Evidence Games: Truth and Commitment[1]

Sergiu Hart[2] Ilan Kremer[3] Motty Perry[4]

October 31, 2016

[1]Previous versions: February 2014; May 2015 (Center for Rationality DP-684); March 2016. The authors thank Maya Bar-Hillel, Elchanan Ben-Porath, Dan Bernhardt, Peter DeMarzo, Kobi Glazer, Ehud Guttel, Johannes Hörner, Vijay Krishna, Rosemarie Nagel, Michael Ostrovsky, David Pérez-Castrillo, Uriel Procaccia, Phil Reny, Tomás Rodríguez-Barraquer, Ariel Rubinstein, Amnon Schreiber, Andy Skrzypacz, Rani Spiegler, Francesco Squintani, Yoram Weiss, and David Wettstein, for useful comments and discussions. We also thank the anonymous referees and the coeditor for their very careful reading and helpful comments and suggestions. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[2]Department of Economics, Institute of Mathematics, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. Research partially supported by an Advanced Investigator Grant of the European Research Council (ERC). *E-mail*: hart@huji.ac.il *Web site*: http://www.ma.huji.ac.il/hart

[3]Department of Economics, Business School, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem; Department of Economics, University of Warwick. Research partially supported by a grant of the European Research Council (ERC). *E-mail*: kremer@huji.ac.il

[4]Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem; Department of Economics, University of Warwick. *E-mail*: m.m.perry@warwick.ac.uk *Web site*: http://www2.warwick.ac.uk/fac/soc/economics/staff/academic/perry

Abstract

An *evidence game* is a strategic disclosure game in which an informed agent who has some pieces of verifiable evidence decides which ones to disclose to an uninformed principal who chooses a reward. The agent, regardless of his information, prefers the reward to be as high as possible. We compare the setup in which the principal chooses the reward after the evidence is disclosed to the mechanism-design setup where he can commit in advance to a reward policy, and show that under natural conditions related to the evidence structure and the inherent prominence of truth, the two setups yield the *same* outcome.

Contents

Ask someone if they deserve a pay raise. The invariable reply (with very few and, therefore, notable exceptions) is, "Of course." Ask defendants in court whether they are guilty and deserve a harsh punishment, and the again invariable reply is, "Of course not."

So how can reliable information be obtained? How can those who deserve a reward, or a punishment, be distinguished from those who do not? Moreover, how does one determine the right reward or punishment when everyone, regardless of information and type, prefers higher rewards and lower punishments?

These are clearly fundamental questions, pertinent to many important setups. The original focus in the literature was on equilibrium and equilibrium prices. This approach was initiated by George Akerlof (1970), and followed by the large body of work on voluntary disclosure, starting with Sanford Grossman and Oliver Hart (1980), Grossman (1981), Paul Milgrom (1981), and Ronald Dye (1985). A related environment was considered by Jerry Green and Jean-Jacques Laffont (1986), but from a general mechanism-design viewpoint, where one can commit in advance to a policy.

As is well known, commitment is a powerful device.[1] The present paper nevertheless identifies a natural and important class of setups—which includes voluntary disclosure as well as various other models of interest—that we call "evidence games," in which the possibility to commit does *not* matter, namely, the equilibrium and the optimal mechanism coincide. This issue of whether commitment can help was initially addressed by Jacob Glazer and Ariel Rubinstein (2004, 2006) (see also Itai Sher 2011).

An *evidence game* is a standard communication game between an "agent" who is informed and sends a message (that does not affect the payoffs) and a "principal" who chooses the action (call it the "reward"). The two distinguishing features of evidence games are, first, that the agent's private information (the "type") consists of certain pieces of verifiable evidence, and the agent can reveal in his message all this evidence (the "whole truth"),

[1]The reader is encouraged to consult the online Appendix C for additional results, extensions, notes, and discussions.

or only some part of it (a "partial truth").[2] The second feature is that the agent's preference is the same regardless of his type—he always prefers the reward to be as high as possible[3]—whereas the principal's utility, which does depend on the type, is single-peaked—he prefers the reward to be as close as possible to the "right reward."

An essential feature of evidence games is the possibility of revealing the whole truth; the slight inherent advantage of the whole truth is used to select equilibria, which we call *truth-leaning equilibria*. Specifically, these obtain from limits of perturbed games with infinitesimal increases in the agent's utility when telling the whole truth, and in his probability of doing so. Truth-leaning thus amounts to the following two conditions: (i) when the reward for revealing a partial truth is the same as the reward for revealing the whole truth, the agent prefers to reveal the whole truth; and (ii) there is a small positive probability that the whole truth is revealed. These simple conditions are most natural, and they (and variants thereof) have been repeatedly used in the literature. The truth is after all a focal point, and there must be good reasons for *not* telling it. As Mark Twain wrote, "When in doubt, tell the truth," and "If you tell the truth you don't have to remember anything." Truth-leaning turns out to be consistent with the various refinement conditions offered in the literature, and equivalent to many of them (such as the equilibria used in the voluntary disclosure literature).

To see the effect of commitment we consider the two distinct ways in which the interaction between the two players may be carried out. One way is for the principal to decide on the reward only *after* receiving the agent's message; the other way is for the principal to *commit* to a reward policy, which is made known *before* the agent sends his message (i.e., the principal is the Stackelberg leader, which can only help him; this is the mechanism-design setup). Our equivalence result can be stated as follows:

> *In evidence games the truth-leaning equilibria without commitment yield the same (ex-post) payoffs as the optimal mechanisms*

[2] Try to recall the number of job applicants who included rejection letters in their files.

[3] This differs from signaling and screening setups, where costs depend on type, and cheap-talk setups, where utility depends on type.

with commitment.

Simple examples that illustrate the result and the intuition behind it are provided in Section I.

A number of comments are in order. First, the result implies in particular that among all Nash equilibria, the truth-leaning equilibria are optimal, i.e., most preferred by the principal.

Second, the "truth structure" of evidence games (which consists of the partial truth relation and truth-leaning) *guarantees* that commitment cannot yield any advantage. Whereas in the above-mentioned work of Glazer and Rubinstein (2004, 2006) and Sher (2011), the commitment outcome is obtained in some equilibrium of the game, but in general not in its other equilibria—and there is no good reason for the former to be picked out over the latter—in evidence games *all* truth-leaning equilibria yield the commitment outcome.

And third, the fact that commitment is not needed in order to guarantee optimality is a striking feature of evidence games; as we will show, the truth structure is indispensable to this result.

We stated above that evidence games constitute a very naturally occurring environment, which includes a wide range of applications and well-studied setups of much interest. We discuss here only two such applications. The first one deals with voluntary disclosure in financial markets. Public firms enjoy a great deal of flexibility when disclosing information. While disclosing false information is a criminal act, withholding information is allowed in some cases, and is practically impossible to detect in other cases. This has led to a growing literature in financial economics and accounting (see for example Dye 1985 and Hyun Song Shin 2003, 2006) on voluntary disclosure and its impact on asset pricing. The equilibria considered there turn out to be (outcome-equivalent to) truth-leaning equilibria, and so our result implies that the market's equilibrium behavior is in fact optimal: it yields the optimal separation between "good" and "bad" firms (i.e., even with mechanisms and commitments—such as managers' contracts—it is not worthwhile to separate more).

The second application concerns the judicial system. The system (the "principal") commits itself through constitutions, laws, legal doctrines, precedents—which include inter alia rules of evidence. All this affects what evidence the parties (the "agents") provide in court. An essential objective of the judicial criminal system is to induce the optimal amount of separation between the guilty and the innocent and to get as close as possible to the right judgement ("fit the punishment to the crime"). Our result says that the power of these commitments does not, however, go beyond selecting among all equilibria the truth-leaning equilibria—which are most natural in this setup. A case in point is the legal doctrine known as "the right to remain silent." In the United States, this right is enshrined in the Fifth Amendment to the Constitution, and is interpreted to include the provision that adverse inferences cannot be drawn, by the judge or the jury, from the refusal of a defendant to provide information. While the right to remain silent is now recognized in many of the world's legal systems, its above interpretation regarding adverse inference has been questioned and is not universal. The present paper sheds some light on this debate. First, because equilibria in general, and truth-leaning equilibria in particular, entail Bayesian inferences, the equivalence result implies that the same inferences apply to the optimal mechanisms; therefore, adverse inferences should be allowed, and surely not committedly disallowed. Second, truth-leaning may well replace commitment: rather than committing to rules such as the right to remain silent and its offshoots, one may instead strengthen and reinforce the (perceived) advantages of truth-telling. In England, for instance, an additional provision (in the Criminal Justice and Public Order Act of 1994) states that "it may harm your defence if you do not mention when questioned something which you later rely on in court," which may be viewed, on the one hand, as allowing adverse inference, and, on the other, as making the revelation of only partial truth possibly disadvantageous—which is the same as giving an advantage to revealing the whole truth (i.e., truth-leaning).

To summarize the main contribution of the present paper: for the class of *evidence games* that we consider—which model very common and important setups in information economics, setups that lie outside the standard

signaling and cheap-talk literature—we prove the *equivalence* between truth-leaning equilibria without commitment and optimal mechanisms with commitment; moreover, we show that the conditions of evidence games—most importantly, the truth structure—are *indispensable* conditions beyond which this equivalence no longer holds. In a nutshell, the paper *identifies the natural structure of evidence with its associated truth-leaning as the setup that guarantees that commitment cannot yield any advantage.*

The paper is organized as follows. We start with two examples that illustrate the result in Section I, followed by a survey of the relevant literature in Section II. Section III describes the model and the assumptions. The equivalence result is stated in Section IV, and proved in Section V (with one of the proofs relegated to Appendix A). In Appendix B it is shown that our conditions are indispensable for the equivalence result, and the online Appendix C provides additional reuslts, extensions, discussions, and comments.

I Examples

We provide two simple examples that illustrate the equivalence result and explain some of the intuition behind it.

Example 1 A professor negotiates his salary with the dean. The dean would like to set the salary as close as possible to the professor's "value," while the professor would naturally like his salary to be as high as possible. The dean asks the professor if he can provide some evidence of his value (such as whether a recent paper was accepted or rejected, outside offers, and so on). Assume that with probability 50% the professor has no such evidence, in which case his (expected) value is 60, and with probability 50% he does have some evidence. In the latter case it is equally likely that the evidence is positive or negative, which translates to a value of 90 and 30, respectively. Thus there are three professor types: the "no-evidence" type t_0, with probability 50% and value 60, the "positive-evidence" type t_+, with probability 25% and value 90, and the "negative-evidence" type t_-, with probability 25% and value 30. The professor can provide only evidence that he has, but

he may choose which evidence to provide (thus, for example, t_- can either reveal his evidence, or act as if he had no evidence, i.e., as if he were t_0); see the bottom arrows in Figure 1.

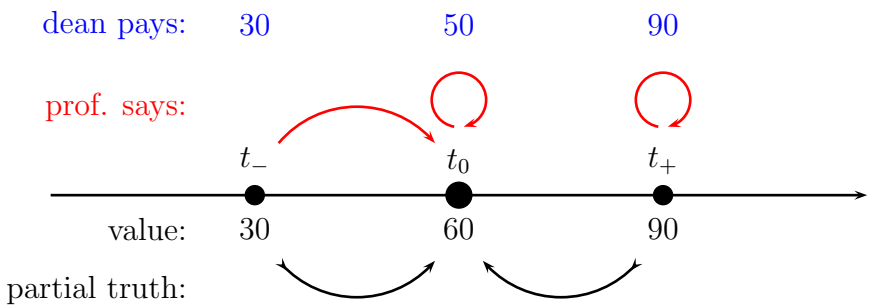


Figure 1: Example 1

Consider first the game setup (without commitment): the professor decides whether to reveal his evidence, if he has any, and then the dean chooses the salary. It is easy to verify (see Appendix C.1) that there is a unique sequential equilibrium, where t_+ reveals his positive evidence and is given a salary of 90 (equal to his value), whereas t_- conceals his evidence and pretends that he has no evidence. When no evidence is presented the dean's optimal response is to set the salary at $50 = (50\%\cdot 60 + 25\%\cdot 30)/(50\% + 25\%)$, i.e., the expected value of the two types that provide no evidence: t_0 and t_-. See the top arrows in Figure 1.

Next, consider the mechanism setup (with commitment): the dean commits to a salary policy (namely, three salaries, denoted by x_+, x_-, and x_0, for those who provide, respectively, positive evidence, negative evidence, and no evidence), and then the professor decides what evidence to reveal. One possibility is of course the above equilibrium: $x_+ = 90$ and $x_- = x_0 = 50$. Can the dean do better by committing? Can he provide incentives to the negative-evidence type t_- to reveal his information? In order to separate and give different salaries to t_- and t_0 , the salary x_- for those who provide negative evidence must be higher than the salary x_0 for those who provide no

evidence (i.e., $x_- > x_0$). Indeed, otherwise (i.e., when $x_- < x_0$) the negative-evidence type t_- will pretend he is t_0 and has no evidence and we are back to the no-separation case. Since the value 30 of t_- is lower than the value 60 of t_0, setting a higher salary for t_- than for t_0 cannot be optimal (indeed, decreasing x_- and/or increasing x_0 is always better for the dean, as it sets the salary of at least one type closer to its value). The conclusion is that an optimal mechanism *cannot separate* t_- from t_0, and so the unique optimal policy is identical to the equilibrium outcome, which is obtained without commitment.[4] □

The following slight variant of Example 1 shows the use of truth-leaning; the requirement of being a sequential equilibrium no longer suffices here.

