson, A. and Grafit, A. (2010) "Authentication of forensic DNA samples," *Forensic science international: genetics*, Vol. 4, pp. 95–103.

Gershkov, A., Szentes, B. (2009) "Optimal voting schemes with costly information acquisition," *Journal of Economic Theory*, Vol. 144, pp. 36–68.

Glazer, J. and Rubinstein, A. (1998) "Motives and implementation: On the design of mechanisms to elicit opinions," *Journal of Economic Theory*, Vol. 79, pp. 157–173.

Granovetter, M. (2017) *Society and economy*, Harvard University Press.

Grice, H.P. (1975) "Logic and conversation," *In Cole, P., Morgan, J.L. (Eds.), Syntax and Semantics, Vol. 3*, Academic Press, New York, pp. 41–58.

Harsanyi, J. (1980) "Rule utilitarianism, rights, obligations and the theory of rational behavior," *Theory and Decision*, Vol. 12, pp. 115–133.

Hett, B. (2014) *Burning the Reichstag. An Investigation into the Third Reich's Enduring Mystery.* Oxford University Press.

Holmström, B. (1982) "Moral hazard in teams," *Bell Journal of Economics*, Vol. 13, pp. 324–340.

Huq, A. (2018) "Terrorism and democratic recession," *University of Chicago Law Review*, Vol. 85, pp. 457–484.

Hurwicz, L. (1973) "The design of mechanisms for resource allocation," *American Economic Review*, Vol. 63, pp. 1–30.

Hurwicz, L. (2007) "But who will guard the guardians?" *Nobel Prize Lecture*.

Lamport, L., Shostak, R. and Pease, M. (1982) "The Byzantine generals problem," *ACM Transactions on Programming Languages and Systems*, Vol. 4, pp. 382–401.

Landes, W. (1971) "An economic analysis of the courts," *The Journal of Law and Economics,* Vol. 14, pp. 61–107.

Levine, D., and Modica, S. (2016) "Peer discipline and incentives within groups," *ACM Transactions on Programming Languages and Systems*, Vol. 4, pp. 382–401.

Lutjens, R. (2016) "Burning the Reichstag: An investigation into the Third Reich's enduring mystery by Benjamin Hett (review)," *German Studies Review*, Vol. 39, pp. 411–412.

Matsushima, H. (2008) "Role of honesty in full implementation," *Journal of Economic Theory*, Vol. 139, pp. 353–359.

Milgrom, P., North, D., and B. Weingast (1990) "The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs," *Economics & Politics*, Vol. 2, pp. 1–23.

Myerson, R. (2006) "Federalism and incentives for success of democracy," *Quarterly Journal of Political Science*, Vol. 1, pp. 3–23.

Myerson, R. (2009) "Fundamental theory of institutions: a lecture in honor of Leo Hurwicz," *Review of Economic Design*, Vol. 13, p. 59–75.

Pei, H.D., and Strulovici, B. (2020) "Crime entanglement, deterrence, and witness credibility," *Working Paper*, Northwestern University.

Rahman, D. (2012) "But who will monitor the monitor?," *American Economic Review*, Vol. 102, pp. 2767–2797.

Roemer, J. (2019) *How we cooperate: A theory of Kantian optimization*, Yale University Press.

Smith, L., Sørensen, P., and Tian, J. (2020) "Informational herding, optimal experimentation, and contrarianism" *Forthcoming, Review of Economic Studies.*

Strulovici, B. (2020) "Learning and Corruptive Bargaining along Monitoring Chains," *Working Paper*, Northwestern University.

Tirole, J. (2006) *The theory of corporate finance.* Princeton University Press.

A Proof of Theorem 1

For any $i, k \geq 1$, let $F_i^k = \Pr(|S_i| \geq k \mid m_1^{i-1})$ denote the probability that there remain at least k signals to discover at the beginning of round i given past reports m_1^{i-1}. The prior probability F^k that the initial sequence S contains at least k signals satisfies $F^k = F_1^k$.

A.1 Gist of the Proof

To understand how the proof works, suppose first that S contains at most one signal, i.e., that $F_i^2 = 0$ for all i. We will show by contradiction that no agent ever works in equilibrium. When $F_i^2 \equiv 0$, the probability F_i^1 that there is a signal to discover in round i is decreasing in i path by path.[45]

Agent i works only if two conditions hold: (i) the probability F_i^1 that a signal remains is high enough—above a cutoff $\underline{F}^1$ provided by Lemma 1, below—(ii) the probability that some agent $j > i$ works after i has worked is high enough.

This creates a tension: On the one hand, the more likely agent i is to work, the larger the expected drop from F_i^1 to F_{i+1}^1. On the other hand, an agent works only if subsequent agents also work with sufficient probability, which requires that F_j^1 stay above $\underline{F}^1$. This dynamic is impossible as F_j^1 must go down by a non trivial amount but becomes squeezed above $\underline{F}^1$.

To formalize this tension, consider any history up to round i and let M_i^+ denote the set of messages m_i such that some $j > i$ works with positive probability. We wish to capture the following intuition: the probability of M_i^+ conditional on i working cannot be too small,

[45] Intuitively, either i shirked, in which case his message is uninformative, or he worked, in which case he either found the only signal, and there is nothing left, or he found nothing, which makes agents more pessimistic that there is a signal left to be found.

which implies that the likelihood that i has worked conditional sending a message m_i in M_i^+ cannot be too small, either. Conditionally on M_i^+, the average drop in F_i^1 must be somewhat in proportion to the probability that i works. This intuition is captured by the following *Discovery-Belief (DB)* equation, proved by Proposition 2. Let $\beta_i = \Pr_i(i \text{ discovers a signal})$ and suppose that $F_i^2 = 0$. Then, there exists $Q > 0$ s.t.

$$\beta_i \leq Q \frac{\mathbb{E}_i \left[(F_i^1 - F_{i+1}^1(m_i)) \mathbb{1}_{m_i \in M_i^+} \right]}{F_i^1}.$$

The numerator captures i's expected impact on beliefs *conditional on sending a message that yields an informative continuation equilibrium* ($m_i \in M_i^+$). The denominator exceeds $\underline{F}^1$.

Consider any informative equilibrium and let $\underline{F} = \inf\{F_j^1 : j \text{ works with positive proba}\}$, i.e., the infimum belief over all informative continuation equilibria. To build a contradiction, we choose an informative continuation equilibrium starting from some round i such that i works and $F_i^1 \leq \underline{F} + \varepsilon$. Such a continuation equilibrium must exist by definition of $\underline{F}$. Summing the DB equation over $j > i$ yields

$$\Pr{}_i(\ \exists j > i \text{ who discovers a signal}) \leq \frac{Q}{\underline{F}^1} \mathbb{E}_i \left[(F_{i+1}^1 - F_J^1) \right] \tag{2}$$

where J (random, possibly infinite) indexes the last person who works.

We have $F_J^1 \geq \underline{F}$ and, by monotonicity, $F_{i+1}^1 \leq F_i^1 \leq \underline{F} + \varepsilon$. From (2), the probability that some $j > i$ discovers a signal must thus be of order ε, which is too small to incentivize i and yields the desired contradiction. □

Challenges

The gist of the proof relied on two facts: $F_i^2 = 0$ and F_i^1 is decreasing. In general, all beliefs F_i^k can be positive and nonmonotonic.[46] The proof must address several challenges:

1. Suppose we wish to show that i never works if F_i^2 is arbitrarily small. Proposition 2 holds as long as F_i^2 is sufficiently small relative to F_i^1. To apply Proposition 2 to all $j > i$ we need to show that F_j^2 remains small for all $j > i$. Doob's martingale inequality (Lemma 4) allows us to show that this event, denoted $\mathcal{A}$ in the proof, is highly likely.

2. However, Doob's martingale inequality holds with respect to the filtration generated by public reports, whereas agent i's incentives are driven by the probability of event $\mathcal{A}$

[46] For example, agent i could send a message suggesting that previous agents have shirked and, hence, there are more signals to discover. Or i could send a message that is positively correlated with the existence of many other signal to discover, i.e., the equivalent of uncovering a "gold mine" of signals.

conditional on i working. This probability can be very different from the unconditional probability at the beginning of round i, especially if the probability that i works is very small. This issue is addressed by Lemma 5.

3. The gist of the proof selected i such that F_i^1 was close to the infimum $\underline{F}$. When there are more than two signals, we would ideally like to choose i such that (i) F_i^2 is arbitrarily small *and* (ii) F_i^1 is close to $\underline{F}$, but we cannot impose both conditions. To address this, the proof uses a *boundary function* mapping $f \mapsto \mathcal{F}^1(f)$, which defines the infimum of beliefs F_i^1 over informative continuation equilibria *such that* $F_i^2 \leq f$, *for* $f \in [0,1]$.

4. To exploit Equation (2), the gist of the proof used that F_{i+1}^1 was close to $\underline{F}$. This is generally false if F_i^1 is nonmonotonic. This issue is addressed by Lemma 6.

A.2 Discovery-Belief Equation

The section presents three lemmas leading to Proposition 2, which contains the discovery-belief equation. Lemma 1 states that an agent works only if the probability F_i^1 that there remains at least one signal to discover is high enough. Lemma 2 states that when an agent works and finds nothing, he cannot be much better off, ex post, than if he had shirked instead, especially if F_i^1 is close to 1, which means that the working agent's lack of a sdiscovery is likely attributable to bad luck. Lemma 3 computes an upper bound on an agent's expected benefit from working relative to shirking. This upper bound is expressed in terms of the probability that a working agent produces a message that triggers an informative continuation equilibrium. This set, M_i^+, plays a key role in the analysis. All three lemmas and Proposition 2 are proved in Appendix C.

We make two simplifications throughout the proof, which are without loss of generality. First, agents' decisions are invariant with respect to a uniform translation in their gross utility functions. We can therefore assume that these functions all take values in some interval $[0, R]$. Moreover, since R is an upper bound on payoffs, it can be increased to guarantee that $R > c$, which is assumed from now on.

Let γ_i denote the probability that i works given m_1^{i-1}.

LEMMA 1 $\gamma_i > 0$ *only if* $F_i^1 \geq \frac{c}{R} > 0$.

Given any round i and report history m_1^{i-1} such that $\gamma_i > 0$, let:

- V_i^* denote i's maximal expected gross utility if he shirks, where the maximum is taken over all possible messages $m_i \in M$ that i can send after shirking;
- $f_i^0 = 1 - F_i^1$ denote the probability that $S_i = \emptyset$ at the beginning of round i.

