Can Society Function Without Ethical Agents?
An Informational Perspective*

Bruno Strulovici
Northwestern University

January 27, 2020

Abstract

Many facts are learned through the intermediation of individuals with special access to information, such as law enforcement officers, officials with a security clearance, or experts with specific knowledge. This paper considers whether societies can learn about such facts when information is cheap to manipulate, produced sequentially, and these individuals are devoid of ethical motive. The answer depends on an “information attrition” condition pertaining to the amount of evidence available which distinguishes, for example, between reproducible scientific evidence and the evidence generated in a crime.

*This project has benefited from numerous conversations and comments, particularly from Alex Frankel, Simone Galperti, Meg Meyer, Francisco Poggi, Doron Ravid, Larry Samuelson, Ron Siegel, Ludvig Sinander, Boli Xu, and Richard Zeckhauser. Early 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, the University of Chicago, Cambridge University INET Economic Theory Conference (2016), Penn State University, Cornell University, Arizona State University, the North American Meeting of the Econometric Society (2017), Stanford University, Duke University, Northwestern University, Bocconi’s workshop on Advances in Information Economics, Kyoto University, Tokyo University (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), and the RI/PE conference at UC San Diego. Part of this research was 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 a 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. 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 not ethical in the sense that their preferences do not directly depend on the fact to be learned. For example, a prosecutor who cares only about convicting a defendant, regardless of his guilt, is non-ethical according to this definition, while a prosecutor who cares about a defendant’s non-ethical

---

1One 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, not unlike 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).
actual guilt may be ethical.²

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 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”, tests 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, 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.

²Caring about the fact to be learned is only a necessary condition for being “ethical.” For instance, a prosecutor who wants to convict innocent defendants and acquit guilty ones can obviously not be called ethical. This observation has no bearing on the formal analysis of this paper, which focuses on truth-independent preferences. The case of mediated learning with ethical agents is discussed in Sections 2.2 and 4.

³Other 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 particle physics and quantum mechanics, signal acquisition is subject to the observer effect, whereby the observation of a phenomenon necessarily affects this phenomenon.
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 pressure from some sport governing instance), it can also do this at no cost. 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 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 feasible only if the agencies in charge of the tests care directly about the truth, i.e., only if they have ethical preferences.\footnote{In practice, the same laboratory is often in charge of storing and testing both samples, which may lead to even more direct incentives problems, as in case of nationally organized doping schemes.} When agencies are unethical, mediated learning is infeasible regardless of the agencies’ material incentives.\footnote{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 monitoring structures.}

1.4 Reproducible Evidence and Incentive Design

Some questions 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 previous unraveling argument no longer applies and, as this paper shows (Theorem 2), mediated learning by unethical 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 information 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.

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. Knowledge based on reproducible evidence can also, given the proper incentives, be protected from learning failures caused by unethical agents.\footnote{The theory can be used to predict a correlation between the establishment of truthfulness oaths and contexts where information suffers from attrition. For instance, it may explain why formal oaths are present in court, but not in physical sciences.}

\section{1.5 Uncertain Attrition}

To assess the scope of the argument illustrated by the blood-testing example, this paper considers an analytical framework based on sequential learning through public reports that generalizes the argument in three important ways. First, the supply of evidence available in a case may be unknown a priori, even to investigators. Second, investigators may not know how past investigators have affected the evidence available in the case. Third, the number of potential investigators may be large and a priori unknown. For example, suppose that a criminal has left behind ten pieces of forensic evidence, which may be discovered through some costly investigative work. Investigators may not know a priori the exact number of pieces left behind. Moreover, if an investigator inherits the case from a previous investigator, he may not know how diligent the previous investigator has been and, hence, what fraction of the evidence has already been discovered, fabricated, or destroyed.

Therefore, each investigator faces both exogenous and endogenous uncertainty concerning the amount of evidence that remains to discover when the investigator is called to action. 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 \ref{thm1} and \ref{thm2} and Theorem \ref{thm3} for the extension to witnesses).

When the necessary conditions are violated, mediated learning by unethical agents is impossible in a strong sense: even when i) there is an unbounded sequence of potential investigators, and ii) one has full discretion to design these investigators’ incentives and...
there is no commitment or agency problem affecting the administration of these incentives, there does not exist any equilibrium in which at least one of the investigators provides an informative report with positive probability.

1.6 Virtual Attrition Paradox. Application to Human Subjects

The necessary conditions are violated—and mediated learning is impossible—whenever information consists of disposable signals whose number has bounded support, no matter how large the bound. For example, suppose that there are initially 1000 signals. Successful learning requires at least one agent to use one of the signals, which can be incentivized only if a subsequent agent uses (with positive probability) a second signal, which can be incentivized only if a third agent uses a third signal with positive probability, and so on. For successful learning to be incentivized, society must therefore, with positive probability, reach a point at which i) the probability that there remain 2 or more signals is arbitrarily low, ii) the probability that some agent makes an informative report is positive. This is essentially the situation problem faced by the two laboratories in the blood-testing example, and we know that these two conditions are incompatible. Mediated learning must therefore 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.

This yields the following paradox: when the supply of disposable signals is bounded, mediated learning fails because agents anticipate that signals must eventually become scarce with positive probability if mediated learning is successful. This anticipation destroys agents’ incentives to seek signals in the first place, implying that the scarcity never actually materializes: it remains virtual and the supply of evidence remains intact even as mediated learning fails entirely.

To appreciate the logical consequences of this paradox, consider a research question that must be investigated through human subjects, such as a question in experimental psychology. The experiment can be run only once on any given subject: exposure to the experiment affects a subject’s awareness, knowledge, and response to future experiments, as can be most vividly intuited from experiments such as Zimbardo’s 1971 Stanford’s prison experiment.

\footnote{In herding models (Bikhchandani, Hirshleifer, and Welch (1992) and Banerjee (1992)), social learning can fail after so much public information has already been acquired that agents forgo their own information. Here, by contrast, learning fails entirely from the first agent onward.}
In these cases, the more experiments are performed, the fewer the number of subjects available for further experiments. Even when the pool of potential subjects is a large, virtual attrition imposes a limited scope for experimental replication and, hence, for incentivizing researchers’ truthful behavior. Remarkably, Zimbardo later admitted to largely influencing the participants of his prison experiment, which largely reduces the experiment’s scientific value.

---

8 Zimbardo (2007, p. 55) admits to directing the guards to handle the prisoners in the following terms: “We can create boredom. We can create a sense of frustration. We can create fear in them, to some degree. We can create a notion of the arbitrariness that governs their lives, which are totally controlled by us, by the system, by you, me . . . They’ll have no privacy at all, there will be constant surveillance—nothing they do will go unobserved. They will have no freedom of action. They will be able to do nothing and say nothing that we don’t permit. We’re going to take away their individuality in various ways . . . In general, what all this should create in them is a sense of powerlessness. We have total power in the situation. They have none.”
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 are purely motivated by unethical motives and are collectively aware of this.

To see how information attrition affects the view 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. 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. In a case of political 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.

9Blood tests could in principle be certified by a third party, who would check whether 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.

10The same observations hold with regard 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 strongest 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.htm). In 2005, the CIA’s Director of Operations at the time ordered the destruction of all interrogation tapes of Abu Zaydah and Abd al-Rahim al-Nashiri that featured “enhanced” interrogations (“Tapes by C.I.A. Lived
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 specific sense given above, they can reject official statements and journalistic reports about the case and maintain strong disagreements.

The theory provides an explanation for why reasonable citizens may remain divided on subjects that are vulnerable to information attrition. The absence of trust in investigative institutions can thus perpetuate polarization. The theory thus provides a specific mechanism to explain why eroding trust in institutions harms social cohesion.

2.2 Ethical Necessity

The arguments 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. 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 institutions but ethically with respect to the rules of the institutions. Such an assumption amounts to a “heroic dichotomy” that, at the very least, deserves closer inspection.

By focusing on mediated learning, this paper provides a tractable framework with a sequential structure to study ethical necessity. Since mediated learning is involved in the resolution of criminal cases, political corruption, and the determination of other possible violation of institutions, successful mediated learning is a necessary condition for society to

---

11 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 work, this paper does not study ethical behavior per se, but the necessity of ethical behavior for society to function. Self-interest remains the default assumption in most economic models and is deeply rooted in the discipline. For example, Edgeworth (1881) observed that “self-interest is the first principle of pure economics.”

12 One may consider alternative designs. For example, a central agency may ask several intermediaries to seek and report the truth simultaneously and independently of one another, although the incentives of such a centralized agency must also be scrutinized, bringing us back to a sequential structure. The parallel monitoring design and other designs are discussed in Section 4.
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. Hurwicz (2007) explicitly discusses the existence of “intervenors,” which he defines as ethical monitors in a monitoring-the-monitor problem. Hurwicz 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. He 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 infinite monitoring hierarchy is folded in loops going through the three agents, which Hurwicz does not formally analyze, omits the possibility of corruption across monitors, studied in Strulovici (2019).

Holmström (1982) shares a similar concern in the conclusion of his analysis of moral hazard in teams. He notes 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?”