Example 2 Replace the positive-evidence type of Example 1 by two types: a (new) positive-evidence type t_+ with value 102 and probability 20%, and a "medium-evidence" type $t_\pm$ with value 42 and probability 5%. The type $t_\pm$ has two pieces of evidence: one is the same positive evidence that t_+ has, and the other is the same negative evidence that t_- has (for example, an acceptance decision on one paper, and a rejection decision on another). Thus, $t_\pm$ may pretend to be any one of the four types $t_\pm, t_+, t_-$, or t_0. In the sequential equilibrium that is similar to that of Example 1, types t_+ and $t_\pm$ both provide positive evidence and get the salary $x_+ = 90$, and types t_0 and t_- provide no evidence and get the salary $x_0 = 50$ (the salaries are equal to the corresponding expected values). It is not difficult to see that this is also the optimal mechanism outcome.

Now, however, the so-called "uninformative equilibrium" (also known as "babbling equilibrium") where the professor, regardless of his type, never provides any evidence, and the dean ignores any evidence that might be

[4]By contrast, t_+ is separated from t_0, because the value of t_+ is *higher*. In general, separation of types that have more evidence from types that have less evidence can occur in an optimal mechanism *only* when the former have higher values than the latter (since someone with more evidence can pretend to have less evidence, but not the other way around). In short, *separation requires that more evidence be associated with higher value*. See Corollary 4 for a formal statement of this property, which is at the heart of our argument.

provided and sets the salary to the average value of 60—which is worse for the dean, as it yields no separation between the types—is also a sequential equilibrium. This equilibrium is supported by the dean's belief that it is much more probable that the out-of-equilibrium positive evidence is provided by $t_{\pm}$ rather than by t_{+}; such a belief, while possible in a sequential equilibrium, appears hard to justify. The uninformative equilibrium is *not*, however, a truth-leaning equilibrium, as truth-leaning implies that the out-of-equilibrium message t_{+} is used infinitesimally by type t_{+} (for which it is the whole truth), and so the reward there must be set to 102, the value of t_{+}. □

Now, the simplicity of the above examples may be misleading, as in general the equilibria can be quite complex and involve no easy unravelings and thresholds (e.g., Examples 7 and 10 in Appendix B, where the agent's strategy must be mixed). Finally, for a simple illustration of how commitment may yield outcomes that are strictly better than anything that can be achieved without it, see Example 3 in Appendix B.1: it is a slight variant of the above examples but with the professor's utility depending on type (and so it does *not* belong to the class of evidence games).

II Related Literature

There is an extensive and insightful literature addressing the interaction between a principal who takes a decision but is uninformed and an agent who is informed and communicates information, either explicitly (through messages) or implicitly (through actions). Separation between different types of the agent may indeed be obtained when the types have different utilities or costs (as in signaling, screening, and cheap-talk setups).

When different types have different possible actions—such as different sets of messages—separation may be obtained even when the agent's utility and cost are the same regardless of his information. Grossman and Hart (1980), Grossman (1981), and Milgrom (1981), who initiated the "voluntary disclosure" literature, showed that unraveling obtains as a result of it being

commonly known that the agent is fully informed.

Disclosure in financial markets by public firms is a prime example of voluntary disclosure. This has led to a growing literature in accounting and finance. Dye (1985) and Woon-Oh Jung and Young Kwon (1988) study disclosure of accounting data. These are the first papers where it is no longer assumed that the agent (in this case, the firm, or, more precisely, the firm's manager) is known to be fully informed. They consider the case where the information is one-dimensional, and show that there is no longer unraveling in equilibrium. Shin (2003, 2006), Ilan Guttman, Ilan Kremer, and Andrzej Skrzypacz (2014), and Suil Pae (2005) consider an evidence structure in which information is multi-dimensional. Since such models typically possess multiple equilibria, these papers focus on what they view as the more natural equilibrium. The selection criteria that they employ are model-specific. However, it may be verified that all these selected equilibria are in fact truth-leaning equilibria, and so truth-leaning turns out to be a natural way to unify all these criteria.

In the mechanism-design framework where the principal commits to a reward policy before the agent's message is sent, Green and Laffont (1986) were the first to consider the setup where types differ in the sets of possible messages that they can send. They show that a necessary and sufficient condition for the revelation principle to hold for any payoff functions is that the message structure be transitive (they call this the "nested range condition")—which is satisfied by the voluntary disclosure models, as well as by our more general evidence games. Elchanan Ben-Porath and Bart Lipman (2012), Navin Kartik and Olivier Tercieux (2012), and Frederic Koessler and Eduardo Perez-Richet (2014) characterize the social choice functions that can be implemented when agents can also supply hard proofs about their types. Our social objective can be viewed as maximizing the fit between types and rewards.

The issue of comparing equilibria and mechanisms originated in Glazer and Rubinstein (2004, 2006). They analyze the optimal mechanism-design problem for general type-dependent message structures, with the principal taking a binary decision of "accepting" or "rejecting"; the agent, regardless

of his type, prefers acceptance to rejection. They show that the resulting optimal mechanism can be supported as an equilibrium outcome; Sher (2011) extended the result to the case in which the decision is no longer binary, provided that the principal's payoff is concave. By comparison, our paper shows that, in the framework of an agent with type-independent utility, the addition of the "truth structure" of evidence games—by which we mean the partial truth relation together with the truth-leaning behavior—yields the stronger result of the equivalence between the resulting equilibria and optimal mechanisms.[5] Finally, an example where commitment does not help in a disclosure game is included in Sourav Bhattacharya and Arijit Mukherjee (2013).

III The Model

There are two players, an *agent* "A" and a *principal* "P." The agent's information is his *type* t, which belongs to a finite set T, and is chosen according to a given probability distribution $p = (p_t)_{t \in T}$ in $\Delta(T)$, the set of probability distributions on T, with $p_t > 0$ for all $t \in T$. The agent knows the realized type t in T, whereas the principal knows only the distribution p but not the realized type.

The general structure of the interaction is that the agent sends a *message*, which consists of a type s in T, and the principal chooses an *action*, which is a real number x in $\mathbb{R}$. The message is costless: it does not affect the payoffs of the agent and the principal. An interpretation to keep in mind is that the type corresponds to the (verifiable) evidence that the agent possesses, and the message corresponds to the evidence that he reveals.

III.A Payoffs and Single-Peakedness

A fundamental assumption of the model (which distinguishes it from the signaling and cheap-talk setups) is that all the types of the agent have the

[5] Our companion paper Hart, Kremer, and Perry (2016) that deals with randomized rewards discusses in detail the connections to the work of Glazer–Rubinstein and Sher.

same preference, which is strictly increasing in x (and does not, as already stated, depend on the message sent). Without loss of generality (only the ordinal preference matters here) we assume that the agent's payoff is x itself, and refer to x as the *reward* (to the agent).

As for the principal, his utility does depend on the type t, but, again, not on the message s; thus, let $h_t(x)$ be the principal's utility for type $t \in T$ and reward $x \in \mathbb{R}$ (and any message $s \in T$). For every probability distribution $q = (q_t)_{t \in T} \in \Delta(T)$ on the set of types T—think of q as a "belief" on types—the expected utility of the principal is given by $h_q(x) := \sum_{t \in T} q_t \, h_t(x)$ for each $x \in \mathbb{R}$. The functions h_t are assumed to be *differentiable* and to satisfy:

(SP) *Single-Peakedness.* For every $q \in \Delta(T)$ the principal's expected utility $h_q(x)$ is a single-peaked function of the reward x.

A differentiable real function $f : \mathbb{R} \to \mathbb{R}$ is *single-peaked* if there exists a point $v \in \mathbb{R}$ such that $f'(v) = 0$; $f'(x) > 0$ for $x < v$; and $f'(x) < 0$ for $x > v$. Thus f has a global maximum at v, is strictly increasing for $x \leq v$, and strictly decreasing for $x \geq v$.

Condition (SP) requires all functions h_t, as well as all their weighted averages, to be single-peaked. Let $v(t)$ and $v(q)$ denote the single peaks of h_t and h_q, respectively. Then $v(t)$ is the reward that the principal views as most fitting ("ideal") for type t; or, the "value" to the principal of t (as in the examples in Section I). Similarly, $v(q)$ is the ideal reward, or the value, when the types are distributed according to q.

Some instances where the single-peakedness condition (SP) holds are:

• *Basic example: Quadratic loss.* Each h_t is the quadratic distance from the ideal point: $h_t(x) = -(x - v(t))^2$. In this case, common in much of the literature, the peak of h_q is easily seen to be the expectation with respect to q of the peaks $v(t)$; i.e., $v(q) = \sum_{t \in T} q_t \, v(t)$.

• *Strict concavity.* Each h_t is a strictly concave function that attains its (unique) maximum at a finite point (which implies that the same holds for any weighted averages of such functions). For instance, h_t is the negative of some distance (not necessarily quadratic) from the ideal point $v(t)$.

• *Monotonic transformations.* Apply a strictly increasing transformation to the variable x, which preserves (SP) (but not concavity).

• Treat types differently, such as making different h_t more or less sensitive to the distance from the corresponding ideal point $v(t)$; e.g., $h_t(x) = -c_t|x - v(t)|^{\gamma_t}$ (with $c_t > 0$ and $\gamma_t > 1$, so as to get strict concavity). Also, the penalties for underestimating vs. overestimating the desired ideal point may be different: take the function h_t to be asymmetric around $v(t)$.

We assume here that there are no further randomizations on the reward x. In case lotteries on x are allowed, the above single-peakedness condition is no longer sufficient and needs to be adapted; we analyze this in the companion paper Hart, Kremer, and Perry (2016). When all the functions h_t are concave, the restriction to pure rewards is easily seen to be without loss of generality: replace every lottery by its expectation.

We conclude with a useful property of single-peakedness.

In-betweenness property of the peaks. Let $x_0 := \min_{t \in T} v(t)$ and $x_1 := \max_{t \in T} v(t)$; because all the functions $h_t(x)$ are strictly increasing for $x \leq x_0$ and strictly decreasing for $x \geq x_1$, the peaks $v(q)$ for all $q \in \Delta(T)$ satisfy $x_0 \leq v(q) \leq x_1$. More generally, if q is a weighted average of $q_1, q_2, ..., q_n$ in $\Delta(T)$, i.e., $q = \sum_{i=1}^n \lambda_i \, q_i$ with $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i > 0$ for all i, then

$$\min_{1 \leq i \leq n} v(q_i) \leq v(q) \leq \max_{1 \leq i \leq n} v(q_i) \tag{1}$$

(indeed, all the functions $h_{q_i}(x)$, and hence also $h_q(x) = \sum_{i=1}^n \lambda_i \, h_{q_i}(x)$, are strictly increasing for $x \leq \min_i v(q_i)$ and strictly decreasing for $x \geq \max_i v(q_i)$). In particular, if T is partitioned into disjoint nonempty subsets $T_1, T_2, ..., T_n$ then $\min_{1 \leq i \leq n} v(T_i) \leq v(T) \leq \max_{1 \leq i \leq n} v(T_i)$, where $v(T)$ stands for $v(p)$ and $v(T_i)$ for $v(p|T_i)$ (with p the prior and $p|T_i$ the conditional of p given T_i). The rewards may thus be restricted to the compact interval $X = [x_0, x_1]$ that contains all the peaks: any reward x outside X is strictly dominated for the principal (by x_0 when $x < x_0$ and by x_1 when $x > x_1$).

III.B Evidence and Truth

The agent's message may be only partially truthful and he need not reveal everything that he knows; however, he cannot transmit false evidence, as any evidence disclosed is assumed to be verifiable. Thus, the agent must "tell the truth and nothing but the truth," but not necessarily "the whole truth."

Let E be the set of (verifiable) pieces of evidence. A type t is identified with a subset E_t of E, namely, the set of pieces of evidence that the agent of type t can provide (e.g., prove in court). The possible messages of t are then either to provide all the evidence that he has (E_t , "the whole truth"), or to pretend to be another type s with less evidence (i.e., $E_s \subseteq E_t$) and provide only the pieces of evidence in E_s (a "partial truth").[6] Thus the set of possible messages of the agent when the type is t, which we denote by $L(t)$, is identified with the set of types that have less evidence (in the weak sense) than t, i.e., $L(t) := \{s \in T : E_s \subseteq E_t\}$. This is immediately seen to entail two conditions:

(L1) $t \in L(t)$ for every type $t \in T$;

(L2) if $s \in L(t)$ and $r \in L(s)$ then $r \in L(t)$.

(L1) says that revealing the whole truth is always possible: t can always say t. (L2) is a transitivity condition: if s has less evidence than t and r has less evidence than s, then r has less evidence than t; that is, if t can say s and s can say r then t can also say r. These conditions are standard; see for instance Green and Laffont (1986), Jesse Bull and Joel Watson (2007), and Appendix C.4. From now on we abstract away from any specific setup and just assume (L1) and (L2).

Remark. A type t thus has two characteristics: his value to the principal (expressed by the function h_t and its peak $v(t)$) and the evidence that he can provide (expressed by $L(t)$). We emphasize that *no relation is assumed*

[6]The restriction that messages correspond to undetectable deviations (i.e., possible types) is without loss of generality: Proposition 7 in Appendix C.4 shows that the equivalence result continues to hold when additional messages are allowed.

between value and evidence; in particular, having more evidence need not be associated with having a higher (or lower) value.

III.C Game and Equilibria

We start by considering the *game* Γ where the principal moves after the agent (and cannot commit to a policy). First, the type $t \in T$ is chosen according to the probability measure $p \in \Delta(T)$, and revealed to the agent but not to the principal. The agent then sends to the principal one of the possible messages s in $L(t)$. Finally, after receiving the message s, the principal decides on a reward $x \in \mathbb{R}$.