LEMMA 2 *Suppose that $\lambda < 1$. If i works, finds nothing, and sends message m_i, his expected gross utility $V_i^w(\emptyset, m_i)$ satisfies*

$$V_i^w(\emptyset, m_i) \leq V_i^* + \frac{f_i^0 R}{1-\lambda}.$$

For any round i and $\tilde{M}_i \subset M$, we consider the following probabilities conditional on history m_1^{i-1}:

- $g_i(\tilde{M}_i)$: probability that i finds a signal and sends a message in $\tilde{M}_i$ conditional on working;
- $d_i(\tilde{M}_i)$: probability that i finds no signal and sends a message in $\tilde{M}_i$ conditional on working;

Also let M_i^+ denote the set of messages m_i followed by an informative continuation equilibrium at round $i+1$ given the report history m_1^{i-1} until round i.

LEMMA 3 *Agent i's expected gross utility conditional on working has the following upper bound. If $\lambda < 1$, then*

$$V_i^w \leq V_i^* + d_i(M_i^+)\frac{f_i^0 R}{1-\lambda} + g_i(M_i^+)R.$$

If $\lambda = 1$, then

$$V_i^w \leq V_i^* + d_i(M_i^+)R + g_i(M_i^+)R.$$

For each round i, we introduce the following variables.

- β_i: probability that i discovers a signal conditional on report history m_1^{i-1} (before observing whether i works, i.e., viewed from the beginning of round i);
- $F_{i+1}^k(m_i)$: probability that there remain at least k signals at the beginning of round $i+1$ given reports $m_1^i = (m_1^{i-1}, m_i)$.

PROPOSITION 2 *Let $C(\lambda) = 2R/(c\lambda(1-\lambda))$ if $\lambda < 1$ and $C(1) = 2R/c$. The following inequality holds for all constants $C \geq C(\lambda)$, round i, and integer $k \geq 1$ such that $F_i^k > CF_i^{k+1}$:*

$$\beta_i \leq C\frac{\mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+}\right]}{F_i^k - CF_i^{k+1}}. \tag{3}$$

A.3 Proof of Theorem 1

The proof proceeds by induction on k. Lemma 1 already proves the claim for $k = 1$ with $\underline{F}^1 = c/R$. Now suppose that the claim holds for some $k \geq 1$: there exists a threshold $\underline{F}^k > 0$ such that any informative continuation equilibrium in round i satisfies $F_i^k \geq \underline{F}^k$.[47] We will show that there is a constant $\underline{F}^{k+1} > 0$ such that a continuation equilibrium can be informative only if $F_i^{k+1} \geq \underline{F}^{k+1}$.

For any $f \in [0,1]$, let $\mathcal{F}^k(f) = \inf\{F_i^k \in [0,1] : \gamma_i > 0, F_i^{k+1} \leq f\}$ where the infimum is taken over all on-path histories and all equilibria of the game, and is by convention equal to 1 if no informative equilibrium exists for which $F_i^{k+1} \leq f$. By construction, $\mathcal{F}^k(f)$ is nonincreasing in f.

Let $\underline{F} = \inf\{f : \mathcal{F}^k(f) < 1\}$. $\underline{F}$ is the smallest value of F_i^{k+1} for which an informative continuation equilibrium exists (or the infimum of such values, if the infimum is not achieved).

Our objective is to show that $\underline{F} > 0$. Let $\bar{F}^k = \lim_{\omega \searrow \underline{F}} \mathcal{F}^k(\omega)$, i.e., the right limit of $\mathcal{F}^k(\cdot)$ at $\underline{F}$. This limit is guaranteed to exist because $\mathcal{F}^k(\cdot)$ is nonincreasing. Roughly put, $\bar{F}^k$ is the smallest probability that there remain at least k signals among informative equilibria that achieve the smallest possible probability that there remain at least $k+1$ signals.

Also let $\varepsilon > 0$ denote any small constant such that $\hat{F}^k = \mathcal{F}^k(\underline{F} + G\varepsilon) \geq \frac{\bar{F}^k}{1+\eta}$, where $G > 0$ is a large constant and $\eta > 0$ is a small constant determined at the end of the proof independently of k.[48] Since $\bar{F}^k$ is the right limit of $\mathcal{F}^k(\cdot)$ at $\underline{F}$, such an ε exists. Moreover, since all $\varepsilon' \in (0, \varepsilon)$ also satisfy the condition, we can choose ε so that

$$\varepsilon \leq \frac{1}{2}\left(\frac{\hat{F}^k}{G}\right)^2. \tag{4}$$

By definition of $\underline{F}$, there exists at least one informative continuation equilibrium for which $F_i^{k+1} \in [\underline{F}, \underline{F} + \varepsilon]$. Moreover, by definition and monotonicity of $\mathcal{F}^k(\cdot)$, there exists an informative continuation equilibrium among those for which $F_i^k \leq \bar{F}^k + \eta\hat{F}^k$.

Consider such a continuation equilibrium, and suppose by contradiction that

$$\underline{F} = 0. \tag{5}$$

[47] Theorem 1, which was stated for round 0, also applies to all continuation equilibria: the thresholds $\{\underline{F}^k\}_{k\geq 1}$ depend only on the parameters (R, P, c, λ), which are constant throughout the game.

[48] For example, if $\lambda < 1$ one can choose G and η such that $\sqrt{G} = 128R^3/(\lambda^2(1-\lambda)c^3)$ and $\eta = 1/\sqrt{G}$, as explained at the end of this proof.

Let $\mathcal{A}$ denote the event that $F_j^{k+1} \leq G\varepsilon$ for all $j \geq i$.

LEMMA 4 *i assigns probability at least $1 - 1/G$ to $\mathcal{A}$.*

Proof. For $j \geq i$, let $\bar{F}_j^{k+1}$ denote the probability assigned at the beginning of round j (i.e., conditional on m_1^{j-1}) that there remain at least $k+1$ signals *at the beginning of round i* (fixed). The process $\{\bar{F}_j^{k+1}\}_{j \geq i}$ is nonnegative and bounded above by 1, and it is a martingale by the law of iterated expectations and the fact that j's filtration grows finer as j increases. Moreover, $\bar{F}_i^{k+1} = F_i^{k+1} \leq \varepsilon$.

Doob's martingale inequality therefore implies that for any $J \geq i$, $\Pr(\max_{i \leq j \leq J} \bar{F}_j^{k+1} \geq G\varepsilon) \leq \frac{\mathbb{E}_i[\bar{F}_i^{k+1}]}{G\varepsilon} \leq 1/G$. The event $\bar{\mathcal{A}}_\infty$ defined by $\{\max_{j \geq i} \bar{F}_j^{k+1} \leq G\varepsilon\}$ is the intersection of the events $\bar{\mathcal{A}}_J = \{\max_{i \leq j \leq J} \bar{F}_j^{k+1} \leq G\varepsilon\}$. Therefore, $\Pr(\bar{\mathcal{A}}_\infty) = \lim_{J \to \infty} \Pr(\bar{\mathcal{A}}_J) \geq 1 - 1/G$.

Finally, we note that $\bar{F}_j^{k+1} \geq F_j^{k+1}$ for all $j \geq i$, because the true number of remaining signals only decreases over time and thus whatever signals remained at the beginning of round i must have contained the signals that remain at the beginning of round j, so $\Pr(\mathcal{A}) \geq \Pr(\bar{\mathcal{A}}_\infty) \geq 1 - 1/G$. ■

Conditional on $\mathcal{A}$, $F_j^{k+1} \leq G\varepsilon$ for all $j \geq i$. Moreover, if round j belongs to an informative continuation equilibrium, we have $F_j^k \geq \mathcal{F}^k(F_j^{k+1}) \geq \mathcal{F}^k(G\varepsilon) = \hat{F}^k$. Given any positive constant C, this implies that for informative continuation equilibrium starting in round j,

$$F_j^k - CF_j^{k+1} \geq \hat{F}^k - CG\varepsilon \geq \hat{F}^k/2 > 0 \tag{6}$$

provided that $G \geq C$, where the last weak inequality comes from (4). The condition $G \geq C$ will be satisfied by setting $C = C(\lambda)$, where $C(\lambda)$ is defined in Proposition 2, and then choosing G large enough. (A specific value of G is given at the end of the proof.)

Let $\mathcal{A}_j$ denote the event that $F_l^{k+1} \leq G\varepsilon$ for all integers l such that $i \leq l \leq j$. The events $\{\mathcal{A}_j\}_{j \geq i}$ form a decreasing sequence that converges to $\mathcal{A}$ as $j \to \infty$. Moreover, $\mathcal{A}_j$ is measurable with respect to the information available at the beginning of round j.

Proposition 2 implies, conditional on event $\mathcal{A}_j$, that:

$$\beta_j \leq \mathbb{E}_j \left[\frac{C(F_j^k - F_{j+1}^k(m_j))\mathbb{1}_{m_j \in M_j^+}}{F_j^k - CF_j^{k+1}} \right]. \tag{7}$$

Let $\mathcal{B}_j$ denote the event that $m_j \in M_j^+$. $\mathcal{B}_j$ is adapted to m_1^j, i.e., known at the beginning of round $j+1$. Notice that if $\mathcal{B}_j$ does *not* occur, it means by definition that no $l > j$ ever works

(i.e., m_1^j is followed by an *un*informative continuation equilibrium). This implies that the sequence of events $\{\mathcal{B}_j\}_{j\geq i}$ is decreasing path by path and, hence, that the sequence $\{\mathbb{1}_{\mathcal{B}_j}\}_{j\geq i}$ is nonincreasing.