Ethical necessity is implicit in Becker and Stigler’s (1974) study of wrongdoing and malfeasance by enforcement officers. The authors assume the existence of exogenous signals about the behavior of enforcement officers. The existence of a reliable second level of monitoring is thus taken for granted, which raises the question of how these signals are generated.

---

13 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 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, this uniquely selects the optimal equilibrium.

14 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.

15 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.

16 Milgrom, North, and Weingast (1990) explicitly study the enforcement of market institutions. They
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. They must rely on experts and officials to access and correctly interpret archives, artifacts, and other sources of information. Moreover, information attrition arises both exogenously (e.g., witnesses die) and endogenously (e.g., documents may be destroyed on purpose).

To give a concrete example, consider the fire of the German parliamentary building (Reichstag) that took place on February 27, 1933. The importance of this event can hardly be overstated. The Nazi party, which had lost seats in the previous parliamentary election, claimed that the fire had been caused by communists and used this excuse to pressure president Hindenburg into imposing martial law on Germany (the “Reichstag fire decree”) and arrest and weaken communists, which allowed the Nazi’s ability to form a majority coalition parliament 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. Historians accepted Tobias’ version of the event until 2001, when two historians studying Gestapo archives came to the conclusion 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 person to consider the role of law merchants to discipline trade in medieval Europe, and they do consider some of the law merchants’ incentives to lie and take bribes.

17This point is most obvious for events that took place before citizens’ lifetime. Even for contemporaneous events, aggregating information and forming a global picture of an event is a highly complex task that requires expertise, time, and access to documentation. This point has 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 but construes a completely erroneous version of the battle.

18The articles were published in Der Spiegel under the title „Stehen Sie auf, van der Lubbe!,” in 1959 and 1960, in issues 43 to 52.
individual to set the fire. In 2019, an affidavit written in 1955 by former SA Hans-Martin Lennings discovered in Tobias’ files was 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. 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 involving 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.

2.4 Trust and Observability of Ethical Agents

The arguments developed so far assume that agents are known to be unethical. 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 unethical, for two reasons: First, the findings

19 Tobias is suspected of having tried to protect former Nazi officers after the war and had some incentive to dissimulate the Nazis’ role in the Reichstag fire.

20 In 1933 another SA, Adolf Rall, confessed to the prosecutors that the SA had set the fire and tried to blame it on the Communists. However, Hall was murdered by the Gestapo in the same year and his confession was suppressed.
of the ethical agent are (perhaps, wrongly) believed to be uninformative. Second, precisely because the 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.

The analysis thus provides a specific mechanism for why eroding the trust in institutions can be particularly 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.\footnote{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, unethical, 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.}

It may be empirically difficult to distinguish between agents who have ethical preferences and agents who merely behave like ethical agents. The belief in ethical behavior can be self-fulfilling.
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 \( \Sigma \). The sequence \( S \) and its length \( \tilde{K} \leq \infty \) are stochastic.

In each round \( i \geq 1 \), a new agent arrives and makes two decisions: First, \( i \) chooses between privately seeking a signal (“working”) at cost \( c > 0 \) and shirking. Second, \( i \) chooses a public message \( m_i \) to send from some finite message space \( M \). \( i \) 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 \( i \) 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 \( s_i \).\(^{22}\) With probability \( 1 - \lambda \), \( i \) discovers no signal. If \( S_i \) is empty, \( i \) surely discovers no signal.

For simplicity we focus on disposable signals and do not explicitly 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 \(^3\)).\(^{24}\) The extension is more complex when some agents, such as witnesses, discover signals for free. This extension is considered 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 previous reports \( m_1^{i-1} = (m_1, \ldots, m_{i-1}) \).

In round \( i + 1 \), \( S_{i+1} = S_i \) if \( i \) discovered no signal, and \( S_{i+1} \) is a subsequence of \( S_i \) with length \( |S_{i+1}| = |S_i| - 1 \) if \( i \) discovers an element of \( S_i \).

Let \( m = (m_1, m_2, \ldots) \) denote the entire stream of reports. The realized utility of agent \( i \)

\(^{22}\)The subscript \( i \) denotes the signal discovered in round \( i \) (if any), which should not be confused with the ordering of signals in the sequence \( S \), which is indexed by a superscript.

\(^{23}\)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 \( \lambda \) of discovering a signal is nondecreasing in the length of \( S_i \).

\(^{24}\)The possibility result Theorem \(^2\) can be interpreted as allowing the exogenous decay of evidence.
is given by

$$U_i = V_i(m, \omega) - c\mathbf{1}_i \text{ works} \quad (1)$$

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

We will say that agent $i$ is unethically if the function $V_i$ is independent of $\omega$, which captures the idea that $i$ does not care directly about the truth. 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 whether $V_i$ is an exogenous utility function, or whether it can be influenced or even designed by a regulator or social planner.

To illustrate the various forms that $V_i$ may take, note that $i$ could be punished if his report is contradicted by subsequent investigators, and/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. 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 $a(m) \in \{\text{‘guilty’, ‘not guilty’}\}$ depends on the statements of all agents involved in the case. If $i$ is unethical, his utility may be modeled by $V_i(m) = R \times \mathbf{1}_{a(m) = \text{‘guilty’}}$, as in Landes (1971). In general, $i$’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 $\tau$, in which case $i$’s utility depends only on $(m_1, \ldots, m_\tau)$. This and similar variations, such as randomizing the order of investigators or how investigator $j$’s report affects $i$’s utility, is easily encompassed by the model.

Agents have a common prior about the distribution of $S$. The equilibrium concept is (weak) Perfect Bayesian Equilibrium.\textsuperscript{26}

\textsuperscript{25}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

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

\textsuperscript{26}The impossibility results in this paper hold for stronger concepts of equilibrium, such as sequential equilibrium. Reciprocally, the equilibrium constructed to prove the positive result, Theorem \textsuperscript{2}, is a sequential equilibrium.
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. Mediated learning is **infeasible** if all equilibria are uninformative.

Uncovering any modicum of information about \( S \) (and, hence, \( \omega \)) 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** Given any parameters \( R, P, \) and \( \lambda \), there exist thresholds \( F^k > 0 \) for \( k \geq 1 \) such that mediated learning is infeasible whenever \( F^k < F^k \) for at least one \( k \geq 1 \).

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 conditions, regardless of the parameters \( P, R, \) and \( \lambda \). Having an unbounded support does not suffice, however: the thresholds imply that the survival function \( F^k = Pr(|S| \geq k) \) cannot decrease too fast for mediated learning to be feasible.

Mediated learning can be incentivized when the number of signals is infinite. Moreover generally, provided that the survival function \( F^k \) decreases at most at a geometric rate, there are instances of the model and utility functions \( \{V_i\}_{i \in N} \) for which mediated learning is feasible.

**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 “$L$”\footnote{The construction used to prove Theorem 2 is almost identical for $\rho \in (0,1)$ and $\rho = 1$ and the results are qualitatively unchanged. The key is that there remain signals to discover with probability at each stage of the sequence with a probability that is bounded below away from 0. Moreover, if the number of initial signals is generated according to a geometric distribution and each signal decays at some fixed rate, the construction used to prove Theorem 2 still works because the probability of discovering new signals remains uniformly bounded away from zero.}. 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 $\Pr(\tilde{s} = “H” | \omega = H) = \Pr(\tilde{s} = “L” | \omega = L) = \pi \in (1/2, 1)$, and the signals are independently distributed conditional on $\omega$.

For expositional simplicity, let us assume that $\lambda = 1$, which means that every agent who works surely discovers a signal\footnote{The construction used to prove Theorem 2 is very similar if instead $\lambda < 1$. Agents’ message space is chosen to be binary: $m_i \in \{“H”, “L”\}$ for all $i$.}.

In the equilibrium that we consider, it is always in an agent’s interest to follow his signal if he acquired one. Therefore, the relevant pure strategies are to i) work at cost $c > 0$ and report one’s signal, i.e., $m_i = s_i$, or ii) shirk and send some message $m_i \in \{“H”, “L”\}$ at zero cost.

Let $p_0 = P(\omega = H)$ denote the prior about $\omega$ before the investigation process. For any given equilibrium, let $\gamma_i$ denote the probability that $i$ works and $p_i$ denote the probability that $\omega = H$, both conditional on the past reports $m_{i}^{i-1}$.

**Proposition 1** For any thresholds $p_-$ and $p_+$ such that $0 < p_- < p_0 < p_+ < 1$, there exist $P, R > 0$, utility functions $\{V_i\}_{i \geq 1}$ taking values in $[-P, R]$, and thresholds $p, \bar{p}$ such that $0 < p < p_-$ and $1 > \bar{p} > p_+$ for which the following strategy profile constitutes an equilibrium: $\gamma_i = 1$ if $p_i \in (p, \bar{p})$ and $\gamma_i = 0$ otherwise.

According to this proposition, it is possible to learn the state of the world $\omega$ with arbitrary precision as long as the rewards and punishments available to design agent’s utility are high enough. In equilibrium, all agents work with probability 1 until the posterior belief becomes extreme enough, at which point agents stop looking for signals.