A strategy σ of the agent associates with every type $t \in T$ a probability distribution $\sigma(\cdot|t) \in \Delta(T)$ with support included in $L(t)$; i.e., $\sigma(s|t)$, which is the probability that type t sends the message s, satisfies $\sigma(s|t) > 0$ only if $s \in L(t)$. A strategy ρ of the principal assigns to every message $s \in T$ a reward $\rho(s) \in \mathbb{R}$.

A pair of strategies (σ, ρ) constitutes a *Nash equilibrium* of the game Γ if the agent uses only messages that maximize the reward, and the principal sets the reward to each message optimally given the distribution of types that send that message. That is, for every message $s \in T$ let $\bar{\sigma}(s) := \sum_{t \in T} p_t \, \sigma(s|t)$ be the probability that s is used; if $\bar{\sigma}(s) > 0$ let $q(s) \in \Delta(T)$ be the conditional distribution of types that chose s, i.e., $q_t(s) := p_t \, \sigma(s|t)/\bar{\sigma}(s)$ for every $t \in T$ (this is the posterior probability of type t given the message s), and $q(s) = (q_t(s))_{t \in T}$. Thus, the equilibrium conditions for the agent and the principal are, respectively:

(A) for every type $t \in T$ and message $s \in T$: if $\sigma(s|t) > 0$ then $\rho(s) = \max_{s' \in L(t)} \rho(s')$;

(P) for every message $s \in T$: if $\bar{\sigma}(s) > 0$ then $h_{q(s)}(\rho(s)) = \max_{x \in \mathbb{R}} h_{q(s)}(x)$ (and so $\rho(s) = v(q(s))$ by the single-peakedness condition).

The *outcome* of a Nash equilibrium (σ, ρ) is the resulting vector of rewards

$\pi = (\pi_t)_{t \in T} \in \mathbb{R}^T$, where, for every type $t \in T$,

$$\pi_t := \max_{s \in L(t)} \rho(s); \tag{2}$$

when the type is t the payoffs are π_t for the agent and $h_t(\pi_t)$ for the principal.

III.D Truth-Leaning Equilibria

As discussed in the introduction, evidence games may have many equilibria; we are interested in those in which truth enjoys a certain prominence. This is expressed in two ways. First, if it is optimal for the agent to reveal the whole truth, then he prefers to do so (this holds for instance when the agent has a "lexicographic" preference: he always prefers a higher reward, but if the reward is the same whether he tells the whole truth or not, he prefers to tell the whole truth). Second, there is an infinitesimal probability that the whole truth is revealed (which happens, for example, when the agent is not strategic and instead always reveals his information; or, when there are "trembles," such as a slip of the tongue, or of the pen, or a document that is attached by mistake, or the surfacing of an unexpected piece of evidence).

To formalize this we use a standard limit-of-small-perturbations approach. Specifically, given $\varepsilon_t > 0$ and $\varepsilon_{t|t} > 0$ for every $t \in T$ (denote such a collection of ε-s by $\boldsymbol{\varepsilon}$), let $\Gamma^{\boldsymbol{\varepsilon}}$ denote the following perturbation of the game Γ. First, the agent's payoff increases by ε_t when the type is t and the message s is equal to the type t; i.e., his payoff is equal to the reward x when $s \neq t$, and to $x + \varepsilon_t$ when $s = t$. Second, the agent's strategy σ is required to satisfy $\sigma(t|t) \geq \varepsilon_{t|t}$ for every type $t \in T$. The agent thus gets an ε_t "bonus" in payoff when he reveals the whole truth, and he must do so with probability at least $\varepsilon_{t|t}$. A Nash equilibrium (σ, ρ) of the original game Γ is *truth-leaning* if it is a limit point of Nash equilibria of $\Gamma^{\boldsymbol{\varepsilon}}$ as all the ε-s converge to 0; i.e., if there are sequences $\varepsilon_t^n \to_{n\to\infty} 0$, $\varepsilon_{t|t}^n \to_{n\to\infty} 0$, and $(\sigma^n, \rho^n) \to_{n\to\infty} (\sigma, \rho)$ such that (σ^n, ρ^n) is a Nash equilibrium of $\Gamma^{\boldsymbol{\varepsilon}^n}$ for every n.

In terms of the original game, truth-leaning turns out to be essentially equivalent to imposing the following two conditions on a Nash equilibrium

(σ, ρ) of Γ:

(A0) for every type $t \in T$: if $\rho(t) = \max_{s \in L(t)} \rho(s)$ then $\sigma(t|t) = 1$;

(P0) for every message $s \in T$: if $\bar{\sigma}(s) = 0$ then $h_s(\rho(s)) = \max_{x \in \mathbb{R}} h_s(x)$ (and so $\rho(s) = v(s)$ by the single-peakedness condition).

Condition (A0) says that when the message t is optimal for type t, it is chosen by t for sure (i.e., if the whole truth is optimal then it is strictly preferred to any other optimal message). Condition (P0) says that, for every message $s \in T$ that is *not used* in equilibrium (i.e., $\bar{\sigma}(s) = 0$), the principal's belief if he were to receive message s would be that it came from type s itself (since there is an infinitesimal probability that type s revealed the whole truth); thus the posterior belief $q(s)$ at s puts probability one on s, and so the principal's optimal response is the peak $v(s)$ of $h_{q(s)} \equiv h_s$. For a rough intuition, (A0) obtains from the positive bonus in payoff, and (P0) from the positive probability of revealing the type (if s is not used then it is not a best reply for s by (A0), and so for no other type by transitivity (L2), which implies that in Γ^ε only s itself uses s with positive probability). We state this formally in Proposition 1, which allows us to conveniently use only (A0) and (P0) in the remainder of the paper.

Proposition 1 *(i) Truth-leaning equilibria exist. (ii) For every truth-leaning equilibrium (σ, ρ) there is an equilibrium (σ', ρ) that satisfies (A0) and (P0) and has the same outcome π as (σ, ρ).*

The proof is relegated to Appendix A. Truth-leaning may thus be viewed as an equilibrium selection criterion (a "refinement"); alternatively, as part of the setup (the actual game being Γ^ε for small ε). In Appendix C.5 we will see that truth-leaning satisfies the requirements of most, if not all, the relevant equilibrium refinements that have been proposed in the literature.

III.E Mechanisms and Optimal Mechanisms

We come now to the second setup, where the principal moves first and *commits* to a reward scheme, i.e., to a function $\rho : T \to \mathbb{R}$ that assigns to every

message $s \in T$ a reward $\rho(s)$. The reward scheme ρ is made known to the agent, who then sends his message s, and the resulting reward is $\rho(s)$ (the principal's commitment to the reward scheme ρ means that he cannot change the reward after receiving the message s).

This is a standard *mechanism-design* framework. The reward scheme ρ is the *mechanism.* Given ρ, the agent chooses his message so as to maximize his reward; thus, the reward when the type is t equals $\pi_t := \max_{s \in L(t)} \rho(s)$. A reward scheme ρ is an *optimal mechanism* if it maximizes the principal's expected payoff

$$H(\pi) = \sum_{t \in T} p_t \, h_t(\pi_t) \tag{3}$$

among all mechanisms.

The assumptions that we have made on the truth structure, i.e., (L1) and (L2), are easily seen to imply that the "Revelation Principle" applies: any mechanism can be implemented by a "direct" mechanism in which it is optimal for each type to be "truthful" and reveal his type; see Green and Laffont (1986), or Appendix C.6. The incentive compatibility constraints are

(IC) $\pi_t \geq \pi_s$ for every $t, s \in T$ with $s \in L(t)$

(indeed, $s \in L(t)$ implies $L(t) \supseteq L(s)$ by the transitivity condition (L2), and so $\pi_t = \max_{r \in L(t)} \rho(r) \geq \max_{r \in L(s)} \rho(r) = \pi_s$). Thus an *optimal mechanism* outcome is a vector $\pi = (\pi_t)_{t \in T} \in \mathbb{R}^T$ that maximizes $H(\pi)$ subject to (IC).

IV The Equivalence Theorem

Our main result is

Theorem 2 (Equivalence Theorem) *There is a unique truth-leaning equilibrium outcome, a unique optimal mechanism outcome, and these two outcomes coincide.*

The intuition is roughly as follows. Consider a truth-leaning equilibrium where a type t pretends to be another type s. Then, first, type s reveals his

type s (had s something better, t would have it as well); and second, the value of s must be higher than the value of t (no one will want to pretend to be worth less than they really are).[7] Thus t and s are *not separated* in this equilibrium, and we claim that they *cannot be separated* in an optimal mechanism either: the only way for the principal to separate them would be to give a *higher* reward to t than to s (otherwise t would pretend to be s), which is not optimal since the value of t is lower than the value of s (decreasing the reward of t or increasing the reward of s would bring the rewards closer to the values). The conclusion is that optimal mechanisms can never separate more than truth-leaning equilibria do (the converse is immediate since whatever can be done without commitment can clearly also be done with commitment).

Remarks. *(a) Outcomes.* The Equivalence Theorem is stated in terms of outcomes—which uniquely determine the (ex-post) payoffs of both the agent and the principal for every type t. While there may be multiple truth-leaning equilibria, this can happen only when both players are indifferent, and then the payoffs are the same (see Appendix B.9).

(b) Tightness of the result. All the assumptions except differentiability are indispensable to the Equivalence Theorem: dropping any single condition yields examples where the result does not hold (see Appendix B). As for differentiability, it is only a convenient technical assumption, as the equivalence result holds also without it (see Appendix C.11).

(c) Constrained Pareto efficiency. In the basic quadratic-loss case, where, as we have seen, $v(q)$ equals the expectation of the values $v(t)$ with respect to q, condition (P) implies that the ex-ante expectation of the rewards, i.e., $\mathbb{E}[\pi_t] = \sum_{t\in T} p_t\,\pi_t$, equals the ex-ante expectation of the values $\mathbb{E}[v(t)] = \sum_{t\in T} p_t\,v(t) = v(T)$ (because $\mathbb{E}[\pi_t|s] = v(q(s)) = \mathbb{E}[v(t)|s]$ for every message s that is used; take expectation over s). Therefore all Nash equilibria yield to the agent the same ex-ante expected payoff $\mathbb{E}[\pi_t] = v(T)$ (they differ ex post, however, in the way this amount is split among the

[7] However reasonable these conditions may seem, they need *not* hold for equilibria that are not truth-leaning.

types). Since, by the Equivalence Theorem, the truth-leaning equilibria maximize the principal's ex-ante expected payoff, it follows that the truth-leaning equilibria are constrained Pareto efficient (i.e., ex-ante Pareto efficient among all equilibria).

V Proof of the Equivalence Theorem

The proof proceeds as follows. We start with some useful and interesting properties of truth-leaning (Section V.A), and then prove that the outcome of any truth-leaning equilibrium outcome is an optimal mechanism outcome, which is moreover unique (Section V.B). Together with the existence of truth-leaning equilibria (Proposition 1(i) in Section III.D) this yields the result.

V.A Preliminaries

Proposition 3 *Let (σ, ρ) be an equilibrium that satisfies (A0) and (P0), let π be its outcome, and let $S := \{t \in T : \bar{\sigma}(t) > 0\}$ be the set of messages used in equilibrium. Then*

$$t \in S \quad \Leftrightarrow \quad \sigma(t|t) = 1 \quad \Leftrightarrow \quad v(t) \geq \pi_t = \rho(t)\,; \quad \textit{and} \tag{4}$$

$$t \notin S \quad \Leftrightarrow \quad \sigma(t|t) = 0 \quad \Leftrightarrow \quad \pi_t > v(t) = \rho(t)\,. \tag{5}$$

Thus, the reward $\rho(t)$ assigned to message t never exceeds the peak $v(t)$ of type t. Moreover, each type t that reveals the whole truth gets an outcome that is at most his value (i.e., $\pi_t \leq v(t)$), whereas each type t that does not reveal the whole truth gets an outcome that exceeds his value (i.e., $\pi_t > v(t)$). This may sound strange at first. The explanation is that the lower-value types are the ones that have the incentive to pretend to be a higher-value type, and so each message t that is used is sent by t as well as by "pretenders" of lower value. In equilibrium, this effect is taken into account by the principal by rewarding messages at their true value or less.

Proof. If $t \in S$, i.e., $\sigma(t|t') > 0$ for some t', then t is a best reply for type t', and hence also for type t (because $t \in L(t) \subseteq L(t')$ by (L1), (L2), and

$t \in L(t')$); (A0) then yields $\sigma(t|t) = 1$. This proves the first equivalence in (4) and in (5).

If $t \notin S$ then $\pi_t > \rho(t)$ (since t is not a best reply for t) and $\rho(t) = v(t)$ by (P0), and hence $\pi_t > v(t) = \rho(t)$.

If $t \in S$ then $\pi_t = \rho(t)$ (since t is a best reply for t); put $\alpha := \pi_t = \rho(t)$. Let $t' \neq t$ be such that $\sigma(t|t') > 0$; then $\pi_{t'} = \rho(t) \equiv \alpha$ (since t is optimal for t'); moreover, $t' \notin S$ (since $\sigma(t|t') > 0$ implies $\sigma(t'|t') < 1$), and so, as we have just seen above, $v(t') < \pi_{t'} \equiv \alpha$. If we also had $v(t) < \alpha$, then the in-betweenness property (1) would yield $v(q(t)) < \alpha$ (because the support of $q(t)$, the posterior after message t, consists of t together with all $t' \neq t$ with $\sigma(t|t') > 0$). But this contradicts $v(q(t)) = \rho(t) \equiv \alpha$ by the principal's equilibrium condition (P). Therefore $v(t) \geq \alpha \equiv \pi_t = \rho(t)$.