Since $\mathcal{A}_j$ is measurable with respect to m_1^{j-1}, equations (6) and (7) imply that for $j \geq i+1$

$$\mathbb{E}_{i+1}[\mathbb{1}_{\mathcal{A}_j}\beta_j] \leq \mathbb{E}_{i+1}\left[\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}\frac{2C(F_j^k - F_{j+1}^k)\mathbb{1}_{\mathcal{B}_j}}{\hat{F}^k}\right]\right].$$

The law of iterated expectations then implies that

$$\mathbb{E}_{i+1}[\mathbb{1}_{\mathcal{A}_j}\beta_j] \leq \mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}_j}\frac{2C(F_j^k - F_{j+1}^k)\mathbb{1}_{\mathcal{B}_j}}{\hat{F}^k}\right] = \frac{2C}{\hat{F}^k}\mathbb{E}_{i+1}[\mathbb{1}_{\mathcal{A}_j}(F_j^k - F_{j+1}^k)\mathbb{1}_{\mathcal{B}_j}]. \tag{8}$$

Summing (8) over all $j \geq i+1$, we obtain

$$\mathbb{E}_{i+1}\left[\sum_{j\geq i+1}\mathbb{1}_{\mathcal{A}_j}\beta_j\right] \leq \frac{2C}{\hat{F}^k}\mathbb{E}_{i+1}\left[\sum_{j\geq i+1}\mathbb{1}_{\mathcal{A}_j}\mathbb{1}_{\mathcal{B}_j}(F_j^k - F_{j+1}^k(m_j))\right]. \tag{9}$$

Since the indicator functions on the right-hand side are nonincreasing in j for $j \geq i+1$, path by path, there must be a first (possibly infinite) round J for which the product of indicators is zero:

$$J = \inf\{j \geq i+1 : \mathbb{1}_{\mathcal{A}_j}\mathbb{1}_{\mathcal{B}_j} = 0\}$$

with the convention that $J = +\infty$ if the set is empty. In words, J is either the first round in which F_j^{k+1} exceeds $G\varepsilon$, or the *last* round at which the continuation equilibrium is informative (which implies that $\gamma_j = 0$ for all $j > J$), whichever occurs first.[49]

Consider first the paths for which J is finite. The argument of $\mathbb{E}_{i+1}[\cdot]$ on the right-hand side of (9) then reduces to

$$\sum_{j=i+1}^{J-1}(F_j^k - F_{j+1}^k) = F_{i+1}^k - F_J^k.$$

Consider now paths for which J is infinite. In this case, the argument of $E_{i+1}[\cdot]$ on the right-hand side of (9) is equal to

$$\sum_{j=i+1}^{\infty}(F_j^k - F_{j+1}^k) = \lim_{\tilde{J}\to\infty}\sum_{j=i+1}^{\tilde{J}}(F_j^k - F_{j+1}^k) = \lim_{\tilde{J}\to\infty}\{F_{i+1}^k - F_{\tilde{J}+1}^k\} = F_{i+1}^k - \lim_{\tilde{J}\to\infty}F_{\tilde{J}}^k.$$

[49] J is not a stopping time with respect to the filtration $\{\mathbb{F}_j\}_{j\geq i+1}$ generated by public histories $\{m_1^{j-1}\}_{j\geq i+1}$: at the beginning of any round j at which j works with positive probability, it is unknown whether j will be the last round in which the agent works. The proofs below do not use the optional sampling theorem or the strong Markov property.

Notice that the limit is well defined because F_j^k is a nonnegative supermartingale with respect to j.[50] We will call this limit F_J^k to be consistent with the case in which J is finite.

Combining these observations with (9), we conclude that

$$\mathbb{E}_{i+1}\left[\sum_{j>i} \mathbb{1}_{\mathcal{A}_j}\beta_j\right] \leq \frac{2C}{\hat{F}^k}\left(F_{i+1}^k - \mathbb{E}_{i+1}[F_J^k]\right).$$

We have $\beta_j = \gamma_j F_j^1 \lambda$. Moreover, Lemma 1 shows that $\gamma_j > 0$ only if $F_j^1 \geq c/R$. This implies[51] that $\gamma_j \leq g\beta_j$ where $g = \frac{R}{\lambda c}$ and hence that

$$\mathbb{E}_{i+1}\left[\sum_{j\geq i+1} \mathbb{1}_{\mathcal{A}_j}\gamma_j\right] \leq \frac{2Cg}{\hat{F}^k}\left(F_{i+1}^k - \mathbb{E}_{i+1}[F_J^k]\right). \tag{10}$$

Let $\mathcal{Z}$ denote the event that at least one agent $j > i$ works and $\pi_{i+1}(m_i) = \Pr_{i+1}(\mathcal{A} \cap \mathcal{Z})$, i.e., the probability that $\mathcal{A}$ and $\mathcal{Z}$ both occur conditional on m_1^i. We have

$$\begin{aligned}
\pi_{i+1}(m_i) &= \Pr_{i+1}\left(\mathbb{1}_{\mathcal{A}}\sum_{j>i}\mathbb{1}_j \text{ works} \geq 1\right) \\
&\leq \mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}}\sum_{j>i}\mathbb{1}_j \text{ works}\right] \\
&= \sum_{j>i}\mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}}\mathbb{1}_j \text{ works}\right] \\
&\leq \sum_{j>i}\mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}_j}\mathbb{1}_j \text{ works}\right] \\
&= \sum_{j>i}\mathbb{E}_{i+1}\left[\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}\mathbb{1}_j \text{ works}\right]\right] \\
&= \sum_{j>i}\mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}_j}\mathbb{E}_j\left[\mathbb{1}_j \text{ works}\right]\right] \\
&= \sum_{j>i}\mathbb{E}_{i+1}\left[\mathbb{1}_{\mathcal{A}_j}\gamma_j\right].
\end{aligned}$$

[50] The argument is similar to the one used to prove Lemma 4. For any fixed j, let $\bar{F}_l^{k+1}$ denote the probability assigned by $l \geq j$ to there remaining at least $k+1$ signals *at the beginning of round* j. The process $\{\bar{F}_l^{k+1}\}_{l\geq j}$ is a martingale in j, by the law of iterated expectations and the fact that that l's filtration grows finer as l increases. Moreover, $\bar{F}_l^{k+1} \geq F_l^{k+1}$ for all $l \geq j$, because the actual number of remaining signals only decreases over time. Therefore, we have $F_j^{k+1} = \bar{F}_j^{k+1} = E_j[\bar{F}_{j+1}^{k+1}] \geq E_j[F_{j+1}^{k+1}]$. This, together with the fact that F_l^{k+1} is uniformly bounded and measurable with respect to the information at the beginning round l, shows that it is a supermartingale.

[51] The inequality clearly holds if $\gamma_j = 0$.

The first equality comes from the fact that $\mathcal{Z}$ is identical to the event $\{\sum_{j>i} \mathbb{1}_j \text{ works} \geq 1\}$. The first inequality comes from the fact that $\mathbb{1}_{\mathcal{A}} \sum_{j>i} \mathbb{1}_j \text{ works}$ is nonnegative and integer valued, so that its expectation exceeds the probability that it is strictly positive. The second equality is an application of Tonelli's theorem. The second inequality comes from the fact that $\mathcal{A} \subset \mathcal{A}_j$ and, hence, $\mathbb{1}_{\mathcal{A}} \leq \mathbb{1}_{\mathcal{A}_j}$. The next equality comes from the law of iterated expectations and the next one comes from the fact that $\mathcal{A}_j$ is measurable with respect to the information at the beginning of round j. The last equality holds by definition of γ_j.

From (10) and Tonelli's theorem, this implies that

$$\pi_{i+1}(m_i) \leq \frac{2Cg}{\hat{F}^k} \left(F_{i+1}^k - \mathbb{E}_{i+1}[F_J^k] \right). \tag{11}$$

Let $F_i^{k,r}(m_i)$ denote the probability that there are at least k signals left conditional on i working and reporting m_i, and for any signal s_i let $F_i^{k,w}(s_i)$ denote the probability that there are at least k signals left conditional on i working and discovering s_i. $F_i^{k,w}(s_i)$ represents i's belief after discovering s_i, while $F_i^{k,r}(m_i)$ represents what $i+1$ would believe conditional on knowing that i worked and observing report m_i.

The following lemmas are proved in Appendices C.5 and C.6. Let $N_i = \{m_i : F_i^{k,r}(m_i) > (1+\eta)F_i^k\}$.

LEMMA 5 *(i)* $\gamma_i(N_i) \leq \frac{1}{2\eta G^2}$, *(ii) for* $m_i \notin N_i$, $\pi_{i+1}(m_i) \leq \frac{2Cg}{\hat{F}^k} \left((1+\eta)F_i^k - \mathbb{E}_{i+1}[F_J^k] \right)$.

Let $T_i = \{m_i : F_{i+1}^{k+1}(m_i) \geq \sqrt{G}\varepsilon\}$.

LEMMA 6 *(i)* $\Pr_{i+1}(\mathcal{A}) \geq 1 - 1/\sqrt{G}$ *for all* $m_i \notin T_i$, *(ii)* $\gamma_i(T_i) \leq 1/\sqrt{G}$.

Let V^w denote i's expected gross utility if he works and $V^w(m_i)$ denote his expected gross utility conditional on working and reporting m_i. We have

$$\begin{aligned} V^w &= \sum_{i \in M_i} \gamma_i(m_i) V^w(m_i) \\ &\leq (\gamma_i(N_i) + \gamma_i(T_i))R + \sum_{m_i \notin N_i \cup T_i} \gamma_i(m_i) V^w(m_i) \\ &\leq \left(\frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} \right) R + \sum_{m_i \notin N_i \cup T_i} \gamma_i(m_i) V^w(m_i) \end{aligned} \tag{12}$$

Moreover,

$$V^w(m_i) = \Pr(\mathcal{Z}|m_1^i) V^w(m_i|\mathcal{Z}) + \Pr(\mathcal{Z}^c|m_1^i) V^w(m_i|\mathcal{Z}^c). \tag{13}$$

Conditional on m_1^i, the event $\mathcal{Z}$ is independent of how m_i was produced (i.e., whether m_i was obtained by work or fabrication). Indeed, as long as no one works, the distribution of reports m_j made by agents following i depends only on m_1^i, not on the signals that remain to be discovered in the case. And as soon as someone works, then by definition $\mathcal{Z}$ has occurred. Thus, what triggers the event $\mathcal{Z}$ (whenever it occurs) is a sequence of uninformative (until $\mathcal{Z}$ occurs) reports m_j for agents following i, whose probability distribution is completely pinned down by m_1^i.

From the previous lemmas, we have $\Pr(\mathcal{A} \cap \mathcal{Z}|m_1^i) \leq \frac{2Cg}{\hat{F}^k}\left((1+\eta)F_i^k - \mathbb{E}_{i+1}[F_J^k]\right)$ for all $m_i \notin N_i$ and $\Pr(\mathcal{A}^c|m_1^i) \leq 1/\sqrt{G}$ for all $m_i \notin T_i$. Letting $\hat{M}_i = M_i \setminus (N_i \cup T_i)$, this implies that

$$\Pr(\mathcal{Z}|m_1^i) = \Pr(\mathcal{Z} \cap \mathcal{A}|m_1^i) + \Pr(\mathcal{Z} \cap \mathcal{A}^c|m_1^i) \leq \frac{2Cg}{\hat{F}^k}\left(F_i^k(1+\eta) - \mathbb{E}_{i+1}[F_J^k]\right) + 1/\sqrt{G} \quad (14)$$

for all $m_i \in \hat{M}_i$.