Since $p_i$ is a martingale and each signal has the same level of informativeness, $p_i$ must exit $[p, \bar{p}]$ 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 \( \bar{p} \) 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

Consider now the addition of witnesses who differ from previous agents in two aspects:

- Witnesses discover a signal for free.
- They are subject to 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 \( |S_{i+1}| = |S_i| - 1 \).

Given past reports \( m_i^{i-1} \), the probability that \( i \) is a witness is \( \varphi_i = 0 \) if \( S_i = \emptyset \) and \( \varphi_i \in [0, 1] \) if \( S_i \neq \emptyset \). Otherwise, \( i \) is an investigator.[29]

In principle, a witness’ signal could be informative about the number of signals in the case, and affect everyone’s belief about the existence of information attrition.[30] This would render moot any systematic analysis based on initial information attrition.

We rule out this possibility and focus on the case in which a witness’ signal never increases.

---

[29] For example, there could be a fixed subset \( N \subset \mathbb{N} \) of rounds such that \( \varphi_i = 1_{i \in N} \) (this requires that \( S \) contains sufficiently many signals). 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 need not be known ex ante and may depend on past reports. We model this in an economical fashion by allowing \( \varphi_i \) to depend arbitrarily on past reports and calendar time.

[30] For example, \(|S|\) could be a finite or infinite with equal probability, and a witness’ signal could reveal that the latter is true, in which case mediated learning may be feasible, as in Theorem 2.
one’s 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 $S^\infty$ 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 $S^\infty$. 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.

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 independently distributed from one another and from all other variables in the model. They have density functions $\{f_{i,m_i}\}_{m_i \in M_i}$ that are bounded above by some constant $\bar{f}$.

The bound $\bar{f}$ plays a role for witnesses that is similar to the ratio $1/c$ for investigators. Intuitively, the opportunity cost for $i$ to send message $m_i$ instead of $m'_i$ (or vice versa) is of order $1/\bar{f}$, as explained in Lemma 8.

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.

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

19
4 Discussion

Learning through intermediaries fails if i) intermediaries do not care about the truth, ii) information is subject to attrition, iii) there is no exogenous, public revelation of the truth, 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 question of interest, and despite the fact that incentives offered to the agents can be administered without any further agency problem: whatever rewards and punishments are promised to the agents can be enforced.

This is not to say that mediated learning fails in practice. The 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 not disposable. Consider some video footage related to a crime or the weapon of a crime. These items can be revisited numerous times without being altered and resemble the case of reproducible evidence.

There are some challenges with this kind of evidence. First, the evidence need not be “hard.” For example, video footage could be fabricated as exemplified by the emergence of baffling deepfakes. In this case, the signals that investigators get is only as reliable as the evidence that they examine in the first place. Second, the evidence could deteriorate over time.

The case of DNA evidence illustrates this. The amount of usable DNA samples on a crime scene is finite, and thus subject to virtual attrition. 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. However, DNA samples can also be synthesized, i.e., literally “fabricated,” to match any desired DNA profile. A laboratory

\[\text{Frumkin, Wasserstrom, Davidson, and Grafit (2010) show the possibility of creating saliva or blood samples with the desired DNA. The authors, as well as subsequent work by other researchers, show that}\]
can store artificial DNA, or even more simply, it can store DNA samples taken outside the crime scene. Ultimately, agents who come later in the investigation can either test the DNA material that previous laboratories have left them, which may be the outcome of manipulation, or look for 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 argument. First, agents could be asked to investigate and report their findings simultaneously, which will be called parallel monitoring. Second, agents could investigate past investigators. Third, in some cases such as criminal cases, agents could be incentivized by the perspective that whoever committed the crime (and, hence, can assess whether the mediated learning findings are accurate) responds differently to the findings. Finally, the same agents could be called to make statements repeatedly over time. These possibilities are examined in turn.

1. Parallel monitoring and weak implementation with a centralized authority

In applications such as the blood-testing example and 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. 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 are match and punished if they don’t.

identifying methylation patterns in DNA samples can help distinguish synthetic and natural DNA, although such identification is challenging.\textsuperscript{32}

\textsuperscript{32}The academic refereeing process has reporting features that resemble this parallel-monitoring design, although the incentives for referees are different and arguably more subtle.
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, 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 resurrests the 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, particularly if laboratories can coordinate their response.

In some cases, parallel monitoring may be difficult to implement: it may be difficult, for instance, to send two investigators on a crime scene, interrogate witness, and collect evidence, without the investigators being able to communicate and possibly coordinate.

Last but not least, successful parallel monitoring can only exist as one of several equilibria, many of which are uninformative. Indeed, even without communication, there are always equilibria in which monitors coordinate on a predetermined report. Among these equilibria, successful parallel monitoring can be non-robust. Even small amounts of strategic uncertainty about other monitors can suffice to destroy the equilibrium.

2. Monitoring the Monitor

In the applications discussed so far, each agent makes a statement about a particular question (e.g., whether an athlete used a banned substance, or who committed a crime). The agents themselves could be investigated. In this case, the subject of study of each agent evolves over time: agent 1 investigates the initial question, agent 2 investigates agent 1’s treatment of the initial question, agent 3 investigates agent 2’s investigation and so on. Such a sequence is called a “monitoring hierarchy.” In his analysis of monitoring, Hurwicz (2007) considers a similar 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 allow an infinite sequence of distinct agents, each of which is tasked with investigating the previous agent in the sequence, another issue emerges: what if an agent who discovers 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

---

This idea is explored in ongoing joint work with Harry Pei.
of corruption among nearby agents. In a separate paper, I show that even this local form of corruption may suffice to unravel agents’ incentives (Strulovici (2019)). Intuitively, the tension comes from two considerations. First, in order for a monitor to accurately report his finding about the previous agent, the monitor’s punishment if he lies and makes a wrongful accusation must be higher than his reward for claiming to have discovered wrongdoing by the previous monitor. Second, a monitor must get a higher reward if he finds evidence against the agent that he was investigating, relative to the punishment that this agent gets when punished. Otherwise, there is a transfer form the guilty agent to the monitor that improves both agents’ payoff. Combined, these observations imply that the chain of rewards is unbounded and, hence, unsustainable.

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 inaccurate findings have on citizens’ future behavior.

In particular, consider a dynamic setting in which a citizen’s decision to commit crime depends on whether this citizen’s past actions were accurately called by past investigators: a citizen who was wrongfully accused of committing crime in the past or who was wrongfully acquitted for crimes that he did commit is more likely to commit crime in the future. In this setting, investigators may have an 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: the public history consists of past official findings and a citizen’s private history consists of what the citizen actually did. In order for private histories to affect play, citizens must therefore be made indifferent between committing crime and abstaining from it, a knife-edge condition which is violated if, for instance, citizens are subject to small private shocks affecting their benefits from committing crime.\(^\text{34}\)

4. Alternating statements

Finally, one could ask a fixed set of agents to take turns repeatedly to investigate and

\(^{34}\)The argument is somewhat similar to the section on witnesses in this paper.
report on the question of interest. The key difference with the earlier analysis is that agents now have private information about what they did in the past, which gives them an advantage to interpret the declarations of other agents. For example, consider two agents taking turns. If an agent has discovered some disposable signals, he knows that the other agent has 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 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 improve 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 to recent versions of Walter Williams’ Journalist’s Creed (Farrar (1998)) also aim to elicit ethical behavior. This paper suggests a positive correlation between the need of oaths in various professions and the severity of information attrition problems in these professions.

Even if a limited fraction of agents is swayed by such oaths, this may in principle suffice for incentivizing truthful behavior by other agents. Studying the precise mechanism through which ethical agents can incentivize mediated learning is beyond the scope of this paper.35

Another approach is to design 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

35One obvious way is that some agent appearing in some round N > 1 is commonly known to be ethical. This agent’s signal is then similar to an exogenous arrival of reliable information and can be used to incentivize all agents coming in rounds i < N.
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. The theory also emphasizes that violations of the rules of capitalism (or of any 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 a economic or political system and the incentives needed to respect the system itself.

A final angle to attack mediated learning failures is to democratize the learning the 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 and such pools are conceivable when the expertise required of intermediaries is limited. The institution of popular juries may be viewed as one such attempt, which trades off intermediaries’ expertise with their representativeness of a more global and diffuse body of stakeholders.

---

36A 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)). Compared to the literature, this paper is concerned with both moral hazard and adverse selection since the object of mediated learning is to elicit effort from investigators as well as truthful revelation of investigators’ findings.


A Proof of Theorem 1

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

A.1 Preliminary Results

Since agents’ decisions are invariant with respect to a uniform translation in their gross utility we assume without loss of generality that their gross utility takes values in some interval $[0, R]$. Moreover, since $R$ is an upper bound on payoffs, it can always be increased to guarantee that $R > c$, which we assume from now on without loss of generality.

Let $\gamma_i$ denote the probability that $i$ works given $m_i^{i-1}$.

Lemma 1 $\gamma_i > 0$ only if $F^1_i \geq \frac{c}{R} > 0$.