Thus we have shown that $t \notin S$ and $t \in S$ imply contradictory statements ($\pi_t > v(t)$ and $\pi_t \leq v(t)$, respectively), which yields the second equivalence in (4) and in (5). ■

Corollary 4 *Let (σ, ρ) be an equilibrium that satisfies (A0) and (P0). If $\sigma(s|t) > 0$ for $s \neq t$ then $v(s) > v(t)$.*

Proof. $\sigma(s|t) > 0$ implies $s \in S$ and $t \notin S$, and thus $v(s) \geq \rho(s)$ by (4), $\pi_t > v(t)$ by (5), and $\rho(s) = \pi_t$ because s is a best reply for t. ■

Thus, no type will ever pretend to be a lower-valued type (this does not, however, hold for equilibria that are *not truth-leaning*, e.g., the uninformative equilibrium in Example 2 in Section I).

V.B From Equilibrium to Mechanism

This section proves that any truth-leaning equilibrium outcome is an optimal mechanism outcome and, moreover, that the latter is unique. We first deal with a special case where there is no separation, and then show how a truth-leaning equilibrium yields a decomposition into instances of this special case.

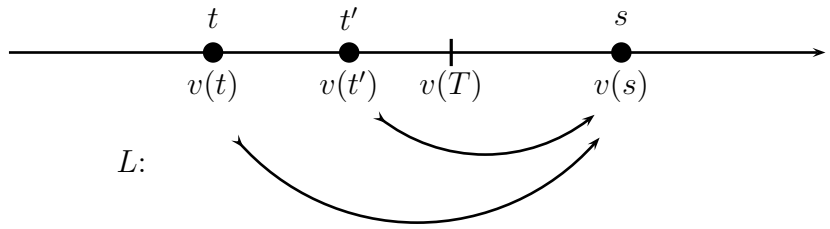
Proposition 5 *Assume that there is a type $s \in T$ such that $s \in L(t)$ for every t. If $v(t) < v(T)$ for every $t \neq s$ then the outcome π^* with $\pi^*_t = v(T)$*

for all $t \in T$ is the unique optimal mechanism outcome; i.e.,

$$\sum_{t \in T} p_t \, h_t(\pi_t) \leq \sum_{t \in T} p_t \, h_t(\pi_t^*) \tag{6}$$

for every incentive-compatible π, with equality if and only if $\pi_t = \pi_t^* = v(T)$ for all $t \in T$.

Thus every type can pretend to be s, and so s has the least amount of evidence (e.g., no evidence at all). The condition $v(t) < v(T)$ for every $t \neq s$ implies that $v(T) \leq v(s)$ by in-betweenness (1), and so $v(t) < v(s)$ for every $t \neq s$; see Figure 2. To get some intuition, consider the simplest case of only



IC: $\pi_s \leq \pi_t, \quad \pi_s \leq \pi_{t'}$

Figure 2: Proposition 5

two types, say, $T = \{s, t\}$. Because the (IC) constraint $\pi_t \geq \pi_s$ goes in the opposite direction of the peaks' inequality $v(t) < v(s)$, it follows that the maximum of $H(\pi) = p_s h_s(\pi_s) + p_t h_t(\pi_t)$ subject to $\pi_t \geq \pi_s$ is attained only when π_t and π_s are equal. Indeed, if $\pi_t > \pi_s$ then we must have $\pi_t > v(t)$ or $\pi_s < v(s)$, and so decreasing π_t or increasing π_s brings it closer to the corresponding peak, and hence increases the value of H. Thus $\pi_t = \pi_s = x$ for some x, and then the maximum is attained when x equals the peak of $h_p(x) = p_s h_s(x) + p_t h_t(x)$, i.e., when $x = v(T)$.

Proof. First, $v(t) < v(T)$ for all $t \neq s$ implies by in-betweenness (1) that $v(R) \geq v(T)$ for every set $R \subseteq T$ that contains s. Next, let π maximize $H(\pi)$

subject to the (IC) constraints; we will show that π must equal π^* (which satisfies all (IC) constraints, as equalities).

Put $\alpha := \min_t \pi_t$ and $R := \{r \in T : \pi_r = \alpha\}$. Because one may change the common value of π_r for all $r \in R$ to any α' close enough to α so that all (IC) inequalities continue to hold (specifically, $\alpha' \leq \beta$ where $\beta := \min_{t \notin R} \pi_t > \alpha$), the optimality of π implies that α must maximize $\sum_{t \in R} p_t h_t(x) = p(R) h_R(x)$, and so $\alpha = v(R)$. But R contains s (because the (IC) constraints include $\pi_s \leq \pi_t$ for all $t \neq s$), and so $\alpha = v(R) \geq v(T)$. Therefore $H(\pi) = \sum_t p_t h_t(\pi_t) \leq \sum_t p_t h_t(\alpha) = h_T(\alpha) \leq h_T(v(T)) = \sum_t p_t h_t(\pi^*_t) = H(\pi^*)$ (the first inequality because $\pi_s = \alpha$, and for $t \neq s$ the function $h_t(x)$ decreases after its peak $v(t)$ and $\pi_t \geq \alpha \geq v(T) > v(t)$; the second inequality because $h_T(x)$ decreases after its peak $v(T)$ and $\alpha \geq v(T)$). Moreover, all the above functions are strictly decreasing after their peaks, and so to get equalities throughout we must have $\pi_t = \alpha = v(T)$ for all t, i.e., $\pi = \pi^*$. ■

Proposition 6 *Let π^* be a truth-leaning equilibrium outcome; then π^* is the unique optimal mechanism outcome.*

Proof. Let (σ, ρ) be an equilibrium that satisfies (A0) and (P0) and has outcome π^* (by Proposition 1). Because π^* satisfies (IC) by (L2), we need to show that $H(\pi^*) > H(\pi)$ for every $\pi \neq \pi^*$ that satisfies (IC).

Let $S := \{s \in T : \bar{\sigma}(s) > 0\}$ be the set of messages that are used in the equilibrium (σ, ρ), and, for each $s \in S$, let $T_s := \{t \in T : \sigma(s|t) > 0\}$ be the set of types that play s. For every $t \neq s$ in T_s we then have $s \in L(t)$ and $t \notin S$ (because $\sigma(s|t) < 1$ implies $\sigma(t|t) < 1$), and so $v(t) = \rho(t) < \pi^*_t = \pi^*_s = \rho(s) = v(q(s))$ (by (5) and (4) in Proposition 3, and the principal's equilibrium condition (P)). We can therefore apply Proposition 5 to the set of types T_s with the distribution $q(s)$ as prior, to get (6) for every π that satisfies (IC), with equality only if $\pi_t = \pi^*_t$ for every $t \in T_s$.

For any $\pi \in \mathbb{R}^T$, the principal's payoff $H(\pi)$ can be decomposed as

$$H(\pi) = \sum_{t \in T} p_t \, h_t(\pi_t) = \sum_{s \in S} \bar{\sigma}(s) \sum_{t \in T_s} q_t(s) \, h_t(\pi_t). \tag{7}$$

Multiplying (6) by $\bar{\sigma}(s) > 0$ and summing over $s \in S$ therefore yields $H(\pi) \leq H(\pi^*)$ for every π that satisfies (IC) (use (7) for both π and π^*). Moreover, to get equality we need equality in (6) for each $s \in S$; that is, $\pi_t = \pi_t^*$ for every $t \in \cup_{s \in S} T_s = T$. ■

References

Akerlof, George A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84, 488–500.

Ben-Porath, Elchanan and Lipman, Bart (2012), "Implementation with Partial Provability," *Journal of Economic Theory* 147, 1689–1724.

Bhattacharya, Sourav and Mukherjee, Arijit (2013), "Strategic Information Revelation when Experts Compete to Influence, " *RAND Journal of Economics* 44, 522–544.

Bull, Jesse and Watson, Joel (2007), "Hard Evidence and Mechanism Design, *Games and Economic Behavior* 58, 75–93.

Crawford, Vincent P. and Sobel, Joel (1982), "Strategic Information Transmission," *Econometrica* 50, 1431–1451.

Dye, Ronald A. (1985), "Strategic Accounting Choice and the Effect of Alternative Financial Reporting Requirements," *Journal of Accounting Research* 23, 544–574.

Glazer, Jacob and Rubinstein, Ariel (2004), "On Optimal Rules of Persuasion," *Econometrica* 72, 1715–1736.

Glazer, Jacob and Rubinstein, Ariel (2006), "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach," *Theoretical Economics* 1, 395–410.

Goltsman, Maria, Hörner, Johannes, Pavlov, Gregory, and Squintani, Francesco (2009), "Mediation, Arbitration and Negotiation," *Journal of Economic Theory* 144, 1397–1420.

Green, Jerry R. and Laffont, Jean-Jacques (1986), "Partially Verifiable Information and Mechanism Design," *The Review of Economic Studies* 53, 447–456.

Grossman, Sanford J. (1981), "The Informational Role of Warranties and Private Disclosures about Product Quality," *Journal of Law and Economics* 24, 461–483.

Grossman, Sanford J. and Hart, Oliver (1980), "Disclosure Laws and Takeover Bids," *Journal of Finance* 35, 323–334.

Guttman, Ilan, Kremer, Ilan, and Skrzypacz, Andrzej (2014), "Not Only What but also When: A Theory of Dynamic Voluntary Disclosure," *American Economic Review,* forthcoming.

Hart, Sergiu, Kremer, Ilan, and Perry, Motty (2016), "Evidence Games with Randomized Rewards," working paper.

Kartik, Navin and Tercieux, Olivier (2012), "Implementation with Evidence," *Theoretical Economics* 7, 323–355.

Jung, Woon-Oh and Kwon, Young K. (1988), "Disclosure When the Market Is Unsure of Information Endowment of Managers," 26, 146–153.

Krishna, Vijay and Morgan, John (2007), "Cheap Talk," in *The New Palgrave Dictionary of Economics,* 2nd Edition.

Koessler, Frederic and Perez-Richet, Eduardo (2014), "Evidence Based Mechanisms," working paper.

Milgrom, Paul R. (1981), "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics* 12, 350–391.

Pae, Suil (2005), "Selective Disclosures in the Presence of Uncertainty About Information Endowment," *Journal of Accounting and Economics* 39, 383–409.

Sher, Itai (2011), "Credibility and Determinism in a Game of Persuasion," *Games and Economic Behavior* 71, 409–419.

Shin, Hyun Song (2003), "Disclosures and Asset Return," *Econometrica* 71, 105–133.

Shin, Hyun Song (2006), "Disclosures Risk and Price Drift," *Journal of Accounting Research* 44, 351–379.

A Appendix: Proof of Proposition 1

We prove here Proposition 1 in Section V.A: first, the existence of truth-leaning equilibria, and second, their payoff equivalence to equilibria that satisfy (A0) and (P0). The former is a standard fixed-point proof, while the latter turns out to be somewhat more delicate than the intuitive arguments in Section V.A may suggest; in particular, it uses the differentiability of the functions[8] h_t.

Proof of Proposition 1. *(i) Existence.* First, a standard fixed-point argument shows that the game $\Gamma^{\boldsymbol{\epsilon}}$ possesses a Nash equilibrium. Let $\Sigma^{\boldsymbol{\epsilon}}$ be the set of strategies of the agent in $\Gamma^{\boldsymbol{\epsilon}}$; then $\Sigma^{\boldsymbol{\epsilon}}$ is a compact and convex subset of $\Delta(T)^T$. Every σ in Σ^{ε} uniquely determines the principal's best reply $\rho \equiv \rho^{\sigma}$ by $\rho^{\sigma}(s) = v(q(s))$ for every $s \in T$ (cf. (P); in $\Gamma^{\boldsymbol{\epsilon}}$ every message is used: $\bar{\sigma}(s) \geq \varepsilon_s p_s > 0$). The mapping from σ to ρ^{σ} is continuous: the posterior $q(s) \in \Delta(T)$ is a continuous function of σ (because $\bar{\sigma}(s)$ is bounded away from 0), and $v(q)$ is a continuous function of q (by the Maximum Theorem together with the single-peakedness condition (SP), which gives the uniqueness of the maximizer). The set-valued function Φ that maps each $\sigma \in \Sigma^{\boldsymbol{\epsilon}}$ to the set of all $\sigma' \in \Sigma^{\boldsymbol{\epsilon}}$ that are best replies to ρ^{σ} in $\Gamma^{\boldsymbol{\epsilon}}$ is therefore upper hemicontinuous, and a fixed point of Φ, whose existence is guaranteed by the Kakutani fixed-point theorem, is precisely a Nash equilibrium of $\Gamma^{\boldsymbol{\epsilon}}$.

Second, the strategy sets of the two players are compact (for the principal, see in-betweenness in Section III.A), and so limit points of Nash equilibria of $\Gamma^{\boldsymbol{\epsilon}}$—i.e., truth-leaning equilibria of Γ—exist (it is immediate to verify that any limit point of Nash equilibria of $\Gamma^{\boldsymbol{\epsilon}}$ is a Nash equilibrium of Γ, i.e., satisfies (A) and (P)).

(ii) (A0) and (P0). Let (σ, ρ) be a truth-leaning equilibrium, given by sequences $\varepsilon_t^n \to_n 0^+$, $\varepsilon_{t|t}^n \to 0^+$, and $(\sigma^n, \rho^n) \to_n (\sigma, \rho)$ such that (σ^n, ρ^n) is a Nash equilibrium in $\Gamma^{\boldsymbol{\epsilon}^n}$ for every n (which is easily seen to imply that (σ, ρ) is a Nash equilibrium of Γ, i.e., that (A) and (P) hold).