Conditional on $\mathcal{A}$, $F_j^{k+1} \leq G\varepsilon$ for all $j \geq i$. By definition of J, all continuation equilibria until round J included are informative, which implies that $F_j^k \geq \mathcal{F}^k(F_j^{k+1})$ for all $j \leq J$. Since $\mathcal{F}^k(\cdot)$ is nonincreasing, this implies that $F_j^k \geq \mathcal{F}^k(G\varepsilon) = \hat{F}^k$ for all $j \leq J$.

We thus have for $m_i \in \hat{M}_i$

$$\begin{aligned}
\mathbb{E}_{i+1}F_J^k &= \Pr_{i+1}(\mathcal{A})\ \mathbb{E}_{i+1}[F_J^k|\mathcal{A}] + \Pr_{i+1}(\mathcal{A}^c)\ \mathbb{E}_{i+1}[F_J^k|\mathcal{A}^c] \\
&\geq \Pr_{i+1}(\mathcal{A})\mathbb{E}_{i+1}[F_J^k|\mathcal{A}] \\
&\geq (1 - 1/\sqrt{G})\hat{F}^k.
\end{aligned}$$

By construction, $F_i^k \leq \bar{F}^k + \hat{F}^k\eta$ and $\hat{F}^k \geq \bar{F}^k - \hat{F}^k\eta$. Therefore, $F_i^k - \hat{F}^k \leq (\bar{F}^k + \hat{F}^k\eta) - (\bar{F}^k - \hat{F}^k\eta) = 2\eta\hat{F}^k$. Letting $B = 8Cg$, (14) then implies (for $\eta \leq 1$, which we assume) that for all m_i in $\hat{M}_i$

$$\Pr(\mathcal{Z}|m_1^i) \leq B\eta + \frac{B}{2\sqrt{G}} + 1/\sqrt{G}. \quad (15)$$

For each m_i, let $V_i^f(m_i|\mathcal{Z}^c)$ denote i's expected gross utility if he sends message m_i conditional on no $j > i$ working and $V^{f,*}$ denote the maximizer of $V_i^f(m_i|\mathcal{Z}^c)$ over all messages $m_i \in \hat{M}_i$. Notice that i's expected gross utility conditional on m_i and no $j > i$ working does not depend on whether i worked or shirked: either way, the subsequent reports $\{m_j\}_{j>i}$ are independent of the signals that remain to be discovered. Therefore, i's conditional expected gross utilities satisfy $V^w(m_i|\mathcal{Z}^c) = V_i^f(m_i|\mathcal{Z}^c)$.

Combining these observations with (13) and (15), we obtain

$$V^w(m_i) \leq \left(B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}}\right) R + V^{f,*},$$

for $m_i \in \hat{M}_i$. Combining this with (12) yields

$$V^w \leq \left(\frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}}\right) R + \left(B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}}\right) R + V^{f,*}.$$

i's utility from working thus satisfies

$$U^w \leq \left(\frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} + B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}}\right) R + V^{f,*} - c. \tag{16}$$

If i sends a message $m_i^* \in \hat{M}_i$ that achieves $V^{f,*}$, his utility U^f satisfies

$$\begin{aligned} U^f &\geq \Pr(\mathcal{Z}|m_1^{i-1}, m_i^*) \times 0 + \Pr(\mathcal{Z}^c|m_1^{i-1}, m_i^*) V^{f,*} \\ &\geq \left(1 - B\eta - \frac{B}{2\sqrt{G}} - \frac{1}{\sqrt{G}}\right) V^{f,*} \\ &\geq V^{f,*} - \left(B\eta - \frac{B}{2\sqrt{G}} - \frac{1}{\sqrt{G}}\right) R \end{aligned}$$

where 0 is used as a lower bound on i's realized gross utility in the first inequality.

Therefore, working is strictly suboptimal if

$$\left(\frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} + B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}}\right) R + V^{f,*} - c < V^{f*} - \left(B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}}\right) R$$

or

$$\frac{c}{R} > \frac{1}{2\eta G^2} + \frac{3}{\sqrt{G}} + 2B\eta + \frac{B}{\sqrt{G}}. \tag{17}$$

This inequality is always satisfied as long as η is small enough and ηG is large enough. In particular, since $B = 8Cg$, $g = R/\lambda c$, and C can be taken to equal $2R/c$ if $\lambda = 1$ and $2R/(c\lambda(1-\lambda))$ as noted in Proposition 2, the inequality is satisfied if $\eta = 1/\sqrt{G}$ and $\sqrt{G} = 128R^3/c^3$ for $\lambda = 1$ and $\sqrt{G} = 128R^3/(c^3\lambda^2(1-\lambda))$ if $\lambda < 1$ (recalling that $R > c$), in which case each term on the right-hand side of (17) is less than $c/4R$ with some strict inequalities. ■

B Proof of Theorem 2

Suppose first that $\lambda = 1$ and $\rho = 1$, which means that any agent who works finds a signal with probability 1. We construct compensation functions for which the strategy profile proposed constitutes an equilibrium. Under this strategy profile, as long as p_i lies in $(\underline{p}, \bar{p})$ all agents work and truthfully report their signal about ω. Moreover, given the symmetric signal structure, p_i depends only on the number of "H" and "L" signals as long as all agents $j < i$ work with probability 1. Therefore, the set of equilibrium posteriors forms a grid $\{q^k\}$ containing $\hat{p}$ and containing a single point on each side of $(\underline{p}, \bar{p})$. Let $q^0 \leq \underline{p} < q^1, \ldots, \hat{p}, \ldots, q^N < \bar{p} < q^{N+1}$ denote this grid. Along the candidate equilibrium, the belief p_i evolves on this grid until it hits either q^0 or q^{N+1}, after which the investigation stops.

Let J denote the last investigator who works: we have $p_J \in \{q^1, q^N\}$ and $p_{J+1} \in \{q^0, q^{N+1}\}$. Also let $\tilde{p} = p_{J+1}$ denote the value of the belief when learning stops under the candidate equilibrium.

We construct utility functions in which an investigator's compensation depends only on his report and on the posterior $\tilde{p}$.

For any i such that $p_i = q^k \in (\underline{p}, \bar{p})$, if i reports "H", he gets a reward $R_H^k \geq 0$ if $\tilde{p} = q^{N+1}$ and a punishment $P_L^k \leq 0$ if $\tilde{p} = q^0$. If i reports "L", he gets $R_L^k \geq 0$ if $\tilde{p} = q^0$ and $P_L^k \leq 0$ if $\tilde{p} = q^{N+1}$.

For any p, q on the grid, let $\pi(p, q)$ denote the probability that the belief sequence ends with $\tilde{p} = q^{N+1}$, i.e., exits $(\underline{p}, \bar{p})$ through $\bar{p}$, from the perspective of an agent who assigns probability p to ω, but the prior used by investigators is $p_0 = q$. That is, $\pi(p, q)$ is the probability that an individual with prior p assigns to the sequence p_i converging to q^{N+1} in equilibrium given that the public belief, which serves as the state variable for the equilibrium, starts at q.

If i sends report "H" starting from prior $p_i = q^k$, he assigns a probability $\pi(q^k, q^{k+1})$ to the public belief converging to q^{N+1}. If i works and receives report "H'', his belief about the continuation equilibrium is $\pi(q^{k+1}, q^{k+1})$. Similarly, if i sends "L'', his belief is $\pi(q^k, q^{k-1})$ whereas if he works and reports "L'' his belief is $\pi(q^{k-1}, q^{k-1})$. It is straightforward to verify the inequalities

$$\pi(q^{k+1}, q^{k+1}) > \pi(q^k, q^{k+1}) \tag{18}$$

and

$$\pi(q^{k-1}, q^{k-1}) < \pi(q^k, q^{k-1}), \tag{19}$$

for all $k \in [2, N-1]$. The strictness of the inequalities comes from the fact that conditional on the true state ω, the dynamic of $\{p_j\}_{j \geq i+1}$ starting any given value of p_{i+1} is strictly increasing in ω in FOSD, as is easily checked. Therefore, the probability of hitting q^{N+1} before q^0 is strictly increasing in the belief p_i that the state is high.

For $k = 1$, the investigation stops if i reports "L" so (19) holds as an equality, but (18) is still strict, because this report triggers further investigation. The reverse is true for $k = N$: (18) only holds as an equality while (19) is strict.

If i shirks, his maximal utility is

$$\max\{\pi(q^k, q^{k+1})R_H^k + (1 - \pi(q^k, q^{k+1}))P_H^k; \pi(q^k, q^{k-1})P_L^k + (1 - \pi(q^k, q^{k-1}))R_L^k\}. \tag{20}$$

The left argument is i's expected payoff if he sends "H", and the right one is his payoff if he sends "L". Since i can send either message at no cost, his best payoff from fabrication is the maximum of these two terms. If i works, he gets

$$z^k[\pi(q^{k+1}, q^{k+1})R_H^k + (1 - \pi(q^k, q^{k+1}))P_H^k] + (1 - z^k)[\pi(q^{k-1}, q^{k-1})P_L^k + (1 - \pi(q^{k-1}, ^{k-1}))R_L^k] \tag{21}$$

where z^k is the probability of receiving signal "H" given belief q^k, and is equal to $z^k = \Pr(\text{“}H''|q^k) = q^k\pi + (1 - q^k) \times (1 - \pi)$.

Working is optimal for i if (21) exceeds (20) by at least c.

This condition is obtained as follows: set $P_H^k = P_L^k = -Q$ where Q is a strictly positive constant, and let $R_H^k = Q\frac{1-\pi(q^k,q^{k+1})}{\pi(q^k,q^{k+1})}$ and $R_L^k = Q\frac{\pi(q^k,q^{k-1})}{1-\pi(q^k,q^{k-1})}$. This guarantees that i's expected payoff from fabrication is zero, regardless of the outcome. From (18) and (19), his payoff from working is of order Q and thus exceeds c, for Q high enough.[52]

If $k = 1$ or N, there is one signal that i can send after working which yields a payoff of order Q, while the other signal yields 0. The signal associated with a positive payoff arises with a probability that is bounded away from 0, since p_i lies in $(\underline{p}, \bar{p})$.

Moreover this scheme is feasible as long as the maximal reward R and and punishment P respectively exceed $\sup\{R_\theta^k : \theta \in \{L, H\}, k \in \{1, \ldots, N\}\}$ and Q.

The proof easily generalizes to $\lambda < 1$ and $\rho < 1$. See Appendix C.7.

[52] To see this, let $\bar{\pi}$ denote a strictly positive lower bound on all inequalities (18) and (19) over all k's whenever they hold strictly. Then, the gain from working is of order $Q\bar{\pi}$.