Proof. Let $U^s$ denote $i$’s expected utility if he shirks and adopts the message distribution that he would follow in equilibrium if he had worked and found nothing. Let $U^w$ denote $i$’s expected utility if he works. We will show that working is strictly suboptimal if $F^1_i < c/R$, which will prove the lemma.

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

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

Conditional on $S_i = \emptyset$, $i$’s expected gross utility if he works is the same as his expected gross utility $V^{s,\emptyset}_i$ 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, $V^{w,\emptyset}_i = V^{s,\emptyset}_i$. Furthermore, we also have $V_i \leq R$ since $R$ is the maximum possible gross utility.

Therefore, $i$’s net utility $U^w_i$ from working, including the cost of working, satisfies

$$U^w_i \leq F^1_i R + (1 - F^1_i) V^{s,\emptyset}_i - c.$$
Similarly, \( i \)'s utility \( U_i^s \) from shirking satisfies
\[
U_i^s \geq F_i^1 \times 0 + (1 - F_i^1)V_i^{s, \emptyset},
\]
since 0 is a lower bound on \( i \)'s realized gross utility.

From these expressions, shirking strictly dominates working if \( F_i^1 R - c < 0 \). ■

Given any round \( i \) and past reports \( m_i^{i-1} \) such that \( \gamma_i > 0 \), let

- \( V_i^s \) denote \( i \)'s maximal expected gross utility if he shirks, where the maximum is taken over all possible messages \( m_i \in M \).
- \( V_i^{w, \emptyset} \) denote \( i \)'s expected gross utility if he works and finds nothing.
- \( f_i^0 = 1 - F_i^1 \) be the probability that \( S_i = \emptyset \) at the beginning of round \( i \) (i.e., conditional on \( m_i^{i-1} \)).

**Lemma 2** Suppose that \( \lambda < 1 \). If \( i \) works and finds nothing and sends message \( m_i \), his expected gross utility \( V_i^{w, \emptyset}(\emptyset, m_i) \) satisfies
\[
V_i^{w, \emptyset}(\emptyset, m_i) \leq V_i^s + \frac{f_i^0 R}{1 - \lambda}
\]
for all \( m_i \).

**Proof.** For any sequence \( S'' \) of signals, let \( \Delta_i(S'') \) denote the probability that \( S_i = S'' \) conditional on past messages \( m_i^{i-1} \) and \( \Delta_i^{\emptyset}(S'') \) denote the probability that \( S_i = S'' \) conditional on \( i \) working and finding nothing. Bayes' rule 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
\[
\Delta_i^\emptyset(S'') - \Delta_i(S'') = -\frac{\lambda f_i^0 \Delta_i(S'')}{(1 - f_i^0)(1 - \lambda) + f_i^0} \tag{2}
\]
for \( S'' \neq \emptyset \) and
\[
\Delta_i^\emptyset(\emptyset) - \Delta_i(\emptyset) = \frac{\lambda(1 - f_i^0) \Delta_i(\emptyset)}{(1 - f_i^0)(1 - \lambda) + f_i^0} \tag{3}
\]
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 \( m_i^1 \) and \( S_{i+1} \) entirely determine the distribution of reports \( m_j \) for
\( j > i \). Therefore, \( V_i(m_i, S'') \) is the same regardless of whether \( i \) worked or shirked. Agent \( i \)'s expected gross utility conditional on i) working, ii) finding no signal, and iii) producing message \( m_i \), is

\[
V_i^w(\emptyset, m_i) = \sum_{S'' \in S} V_i(m_i, S'') \Delta_1^0(S'')
\]

whereas if he shirks and sends message \( m_i \), his expected gross utility is

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

because \( i \) has learned nothing from shirking and thus holds the same belief as his prior belief at the beginning of round \( i \). Therefore,

\[
V_i^w(\emptyset, m_i) - V_i^s(m_i) = \sum_{S'' \in S} V_i(m_i, S'') (\Delta_i^0(S'') - \Delta_i(S'')).
\]  \( (4) \)

Since \( V_i(m_i, S'') \in [0, R] \) for all \( m_i \) and \( S'' \), combining (4) with (2) and (3) yields

\[
V_i^w(\emptyset, m_i) - V_i^s(m_i) \leq R \Delta_i(\emptyset) \frac{\lambda (1 - f_0^0)}{(1 - f_0^0)(1 - \lambda) + f_0^0}.
\]

Since \( \lambda \leq 1 \) and \( \Delta_i(\emptyset) = f_0^0 \), the denominator is bounded below by \( 1 - \lambda \), while the numerator is bounded above by \( R f_0^0 \). This yields

\[
V_i^w(\emptyset, m_i) \leq V_i^s(m_i) + f_0^0 \frac{R}{1 - \lambda} \leq V_i^s + f_0^0 \frac{R}{1 - \lambda},
\]

which proves the lemma. Intuitively, if \( f_0^0 \) is negligible relative to \( (1 - \lambda) \), \( i \)'s expected gross utility after working and finding nothing is barely higher than if he had simply shirked because finding nothing in this case simply reveals that he was unlucky.  \( \square \)

For each round \( i \) and subset \( \bar{M}_i \subset M \) of messages, we define several probabilities conditional on \( m_{i-1}^i \):

- \( \gamma_i(\bar{M}_i) \): probability that \( i \) produces a report in \( \bar{M}_i \) conditional on working
- \( g_i(\bar{M}_i) \): probability that \( i \) finds a signal and produces a report in \( \bar{M}_i \) conditional on working
- \( d_i(\bar{M}_i) \): probability that \( i \) finds no signal and produces a report in \( \bar{M}_i \) conditional on working.

\( 37 \) This is because \( V_i(m_i, S'') \) is computed conditional on \( S_{i+1} = S'' \). Of course, the distribution of \( S_{i+1} \) generally depends on whether \( i \) works or shirks.
We also let $M_i^+$ denote the set of messages $m_i$ that are followed by an informative continuation equilibrium at round $i + 1$.

**Lemma 3** 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

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

**Proof.** For each $m_i \in M$, let $V^w(m_i)$ denote $i$’s expected gross utility conditional on working and sending message $m_i$. Also recall that $V^s(m_i)$ denotes $i$’s expected gross utility conditional on shirking and sending message $m_i$. Finally, let $M_i^-$ denote the set of messages $m_i$ after which no $j > i$ ever works, so that $M = M_i^+ \cup M_i^-$ and $M_i^+ \cap M_i^- = \emptyset$.

We have

$$V_i^w = \sum_{m_i \in M_i^-} \gamma_i(m_i) V_i^w(m_i) + \sum_{m_i \in M_i^+} \gamma_i(m_i) V_i^w(m_i).$$ (5)

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. 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 (5) 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 $R$.

Combining these observations yields

$$V_i^w \leq \gamma_i(M_i^-) V_i^* + g_i(M_i^+) R + \sum_{m_i \in M_i^+} d_i(m_i) V_i^w(\emptyset, m_i).$$ (6)

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 $m_i \in M_i^+$, we get

$$\sum_{m_i \in M_i^+} d_i(m_i) V_i^w(\emptyset, m_i) \leq d_i(M_i^+) V_i^* + d_i(M_i^+) \frac{f_i^0 R}{1 - \lambda},$$ (7)
Since $\gamma_i(M^-_i) + d_i(M^+_i) \leq \gamma_i(M^-_i) + \gamma_i(M^+_i) = 1$, combining (6) and (7) proves the lemma when $\lambda < 1$.

If $\lambda = 1$, using that $V_i^w(\emptyset, m_i)$ is bounded above by $R$, equation (6) directly yields the lemma.

For each round $i$, we define another set of probabilities conditional on $m_{1:i-1}$:

- $\beta_i$: probability that $i$ discovers a signal (before observing whether $i$ works, i.e., viewed from the beginning of round $i$);
- $\beta_i(\tilde{M}_i)$: probability that $i$ discovers a signal and sends a message in $\tilde{M}_i$;
- $\alpha_i(\tilde{M}_i)$: probability that $i$ shirks and sends a message in $\tilde{M}_i$;
- $\delta_i(\tilde{M}_i)$: probability that $i$ produces a report in $\tilde{M}_i$ conditional on working and finding no signal;
- $F_{i+1}^k(m_i)$: probability that there remain at least $k$ signals at the beginning of round $i+1$ given reports $m_i = (m_{i-1}, m_i)$.

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

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

**Proof.** We start with the following observation: The right-hand side of (8) is increasing in $C$ over the range of $C$ that satisfy the condition $F_i^k - CF_i^{k+1} > 0$. Therefore, if (8) 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 result 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 $\beta_i > 0$. 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 arises in equilibrium. The right-hand side of (8) is thus equal to zero and (8) is satisfied.

---

Note that $\delta_i(\tilde{M}_i) \geq d_i(\tilde{M}_i)$, where $d_i(\tilde{M}_i)$ was defined before Lemma 3.
Now suppose that $\beta_i > 0$ or, equivalently, $\gamma_i > 0$: $i$ works with positive probability. We consider two cases, distinguished by whether working is likely to lead to the discovery of a signal and, under $i$’s equilibrium strategy, to a message in $M_i^+$.