Let t be such that $\sigma(t|t) < 1$. Then $\sigma(s|t) > 0$ for some $s \neq t$ in $L(t)$, and

[8] See Appendix C.11 for the nondifferentiable case.

so $\sigma^n(s|t) > 0$ for all (large enough) n. In $\Gamma^{\boldsymbol{\varepsilon}^n}$ we thus have: s is a best reply for t, hence $\rho^n(s) \geq \rho^n(t) + \varepsilon_t^n > \rho^n(t)$, hence t is not optimal for any $r \neq t$ (because $t \in L(r)$ implies $s \in L(r)$ by transitivity (L2) of L and s gives to r a strictly higher payoff than t in $\Gamma^{\boldsymbol{\varepsilon}^n}$), and thus $\sigma^n(t|s) = 0$. Taking the limit yields:

$$\text{if } \sigma(t|t) < 1 \text{ then } \sigma(t|s) = 0 \text{ for all } s \neq t; \tag{8}$$

this says that if t does not choose t for sure, then no other type chooses t. Moreover, the posterior $q^n(t)$ after message t puts all the mass on t (since $\sigma^n(t|t) \geq \varepsilon_{t|t}^n > 0$ whereas $\sigma^n(t|s) = 0$ for all $s \neq t$), i.e., $q^n(t) = \mathbf{1}_t$, and so $\rho^n(t) = v(q^n(t)) = v(t)$; in the limit:

$$\text{if } \sigma(t|t) < 1 \text{ then } \rho(t) = v(t). \tag{9}$$

This in particular yields (P0), because $\bar{\sigma}(t) = 0$ implies $\sigma(t|t) = 0 < 1$.

To get (A0) we may need to modify σ slightly, as follows. Let $t \in T$ be such that t is a best reply for t (i.e., $\rho(t) = \max_{s \in L(t)} \rho(s)$) but $\sigma(t|t) < 1$. Then $\rho(t) = v(t)$ by (9), and every message $s \neq t$ that t uses, i.e., $\sigma(s|t) > 0$, gives the same reward as message t, and so $v(q(s)) = \rho(s) = \rho(t) = v(t)$. Therefore we define σ' to be identical to σ except that type t chooses only message t; i.e., $\sigma'(t|t) = 1$ and $\sigma'(s|t) = 0$ for every $s \neq t$.

Let $q'(s)$ be the new posterior after a message $s \neq t$ that was used by t (i.e., $\sigma(s|t) > 0$; note that $\bar{\sigma}'(s) \geq p_s > 0$ since $\sigma'(s|s) = \sigma(s|s) = 1$ by (8) applied to s). Let $\alpha := v(q(s)) = v(t)$ (see above); using the differentiability of the functions h_r we will show that the peak of $h_{q'(s)}$ is also at[9] α. Indeed, $q(s)$ is a weighted average of $q'(s)$ and $\mathbf{1}_t$, and so $h_{q(s)}$ is a weighted average of $h_{q'(s)}$ and h_t. The derivatives of $h_{q(s)}$ and h_t both vanish at α, and so the derivative of $h_{q'(s)}$ must also vanish there—thus $v(q'(s)) = \alpha = v(q(s) = v(t)$.

It follows that (σ', ρ) is a Nash equilibrium of Γ: the agent is indifferent between the messages t and s, and the principal maximizes his payoff also at the new posterior $q'(s)$. Clearly (8) and (9), and hence (P0), continue

[9]Example 12 in Appendix C.11 shows that this property need *not* hold without differentiability. The argument below amounts to *strict* in-betweenness; see Appendix C.3.

to hold; moreover, the outcome remains the same. Proceeding this way for every t as needed will in the end yield also (A0). ■

B Appendix: Tightness of the Equivalence Theorem

We will show here that our Equivalence Theorem is tight. First, we show that dropping any single assumption (except for differentiability, which is assumed for convenience; see Appendix C.11) allows examples where the equivalence between optimal mechanisms and truth-leaning equilibria does not hold (Sections B.1 to B.7). Second, we show that the conclusions cannot be strengthened; specifically, truth-leaning equilibria need be neither pure nor unique (Sections B.8 and B.9).

B.1 Agent's Payoffs Depend on Type

We provide a slight variant of the examples of Section I—which can also be easily restated in the standard cheap-talk setup (Vincent Crawford and Joel Sobel 1982) —that shows that the equivalence result may fail when the agent's types do not all have the same preference: commitment strictly helps here.

Example 3 There are only two types of professor, and they are equally likely: t_0, with no evidence and value 60, and t_-, with negative evidence and value 30. As above, the dean wants to set the salary as close as possible to the value, and t_0 wants as high a salary as possible. However, t_- now wants his salary to be as close as possible to 50 (for instance, getting too high a salary would entail duties that he does not like): his utility when he gets salary x is $-(x-50)^2$.

There can be no separation between the two types in equilibrium: when no evidence is provided the salary is between 45 and 60 (the posterior probability of t_-, which depends on his probability of providing no evidence, is at most $1/2$, and so the resulting average of 30 and 60 is at least 45); but any salary in that range is strictly preferred by t_- to 30, which is what he gets when he reveals his evidence. Thus the uninformative equilibrium where no evidence is provided and the salary is set to 45, the average of the two values, is the unique Nash equilibrium.

Consider now the mechanism where the salary policy is to pay 30 when negative evidence is provided, and 75 when no evidence is provided. Since t_- prefers 30 to 75, he will reveal his evidence, and so separation is obtained. The mechanism outcome is better for the dean than the equilibrium outcome (he makes an error of 15 for t_0 only in the mechanism, and an error of 15 for *both* types in equilibrium). Note that the above mechanism requires the dean to *commit* to pay 75 when he gets no evidence; otherwise, after getting no evidence (which happens when the type is t_0), he will want to change his decision and pay 60 instead. In general, commitment is required when implementing reward schemes that are *not ex-post optimal* (our result implies that this does *not* happen in evidence games; the requirement that is *not* satisfied in Example 3 is that the agent's utility be the same for all types). □

Remarks. *(a)* Two optimal mechanisms are as follows: the salaries are set to 30 for negative evidence and 70 for no evidence in the first, and to 40 and 60, respectively, in the second; in both mechanisms, t_-, who is indifferent between revealing and concealing his evidence, reveals it.

(b) Taking the utility of t_0 to be $-(x-80)^2$, which does not affect the example, sets it in the standard Crawford and Sobel (1982) cheap-talk setup. The fact that commitment may be advantageous in cheap-talk games is known; see Vijay Krishna and John Morgan (2007) and Maria Goltsman, Johannes Hörner, Gregory Pavlov, and Francesco Squintani (2009).

B.2 Without Reflexivity (L1)

We provide an example where the condition (L1) that $t \in L(t)$ for all $t \in T$ is not satisfied—some type cannot tell the whole truth and reveal his type—and there is a truth-leaning Nash equilibrium whose payoffs are different from those of the optimal mechanism.

Example 4 The type space is $T = \{0, 2, 4\}$ with the uniform distribution: $p_t = 1/3$ for each $t \in T$. The principal's payoff functions are $h_t(x) = -(x-t)^2$,

and so $v(t) = t$ for all t. Types 0 and 2 have less evidence than type 4, but message 4 is *not* allowed; i.e., $L(0) = \{0\}$, $L(2) = \{2\}$, and $L(4) = \{0, 2\}$.

The unique optimal mechanism outcome is $\pi_0 = v(0) = 0$ and $\pi_2 = \pi_4 = v(\{2, 4\}) = 3$, i.e.,[10] $\pi = (\pi_0, \pi_2, \pi_4) = (0, 3, 3)$.

Truth-leaning entails no restrictions here: types 0 and 2 each have a single message (their type), and type 4 cannot send message 4. There are two Nash equilibria: (i) 4 sends message 2, $\rho(0) = 0$, $\rho(2) = 3$, with outcome $\pi = (0, 3, 3)$ (which is the optimal mechanism outcome); (ii) 4 sends message 0, $\rho(0) = 2$, $\rho(2) = 2$, with $\pi' = (2, 2, 2)$. Note that $H(\pi) > H(\pi')$. □

B.3 Without Transitivity (L2)

We provide an example where (L2) is not satisfied—the "less evidence" relation is not transitive—and there is a truth-leaning equilibrium outcome that is different from the optimal mechanism outcome.

Example 5 The type space is $T = \{0, 2, 4\}$ with the uniform distribution: $p_t = 1/3$ for each $t \in T$. The principal's payoff functions are $h_t(x) = -(x-t)^2$, and so $v(t) = t$ for all t. The allowed messages are $L(0) = \{0, 4\}$, $L(2) = \{2\}$, and $L(4) = \{2, 4\}$. This does not satisfy (L2): type 0 can send message 4 and type 4 can send message 2, but type 0 cannot send message 2.

The unique optimal mechanism is given by[11] the reward scheme $\rho = (0, 3, 0)$, with outcome $\pi = (0, 3, 3)$; indeed, if 2 and 4 are separated then it is best to set $\rho(2) = v(2) = 2$ and $\rho(4) = v(\{0, 4\}) = 2$, yielding the outcome $\pi' = (2, 2, 2)$; and if they are not separated then it is best to set $\rho(2) = v(\{2, 4\}) = 3$ and $\rho(0) = \rho(4) = v(0) = 0$, yielding the outcome $\pi = (0, 3, 3)$; the latter is better: $H(\pi) = -2/3 > -8/3 = H(\pi')$.

There is no equilibrium satisfying (A0) and (P0) with outcome π: type 0 must use 0 (by (A0), because $\rho(0) = \pi_0$), types 2 and 4 must use 2 (because

[10] When writing vectors such as π the coordinates are ordered according to increasing value; thus here we have $\pi = (\pi_0, \pi_2, \pi_4)$ (recall that $v(t) = t$).

[11] While type 0 can send message 4, he *cannot* fully mimic type 4, because he cannot send message 2, which type 4 can. The incentive-compatibility constraints can no longer be written as $\pi_t \geq \pi_s$ for $s \in L(t)$ as in Section III.E; they are $\pi_t = \max\{\rho(s) : s \in L(t)\}$ where $\rho : T \to \mathbb{R}$ is a reward scheme (cf. Green and Laffont 1986).

$\pi_2 = \pi_4 = 3$), but then 4 is unused and so $\rho(4) = v(4) = 4$ (by (P0)), contradicting (P).

Both π and π' are truth-leaning equilibrium outcomes:[12] take $\Gamma^{\boldsymbol{\varepsilon}}$ with $\varepsilon_t = \varepsilon_{t|t} = \varepsilon$ for all t; then π obtains from the limit of[13] $\sigma^\varepsilon(\cdot|0) = (\varepsilon, 0, 1-\varepsilon)$, $\sigma^\varepsilon(\cdot|4) = (0, 1-\varepsilon, \varepsilon)$, and $\rho^\varepsilon = (0, 3-\varepsilon/(2-\varepsilon), 4\varepsilon)$; and π' obtains from the limit of $\sigma^\varepsilon(\cdot|0) = (\varepsilon, 0, 1-\varepsilon)$, $\sigma^\varepsilon(\cdot|4) = (0, 0, 1)$, and $\rho^\varepsilon = (0, 2, 4/(2-\varepsilon)$. □

B.4 Without (A0)

We provide an example of a sequential equilibrium that does not satisfy the (A0) condition of truth-leaning, and whose outcome differs from the unique optimal mechanism outcome.

Example 6 The type space is $T = \{0, 2, 4\}$ with the uniform distribution: $p_t = 1/3$ for each $t \in T$. The principal's payoff functions are $h_t(x) = -(x-t)^2$ (and so $v(t) = t$) for each $t \in T$. Type 0 has less evidence than type 4, who has less evidence than type 2; i.e., $L(0) = \{0\}$, $L(2) = \{0, 2, 4\}$, and $L(4) = \{0, 4\}$.

The unique optimal mechanism outcome is $\pi = (0, 3, 3)$, and in the unique equilibrium that satisfies (A0) and (P0) types 2 and 4 send message 4 (type 0 must send 0) and[14] $\rho = (0, 0, 3)$. There is however another (sequential) equilibrium: type 2 sends message 4 and type 4 sends message 0, and $\rho' = (2, 2, 2)$, with outcome $\pi' = (2, 2, 2)$, which is not optimal ($H(\pi') < H(\pi)$). At this equilibrium (P0) is satisfied (since $\rho'(2) = v(2)$ for the unused message 2), but (A0) is not satisfied (since message 2 is optimal for type 2 but he sends 4). □

B.5 Without (P0)

Example 2 in Section I has an equilibrium (the uninformative equilibrium) that satisfies (A0) but does not satisfy (P0), and its outcome differs from the

[12] Once we go beyond our setup, the outcome equivalence given in Proposition 1 between truth-leaning and (A0)+(P0) need no longer hold.

[13] $\sigma^\varepsilon(\cdot|0) = (\varepsilon, 0, 1-\varepsilon)$ means that $\sigma^\varepsilon(s|0) = \varepsilon, 0, 1-\varepsilon$ for $s = 0, 2, 4$, respectively (the order on types is again increasing in value); similarly for ρ^ε.

[14] By Corollary 4 and Appendix C.8 (b) we may drop 0 from $L(2)$.

unique optimal mechanism outcome. However, that specific equilibrium can be ruled out by requiring the belief of the principal after an unused message to be equal to the conditional probability over the set of types that can send that message. That is, if message t is unused then put $q(t) = p|L^{-1}(t)$, the conditional of the prior p over the set $L^{-1}(t) := \{r \in T : t \in L(r)\}$ of all types r that can send message t, and $\rho(t) = v(q(t)) = v(p|L^{-1}(t))$ (instead of $q(t) = \mathbf{1}_t$ and $\rho(t) = v(t)$ in (P0)). The following example shows that replacing (P0) with this requirement is not enough to get equivalence.