C Remaining Proofs for Theorems 1 and 2

C.1 Proof of Lemma 1

Let V_i^w denote i's expected gross utility if he works. V_i^w may be decomposed in terms of i's expected gross utility $\bar{V}_i$ if he works *and* there exists some signal left to be found, and his expected gross utility $V_i^{w,\emptyset}$ if he works but there is no signal left to be found ($S_i = \emptyset$):

$$V_i^w = F_i^1 \bar{V}_i + (1 - F_i^1) V_i^{w,\emptyset}.$$

Conditional on $S_i = \emptyset$, i's expected gross utility if he works is the same as his expected gross utility $V_i^{s,\emptyset}$ if he shirks and uses the same reporting strategy as he does after working and finding nothing: conditional on i's report (whatever it is), the distribution of reports by subsequent agents is identical since there is no signal left to be found. Therefore, $V_i^{w,\emptyset} = V_i^{s,\emptyset}$. Furthermore, we also have $\bar{V}_i \leq R$ since R is the maximum possible gross utility.

Therefore, i's net utility U_i^w from working, including the cost of working, satisfies

$$U_i^w \leq F_i^1 R + (1 - F_i^1) V_i^{s,\emptyset} - c.$$

Similarly, i's utility U_i^s from shirking satisfies

$$U_i^s \geq F_i^1 \times 0 + (1 - F_i^1) V_i^{s,\emptyset} = (1 - F_i^1) V_i^{s,\emptyset}$$

where the inequality comes from the fact that 0 is a lower bound on i's realized gross utility. Comparing the previous two equations shows that shirking strictly dominates working if $F_i^1 R - c < 0$.

C.2 Proof of Lemma 2

For any sequence S'' of signals, let $\Delta_i(S'')$ denote the probability that $S_i = S''$ conditional on report history m_1^{i-1}, and $\Delta_i^\emptyset(S'')$ denote the probability that $S_i = S''$ conditional on i working and finding nothing. Bayesian updating implies that for any $S'' \neq \emptyset$:

$$\Delta_i^\emptyset(S'') = \Delta_i(S'') \frac{(1 - \lambda)}{(1 - f_i^0)(1 - \lambda) + f_i^0},$$

and for $S'' = \emptyset$:

$$\Delta_i^\emptyset(\emptyset) = \Delta_i(\emptyset) \frac{1}{(1 - f_i^0)(1 - \lambda) + f_i^0}.$$

This implies that

$$\Delta_i^{\emptyset}(S'') - \Delta_i(S'') = -\frac{\lambda f_i^0 \Delta_i(S'')}{(1-f_i^0)(1-\lambda)+f_i^0} \tag{22}$$

for any $S'' \neq \emptyset$, and

$$\Delta_i^{\emptyset}(\emptyset) - \Delta_i(\emptyset) = \frac{\lambda(1-f_i^0)\Delta_i(\emptyset)}{(1-f_i^0)(1-\lambda)+f_i^0}. \tag{23}$$

Let $V_i(m_i, S'')$ denote i's expected gross utility conditional on i producing evidence m_i and on $S_{i+1} = S''$. Notice that $m_1^i = (m_1^{i-1}, m_i)$ and S_{i+1} completely determine the distribution of reports $\{m_j\}_{j>i}$. Therefore, $V_i(m_i, S'')$ is the same regardless of whether i has worked or shirked. Agent i's expected gross utility conditional on (i) working, (ii) finding no signal, and (iii) producing message m_i, is

$$V_i^w(\emptyset, m_i) = \sum_{S'' \in \mathcal{S}} V_i(m_i, S'') \Delta_i^{\emptyset}(S''),$$

whereas his expected gross utility if i shirks and sends message m_i is

$$V_i^s(m_i) = \sum_{S'' \in \mathcal{S}} V_i(m_i, S'') \Delta_i(S'')$$

because i has learned nothing from shirking and thus holds the same belief as his prior belief at the beginning of round i. Combining these expressions, we get

$$V_i^w(\emptyset, m_i) - V_i^s(m_i) = \sum_{S'' \in \mathcal{S}} V_i(m_i, S'')(\Delta_i^{\emptyset}(S'') - \Delta_i(S'')). \tag{24}$$

Since $V_i(m_i, S'') \in [0, R]$ for all m_i and S'', combining (24) with (22) and (23) yields

$$V_i^w(\emptyset, m_i) - V_i^s(m_i) \leq \frac{R\Delta_i(\emptyset)\lambda(1-f_i^0)}{(1-f_i^0)(1-\lambda)+f_i^0}.$$

Since $\lambda < 1$, the denominator is bounded below by $1-\lambda$. Since $\Delta_i(\emptyset) = f_i^0$, the numerator is bounded above by Rf_i^0. This yields

$$V_i^w(\emptyset, m_i) \leq V_i^s(m_i) + f_i^0 \frac{R}{1-\lambda} \leq V_i^* + f_i^0 \frac{R}{1-\lambda},$$

which proves the lemma. Intuitively, this results means that if f_i^0 is negligible relative to $(1-\lambda)$, then i's expected gross utility after working and finding nothing cannot be much higher than if i had shirked, because finding nothing in this case merely reveals that i was unlucky and otherwise conveys little else information.

C.3 Proof of Lemma 3

For each $m_i \in M$, let $V_i^w(m_i)$ denote i's expected gross utility conditional on working and sending message m_i and M_i^- denote the set of messages m_i after which no $j > i$ ever works, so that $M = M_i^+ \cup M_i^-$ and $M_i^+ \cap M_i^- = \emptyset$. Letting $\gamma_i(\tilde{M}_i)$ denote the probability that i sends a message in $\tilde{M}_i$ conditional on working and on m_1^{i-1}, we have:

$$V_i^w = \sum_{m_i \in M_i^-} \gamma_i(m_i) V_i^w(m_i) + \sum_{m_i \in M_i^+} \gamma_i(m_i) V_i^w(m_i). \tag{25}$$

For the first term, note that i's expected utility conditional on reporting m_i and on no $j > i$ ever producing real evidence does not depend on whether i worked or shirked: either way, the distribution of the reports $\{m_j\}_{j>i}$ is independent of the set of signals that remain in the case. Letting, as in the previous lemma, $V_i^s(m_i)$ denote i's expected gross utility conditional on shirking and sending message m_i, we thus have $V_i^w(m_i) = V_i^s(m_i)$ for all $m_i \in M_i^-$. Since $V_i^* = \max_{m_i \in M} V_i^s(m_i)$, the first term in (25) is bounded above by $\gamma_i(M_i^-)V_i^*$.

For the second term, we have $\gamma_i(m_i) = d_i(m_i) + g_i(m_i)$ and

$$\gamma_i(m_i) V_i^w(m_i) \leq d_i(m_i) V_i^w(\emptyset, m_i) + g_i(m_i) R,$$

where we used the fact that i's expected gross utility conditional on working, finding a signal, and reporting m_i is bounded by R.

Combining these observations yields

$$V_i^w \leq \gamma_i(M_i^-) V_i^* + g_i(M_i^+) R + \sum_{m_i \in M_i^+} d_i(m_i) V_i^w(\emptyset, m_i). \tag{26}$$

If $\lambda < 1$, Lemma 2 implies that $V_i^w(\emptyset, m_i) \leq V_i^* + \frac{f_i^0 R}{1-\lambda}$. Summing over all $m_i \in M_i^+$, we get

$$\sum_{m_i \in M_i^+} d_i(m_i) V_i^w(\emptyset, m_i) \leq d_i(M_i^+) V_i^* + d_i(M_i^+) \frac{f_i^0 R}{1-\lambda}. \tag{27}$$

Since $\gamma_i(M_i^-) + d_i(M_i^+) \leq \gamma_i(M_i^-) + \gamma_i(M_i^+) = 1$, combining (26) and (27) proves the lemma when $\lambda < 1$.

If $\lambda = 1$, using in (26) the fact that $V_i^w(\emptyset, m_i)$ is bounded above by R directly proves the lemma.

C.4 Proof of Proposition 2

The right-hand side of (3) is increasing in C over the range of C that satisfy the condition $F_i^k - CF_i^{k+1} > 0$. Therefore, if (3) is satisfied for any C such that $F_i^k - CF_i^{k+1} > 0$, it is also satisfied for any $C' \geq C$ such that $F_i^k - C'F_i^{k+1} > 0$. The proposition thus follows if we show the inequality for $C(\lambda)$.

First, we show that the claim holds if $\beta_i = 0$. In this case, i must shirk with probability 1: if not, Lemma 1 implies that S_i is nonempty with positive probability and, hence, that $\beta_i > 0$. Since i shirks with probability 1, there is no belief update between rounds i and $i+1$. Therefore, $F_i^k = F_{i+1}^k(m_i)$ for any message m_i that i sends in equilibrium. The right-hand side of (3) is thus equal to zero and (3) is satisfied.

Now suppose that $\beta_i > 0$ or, equivalently, that $\gamma_i > 0$: i works with positive probability. We consider two cases, distinguished by the magnitude of the probability $g_i(M_i^+)$ that i finds a signal and sends a message in M_i^+ conditional on working and on history m_1^{i-1}.

Case 1: $g_i(M_i^+) \geq \frac{c}{2R}$**.** By definition, we have

$$\beta_i = \gamma_i \lambda (1 - f_i^0),$$

which implies that $\beta_i \leq \gamma_i$, and

$$\beta_i(M_i^+) = \gamma_i g_i(M_i^+)$$

where $\beta_i(\tilde{M}_i)$ is the probability that i discovers a signal and sends a message in $\tilde{M}_i$. Since $g_i(M_i^+) \geq c/2R$, this implies that

$$\beta_i \leq \gamma_i \leq \frac{2R}{c} \beta_i(M_i^+). \tag{28}$$

Therefore, the desired inequality (3) will follow for $C = 2R/c$ if we prove that $\beta_i(M_i^+)$ is bounded above by $\frac{\mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i)) 1_{m_i \in M_i^+}\right]}{F_i^k - CF_i^{k+1}}$.

For each m_i that i may send in equilibrium and $k \geq 1$, Bayesian updating implies that

$$F_{i+1}^k(m_i) = \frac{F_i^k \left(\alpha_i(m_i) + \gamma_i(1-\lambda)\delta_i(m_i)\right) + \Phi(m_i)}{\alpha_i(m_i) + \gamma_i((1 - F_i^1) + F_i^1(1-\lambda))\delta_i(m_i) + \beta_i(m_i)} \tag{29}$$

where the following probabilities are defined conditional on m_1^{i-1}:

- $\alpha_i(\tilde{M}_i)$: probability that i shirks and sends a message in $\tilde{M}_i$;

- $\delta_i(\tilde{M}_i)$: probability that i sends a message in $\tilde{M}_i$ conditional on working *and* finding no signal;[53]
- $\Phi(m_i)$ is the probability that (i) i works, (ii) i discovers a signal, (iii) i sends report m_i, and (iv) there remain at least k signals at the beginning of round $i+1$.