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

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

so in particular $\beta_i \leq \gamma_i$, and

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

Since $g_i(M_i^+)$ is bounded below by $c/2R$, this implies that

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

Therefore, (8) will follow if we prove that $\beta_i(M_i^+)$ is bounded above by

$$\frac{E_i[(F_i^k - F_i^{k+1}(m_i))]_{m_i \in M_i^+}}{F_i^k - CF_i^{k+1}}$$

for $C = 2R/c$.

For each $m_i$ that arises in equilibrium and $k \geq 1$, Bayes’ rule implies that

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

(10)

where $\Phi(m_i)$ 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 $m_{i-1}^i$: $p_i(m_i)$ is the denominator of (10). Rearranging (10) and simplifying, we have

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

(11)

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

$$F_i^k \beta_i(M_i^+) \leq E_i[(F_i^k - F_i^{k+1}(m_i)) \mathbb{1}_{m_i \in M_i^+}] + \sum_{m_i \in M_i^+} \Phi(m_i).$$

(12)

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

$$\sum_{m_i \in M_i^+} \Phi(m_i) \leq \sum_{m_i \in M_i^+} E_i[\mathbb{1}_{|S_i| \geq k+1} \mathbb{1}_i \text{ works and reports } m_i]$$

$$= E_i[\mathbb{1}_{|S_i| \geq k+1} \mathbb{1}_i \text{ works and reports } m_i \in M_i^+]$$

$$\leq E_i[\mathbb{1}_{|S_i| \geq k+1} \mathbb{1}_i \text{ works}]$$

$$= F_i^{k+1} \gamma_i.$$
noting, for the last equality, that the event that $i$ works, which has probability $\gamma_i$, depends only on $m_i^{t-1}$ and is thus independent of the event $\{|S_t| \geq k + 1\}$ conditional on $m_i^{t-1}$.

Combining this with (12) yields

$$F_i^k \beta(M_i^+) \leq \mathbb{E}_i \left[ (F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+} \right] + F_i^{k+1} \gamma_i.$$  \hspace{1cm} (14)

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 (14) then yields

$$\beta(M_i^+) \leq \frac{1}{F_i^k - 2R/cF_i^{k+1}} \mathbb{E}_i \left[ (F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+} \right].$$  \hspace{1cm} (15)

Combining this with (9) yields (8) for $\lambda < 1$, and by  (13) and $f_i^0 = 1 - F_i^1$ yields

$$F_i^k \beta_i(M_i^+) + \gamma_i F_i^k \delta_i(M_i^+) f_i^0 \lambda \leq \mathbb{E}_i \left[ (F_i^k - F_{i+1}^k(m_i))1_{m_i \in M_i^+} \right] + F_i^{k+1} \gamma_i.$$  \hspace{1cm} (16)

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) and $F_i^k \beta_i(M_i^+) \geq 0$, (16) implies that

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

\textsuperscript{39}Note that the proposition’s assumption that $F_i^k - C(\lambda) F_i^{k+1} > 0$ implies that $F_i^k - 2R/cF_i^{k+1} > 0$, since $C(\lambda) \geq C(1)$ regardless of $\lambda$.  

34
Multiplying the numerator and denominator of the right-hand side 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, (16) implies that

\[ \gamma_i \leq \frac{\mathbb{E}_i \left[ (F^k_i - F^k_{i+1}(m_i))1_{m_i \in M^+_i} \right]}{F^k_i(c/2R) - F^k_{i+1}}. \]

Multiplying the numerator and denominator of the right-hand side by $C(\lambda)$ again yields the result.

\[ \blacksquare \]

A.2 Proof of Theorem 1

The proof proceeds by induction on $k$. Lemma 1 already proves the claim for $k = 1$ with $F^1 = c/R$. Now suppose that the claim holds for some $k \geq 1$: there exists a threshold $F^k > 0$ such that any informative continuation equilibrium in round $i$ satisfies $F^k_i \geq F^k_i$. We will show that there is a constant $F^{k+1} > 0$ such that the continuation equilibrium must also satisfy $F^{k+1}_i \geq F^{k+1}_i$.

For any $\omega \in [0, 1]$, let $F^k(\omega) = \inf\{F^k_i \in [0, 1] : \gamma_i > 0, F^{k+1}_i \leq \omega \}$ 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 $F^{k+1}_i \leq \omega$. $F^k(\omega)$ is by construction nonincreasing in $\omega$.

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

Our objective is to show that $F > 0$. Let $F^k = \lim_{\omega \downarrow F} F^k(\omega)$, i.e., the right limit of $F^k(\cdot)$ at $F$. The limit is well defined by monotonicity of $F^k(\cdot)$. Intuitively, $F^k$ is the smallest probability of having at least $k$ signals left that can be achieved by informative continuation equilibria featuring the smallest possible probability having at least $k + 1$ signals left.

Also let $\varepsilon > 0$ denote any small constant such that $\tilde{F}^k = F^k(F + G\varepsilon) \geq \tilde{F}^k 1/1 + \eta$, where $G > 0$ is a large constant and $\eta > 0$ is a small constant determined at the end of the proof independently of $k$.\footnote{Since $\tilde{F}^k$ is the right-limit of $F^k(\cdot)$ at $\tilde{F}$, such an $\varepsilon$ exists. Moreover, for example, one can choose $G$ and $\eta$ so that $\sqrt{G} = \max\{64R^3/(\lambda c^3), 12R/c\}$ and $\eta = 1/\sqrt{G}$, as explained at the end of this proof.}
since all $\varepsilon' \in (0, \varepsilon)$ also satisfy the condition, we can choose $\varepsilon$ such that

$$\varepsilon \leq \frac{1}{2} \left( \frac{\hat{F}^k}{G} \right)^2.$$  \hfill (17)

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

Consider such a continuation equilibrium, starting at round $i$, say, and suppose by contradiction that

$$F = 0. \hfill (18)$$

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

**Lemma 4** $i$ assigns probability at least $1 - 1/G$ to $A$.

**Proof.** For $j \geq i$, let $\bar{F}_{i+1}^k$ denote the probability assigned at the beginning of round $j$ (i.e., conditional on $m_{i+1}^j$) that there remain at least $k + 1$ signals at the beginning of round $i$ (fixed). The process $\{\bar{F}_{j+1}^k\}_{j \geq i}$ 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, $\bar{F}_{i+1}^k = F_{i+1}^k \leq \varepsilon$.

Doob’s martingale inequality therefore implies that for any $J \geq i$, $Pr(\max_{i \leq j \leq J} \bar{F}_{j+1}^k \geq G\varepsilon) \leq \frac{E_0(\bar{F}_{j+1}^k)}{G\varepsilon} \leq 1/G$. The event $\bar{A}_\infty$ defined by $\{\max_{j \geq i} \bar{F}_{j+1}^k \leq G\varepsilon\}$ is the intersection of the events $\bar{A}_j = \{\max_{i \leq j \leq J} \bar{F}_{j+1}^k \leq G\varepsilon\}$. Therefore, $Pr(\bar{A}_\infty) = \lim_{J \to \infty} Pr(\bar{A}_j) \geq 1 - 1/G$.

Finally, we note that $\bar{F}_{i+1}^k \geq F_{i+1}^k$ 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 $Pr(A) \geq Pr(\bar{A}_\infty) \geq 1 - 1/G$. \hfill ■

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

$$F_j^k - CF_{j+1}^k \geq \hat{F}^k - CG\varepsilon \geq \hat{F}^k / 2 > 0 \hfill (19)$$

provided that $G \geq C$, where the penultimate inequality comes from the bound on $\varepsilon$ (17). The condition $G \geq C$ will be guaranteed by fixing $C$ at the level $C(\lambda)$ provided by Proposition 2.
and then choosing $G$ large enough. (In particular, $G \geq C(\lambda)$. A specific value of $G$ is given at the end of the proof.)

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

From Proposition 2, we have

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

over $A_j$.

Let $B_j$ denote the event that $m_j \in M_j^+$. $B_j$ is adapted to $m_j^{i-1}$, i.e., known at the beginning of round $j + 1$. Notice that if $B_j$ does not occur, it means by definition that no $l > j$ ever works (i.e., $m_j^1$ is followed by an uninformative continuation equilibrium). This implies that the sequence of events $\{B_j\}_{j \geq i}$ is decreasing path by path and, hence, that the sequence $\{1_{B_j}\}_{j \geq i}$ is nonincreasing.