Example 7 The type space is $T = \{0, 3, 10, 11\}$ with the uniform distribution: $p_t = 1/4$ for each t. The principal's payoff functions are $h_t(x) = -(x-t)^2$ (and so $v(t) = t$) for each $t \in T$. Types 10 and 11 both have less evidence than type 0, and more evidence than type 3; i.e., $L(0) = \{0, 3, 10, 11\}$, $L(3) = \{3\}$, $L(10) = \{3, 10\}$, and $L(11) = \{3, 11\}$.

The unique equilibrium that satisfies (A0) and (P0) is mixed: $\sigma(\cdot|0) = (0, 0, 3/7, 4/7)$, all the other types $t \neq 0$ reveal their type, and $\rho = (0, 3, 7, 7)$ (use for instance L' as in Appendix C.8 (b); note that $v(q(10)) = v(q(11)) = v(\{0, 10, 11\}) = 7$). The unique truth-leaning and optimal mechanism outcome is thus $\pi = (7, 3, 7, 7)$.

Consider now the uninformative equilibrium where every type sends message 3 and $\rho = (0, 6, 5, 5.5)$ (note that $\rho(3) = v(T) = 6$); its outcome $\pi' = (6, 6, 6, 6)$ is different from π. This equilibrium satisfies (A0) (because type 3 sends message 3) but not (P0) (for types 10 and 11). However, it does satisfy the alternative condition above: $\rho(0) = v(L^{-1}(0)) = v(0) = 0$, $\rho(10) = v(L^{-1}(10)) = v(\{0, 10\}) = 5$, and $\rho(11) = v(L^{-1}(11) = v(\{0, 11\}) = 5.5$. □

B.6 Without Payoff or Probability Boost

We provide an example where in the perturbed games telling the truth gets no payoff boost or no probability boost, and the resulting outcome differs from the unique optimal mechanism outcome.

Example 8 The type space is $T = \{0, 2, 4, 6\}$ with the uniform distribution: $p_t = 1/4$ for each $t \in T$. The principal's payoff functions are $h_t(x) = -(x-t)^2$ (and so $v(t) = t$) for each $t \in T$. The mapping L is $L(0) = \{0, 4\}$, $L(2) = \{0, 2, 4, 6\}$, $L(4) = \{4\}$, and $L(6) = \{4, 6\}$ (e.g., type 4 has no evidence, type 0 has a piece of negative evidence, type 6 has a piece of positive evidence, and type 2 has both pieces of evidence; this is the same evidence structure as in Example 2 in Section I[15]).

The unique optimal mechanism outcome is $\pi = (2, 4, 2, 4)$, and in the unique equilibrium that satisfies (A0) and (P0) types 0 and 4 send message 4 and types 2 and 6 send message 6.

The uninformative equilibrium where every type uses message 4 and the outcome is $\pi' = (3, 3, 3, 3)$ (with $H(\pi') = -5 < -4 = H(\pi)$) is the limit of Nash equilibria $(\sigma^\varepsilon, \rho^\varepsilon)$ of $\Gamma^{\boldsymbol{\varepsilon}}$ with $\varepsilon_6 = 0$ and all other ε_t and $\varepsilon_{t|t}$ equal to ε, as follows: $\sigma^\varepsilon(0|0) = \sigma^\varepsilon(2|2) = \sigma^\varepsilon(6|6) = \varepsilon$, $\sigma^\varepsilon(6|2) = \varepsilon(6 - 5\varepsilon)/(2 + \varepsilon)$, and with the remaining probabilities every type uses 4; and $\rho^\varepsilon = (0, 2, 3 - 4\varepsilon/(2 - \varepsilon), 3 - 4\varepsilon/(2 - \varepsilon))$.

If we instead take $\varepsilon_{6|6} = 0$ and all other $\varepsilon_{t|t}$ and ε_t to be equal to ε, then the Nash equilibria of $\Gamma^{\boldsymbol{\varepsilon}}$ with $\sigma^\varepsilon(0|0) = \sigma^\varepsilon(2|2) = \varepsilon$, $\sigma^\varepsilon(4|0) = \sigma^\varepsilon(4|2) = 1 - \varepsilon$, $\sigma^\varepsilon(4|4) = \sigma^\varepsilon(4|6) = 1$, and $\rho^\varepsilon(0) = 0$, $\rho^\varepsilon(2) = 2$, $\rho^\varepsilon(4) = (6 - \varepsilon)/(2 + \varepsilon) \geq \rho^\varepsilon(6)$ (message 6 is unused) again yield π' in the limit. □

B.7 Without (SP)

We provide an example where one of the functions h_t is not single-peaked and all the Nash equilibria yield an outcome that is strictly worse for the principal than the optimal mechanism outcome.

Example 9 The type space is $T = \{1, 2\}$ with the uniform distribution, i.e., $p_t = 1/2$ for $t = 1, 2$. The principal's payoff functions h_1 and h_2 are both strictly increasing for $x < 0$, strictly decreasing for $x > 2$, and piecewise linear[16] in the interval $[0, 2]$ with values at $x = 0, 1, 2$ as follows: $-3, 0, -2$

[15]The only reason that we do not work with Example 2 is that the numbers here are smaller and easier to handle.

[16]The example is not affected if the two functions h_1, h_2 are made differentiable (by smoothing out the kinks at $x = 0, 1$, and 2).

for h_1, and $2, 0, 3$ for h_2. Thus h_1 has a single peak at $v(1) = 1$, whereas h_2 is not single-peaked: its global maximum is at $v(2) = 2$, but it has another local maximum at $x = 0$. Type 2 has less evidence than type 1, i.e., $L(1) = \{1, 2\}$ and $L(2) = \{2\}$.

Consider first the optimal mechanism; the only (IC) constraint is $\pi_1 \geq \pi_2$. Fixing π_1 (in the interval $[0, 2]$), the value of π_2 should be as close as possible to one of the two peaks of h_2, and so either $\pi_2 = 0$ or $\pi_2 = \pi_1$. In the first case the maximum of $H(\pi)$ is attained at $\pi = (1, 0)$, and in the second case, at $\pi' = (2, 2)$ (because 2 is the peak of $h_p = (1/2)h_1 + (1/2)h_2$). Since $H(\pi) = 1 > 1/2 = H(\pi')$, the optimal mechanism outcome is $\pi = (1, 0)$.

Next, we will show that every Nash equilibrium (σ, ρ), whether truth-leaning or not, yields the worse outcome $\pi' = (2, 2)$. Indeed, type 2 can only send message 2, and so the posterior $q(2)$ after message 2 must put at least as much weight on type 2 as on type 1 (i.e., $q_2(2) \geq 1/2 \geq q_1(2)$; recall that the prior is $p_1 = p_2 = 1/2$). Therefore the principal's best reply is always 2 (because $h_{q(2)}(0) < 0$, $h_{q(2)}(1) = 0$, and $h_{q(2)}(2) > 0$). Therefore type 1 will never send the message 1 with positive probability (because then $q(1) = (1, 0)$ and so $\rho(1) = v(1) = 1 < 2$). Thus both types only send message 2, and we get an equilibrium if and only if $\rho(2) = 2 \geq \rho(1)$ (and, in the unique truth-leaning equilibrium, (P0) implies $\rho(1) = v(1) = 1$), resulting in the outcome $\pi' = (2, 2)$, which is not optimal: the optimal mechanism outcome is $\pi = (1, 0)$. □

Thus, the separation between the types—which is better for the principal—can be obtained here *only* with commitment.

B.8 Mixed Truth-Leaning Equilibria

We show here that we cannot restrict attention to pure equilibria: the agent's strategy may well have to be mixed (Example 7 above is another such case); in general, the equilibria can be quite complex and involve no easy unravelings and thresholds.

Example 10 The type space is $T = \{0, 2, 3\}$ with the uniform distribution: $p_t = 1/3$ for all t. The principal's payoff function is $h_t(x) = -(x - t)^2$,

and so $v(t) = t$. Types 2 and 3 both have less evidence than type 0; i.e., $L(0) = \{0, 2, 3\}$, $L(2) = \{2\}$, and $L(3) = \{3\}$.

Let (σ, ρ) be a truth-leaning equilibrium. Only the choice of type 0 needs to be determined. Since $\rho(0) = 0$ whereas $\rho(2) \geq 1 = v(\{0, 2\})$ and $\rho(3) \geq v(\{0, 3\}) = 3/2$, type 0 never chooses 0. Moreover, type 0 must put positive probability on message 2 (otherwise $\rho(2) = 2 > 3/2 = v(\{0, 3\}) = \rho(3)$), and also on message 3 (otherwise $\rho(3) = 3 > 1 = v(\{0, 2\}) = \rho(2)$). Therefore $\rho(2) = \rho(3)$ (since both are best replies for 0), and then $\alpha := \sigma(2|0)$ must solve $2/(1 + \alpha) = 3/(2 - \alpha)$, and hence $\alpha = 1/5$. This is therefore the unique truth-leaning equilibrium; its outcome is $\pi = (5/3, 5/3, 5/3)$. □

B.9 Multiple Truth-Leaning Equilibria

We show here that there need not be a unique truth-leaning equilibrium. Now all truth-leaning equilibria (σ, ρ) coincide in their principal's strategy ρ (which is uniquely determined by the outcome π: Proposition 3 implies that $\rho(t) = \min\{v(t), \pi_t\}$ for all t), but they may differ in their agent's strategies σ. However, this can happen *only* when the agent is indifferent—in which case the principal is also indifferent—which makes the nonuniqueness insignificant. As for optimal mechanisms, while there is a unique direct mechanism with outcome π (namely, the reward policy is π itself, i.e., $\rho(t) = \pi_t$ for all t), there may well be other optimal mechanisms (the reward for a message t may be lowered when there is a message $s \neq t$ in $L(t)$ with $\pi_s = \pi_t$).

An example with multiple truth-leaning equilibria is as follows.

Example 11 Let $T = \{0, 1, 3, 4\}$ with the uniform distribution: $p_t = 1/4$ for all $t \in T$; the principal's payoff functions are $h_t(x) = -(x - t)^2$ (and so $v(t) = t$) for all t, and $L(0) = \{0, 1, 3, 4\}$, $L(1) = \{1, 3, 4\}$, $L(3) = \{3, 4\}$, and $L(4) = \{4\}$ (i.e., a higher t goes with less evidence). The unique optimal mechanism outcome is $\pi_t = v(T) = 2$ for all t, and (σ, ρ) is a truth-leaning Nash equilibrium whenever $\rho(0) = 0$, $\rho(1) = 1$, $\rho(3) = \rho(4) = 2$, $\sigma(\cdot|0) = (0, 0, \alpha, 1 - \alpha)$, $\sigma(\cdot|1) = (0, 0, 1 - 2\alpha, 2\alpha)$, $\sigma(3|3) = 1$, and $\sigma(4|4) = 1$, for any $\alpha \in [0, 1/3]$. □

C Online Appendix: Extensions and Comments

This appendix contains material that could not be included in the streamlined main body of the paper: additional results, extensions, discussions, and comments. Among the more significant results we point out Proposition 7 (the equivalence result when messages need not be types), Section C.10 (the structure of the optimal outcome), and Section C.11 (equivalence without differentiability).

The order throughout is according to the sections in the main body of the paper, followed by the two additional Sections C.10 and C.11.

C.1 Introduction

(a) The importance of being able to commit. Think for instance of the advantage that it confers in bargaining, in oligopolistic competition (Stackelberg vs. Cournot), and also in cheap talk (cf. Example 3—see Remark (b) following it—in Appendix B.1).

(b) Interaction timeline. Interestingly, what distinguishes between "signaling" and "screening" is precisely the two different timelines of interaction that we consider: the agent moves first and the principal responds in signaling, and the principal moves first and the agent responds in screening.

(c) Mark Twain. The quotes are from his *Notebook* (1894). When he writes "truth" it means "the whole truth," since any partial truth requires remembering what was revealed and what wasn't.

(d) Application: medical overtreatment. A third possible application concerns medical overtreatment, which is one of the more serious problems in many health systems in the developed world; see, e.g., Shannon Brownlee (2008). One reason for overtreatment may be fear of malpractice suits; but the more powerful reason is that doctors and hospitals are paid more when overtreating. To overcome this problem one needs to give doctors incentives to provide evidence; the present paper may perhaps help in this direction.

C.2 Examples (Section I)

(a) Example 1. Formally, the dean wants to minimize $(x - v)^2$, where x is the salary and v is the professor's value; the dean's optimal response to any evidence is thus to choose x to be the expected value of the types that provide this evidence. The dean wants the salary to be "right" since, on the one hand, he wants to pay as little as possible, and, on the other hand, if he pays too little the professor may move elsewhere. The same applies when the dean is replaced by the "market."

In every sequential equilibrium the salary of a professor providing positive evidence must be 90 (because the positive-evidence type is the only one who can provide such evidence), and similarly the salary of a professor providing negative evidence must be 30. This shows that the uninformative equilibrium—where the professor, regardless of his type, provides no evidence, and the dean ignores any evidence that might be provided and sets the salary to the average value of 60—is not a sequential equilibrium here. Finally, we note that truth-leaning equilibria are always sequential equilibria.

(b) Example 2. It may be checked that the uninformative equilibrium satisfies all the standard refinements in the literature; cf. Appendix C.5.

This uninformative equilibrium may be eliminated here also by taking the posterior belief at unused messages to be the conditional prior (because the belief at message t_+ would then be $80\% - 20\%$ on t_+ and $t_\pm$); however, this would not suffice in general—see Example 7 in Appendix B.5.

C.3 Payoffs and Single-Peakedness (Section III.A)

(a) Single-peakedness. When going to more general models (e.g., Hart, Kremer, and Perry 2016), single-peakedness of the principal's utilities is taken with respect to the order on rewards that is induced by the agent's preference.