Let $p_i(m_i)$ denote the probability that i produces report m_i conditional on m_1^{i-1}: $p_i(m_i)$ is the denominator of (29). Rearranging (29) and simplifying, we have

$$F_i^k(\beta_i(m_i) + \gamma_i(1-F_i^1)\lambda\delta_i(m_i)) = (F_i^k - F_i^{k+1}(m_i))p_i(m_i) + \Phi(m_i) \tag{30}$$

Since $\gamma_i(1-F_i^1)\lambda\delta_i(m_i) \geq 0$, summing the previous equation over $m_i \in M_i^+$ yields

$$F_i^k\beta_i(M_i^+) \leq \mathbb{E}_i\left[\left(F_i^k - F_{i+1}^k(m_i)\right)\mathbb{1}_{m_i\in M_i^+}\right] + \sum_{m_i\in M_i^+}\Phi(m_i). \tag{31}$$

Since $\Phi(m_i) = \mathbb{E}_i\left[\mathbb{1}_{|S_i|\geq k+1}\ \mathbb{1}_{\text{i works, discovers a signal, and reports } m_i}\right]$, we have

$$\begin{aligned}
\sum_{m_i\in M_i^+}\Phi(m_i) &\leq \sum_{m_i\in M_i^+}\mathbb{E}_i\left[\mathbb{1}_{|S_i|\geq k+1}\ \mathbb{1}_{\text{i works and reports } m_i}\right] \\
&= \mathbb{E}_i\left[\mathbb{1}_{|S_i|\geq k+1}\ \mathbb{1}_{\text{i works and reports } m_i\in M_i^+}\right] \\
&\leq \mathbb{E}_i\left[\mathbb{1}_{|S_i|\geq k+1}\ \mathbb{1}_{\text{i works}}\right] \\
&= F_i^{k+1}\gamma_i,
\end{aligned} \tag{32}$$

noting, for the last equality, that the event that i works, which has probability γ_i, depends only on m_1^{i-1} and is thus independent of the event $\{|S_i| \geq k+1\}$ conditional on m_1^{i-1}.

Combining this with (31) yields

$$F_i^k\beta(M_i^+) \leq \mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i))1_{m_i\in M_i^+}\right] + F_i^{k+1}\gamma_i. \tag{33}$$

Since $g_i(M_i^+) \geq c/2R$, we have $\beta(M_i^+) = \gamma_i g_i(M_i^+) \geq \gamma_i c/2R$. Inequality (33) then yields

$$\beta(M_i^+) \leq \frac{1}{F_i^k - 2R/cF_i^{k+1}}\mathbb{E}_i[(F_i^k - F_{i+1}^k(m_i))1_{m_i\in M_i^+}]. \tag{34}$$

Combining this with (28) yields (3) for $C(1) = 2R/c$.[54] Since $C(\lambda) \geq C(1)$ for all $\lambda \in (0,1]$, the monotonicity noted at the beginning of the proof yields the desired conclusion for $C(\lambda)$.

[53]Note that $\delta_i(\tilde{M}_i) \geq d_i(\tilde{M}_i)$, where $d_i(\tilde{M}_i)$ was defined before Lemma 3.

[54]Note that the proposition's assumption that $F_i^k - C(\lambda)F_i^{k+1} > 0$ implies that $F_i^k - \frac{2R}{c}F_i^{k+1} > 0$ since $C(\lambda) \geq C(1)$ regardless of λ.

Case 2: $g_i(M_i^+) < \frac{c}{2R}$**.** We prove that γ_i is bounded above by the right-hand side of (3) for $C = C(\lambda)$. Since $\gamma_i \geq \beta_i$, this will yield the desired conclusion.

Intuitively, in Case 2 the probability of discovering a signal that, together with i's equilibrium message strategy, triggers subsequent work is too low to incentivize i to work. The only way of incentivizing i to work is therefore for him to signal by his message that he found nothing through work. For this to happen, the probability f_i^0 that there remains no evidence must be high enough. We will use this fact to obtain a bound on γ_i.

From Lemma 3, if $g_i(M_i^+) < c/2R$, i's utility from working is bounded above by

$$U_i^w = V_i^w - c \leq V_i^* + d_i(M_i^+)\frac{f_i^0 R}{1-\lambda} - \frac{c}{2}$$

if $\lambda < 1$, and by

$$U_i^w \leq V_i^* + d_i(M_i^+)R - \frac{c}{2}$$

if $\lambda = 1$. Therefore, working is optimal only if $d_i(M_i^+)f_i^0 \geq c(1-\lambda)/2R$ when $\lambda < 1$ and only if $d_i(M_i^+) \geq c/2R$ when $\lambda = 1$.

Summing (30) over M_i^+ and using (32) and $f_i^0 = 1 - F_i^1$ yields

$$F_i^k \beta_i(M_i^+) + \gamma_i F_i^k \delta_i(M_i^+) f_i^0 \lambda \leq \mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+}\right] + F_i^{k+1}\gamma_i. \tag{35}$$

For $\lambda < 1$, we have $d_i(M_i^+)f_i^0 \geq c(1-\lambda)/2R$. Since $\delta_i(m_i) \geq d_i(m_i)$ for all m_i (by definition of these variables) and $F_i^k \beta_i(M_i^+) \geq 0$, (35) implies that

$$\gamma_i \leq \frac{\mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+}\right]}{F_i^k c\lambda(1-\lambda)/2R - F_i^{k+1}}.$$

Multiplying the numerator and denominator by $C(\lambda)$ yields the result.

For $\lambda = 1$, we have $d_i(m_i) = \delta_i(m_i)f_i^0$ for all m_i and, hence, $\delta_i(M_i^+)f_i^0\lambda = d_i(M_i^+)$, which is greater than $c/2R$ as noted earlier. Therefore, (35) implies that

$$\gamma_i \leq \frac{\mathbb{E}_i\left[(F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+}\right]}{F_i^k(c/2R) - F_i^{k+1}}.$$

Multiplying the numerator and the denominator by $C(1)$ yields the result. ■

C.5 Proof of Lemma 5

(i) Let $S_i' = \{s_i : F_i^{k,w}(s_i) > F_i^k\}$, and, for each s_i, let $\gamma_i'(s_i)$ (resp. $\beta_i'(s_i)$) denote the probability that i discovers s_i given that he works (resp., the probability that i works and discovers s_i). Also let $\gamma_i'(s_i|k+1)$ (resp. $\beta_i'(s_i|k+1)$) denote the same probabilities conditional on $|S_i| \geq k+1$.

We have for all s_i

$$F_i^{k,w}(s_i) = \frac{F_i^{k+1}\gamma_i'(s_i|k+1)}{\gamma_i'(s_i)} = \frac{F_i^{k+1}\beta_i'(s_i|k+1)}{\beta_i'(s_i)}$$

where the second equality comes from $\beta_i'(s_i) = \gamma_i\gamma_i'(s_i)$ and $\beta_i'(s_i|k+1) = \gamma_i\gamma_i'(s_i|k+1)$. Therefore, $F_i^{k,w}(s_i) > F_i^k$ only if $\beta_i'(s_i) < \beta_i'(s_i|k+1)F_i^{k+1}/F_i^k$.

We have $\sum_{s_i \in S_i'} F_i^{k+1}\beta_i'(s_i|k+1) \leq \gamma_i F_i^{k+1}$. Therefore, the probability $\beta_i'(S_i')$ that i works and finds a signal in S_i' satisfies

$$\beta_i'(S_i') = \sum_{s_i \in S_i'} \beta_i'(s_i) \leq \gamma_i \frac{F_i^{k+1}}{F_i^k}.$$

Since $\beta_i'(S_i') = \gamma_i\gamma_i'(S_i')$, we have

$$\gamma_i'(S_i') \leq \frac{F_i^{k+1}}{F_i^k} \leq \frac{\varepsilon}{\hat{F}^k},$$

since $F_i^{k+1} \leq \varepsilon$ and $F_i^k \geq \hat{F}^k$. From (4), the right-hand side is bounded above by $\frac{\hat{F}^k}{2G^2}$.

For any m_i, let $q(m_i)$ denote the probability, conditional on i working and sending report m_i, that i has discovered a signal $s_i \in S_i'$, and let $\sigma(s_i|m_i)$ denote the probability that i discovered s_i given that he worked and reported m_i. We also let $s_i = \emptyset$ denote the event that i did not find anything, $\sigma(\emptyset|m_i)$ denote the probability that i found nothing given that he worked and reported m_i, $F_i^{k,w}(\emptyset)$ denote the probability that there at least k signals conditional on i working and finding nothing. We have

$$\begin{aligned} F_i^{k,r}(m_i) &= \sum_{s_i \in S_i} \sigma(s_i|m_i)F_i^{k,w}(s_i) \\ &= \sum_{s_i \in S_i'} \sigma(s_i|m_i)F_i^{k,w}(s_i) + \sigma(\emptyset|m_i)F_i^{k,w}(\emptyset) + \sum_{s_i \neq \emptyset, s_i \in S_i \setminus S_i'} \sigma(s_i|m_i)F_i^{k,w}(s_i) \end{aligned}$$

By construction, $F_i^{k,w}(s_i) \leq F_i^k$ for all s_i in the last term. Moreover, we have $F_i^{k,w}(\emptyset) \leq F_i^k$, as is easily checked:[55] intuitively, finding nothing always increases the probability that there

[55] Formally, for $k \geq 1$, we have $F_i^{k,w}(\emptyset) = \frac{F_i^k(1-\lambda)}{(1-F_i^1)+F_i^1(1-\lambda)} = F_i^k \frac{1-\lambda}{(1-F_i^1)+F_i^1(1-\lambda)} \leq F_i^k$.

are no signals remaining to be found. Finally, the first term is bounded above by $\sigma(S_i'|m_i) = q(m_i)$. Therefore,

$$\begin{aligned} F_i^{k,r}(m_i) \geq (1+\eta)F_i^k &\Rightarrow q(m_i) + (1-q(m_i))F_i^k \geq (1+\eta)F_i^k \\ &\Rightarrow q(m_i) \geq \eta F_i^k. \end{aligned}$$

To conclude, note that

$$\sum_{m_i} \gamma_i(m_i)q(m_i) = \Pr(s_i \in S_i'|m_1^{i-1}, i \text{ works}) \leq \frac{\hat{F}^k}{2G^2}$$

The left-hand side is bounded below by $\gamma(N_i)\eta F_i^k$. Since $F_i^k \geq \hat{F}^k$, this implies that

$$\gamma(N_i) \leq \frac{1}{2\eta G^2}.$$

(ii) $F_{i+1}^k(m_i)$ is a convex combination[56] of F_i^k and $F_i^{k,r}(m_i)$. This implies that $F_{i+1}^k(m_i) \leq (1+\eta)F_i^k$ for all $m_i \notin N_i$. From (11), this further implies that

$$\pi_{i+1}(m_i) \leq \frac{2Cg}{\hat{F}^k}\left((1+\eta)F_i^k - \mathbb{E}_{i+1}[F_J^k]\right)$$

for all $m_i \notin N_i$.