Since $A_j$ is measurable with respect to $m_j^{i-1}$, (19) implies that for $j \geq i + 1$

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

The law of iterated expectations then implies that

$$\mathbb{E}_{i+1}[1_{A_j}\beta_j] \leq \mathbb{E}_{i+1} \left[ 1_{A_j} \frac{2C(F_j^k - F_{j+1}^k)1_{B_j}}{F_j^k} \right] = \frac{2C}{F_j^k} \mathbb{E}_{i+1}[1_{A_j} (F_j^k - F_{j+1}^k)1_{B_j}].$$

Summing (20) over all $j \geq i + 1$, we obtain

$$\mathbb{E}_{i+1} \left[ \sum_{j \geq i+1} 1_{A_j}\beta_j \right] \leq \frac{2C}{F_k^i} \mathbb{E}_{i+1} \left[ \sum_{j \geq i+1} 1_{A_j}1_{B_j}(F_j^k - F_{j+1}^k(m_j)) \right].$$

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:

$$J = \inf\{j \geq i + 1 : 1_{A_j}1_{B_j} = 0\}$$

with the convention that $J = +\infty$ if the set is empty. In words, $J$ is either the first round in which $F_{j+1}^k$ exceeds $G\varepsilon$, or the last round at which the continuation equilibrium is informative.
(which implies that $\gamma_j \geq 0$ for all $j > J$), whichever occurs first.\footnote{J is not a stopping time with respect to the filtration $\{F_j\}_{j \geq i+1}$ generated by public histories $\{m_{i+1}^{j-1}\}_{j \geq i+1}$: 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.}

Consider first the paths for which $J$ is finite. The argument of $E_{i+1}[\cdot]$ on the right-hand side of (21) 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 where $J$ is infinite. In this case, the argument of $E_{i+1}[\cdot]$ on the right-hand side of (21) is equal to

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

Notice that the limit is well defined because $F_j^k$ is a nonnegative supermartingale.\footnote{The argument is similar to the one used to prove Lemma 4. For any fixed $j$, let $\tilde{F}_l^{k+1}$ denote the probability assigned by $l \geq j$ to there remaining at least $k + 1$ signals at the beginning of round $j$. The process $\{\tilde{F}_l^{k+1}\}_{l \geq j}$ 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, $\tilde{F}_l^{k+1} \geq F_{l+1}^{k+1}$ for all $l \geq j$, because the actual number of remaining signals only decreases over time. Therefore, we have $F_{j}^{k+1} = \tilde{F}_{j}^{k+1} = E_{j}[\tilde{F}_{j+1}^{k+1}] \geq E_j[F_{j+1}^{k+1}]$. 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.}\footnote{The inequality clearly holds if $\gamma_j = 0.$}

For notational consistency, we will call this limit $F_{\cdot}^k$, regardless of whether $J$ is finite.

Combining these observations with (21), we conclude that

$$E_{i+1} \left[ \sum_{j \geq i+1} 1_{A_j \beta_j} \right] \leq \frac{2C}{F^k} (F_{i+1}^k - E_{i+1}[F_j^k]).$$

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\footnote{The inequality clearly holds if $\gamma_j = 0.$} that $\gamma_j \leq g \beta_j$ where $g = \frac{R}{\lambda c}$ and hence that

$$E_{i+1} \left[ \sum_{j \geq i+1} 1_{A_j \gamma_j} \right] \leq \frac{2C g}{F^k} (F_{i+1}^k - E_{i+1}[F_j^k]).$$

(22)
Let $Z$ denote the event that at least one agent $j > i$ works and $\pi_{i+1}(m_i) = Pr_{i+1}(A \cap Z)$, i.e., the probability that $A$ and $Z$ both occur conditional on $m_i$. We have

$$\pi_{i+1}(m_i) = Pr_{i+1}\left(\mathbb{1}_A \sum_{j > i} \mathbb{1}_j \text{ works} \geq 1\right)$$

$$\leq E_{i+1}\left[\mathbb{1}_A \sum_{j > i} \mathbb{1}_j \text{ works}\right]$$

$$= \sum_{j > i} E_{i+1}\left[\mathbb{1}_A \mathbb{1}_j \text{ works}\right]$$

$$\leq \sum_{j > i} E_{i+1}\left[\mathbb{1}_A \mathbb{1}_j \text{ works}\right]$$

$$= \sum_{j > i} E_{i+1}\left[E_j\left[\mathbb{1}_A \mathbb{1}_j \text{ works}\right]\right]$$

$$= \sum_{j > i} E_{i+1}\left[\mathbb{1}_A E_j\left[\mathbb{1}_j \text{ works}\right]\right]$$

$$= \sum_{j > i} E_{i+1}\left[\mathbb{1}_A \gamma_j\right].$$

The first equality comes from the fact that $Z$ is identical to the event \(\sum_{j > i} \mathbb{1}_j \text{ works} \geq 1\). The first inequality comes from the fact that $\mathbb{1}_A \sum_{j > i} \mathbb{1}_j \text{ works}$ 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 $A \subset A_j$ and, hence, $\mathbb{1}_A \leq \mathbb{1}_{A_j}$. The next equality comes from the law of iterated expectations and the next one comes from the fact that $A_j$ is measurable with respect to the information at the beginning of round $j$. The last equality comes from the definition of $\gamma_j$.

From (22) and Tonelli’s theorem, this implies that

$$\pi_{i+1}(m_i) \leq \frac{2Cg}{Fk} \left(F_{i+1}^k - E_{i+1}[F_j^k]\right). \quad (23)$$

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 $m_i$.

Let $N_i = \{m_i : F_i^{k,r}(m_i) > (1 + \eta)F_i^k\}$. 

39
Lemma 5  

i) \( \gamma_i(N_i) \leq \frac{1}{2 \pi \sigma^2} \), ii) for \( m_i \notin N_i \), \( \pi_{i+1}(m_i) \leq 2 C \left( (1 + \eta) F_{i}^k - \mathbb{E}_{i+1}[F_{j}^k] \right) \).

Proof.  

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 \( \sum_{s_i \in S'_i} F_{i}^{k+1}\beta'_i(s_i|k + 1) \leq \gamma_iF_{i}^{k+1} \). 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

\[
\gamma'_i(S'_i) \leq \frac{F_{i}^{k+1}}{F_{i}^k} \leq \frac{\varepsilon}{F_{i}^k},
\]

since \( F_{i}^{k+1} \leq \varepsilon \) and \( F_{i}^k \geq \hat{F}_i^k \). From \((17)\), the right-hand side is bounded above by \( \frac{\hat{F}_i^k}{2\sigma^2} \).

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

\[
F_{i}^{k,r}(m_i) = \sum_{s_i \in S_i} \sigma(s_i|m_i)F_{i}^{k,w}(s_i)
= \sum_{s_i \in S'_i} \sigma(s_i|m_i)F_{i}^{k,w}(s_i) + \sigma(\emptyset|m_i)F_{i}^{k,w}(\emptyset) + \sum_{s_i \notin S_i \setminus S'_i} \sigma(s_i|m_i)F_{i}^{k,w}(s_i)
\]

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.\footnote{Formally, for \( k \geq 1 \), we have \( F_{i}^{k,w}(\emptyset) = \frac{F_{i}^{k}(1-\lambda)}{(1-F_{i}^k)^{1-\lambda}} = F_{i}^k \frac{1-\lambda}{(1-F_{i}^k)^{1-\lambda}} \leq F_{i}^k \).} Intuitively, finding nothing always increases the probability that there
are no signals remaining to be found. Finally, the first term is bounded above by $\sigma(S_i'|m_i) = q(m_i)$. Therefore,

$$F_{i}^{k,r}(m_i) \geq (1 + \eta)F_{i}^{k} \Rightarrow q(m_i) + (1 - q(m_i))F_{i}^{k} \geq (1 + \eta)F_{i}^{k}$$

$$\Rightarrow q(m_i) \geq \eta F_{i}^{k}.$$  

To conclude, note that

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

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 of $F_{i}^{k}$ and $F_{i}^{k,r}(m_i)$, we have $F_{i+1}^{k}(m_i) \leq (1 + \eta)F_{i}^{k}$ for all $m_i \notin N_i$. From (23), this implies that

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

for all $m_i \notin N_i$.  

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) $\gamma_i(T_i) \leq 1/\sqrt{G}$.

**Proof.** 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 $F_{i+1}^{k+1} \leq \varepsilon$ 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
Using Markov’s inequality, this implies that

\[
\Pr_{i} \left[ \text{signals left at the beginning of round } i \right] \leq \sqrt{G} \varepsilon.
\]

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 \( F^{k+1}_i(m_i) \leq \varepsilon \), the law of iterated expectations implies that the probability \( F^{k+1}_i(m_i) \) 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

\[
\mathbb{E}_i[F^{k+1}_i|m_i] = \sum_{m_i \in M_i} \gamma_i(m_i) F^{k+1}_i(m_i) = F^{k+1}_i \leq \varepsilon.
\]

Using Markov’s inequality, this implies that \( \Pr_i(m_i : F^{k+1}_i(m_i) \geq \sqrt{G} \varepsilon | i \text{ works}) \leq \frac{\varepsilon}{\sqrt{G} \varepsilon} = 1/\sqrt{G} \). Since also \( F^{k+1}_i(m_i) \geq F^{k+1}_i(m_i) \), we get \( \gamma_i(T'_i) \leq 1/\sqrt{G} \). Since \( T_i \subset T'_i \), this shows that \( \gamma_i(T_i) \leq 1/\sqrt{G} \).