(b) Averages of single-peaked functions. To get (SP) it does *not suffice* that the functions h_t for $t \in T$ are all single-peaked, since averages of single-peaked functions need not be single-peaked (this is true, however, if the functions h_t are strictly concave). For example, let $\varphi(x)$ be a function that is strictly

increasing for $x < -2$, strictly decreasing for $x > 2$, has a single peak at $x = 2$, and takes the values $0, 3, 4, 7, 8$ at $x = -2, -1, 0, 1, 2$, respectively; in between these points interpolate linearly. Take $h_1(x) = \varphi(x)$ and $h_2(x) = \varphi(-x)$. Then h_1 and h_2 are single-peaked (with peaks at $x = 2$ and $x = -2$, respectively), but $(1/2)h_1 + (1/2)h_2$, which takes the values $4, 5, 4, 5, 4$ at $x = -2, -1, 0, 1, 2$, respectively, has two peaks (at $x = -1$ and $x = 1$). Smoothing out the kinks and making φ differentiable (by slightly changing its values in small neighborhoods of $x = -2, -1, 0, 1, 2$) does not affect the example.

(c) Non-concavity. The single-peakedness condition (SP) goes beyond concavity. Take for example $h_1(x) = -(x^3 - 1)^2$ and $h_2(x) = -x^6$; then h_1 is *not concave* (for instance, $h_1(1/2) = -49/64 < -1/2 = (1/2)h_1(0) + (1/2)h_1(1)$), but, for every $0 \leq \alpha \leq 1$, the function h_α has a single peak, at $\sqrt[3]{\alpha}$ (because $h'_\alpha(x) = -6x^2(x^3 - \alpha)$ vanishes only at $x = 0$, which is an inflection point, and at $x = \sqrt[3]{\alpha}$, which is a maximum).[17]

(d) Strict in-betweenness. The differentiability of the functions h_t is not needed to get in-betweenness (1). Differentiability yields a stronger property, *strict in-betweenness*: both inequalities in (1) are strict when the $v(q_i)$ are not all identical. Indeed, if $v(q_j) < v(q_k)$, then the derivative $h'_q(x) = \sum_i \lambda_i h'_{q_i}(x)$ is positive at $x = y_0 := \min_i v(q_i)$ (because $y_0 < v(q_k)$ and so $h'_{q_k}(y_0) > 0$), and is negative at $x = y_1 := \max_i v(q_i)$ (because $y_1 > v(q_j)$ and so $h'_{q_j}(y_1) < 0$); therefore $v(q) \in (y_0, y_1)$. Example 12 in Appendix C.11 shows that without differentiability these strict inequalities need not hold.

Strict in-betweenness is used (implicitly) only in the final argument in the Proof of Proposition 1 (ii) in Appendix A: if q is the average of q' and q'', and $v(q'') = v(q)$, then necessarily $v(q') = v(q)$.

[17] Alternatively, (SP) holds for the strictly concave $\hat{h}_1(y) = -(y-1)^2$ and $\hat{h}_2(y) = -y^2$; applying the strictly increasing transformation $y = x^3$, which preserves (SP), yields the given h_1 and h_2.

C.4 Evidence and Truth Structure (Section III.B)

(a) Detectable deviations. If t were to provide a subset of his pieces of evidence that did *not* correspond to a possible type s, it would be immediately clear that he was withholding some evidence (think for instance of the professor who provides to the dean *only* the Report of Referee #2). The only undetectable deviations of t are to reveal all the evidence of another possible type s that has fewer pieces of evidence than t (i.e., to pretend to be s).

However, our equivalence result would not change if we were to allow messages that do not correspond to types; see Proposition 7 in (d) below.

(b) Partial order on types. A general approach to the truth and evidence structure starts from a weak partial order[18] "$\rightarrowtail$" on the set of types T, with "$t \rightarrowtail s$" being interpreted as type t having (weakly) more evidence than type s; we will say that "s is a partial truth at t" (or "s is less informative than t"). The set of possible messages of the agent when the type is t, which we denote by $L(t)$, consists of all types that have less evidence than t, i.e., $L(t) := \{s \in T : t \rightarrowtail s\}$. Thus, $L(t)$ is the set of all possible "partial truth" revelations at t, i.e., all types s that t can pretend to be. The reflexivity and transitivity of the partial order $\rightarrowtail$ are immediately seen to be equivalent[19] to conditions (L1) and (L2).

Some natural models for the relation $\rightarrowtail$ are as follows.

(i) Pieces of evidence: As in Section III.B, let E be the set of possible pieces of evidence, and identify each type t with a subset E_t of E; thus, $T \subseteq 2^E$ (where 2^E denotes the set of subsets of E). Put $t \rightarrowtail s$ if and only if $t \supseteq s$; that is, t has every piece of evidence that s has. It is immediate that $\rightarrowtail$ is a weak partial order, i.e., reflexive and transitive.

(ii) Partitions: Let Ω be a set of states of nature, and let $\Lambda_1, \Lambda_2, ..., \Lambda_n$

[18] A *weak partial order* is a binary relation that is reflexive (i.e., $t \rightarrowtail t$ for all t) and transitive (i.e., $t \rightarrowtail s \rightarrowtail r$ implies $t \rightarrowtail r$ for all r, s, t). However, it need not be complete (i.e., there may be t, s for which neither $t \rightarrowtail s$ nor $s \rightarrowtail t$ holds). While for our results we do not need to assume that $\rightarrowtail$ is asymmetric, in most applications it is; moreover, we can always make it asymmetric by identifying any $t \neq t'$ with $t \rightarrowtail t'$ and $t' \rightarrowtail t$ (and then for any s and t, if $s \in L(t)$ then $t \notin L(s)$).

[19] Given L that satisfies (L1) and (L2), putting $t \rightarrowtail s$ iff $s \in L(t)$ yields a weak partial order.

be an increasing sequence of finite partitions of Ω (i.e., Λ_{i+1} is a refinement of Λ_i for every $i = 1, 2, ..., n-1$). The type space T is the collection of all blocks (also known as "kens") of all partitions. Then $t \rightarrowtail s$ if and only if $t \subseteq s$; thus more states ω are possible at s than at t, and so s is less informative than t. For example, take $\Omega = \{1, 2, 3, 4\}$ with the partitions $\Lambda_1 = (1234)$, $\Lambda_2 = (12)(34)$, and $\Lambda_3 = (1)(2)(3)(4)$. There are thus seven types: $\{1, 2, 3, 4\}, \{1, 2\}, \{3, 4\}, \{1\}, \{2\}, \{3\}, \{4\}$ (the first one from Λ_1, the next two from Λ_2, and the last four from Λ_3). Thus type $t = \{1, 2, 3, 4\}$ (who knows nothing) is less informative than type $s = \{1, 2\}$ (who knows that the state of nature is either 1 or 2), who in turn is less informative than type $r = \{2\}$ (who knows that the state of nature is 2); the only thing type t can say is t, whereas type s can say either s or t, and type r can say either r, s, or t. The probability p on T is naturally generated by a probability distribution μ on Ω together with a probability distribution λ on the set of partitions: if t is a ken in the partition Λ_i then $p_t = \lambda(\Lambda_i) \cdot \mu(t)$.

(iii) Signals: Let $Z_1, Z_2, ..., Z_n$ be random variables on a probability space Ω, where each Z_i takes finitely many values. A type t corresponds to some knowledge about the values of the Z_i-s (formally, t is an event in the field generated by the Z_i-s), with the straightforward "less informative" order: s is less informative than t if and only if $t \subseteq s$. For example, the type $s = [Z_1 = 7, 1 \leq Z_3 \leq 4]$ is less informative than the type $t = [Z_1 = 7, Z_3 = 2, Z_5 \in \{1, 3\}]$. (It is easy to see that (i) and (ii) are special cases of (iii).)

(c) General state space. We indicate how a general states-of-the-world setup reduces to our model.

Let $\omega \in \Omega$ be the state of the world, chosen according to a probability distribution $\mathbb{P}$ on Ω (formally, we are given a probability space[20] $(\Omega, \mathcal{F}, \mathbb{P})$). Each state $\omega \in \Omega$ determines the type $t = \tau(\omega) \in T$ and the utilities $U^A(x; \omega)$ and $U^P(x; \omega)$ of the agent and the principal, respectively, for any action (reward) $x \in \mathbb{R}$. The principal has no information, and the agent is informed of the type $t = \tau(w)$. Since neither player has any information beyond the type, we can reduce everything to the set of types T; namely, $p_t = \mathbb{P}[\tau(\omega) = t]$

[20] All sets and functions below are assumed measurable (and integrable when needed).

and $U^i(x;t) = \mathbb{E}\left[U^i(x;\omega)|\tau(\omega) = t\right]$ for $i = A, P$.

For a simple example, assume that the state space is $\Omega = [0,1]$ with the uniform distribution, $U^A(x;\omega) = x$, and $U^P(x;\omega) = -(x-\omega)^2$ (i.e., the "value" in state ω is ω itself). With probability $1/2$ the agent is told nothing about the state (which we call type t_0), and with probability $1/2$ he is told whether ω is in $[0,1/2]$ or in $(1/2,1]$ (types t_1 and t_2, respectively). Thus $T = \{t_0, t_1, t_2\}$, with probabilities $p_t = 1/2, 1/4, 1/4$ and expected values $v(t) = 1/2, 1/4, 3/4$, respectively. This example illustrates the setup where the agent's information is generated by an increasing sequence of partitions (cf. (ii) in the note above), which is useful in many applications (such as the voluntary disclosure setup).

(d) Additional messages. The equivalence result continues to hold if we allow additional messages beyond the set of types T; for instance, a message such as "t_1 *or* t_2" with $t_1 \notin L(t_2)$ and $t_2 \notin L(t_1)$, or a strict subset of the pieces of evidence that one has and that does not correspond to a type.

Let $M \supseteq T$ be the set of possible messages and let $L(t) \subseteq M$ for each $t \in T$ satisfy (L1) and (L2); the latter is now "$s \in L(t)$ and $m \in L(s)$ imply $m \in L(t)$," or, equivalently, "$s \in L(t)$ implies $L(t) \supseteq L(s)$."

Proposition 7 *Assume that the set M of possible messages contains the set of types T and that the mapping L satisfies (L1) and (L2). Then the Equivalence Theorem holds; moreover, replacing $L(t)$ with $L'(t) := L(t) \cap T$ for every $t \in T$ does not change the truth-leaning and optimal mechanism outcome.*

Proof. Consider first optimal mechanisms. The Revelation Principle still applies (because the (IC) constraints remain the same: $\pi_t \geq \pi_s$ for all types $s, t \in T$ with $s \in L(t)$; or, see Theorem 2 in Green and Laffont 1986). But direct mechanisms use only the set of types T as messages, and so $M \backslash T$ is not relevant, and being an optimal mechanism outcome for L and for L' is the same.

Consider next truth-leaning equilibria (note that truth-leaning makes no requirement on $\rho(m)$ for messages $m \notin T$ that are not used). We claim

that none of the messages $m \notin T$ are used in a truth-leaning equilibrium (σ, ρ), i.e., $\bar{\sigma}(m) = 0$ for all $m \notin T$. Indeed, let $m \notin T$; for every type $t \in T$ that uses m, i.e., $\sigma(m|t) > 0$, we get $\pi_t = \rho(m) > \rho(t) = v(t)$ (by (A), (A0), and (P0)). Therefore $\rho(m) > v(q(m))$ by in-betweenness (1), which contradicts (P). Finally, every truth-leaning equilibrium for L' is clearly also a truth-leaning equilibrium for L. ■

(e) Normal evidence. Bull and Watson (2007) consider the notion of "normal evidence," which allows the set of messages M to be arbitrary, and requires that for every type t in T there be a message m_t in $L(t)$ such that for every type s, if $m_t \in L(s)$ then $L(s) \supseteq L(t)$. Assuming that one can choose $m_t \neq m_s$ for[21] all $t \neq s$, we identify each m_t with t, which leads to the case $M \supseteq T$ discussed in (d) above (with normality yielding (L2)). Thus, again, the Equivalence Theorem applies here too.

C.5 Truth-Leaning Equilibria (Section III.D)

(a) Small perturbations. It is easy to check that truth-leaning would not be affected if we were to require that all choices have positive probabilities in Γ^{ϵ}, namely, $\sigma(s|t) \geq \varepsilon_{s|t} > 0$ for every s, t with $s \in L(t)$, provided that $\varepsilon_{s|t}$ for $s \neq t$ is much smaller than $\varepsilon_{t|t}$, i.e., $\varepsilon_{s|t}/\varepsilon_{t|t} \to 0$.

(b) Alternative perturbations. Both conditions of truth-leaning can also be obtained by perturbing *only* the payoff function of the agent. Given a random variable $Z > 0$ whose support is the whole positive line $\mathbb{R}_+$, let Γ^Z be the game where the utility of the agent for reward x, type t, and message s, is x when $s \neq t$, and $x + Z$ when $s = t$ (i.e., revealing the whole truth increases the agent's payoff by Z), and where the realized value of Z is known to the agent, but not to the principal. Now take a sequence Z_n with $\mathbb{E}[Z_n] \to 0$ as $n \to \infty$; then limit points of equilibria of Γ^{Z_n} are truth-leaning equilibria of[22] Γ.

[21] In Bull and Watson (2007) the messages are taken from $M \times T$, and so if $m_t = m_s$ then they are replaced by (m_t, t) and (m_s, s), which are different for $t \neq s$.