C.6 Proof of Lemma 6

(i) If $m_i \notin T_i$, we have $F_{i+1}^{k+1} \leq \sqrt{G}\varepsilon$. Using this inequality in Lemma 4 instead of $F_i^{k+1} \leq \varepsilon$ and repeating its argument applied to round $i+1$, we conclude that $i+1$ assigns probability at least $1 - 1/\sqrt{G}$ to $\mathcal{A}$ whenever $m_i \notin T_i$.

(ii) Let $F_i^{k+1,r}(m_i)$ denote the probability that there are at least $k+1$ signals left at the beginning of round $i+1$ conditional on i working and reporting m_i. F_{i+1}^{k+1} is a convex combination of F_i^{k+1} and $F_i^{k+1,r}(m_i)$. This, together with the fact that $F_i^{k+1} \leq \varepsilon$ and the

[56] $F_{i+1}^k(m_i)$ is the probability that $i+1$ assigns to there being at least k signals left upon observing m_i. If $i+1$ knew that i didn't work and simply sent message m_i, this belief should be F_i^k since m_i conveys no additional information. And if $i+1$ knew that i produced m_i through working and then reporting m_i, his updated belief should be $F_i^{k,r}(m_i)$. Since $i+1$ doesn't observe i's action, in general $F_{i+1}^k(m_i)$ is a convex combination of these two posteriors, where the weights corresponds to the probability assigned by $i+1$ to i fabricating or working conditional on observing m_i. This fact is straightforward to check using Bayesian updating.

definition of T_i, shows that $m_i \in T_i$ only if $F_i^{k+1,r}(m_i) \geq \sqrt{G}\varepsilon$. Let T_i' denote the set of messages m_i for which the last inequality holds. As noted, $T_i \subset T_i'$.

Since i's prior probability that there are at least $k+1$ signals left is $F_i^{k+1} \leq \varepsilon$, the law of iterated expectations implies that the probability $\bar{F}_i^{k+1,r}(m_i)$ that there were at least $k+1$ signals left at the beginning of round i conditional on i working and finding m_i must satisfy

$$\mathbb{E}_i[\bar{F}^{k+1,r}|\text{working}] = \sum_{i \in M_i} \gamma_i(m_i)\bar{F}_i^{k+1,r}(m_i) = F_i^{k+1} \leq \varepsilon.$$

Using Markov's inequality, this implies that $\Pr(m_i : \bar{F}_i^{k+1,r}(m_i) \geq \sqrt{G}\varepsilon|\ i \text{ works}) \leq \frac{\varepsilon}{\sqrt{G}\varepsilon} = 1/\sqrt{G}$. Since also $\bar{F}_i^{k+1,r}(m_i) \geq F_i^{k+1,r}(m_i)$, we get $\gamma_i(T_i') \leq 1/\sqrt{G}$. Since $T_i \subset T_i'$, this shows that $\gamma_i(T_i) \leq 1/\sqrt{G}$.

C.7 Proof of Theorem 2: General case

The argument of Section B extends easily when ρ and/or λ are less than 1.

With $\rho < 1$ and $\lambda = 1$, the informative equilibrium is identical to the one described in Proposition 1 except that learning stops as soon as an agent fails to report evidence, in which case he gets a zero compensation. By construction of the equilibrium in the proof above, shirking and reporting that no evidence was found has the same value as fabricating any other message and can thus be deterred. Since a working agent may find nothing, or the learning process may be interrupted before the belief process exits $(\underline{p}, \bar{p})$, in which case the working agent receives 0, the rewards and punishments must be scaled up by $1/\pi_\rho(q^k)$, where $\pi_\rho(q^k)$ is the probability that the belief process exits $(\underline{p}, \bar{p})$ in equilibrium, given the current belief q^k, so that the expected compensation of a working agent still exceeds the cost c of working.

If $\lambda < 1$ and $\rho = 1$, a working agent may fail to find evidence even when there surely exists some. In this case, we assume once more that the compensation is zero, which deters shirking and reporting the empty message, and scale up all rewards and punishments by $1/\lambda$ to incentive the agent to work, as in the previous paragraph. The belief process will surely exit $(\underline{p}, \bar{p})$ since the amount of evidence is unlimited (only individual agents may be unlucky and find nothing with probability $1-\lambda$).

The case in which both λ and ρ are less than 1 is a convex combination of the previous cases and addressed accordingly.

D Proof of Theorem 3

Without loss of generality, we assume once more that agents' gross utility functions $\{V_i\}_{i\in\mathbb{N}}$ all take values in $[0, R]$.

For any round i, consider the event $\mathcal{Q}_i$ that all past investigators who may have failed to discover signals, given their strategy and reporting history m_1^{i-1}, have indeed failed to discover signals. Conditional on $\mathcal{Q}_i$, the number q_i of signals that have been uncovered until round i is equal to the number of past witnesses plus the number of past investigators whose messages reveal that they have surely discovered signals given m_1^{i-1}. Put differently, q_i is the number of signals that have been surely discovered by round i. Let $\hat{F}_i^k$ denote the probability that $|S_i| \geq k$ conditional on $\mathcal{Q}_i$ and q_i.

We observe that (i) q_i is nondecreasing along any equilibrium path and is strictly increasing whenever a witness arrives, (ii) given the construction of S, the distribution of S_i conditional on m_1^{i-1} and $\mathcal{Q}_i$ is only a function of q_i, and (iii) $\hat{F}_1^k = F_1^k$ for all k.

We will prove that there exist strictly positive thresholds $\{\underline{F}^k\}_{k\geq 1}$ such that an informative continuation equilibrium exists in round i only if $\hat{F}_i^k \geq \underline{F}^k$ for all $k \geq 1$. Applied to $i = 1$, this result implies Theorem 3. The proof uses the following lemma.

LEMMA 7 *Under Assumption 1, the following inequalities hold: (i) $F_i^k \leq \hat{F}_i^k$ for all $i, k \geq 1$ and (ii) Path by path, $\hat{F}_j^k \leq \hat{F}_i^k$ for all $j \geq i$ and $k \geq 1$.*

Proof. Part (i): Let r_i denote the number of signals discovered by round i. We have $r_i \geq q_i$ and

$$\begin{aligned} F_i^k &= \sum_{r\in\{q_i,\ldots,i-1\}} \Pr(r_i = r \mid m_1^{i-1}) \Pr(\tilde{K} \geq r + k \mid \tilde{K} \geq r) \\ &\leq \sum_{r\in\{q_i,\ldots,i-1\}} \Pr(r_i = r \mid m_1^{i-1}) \Pr(\tilde{K} \geq q_i + k \mid \tilde{K} \geq q_i) \\ &= \sum_{r\in\{q_i,\ldots,i-1\}} \Pr(r_i = r \mid m_1^{i-1}) \hat{F}_i^k \\ &= \hat{F}_i^k. \end{aligned}$$

The first equality comes from the independence of $\tilde{K}$ from S^∞: r_i is a sufficient statistic for $\tilde{K}$ given all the information produced before round i, and only to the extent that it reveals that $\tilde{K} \geq r_i$. The inequality comes from the increasing hazard rate condition, which implies

that for any $k \geq 0$, $\Pr(\tilde{K} \geq k+q \mid \tilde{K} \geq q)$ is non-increasing in q.[57] The second equality is due to the equality $\hat{F}_i^k = \Pr(\tilde{K} \geq k+q_i \mid \tilde{K} \geq q_i)$, which again comes from the independence of $\tilde{K}$ and S^∞: the only relevant information about $\tilde{K}$ conditional on $\mathcal{Q}_i$ is the number of signals q_i discovered by round i.

Part (ii) For any $j \geq i$, we have $q_j \geq q_i$. As explained in the proof of Part (i), we have for all $k \geq 1$

$$\begin{aligned}\hat{F}_j^k &= \Pr(\tilde{K} \geq k+q_j \mid \tilde{K} \geq q_j) \\ &\leq \Pr(\tilde{K} \geq k+q_i \mid \tilde{K} \geq q_i) \\ &= \hat{F}_i^k,\end{aligned}$$

where the inequality comes from the increasing hazard rate property (see Footnote 57). ■

The existence of thresholds $\{\underline{F}^k\}_{k\geq 1}$ in Theorem 3 is proved by induction on k. We start with the base case $k=1$ and then prove the induction step.

D.1 Proof for $k = 1$

Suppose that $\hat{F}_i^1 < c/2R$. Combining the two parts of Lemma 7, this implies that $F_j^1 < c/2R$ for all $j \geq i$. Lemma 1 still applies: no investigator $j \geq i$ works because the probability that he finds something is too small to justify the cost of effort, given the maximal reward R.

We now show that if $\hat{F}_i^1$ lies below another threshold, smaller than $c/2R$, witnesses provide no informative message, either.

If i is a witness, we will use the following notation:

- β_i: probability that i produces an informative message given m_1^{i-1};
- M_i^+: set of messages m_i that are followed by an informative continuation equilibrium;
- $\Pr_i(M_i+)$: probability that i produces a message in M_i^+ given m_1^{i-1};
- $\gamma_i(m_i)$: probability that i sends m_i given m_1^{i-1};

[57] See, e.g., Barlow et al. (1963, p. 379). In brief, a random variable X with distribution F has the increasing hazard rate property if and only if the survival distribution $\bar{F} = 1 - F$ is log-concave. This property implies, as is easily checked, that for any $p, q \geq 0$, $\Pr(X \geq p+q \mid X \geq p) = \bar{F}(p+q)/\bar{F}(p)$ is decreasing in p.

- $\gamma_i(m_i|s_i)$: probability that i sends m_i after observing s_i.