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

\[
V^w = \sum_{m_i \in M_i} \gamma_i(m_i) V^w(m_i)
\leq (\gamma_i(N_i) + \gamma_i(T_i)) R + \sum_{m_i \notin N_i \cup T_i} \gamma_i(m_i) V^w(m_i)
\leq \left( \frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} \right) R + \sum_{m_i \notin N_i \cup T_i} \gamma_i(m_i) V^w(m_i)
\]

(24)

Moreover,

\[
V^w(m_i) = \Pr(Z|m_i^1) V^w(m_i|Z) + \Pr(Z^c|m_i^1) V^w(m_i|Z^c).
\]

(25)

Conditional on \( m_i^1 \), the event \( Z \) 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 \( m_i^1 \), not on the signals that remain to be discovered in the case. And as soon as someone works, then by definition \( Z \) has occurred. Thus, what triggers the event \( Z \) (whenever it occurs) is a sequence of uninformative (until \( Z \) occurs) reports \( m_j \) for agents following \( i \), whose probability distribution is completely pinned down by \( m_i^1 \).

From the previous lemmas, we have

\[
\Pr(A \cap Z|m_i^1) \leq \frac{2C \eta}{F^k} \left( (1 + \eta) F^{k}_i - \mathbb{E}_{i+1}[F^{k}_j] \right) \]

for all \( m_i \notin N_i \) and \( \Pr(A^c|m_i^1) \leq 1/\sqrt{G} \) for all \( m_i \notin T_i \). Letting \( M_i = M_i \setminus (N_i \cup T_i) \), this implies that

\[
\Pr(Z|m_i^1) = \Pr(Z \cap A|m_i^1) + \Pr(Z \cap A^c|m_i^1) \leq \frac{2C \eta}{F^k} \left( F^{k}_i (1 + \eta) - \mathbb{E}_{i+1}[F^{k}_j] \right) + 1/\sqrt{G}
\]

(26)
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 \( F_j^k \geq F^k(F_j^{k+1}) \) for all \( j \leq J \). Since \( F^k(\cdot) \) is nonincreasing, this implies that \( F_j^k \geq F^k(G \varepsilon) = \hat{F}_j^k \) for all \( j \leq J \).

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

\[
\mathbb{E}_{i+1}F_j^k = Pr_{i+1}(A)\mathbb{E}_{i+1}[F_j^k|A] + Pr_{i+1}(A^c)\mathbb{E}_{i+1}[F_j^k|A^c]
\geq Pr_{i+1}(A)\mathbb{E}_{i+1}[F_j^k|A]
\geq (1 - 1/\sqrt{G})\hat{F}_j^k.
\]

By construction, \( F_i^k \leq \hat{F}_i^k + \hat{F}_i^k \eta \) and \( \hat{F}_i^k \geq \hat{F}_i^k - \hat{F}_i^k \eta \) and hence \( F_i^k - \hat{F}_i^k \leq (\hat{F}_i^k + \hat{F}_i^k \eta) - (\hat{F}_i^k - \hat{F}_i^k \eta) = 2\hat{F}_i^k \). Letting \( B = 8CG \), (26) then implies (for \( \eta \leq 1/2 \), which we assume) that for all \( m_i \) in \( \hat{M}_i \)

\[
Pr(Z|m_i^1) \leq B\eta + \frac{B}{2\sqrt{G}} + 1/\sqrt{G}.
\]  

(27)

For each \( m_i \), let \( V_{i|C}^j(m_i|Z^c) \) denote \( i \)'s expected gross utility if he sends message \( m_i \) conditional on no \( j > i \) working and \( V_{i|C}^j \) denote the maximizer of \( V_{i|C}^j(m_i|Z^c) \) 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 \( V^w(m_i|Z^c) = V_{i|C}^j(m_i|Z^c) \).

Combining these observations with (25) and (27), we obtain

\[
V^w(m_i) \leq \left( B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}} \right) R + V_{i|C}^j,
\]

for \( m_i \in \hat{M}_i \). Combining this with (24) yields

\[
V^w \leq \left( \frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} \right) R + \left( B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}} \right) R + V_{i|C}^j.
\]

\( i \)'s utility from working thus satisfies

\[
U^w \leq \left( \frac{1}{2\eta G^2} + \frac{1}{\sqrt{G}} + 2B\eta + \frac{B}{2\sqrt{G}} + \frac{1}{\sqrt{G}} \right) R + V_{i|C}^j - c.
\]  

(28)
If $i$ sends a message $m_i^* \in \hat{M}_i$ that achieves $V^{f,*}$, his utility $U^f$ satisfies

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

where 0 is used as a lower bound on $i$’s realized gross utility to derive the second inequality.

Therefore, working is strictly suboptimal if

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

or

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

(29)

This inequality is always satisfied as long as $\eta$ is small enough and $\eta G$ 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 (29) is less than $c/4R$ with some strict inequalities.

\section{B Proof of Theorem 2}

We construct compensation functions for which the strategy profile proposed constitutes an equilibrium. Under this strategy profile, as long as $p_i$ lies in $(\bar{p}, \bar{p})$ all agents work and report an informative signal about $\omega$. 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 $\hat{\rho}$ and containing a single point on each side of $(p, \bar{p})$. Let $q^0 \leq p < q^1, \ldots, \hat{\rho}, \ldots, q^N < \bar{p} < q^{N+1}$ 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 $\hat{\rho} = p_{J+1}$ 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 \( p_i = q^k \in (p, \tilde{p}) \), if \( i \) reports “H”, he gets a reward \( R^k_H \geq 0 \) if \( \tilde{p} = q^{N+1} \) and a punishment \( P^k_L \leq 0 \) if \( \tilde{p} = q^0 \). If \( i \) reports “L”, he gets \( R^k_L \geq 0 \) if \( \tilde{p} = q^0 \) and \( P^k_L \leq 0 \) 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 \( (p, \tilde{p}) \) through \( \tilde{p} \), from the perspective of an agent who assigns probability \( p \) to \( \omega \), 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

\[
\pi(q^{k+1}, q^{k+1}) > \pi(q^k, q^{k+1}) \quad (30)
\]

and

\[
\pi(q^{k-1}, q^{k-1}) < \pi(q^k, q^{k-1}) \quad (31)
\]

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

For \( k = 1 \), the investigation stops if \( i \) reports “L” so \((31)\) holds as an equality, but \((30)\) is still strict, because this report triggers further investigation. The reverse is true for \( k = N \): \((30)\) only holds as an equality while \((31)\) is strict.

If \( i \) shirks, his maximal utility is

\[
\max \{ \pi(q^k, q^{k+1}) R^k_H + (1 - \pi(q^k, q^{k+1})) P^k_H; \pi(q^k, q^{k-1}) P^k_L + (1 - \pi(q^k, q^{k-1})) R^k_L \}. \quad (32)
\]

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

\[
 z^k [\pi(q^{k+1}, q^{k+1}) R^k_H + (1 - \pi(q^k, q^{k+1})) P^k_H] + (1 - z^k) [\pi(q^{k-1}, q^{k-1}) P^k_L + (1 - \pi(q^{k-1}, q^{k-1})) R^k_L]
\]

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

It is optimal for \( i \) to work if the expression in (33) exceeds the expression in (32) by at least \( c \).

This condition is obtained as follows: set \( P^k_H = P^k_L = -Q \) where \( Q \) is a strictly positive constant, and let \( R^k_H = Q^{1-\pi(q^k, q^{k+1})} \) and \( R^k_L = Q^{\pi(q^k, q^{k-1})} \). This guarantees that \( i \)'s expected payoff from fabrication is zero, regardless of the outcome. From (30) and (31), his payoff from working is of order \( Q \) and thus exceeds \( c \), for \( Q \) high enough.

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 \((\bar{p}, \bar{p})\).

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

**General case**

As noted, the argument extends easily to the case in which \( \rho \) and/or \( \lambda \) 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 \((p, \bar{p})\), 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 \((\rho, \bar{p})\) in equilibrium, given the current belief \( q^k \), so that the expected compensation of a working agent still exceeds the cost.

\footnote{To see this, let \( \bar{\pi} \) denote a strictly positive lower bound on all inequalities (30) and (31) over all \( k \)'s whenever they hold strictly. Then, the gain from working is of order \( Q\bar{\pi} \).}
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 again assume 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 \((p, \bar{p})\) 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 \( \lambda \) and \( \rho \) are less than 1 is a convex combination of the previous cases and addressed accordingly.

C Proof of Theorem 3

Without loss of generality, we again assume that agents’ gross utility functions \( \{V_i\}_{i \in \mathbb{N}} \) all take values in \([0, R]\).

For any round \( i \), consider the event \( Q_i \) that all past investigators whose may have failed to discover signals, given the history \( m_i \), have indeed failed to discover signals. Conditional on \( Q_i \), 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 \( m_i \). 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 \( Q_i \) 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 \( m_i \) and \( Q_i \) is only a function of \( q_i \), iii) \( \hat{F}_i^k = F_i^k \) for all \( k \).

We will prove that there exist strictly positive thresholds \( \{\hat{F}_i^k\}_{k \geq 1} \) such that an informative continuation equilibrium exists in round \( i \) only if \( \hat{F}_i^k \geq \hat{F}_i^k \) 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) \( \hat{F}_j^k \leq \hat{F}_i^k \) 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

\[
F^k_i = \sum_{r \in \{q_i, \ldots, i-1\}} \Pr(r_i = r|m_i^{i-1}) \Pr(\tilde{K} \geq r + k | \tilde{K} \geq r) \\
\leq \sum_{r \in \{q_i, \ldots, i-1\}} \Pr(r_i = r|m_i^{i-1}) \Pr(\tilde{K} \geq q_i + k | \tilde{K} \geq q_i) \\
= \sum_{r \in \{q_i, \ldots, i-1\}} \Pr(r_i = r|m_i^{i-1}) \tilde{F}^k_i \\
= \tilde{F}^k_i.
\]

The first equality comes from the independence of \(\tilde{K}\) from \(S^\infty\): \(r_i\) 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 \(k \geq 0\), \(\Pr(\tilde{K} \geq k + q | \tilde{K} \geq q)\) is non-increasing in \(q\). The second equality is due to the equality \(\tilde{F}^k_i = \Pr(\tilde{K} \geq k + q_i | \tilde{K} \geq q_i)\), which again comes from the independence of \(\tilde{K}\) and \(S^\infty\): the only relevant information about \(\tilde{K}\) conditional on \(Q_i\) 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 \(k \geq 1\)

\[
\tilde{F}^k_j = \Pr(\tilde{K} \geq k + q_j | \tilde{K} \geq q_j) \\
\leq \Pr(\tilde{K} \geq k + q_i | \tilde{K} \geq q_i) \\
= \tilde{F}^k_i,
\]

where the inequality comes from the increasing hazard rate property (see Footnote 47).

The existence of thresholds \(\{F^k\}_{k \geq 1}\) in the result mentioned above is proved by induction on \(k\). We start with the base case \(k = 1\) and then prove the induction step.

C.1 Base case: Proof of the result for \(k = 1\)

Suppose that \(\tilde{F}^1_i < c/2R\). Combining the two parts of Lemma 7, this implies that \(F^1_j < c/2R\) for all \(j \geq i\). Lemma 1 still applies: no investigator \(j \geq i\) works because the probability that

---

47See, 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 \(p, q \geq 0\), \(\Pr(X \geq p + q | X \geq p) = \bar{F}(p + q)/\bar{F}(p)\) is decreasing in \(p\).
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:

- $\beta_i$: probability that $i$ produces an informative message given $m_1^{i-1}$;
- $M_i^+$: set of messages $m_i$ which are followed by an informative continuation equilibrium (as in the case of investigators);
- $Pr_i(M_i^+)$: 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 $m_1^{i-1}$;
- $\gamma_i(m_i|s_i)$: 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 $f_\ell$ that are bounded above by $\bar{f}$. For any $\varepsilon \geq 0$, let $E_\varepsilon^L$ denote the event that $\exists \ell, \ell' \leq L$ such that $|Y_\ell - Y_{\ell'}| \leq \varepsilon$. Then

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

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

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

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

$$Pr(E_\varepsilon^L) \leq Pr(E_\varepsilon^{L-1}) + \sum_{\ell \leq L-1} Pr(|Y_\ell - Y_L| \leq \varepsilon)$$

$$\leq (L-1)(L-2)\bar{f}\varepsilon + (L-1) \times 2\varepsilon\bar{f}$$

$$= L(L-1)\bar{f}\varepsilon,$$

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 exists a threshold $F^1 \in (0, c/2R)$ such that $\beta_i > 0$ in some witness round $i$ only if $\hat{F}_i^1 \geq F^1$. 

49
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 \( m_i^{i-1} \).

If \( i \) is a witness and has preference \( \epsilon \), his message is said to be \( \epsilon \)-informative if there exist two signals \( s_i \neq s_i' \) such that the equilibrium distributions of \( m_i \) conditional on \( i \) getting signals \( s_i \) and \( s_i' \) given \( i \)'s preference \( \epsilon \), are different across the 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 \( \epsilon \) for which i’s message is \( \epsilon \)-informative has positive probability.

For any equilibrium and round \( i \) in which \( i \) is a witness, let \( \nu_i \) denote the probability that \( i \)'s preference shock \( \epsilon_i \) is such that \( i \)'s message is \( \epsilon_i \)-informative.

For any \( F < c/2R \), let \( \nu(F) \) denote the supremum of \( \nu_i \) over all witness rounds \( i \) of all equilibria such that \( \hat{F}_i^{i} \leq F \). 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

\[
U_i(m_i; s_i) = z(m_i)\mathbb{E}_i[V_i(m)|s_i, m_i^1] + (1 - z(m_i))V_i(m_i) + \epsilon_i(m_i)
\]

(34)

where

\[
V_i(m_i) = \mathbb{E}_i[V_i(m)|m_i^1, \text{no } j > i \text{ produces an informative message}].
\]

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

Given \( \epsilon_i \), \( i \) 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 (34) and the fact that \( V_i(m) \in [0, R] \), this is possible only if

\[
|\epsilon_i(m_i) + V_i(m_i) - \epsilon_i(m_i') + V_i(m_i')| \leq R(z(m_i) + z(m_i'))
\]
The random variables \( Y_\ell = \epsilon_i(m_\ell) + \bar{V}_i(m_\ell) \) satisfy the assumptions of Lemma 8. Letting \(|M|\) denote the cardinality of the message space \( M \), we thus have

\[
Pr(i \text{ sends an informative message}) \leq |M|^2 \bar{f} R(z(m_i) + z(m'_i)) \tag{35}
\]

Since no investigator \( j \geq i \) works when \( \hat{F}_i \leq c/2R \), we have for any message \( m_i \)

\[
z(m_i) \leq \sum_{j \geq 1} Pr(\text{there are } j \text{ witnesses in the sequence after round } i, \text{ given message } m_i) j \nu(F). \tag{36}
\]

Indeed, by definition of \( \nu(F) \) and the fact that \( \hat{F}_j \leq F \) 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 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 (36) 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 (35), we obtain

\[
Pr(i \text{ sends an informative message} \mid \hat{F}_i \leq F) \leq 2|M|^2 \bar{f} R \nu(F) F/(1 - F)^2 \tag{37}
\]

Taking the supremum of the left-hand side over all witness rounds \( i \) and equilibria such that \( \hat{F}_i \leq F \), 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
and starts at zero. Therefore, we conclude that there exists a threshold $F^1 > 0$ such that $\nu(F) = 0$ for all $F \leq F^1$.

\[\]

\section*{C.2 Induction step}

Suppose that there exist strictly positive thresholds $\{F^{k'}\}_{k' \in \{1, \ldots, k\}}$ such that a continuation equilibrium starting at round $i$ is informative only if $\hat{F}^{k'}_i \geq F^{k'}$ 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 $\hat{F}^{k+1}_i \leq \varepsilon$ and obtain an impossibility for $\varepsilon$ small enough.

Consider any $\varepsilon < F^k \times F^1$ and any informative continuation equilibrium starting at round $i$ such that $\hat{F}^{k+1}_i \leq \varepsilon$. Then, all continuation equilibria are uninformative as soon as the 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

\[
\hat{F}^{k}_{j+1}(m_j) = \Pr(\tilde{K} \geq k + q_j + 1 \mid \tilde{K} \geq q_j + 1)
\]

\[
= \frac{\Pr(\tilde{K} \geq k + q_j + 1)}{\Pr(\tilde{K} \geq q_j + 1)}
\leq \frac{\Pr(\tilde{K} \geq k + q_i + 1)}{\Pr(\tilde{K} \geq q_i + 1)}
\]

\[
= \frac{\Pr(\tilde{K} \geq k + q_i + 1)}{\Pr(\tilde{K} \geq q_i + 1)} \times \frac{\Pr(\tilde{K} \geq q_i)}{\Pr(\tilde{K} \geq q_i + 1)}
\]

\[
= \hat{F}^{k+1}_i \hat{F}^1_i,
\]

where the inequality comes for the monotone hazard rate assumption (see Footnote 47) and the fact that $q_j \geq q_i$. Since the continuation equilibrium from round $i$ is informative, we must have $\hat{F}^1_i \geq F^1$. Therefore, $\hat{F}^{k+1}_i \leq \varepsilon < F^k F^1$ implies that

\[
\hat{F}^{k}_{j+1} < F^k
\]

which shows, by the 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.\footnote{Formally, the argument is similar to the proof of Lemma 9 except that here $z(m_i) = 0$ regardless of the}
Therefore, if \( \hat{F}^{k+1}_i \leq \varepsilon \), 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 \( F^{k+1}_i \) exceeds the \( k+1 \)-threshold given by Theorem 1, which we denote here \( \tilde{F}^{k+1} \). Since \( F^{k+1}_i \leq \hat{F}^{k+1}_i \) by Part i) of Lemma 7, letting \( \bar{F}^{k+1}_i = \min\{\tilde{F}^{k+1}_i, F^k E^1_i\} \) we conclude that no informative continuation equilibrium exists in round \( i \) if \( \hat{F}^{k+1}_i \leq \bar{F}^{k+1}_i \). This concludes the induction step.