LEMMA 8 *Consider $L \geq 2$ pairwise independent random variables $\{Y_\ell\}$ with non-atomic distributions over $\mathbb{R}$ and densities f_ℓ that are bounded above by $\bar{f}$. For any $\varepsilon \geq 0$, let E^L_ε denote the event that $\exists \ell, \ell' \leq L$ such that $|Y_\ell - Y_{\ell'}| \leq \varepsilon$. Then*

$$\Pr(E^L_\varepsilon) \leq L(L-1)\bar{f}\varepsilon.$$

Proof. The result is proved by induction on $L \geq 2$. For $L = 2$, we have

$$\Pr(|Y - Y'| \leq \varepsilon) = \int_{\mathbb{R}} f_Y(x) F_{Y'}[x-\varepsilon, x+\varepsilon] dx \leq 2\varepsilon\bar{f} \int_{\mathbb{R}} f_Y(x)dx = 2\varepsilon\bar{f}.$$

Now suppose that the claim holds for $L-1$. Notice that the event E^L_ε is the union of L events: the event E^{L-1}_ε concerning the first $L-1$ random variables, and, for each $\ell \leq L-1$, the event $E^{\ell,L}$ that the L^{th} random variable lies within ε of the ℓ^{th} random variable. Therefore,

$$\begin{aligned}
\Pr(E^L_\varepsilon) &\leq \Pr(E^{L-1}_\varepsilon) + \sum_{\ell \leq L-1} \Pr(|Y_\ell - Y_L| \leq \varepsilon) \\
&\leq (L-1)(L-2)\bar{f}\varepsilon + (L-1) \times 2\varepsilon\bar{f} \\
&= L(L-1)\bar{f}\varepsilon,
\end{aligned}$$

where the second inequality comes from the induction hypothesis and the fact that $\Pr(|Y_\ell - Y_L| \leq \varepsilon) \leq 2\bar{f}\varepsilon$, as shown in the first step of the induction for $L = 2$. ■

LEMMA 9 *There is a threshold $\underline{F}^1 \in (0, c/2R)$ such if i a witness, then $\beta_i > 0$ only if $\hat{F}^1_i \geq \underline{F}^1$.*

Proof. Recall that if i is a witness, his message m_i is informative (in equilibrium) if it is statistically dependent of S conditional on m_1^{i-1}. Say that i's message is *ϵ-informative* if whenever i has preference ϵ there exist two signals $s_i \neq s_i'$ such that the equilibrium distributions of m_i conditional on i getting signals s_i versus s_i' are different across these two signals.

The following observation is straightforward to prove.

OBSERVATION 1 *i's message is informative if and only if the set of preference shocks ϵ for which i's message is ϵ-informative has positive probability.*

For any equilibrium and history up to some round i in which i is a witness, let ν_i denote the probability that i's preference shock ϵ_i is such that i's message is ϵ_i-informative.

For any $F < c/2R$, let $\nu(F)$ denote the supremum of ν_i over all witness rounds i of all equilibria such that $\hat{F}_i^1 \leq F$. We will show that $\nu(F) = 0$ for all F below some strictly positive threshold.

Consider such an equilibrium. For any witness round i and message m_i, let $z(m_i)$ denote the probability that at least some $j > i$ produces an informative message following message m_i.

i's expected utility if he receives signal s_i and sends message m_i is given by

$$U_i(m_i; s_i) = z(m_i)\mathbb{E}_i[V_i(m) \mid s_i, m_1^i] + (1 - z(m_i))\underline{V}_i(m_i) + \epsilon_i(m_i) \tag{36}$$

where

$$\underline{V}_i(m_i) = \mathbb{E}_i[V_i(m) \mid m_1^i, \text{no } j > i \text{ produces an informative message}].$$

Notice that $\underline{V}_i(m_i)$ does not depend on the signal s_i since this signal is payoff irrelevant whenever no $j > i$ produces an informative message.

Given ϵ_i, i sends an informative signal only if there exist $m_i \neq m_i'$ and signals $s_i \neq s_i'$ such that

$$U_i(m_i; s_i) \geq U_i(m_i'; s_i)$$

and

$$U_i(m_i'; s_i') \geq U_i(m_i; s_i')$$

From (36) and the fact that $V_i(m) \in [0, R]$, this is possible only if:

$$|\epsilon_i(m_i) + \underline{V}_i(m_i) - \epsilon_i(m_i') + \underline{V}_i(m_i')| \leq R(z(m_i) + z(m_i')).$$

The random variables $Y_\ell = \epsilon_i(m_\ell) + \underline{V}_i(m_\ell)$ satisfy the assumptions of Lemma 8. Letting $|M|$ denote the cardinality of the message space M, we thus have

$$\Pr(i \text{ sends an informative message}) \leq |M|^2 \bar{f} R(z(m_i) + z(m_i')) \tag{37}$$

Since no investigator $j \geq i$ works when $\hat{F}_i^1 < c/2R$, we have for any message m_i:

$$z(m_i) \leq \sum_{j \geq 1} \Pr(\text{there are } j \text{ witnesses in the sequence after round } i, \text{ given message } m_i) j\nu(F). \tag{38}$$

Indeed, by definition of $\nu(F)$ and the fact that $\hat{F}_j^1 \leq F$ for all $j \geq i$, a witness provides an informative signal with probability at most $\nu(F)$. Thus $j\nu(F)$ is an upper bound on the probability that at least one witness provides an informative signal given there are j such witnesses.

The probability that at least j witnesses come after round i is bounded above by F^j, where j is an exponent (not a superscript): To show this for $j = 1$, notice that by Lemma 7, $\hat{F}_{i+1}^1 \leq F$, and the probability that there is at least one witness after round i is bounded above by the probability F_{i+1}^1 that there is at least one more signal, which is less than $\hat{F}_{i+1}^1$ again by Lemma 7. For $j = 2$, note that conditional on the first witness arriving, Lemma 7 implies that the probability that a second witness arrives is again by bounded by F since the probability that there remains another signal is bounded by F, and a witness can arise only if such a signal exists. By induction, this shows that the probability of having at least j witnesses and, hence, the probability of having exactly j witnesses, are bounded above by F^j. Combining this with (38) and using the standard formula $\sum_{j\geq 1} jx^j = x/(1-x)^2$ for all $x \in (0,1)$, we get

$$z(m_i) \leq \sum_{j\geq 1} F^j j\nu(F) = \frac{F\nu(F)}{(1-F)^2},$$

Combining this with (37), we obtain

$$\Pr(i \text{ sends an informative message} \mid \hat{F}_i^1 \leq F) \leq 2|M|^2 \bar{f} R \nu(F) \frac{F}{(1-F)^2}. \tag{39}$$

Taking the supremum of the left-hand side over all witness rounds i and equilibria such that $\hat{F}_i^1 \leq F$, we obtain

$$\nu(F) \leq \frac{2\nu(F)|M|^2 \bar{f} R F}{(1-F)^2}.$$

For $2F/(1-F)^2 \leq 1/|M|^2 \bar{f} R$, this relation is possible only if $\nu(F) = 0$, because the function on the right-hand side is a contraction of $\nu(F)$. The function $F/(1-F)^2$ is increasing on $[0,1)$ and starts at zero. Therefore, we conclude that there exists a threshold $\underline{F}^1 > 0$ such that $\nu(F) = 0$ for all $F \leq \underline{F}^1$. ■

D.2 Induction Step

Suppose that there exist strictly positive thresholds $\{\underline{F}^{k'}\}_{k' \in \{1,\ldots,k\}}$ such that a continuation equilibrium starting at round i is informative only if $\hat{F}_i^{k'} \geq \underline{F}^{k'}$ for all $k' \leq k$. We will

show that a similar condition holds for $k+1$. The proof works by contradiction: we will suppose that for all $\varepsilon \in (0,1)$, there exists an informative continuation equilibrium such that $\hat{F}_i^{k+1} \leq \varepsilon$ and obtain an impossibility for ε small enough.

Consider any $\varepsilon < \underline{F}^k \times \underline{F}^1$ and any informative continuation equilibrium starting at round i such that $\hat{F}_i^{k+1} \leq \varepsilon$. Then, all continuation equilibria are uninformative as soon as some witness $j \geq i$ arrives. To see this, suppose that some witness arrives at round $j \geq i$. We have for any message m_j sent by this witness:

$$\begin{aligned}
\hat{F}_{j+1}^k(m_j) &= \Pr(\tilde{K} \geq k + q_j + 1 \mid \tilde{K} \geq q_j + 1) \\
&= \frac{\Pr(\tilde{K} \geq k + q_j + 1)}{\Pr(\tilde{K} \geq q_j + 1)} \\
&\leq \frac{\Pr(\tilde{K} \geq k + q_i + 1)}{\Pr(\tilde{K} \geq q_i + 1)} \\
&= \frac{\Pr(\tilde{K} \geq k + q_i + 1)}{\Pr(\tilde{K} \geq q_i)} \times \frac{\Pr(\tilde{K} \geq q_i)}{\Pr(\tilde{K} \geq q_i + 1)} \\
&= \frac{\hat{F}_i^{k+1}}{\hat{F}_i^1}
\end{aligned}$$

where the inequality comes for the monotone hazard rate assumption (see Footnote 57) and the fact that $q_j \geq q_i$. Since the continuation equilibrium from round i is informative, we must have $\hat{F}_i^1 \geq \underline{F}^1$. Therefore, $\hat{F}_i^{k+1} \leq \varepsilon < \underline{F}^k \underline{F}^1$, which implies that

$$\hat{F}_{j+1}^k < \underline{F}^k$$

and shows by induction hypothesis that all continuation equilibria are uninformative from round $j+1$ onwards.

Moreover, the first witness, j, knowing that continuation equilibria are uninformative regardless of his message, has no incentive to send an informative message.[58]

Therefore, if $\hat{F}_i^{k+1} \leq \varepsilon$, the only agents who may send informative messages are the investigators arriving between round i and the arrival of the first witness. The situation is therefore almost identical to the setting of Theorem 1, in the absence of witnesses, except that any sequential learning activity is interrupted at the apparition of the first witness. We know from Theorem 1 that such equilibria can be informative only if F_i^{k+1} exceeds the $(k+1)^{th}$-threshold given by Theorem 1, which we denote here $\tilde{F}^{k+1}$. Since $F_i^{k+1} \leq \hat{F}_i^{k+1}$ by Part (i) of

[58] Formally, the argument is similar to the proof of Lemma 9 except that here $z(m_i) = 0$ regardless of the message.

Lemma 7, we conclude, letting $\underline{F}^{k+1} = \min\{\tilde{F}_i^{k+1}, \underline{F}^k \underline{F}^1\}$, that no informative continuation equilibrium exists in round i if $\hat{F}_i^{k+1} \leq \underline{F}^{k+1}$. This concludes the induction step. ■