[22] The condition that the support of Z is all of $\mathbb{R}_+$ is too strong; it suffices that there is positive probability that Z takes some value larger than, say, $x_1 - x_0$, where the interval

(c) Refinements. Truth-leaning is consistent with all standard refinements in the literature. Indeed, they all amount to certain conditions on the principal's belief (which determines the reward) after an out-of-equilibrium message. Now the information structure of evidence games implies that in any equilibrium the payoff of a type s is minimal among all the types t that can send the message s (i.e., $\pi_s \leq \pi_t$ for every t with $s \in L(t)$). Therefore, if message s is not used in equilibrium (i.e., $\bar{\sigma}(s) = 0$), then the out-of-equilibrium belief at s that it was type s itself that deviated is allowed by all the standard refinements, specifically, the intuitive criterion, the D1 condition, universal divinity, and the never-weak-best-reply criterion (Elon Kohlberg and Jean-François Mertens 1986, Jeffrey Banks and Sobel 1987, In-Koo Cho and David Kreps 1987). However, these refinements may not eliminate equilibria such as the uninformative equilibrium of Example 2 in Section I (see also Example 7 in Appendix B.5); only truth-leaning does.[23] The no-incentive-to-separate (NITS) condition (Ying Chen, Kartik, and Sobel 2008), which requires the payoff of the lowest type to be no less than its value (which is what the principal would pay if he knew the type), is satisfied in our setup by all equilibria (because $\pi_s \geq \min_{t \in T} v(t)$ for every s; see the last sentence in Section III.A).

(d) Voluntary disclosure. In most of the voluntary disclosure literature the equilibrium is unique; when it is not, e.g., Shin (2003), the selected equilibrium ("sanitizing equilibrium") turns out to yield the same outcome as the truth-leaning equilibrium (we will show this in Proposition 8 below). As a consequence of our Equivalence Theorem, the resulting outcome is thus also the optimal mechanism outcome, and so the separation that is obtained in the voluntary disclosure literature is the optimal separation.

The setup of Shin (2003) can be summarized as follows. The principal minimizes the quadratic loss (and so we are in the basic setup); a type is $t = (s, f)$ where s and f are nonnegative integers with $s + f \leq N$ (for a fixed

$[x_0, x_1]$ contains all the peaks $v(t)$; see the last paragraph in Section III.A.

[23] Interestingly, if we consider the perturbed game where the agent's payoff is increased by $\varepsilon_t > 0$ when type t reveals the type, but the strategy is *not* required to satisfy $\sigma(t|t) > 0$, the refinements D1, universal divinity, and the never-weak-best-reply criterion (but not the intuitive criterion) yield in the limit the (P0) condition, and thus truth-leaning (we thank Phil Reny for this observation).

N); the value $v(s,f)$ of type (s,f) is decreasing in f, and the expected value $\bar{v}(s)$ of the set $T_s := \{(s,f) : 0 \leq f \leq N-s\}$ is increasing in s; finally, the partial truth mapping is $(s',f') \in L(s,f)$ if and only if $s' \leq s$ and $f' \leq f$.

The "sanitizing" equilibrium which Shin (2003) has chosen to study is given by: each type (s,f) sends the message $(s,0)$, and the rewards are $\rho(s,0) = \bar{v}(s)$ and $\rho(s,f) = v(s,N-s)$ for $f > 0$ (thus the equilibrium is supported by the not very reasonable belief that any out-of-equilibrium message (s,f) with $f > 0$ is sent by the type with the lowest value $(s,N-s)$). This is in general not a truth-leaning equilibrium (because, for instance, $v(s,1)$ may well be higher than $\bar{v}(s)$, and then (P0) cannot hold). However, there is always a truth-leaning equilibrium with the same outcome π^*, namely, $\pi^*_{s,f} = \bar{v}(s)$ for every (s,f), defined as follows. For every s let $k \equiv k_s$ be such that $v(s,k) \geq \bar{v}(s) > v(s,k+1)$; then each type (s,f) with $f \leq k$ sends the message (s,f) (i.e., reveals the type), whereas each type v(s,f) with $f \geq k+1$ sends the message (s,j) for $j = 0,1,...,k$ with probability $\lambda_j = p_{(s,j)}(v(s,j) - \bar{v}(s))/\sum_{i=0}^{k} p_{(s,i)}(v(s,i) - \bar{v}(s))$. The rewards are $\rho(s,f) = \bar{v}(s)$ for $f \leq k$ and $\rho(s,f) = v(s,f)$ for $f \geq k+1$. Thus for every s the messages used in equilibrium are (s,f) for all $f \leq k$, and they all yield the same reward $\bar{v}(s)$. It is straightforward to verify that this constitutes a truth-leaning equilibrium (for (P), use $\sum_{i=0}^{k} p_{(s,i)}(v(s,i) - \bar{v}(s)) = \sum_{i=k+1}^{N-s} p_{(s,i)}(\bar{v}(s) - v(s,i))$, because $\bar{v}(s)$ is the mean of the $v(s,f)$), and the outcome is π^*. We have thus shown:

Proposition 8 *In the voluntary disclosure model of Shin (2003), the "sanitizing" equilibrium outcome is the unique truth-leaning outcome, and thus also the unique optimal mechanism outcome.*

Appendix C.10 provides an alternative proof.

C.6 Mechanisms and Optimal Mechanisms (Section III.E)

(a) Green and Laffont. Green and Laffont (1986) show that, given (L1), condition (L2) is necessary and sufficient for the Revelation Principle to apply to *any* payoff functions of the agent. We need only the sufficiency part, which can be easily seen directly. Let ρ be a reward function; when

the type is t the agent's payoff is $\pi_t := \max_{r \in L(t)} \rho(r)$, and the principal's payoff is[24] $h_t(\pi_t)$. If t can pretend to be s, i.e., $s \in L(t)$, then $L(t) \supseteq L(s)$ by transitivity (L2), and thus $\pi_t \geq \pi_s$, which yields the incentive-compatibility constraints (IC). Conversely, any $\pi \in \mathbb{R}^T$ satisfying (IC) can be implemented by (L1) with a direct mechanism, namely, $\rho(t) = \pi_t$ for every t.

(b) Truth-leaning mechanisms. Truth-leaning does not affect optimal mechanisms, because a direct mechanism where the agent always reveals his type is clearly truth-leaning (moreover, in the limit-of-perturbations approach, it is not difficult to show that incentive-compatible mechanisms with and without truth-leaning yield payoffs that are the same in the limit).

(c) Existence and uniqueness of optimal mechanisms. It is immediate to see that an optimal mechanism exists, because the function H is continuous and the rewards π_t can be restricted to a compact interval X (see Section III.A). Uniqueness of the optimal mechanism outcome is not, however, straightforward (unless the principal's payoff functions h_t, and thus H, are all strictly concave—which we do not assume).

C.7 Proof (Section V)

(a) Our proof concludes that the (unique) optimal mechanism outcome can be obtained by a truth-leaning equilibrium indirectly (truth-leaning equilibria exist, and their outcomes coincide with the unique optimal mechanism outcome). A direct proof is presented in our companion paper Hart, Kremer, and Perry (2016): a (truth-leaning) equilibrium is constructed from an optimal mechanism using Hart and Kohlberg's (1974) extension of Philip Hall's marriage theorem (Hall 1935, Paul Halmos and Herbert Vaughn 1950).

C.8 Proof: Preliminaries (Section V.A)

(a) Full revelation when value increases with evidence. Corollary 4 implies that in the case where evidence always has positive value—i.e., if t has more

[24] Therefore in our setup the payoffs are not affected by how the agent breaks ties (an issue that arises in general mechanism setups).

evidence than s then the value of t is at least as high as the value of s (that is, $s \in L(t)$ implies $v(t) \geq v(s)$)—the (unique) truth-leaning equilibrium is fully revealing (i.e., $\sigma(t|t) = 1$ for every type t).

(b) Irrelevant messages. One may drop from $L(t)$ every $s \neq t$ with $v(s) \leq v(t)$; this affects neither the truth-leaning equilibrium outcomes (by Corollary 4) nor, by our Equivalence Theorem, the optimal mechanism outcomes; it amounts to replacing each $L(t)$ with its subset $L'(t) := \{s \in L(t) : v(s) > v(t)\} \cup \{t\}$. Note that L' also satisfies (L1) and (L2).

We provide an alternative proof of this statement that deals directly with mechanisms, and has the further advantage that instead of (SP), it uses only the weaker assumption that every function h_t is single-peaked (and not necessarily differentiable).

Let (IC') denote the incentive constraints given by L' (i.e., $\pi_t \geq \pi_s$ for all s, t with $s \in L'(t)$).

Proposition 9 *Assume that all the functions h_t are single-peaked (and not necessarily differentiable). Then π^* maximizes $H(\pi)$ subject to the (IC') constraints if and only if π^* maximizes $H(\pi)$ subject to the (IC) constraints.*

Proof. Since (IC') is a subset of the (IC) constraints, it suffices to show that if π^* maximizes $H(\pi)$ subject to (IC') then π^* satisfies all (IC) constraints.

Assume by way of contradiction that there are s, t such that $s \in L(t)$ but $\pi^*_t < \pi^*_s$; because π^* satisfies (IC'), we must have $v(s) \leq v(t)$. Among all pairs s, t as above, choose one where the difference $v(t) - v(s)$ (which is nonnegative) is minimal. Fix s and t. We have:

(i) All the (IC') constraints of the form $\pi_u \geq \pi_t$ for some u are not binding at π^*; i.e., $\pi^*_u > \pi^*_t$ for every u with $t \in L'(u)$.

Proof. If $\pi_u \geq \pi_t$ is an (IC') constraint then $t \in L(u)$ and $v(t) > v(u)$, and so $s \in L(u)$ by transitivity. If $\pi^*_u = \pi^*_t$ then $\pi^*_s > \pi^*_t = \pi^*_u$ and so $\pi_u \geq \pi_s$ cannot be an (IC') constraint; thus $s \notin L'(u)$, and so $v(s) \leq v(u)$. Hence $0 \leq v(u) - v(s) < v(t) - v(s)$, which contradicts the minimality of $v(t) - v(s)$.

(ii) $\pi^*_t \geq v(t)$.

Proof. If $\pi_t^* < v(t)$ then π_t^* lies in the region where h_t strictly increases, and so slightly increasing π_t^* (which can be done by (i)) increases the objective function H; this contradicts the optimality of π^*.

(iii) All the (IC') constraints of the form $\pi_s \geq \pi_r$ for some r are not binding at π^*; i.e., $\pi_s^* > \pi_r^*$ for every $r \in L'(s)$.

Proof. If $\pi_s \geq \pi_r$ is an (IC') constraint then $r \in L(s)$ and $v(r) > v(s)$, and so $r \in L(t)$ by transitivity. If $\pi_s^* = \pi_r^*$ then $\pi_t^* < \pi_s^* = \pi_r^*$ and so $\pi_t \geq \pi_r$ cannot be an (IC') constraint; thus $r \notin L'(t)$, and so $v(r) \leq v(t)$. Hence $0 \leq v(t) - v(r) < v(t) - v(s)$, which contradicts the minimality of $v(t) - v(s)$.

(iv) $\pi_s^* \leq v(s)$.

Proof. If $\pi_s^* > v(s)$ then π_s^* lies in the region where h_s strictly decreases, and so slightly decreasing π_s^* (which can be done by (iii)) increases the objective function H; this contradicts the optimality of π^*.

From (ii) and (iv) we get $v(t) \leq \pi_t^* < \pi_s^* \leq v(s)$, contradicting $v(s) \leq v(t)$). ■

C.9 From Equilibrium to Mechanism (Section V.B)

(a) Generalizing Propositions 5 and 6. The strict inequalities $v(t) < v(T)$ for every $t \neq s$ are used in the Proof of Proposition 5 to get, by in-betweenness (1), $v(R) \geq v(T)$ for any R that contains s; for their other use, to imply that $h_t(x)$ for $t \neq s$ is strictly decreasing for $x \geq v(T)$, the weak inequalities $v(t) \leq v(T)$ suffice. We thus get the following variant of Proposition 5:

Proposition 10 *Assume that there is a type $s \in T$ such that $s \in L(t)$ for every t. If*[25]

(i) $v(t) \leq v(T)$ for every $t \neq s$; and

(ii) $v(R) \geq v(T)$ for every R that contains s (i.e., $s \in R$),

then the outcome π^* with $\pi_t^* = v(T)$ for all $t \in T$ is the unique optimal mechanism outcome.[26]

[25]Condition (i) is equivalent to "$v(Q) \leq v(T)$ for every Q not containing s" (because $v(Q) \leq \max_{t \in Q} v(t)$ by in-betweenness (1)). Also, (i) and (ii) may be elegantly rewritten as $\max_{Q: s \notin Q} v(Q) \leq \min_{R: s \in R} v(R)$ (because by in-betweenness we have $v(T \backslash R) \leq v(T) \leq v(R)$ for every R that contains s, and so $v(T) = \min_{R: s \in R} v(R)$).

[26]When $L(s) = \{s\}$ and $L(t) = \{t, s\}$ for every $t \neq s$, conditions (i) and (ii) are also

This yields the following generalization of Proposition 6:

Proposition 11 *Let (σ, ρ) be a Nash equilibrium that satisfies, for every message s that is used (i.e., $\bar{\sigma}(s) > 0$),*

(i) $v(t) \leq v(q(s))$ for every $t \neq s$ that plays s (i.e., $\sigma(s|t) > 0$); and

(ii) $v(q(s)|R) \geq v(q(s))$ for every R that contains s (i.e., $s \in R$).

Then the outcome π^* of (σ, ρ) is the unique optimal mechanism outcome.

Proof. As in the Proof of Proposition 6, use the decomposition induced by (7) and then, for each s with $\bar{\sigma}(s) > 0$, apply Proposition 10 to $T_s := \{t : \sigma(s|t) > 0\}$ with prior $q(s)$. ■

These results are useful in the nondifferentiable case (see Appendix C.11).

C.10 The Optimal Outcome

We provide here results on the structure of optimal mechanisms and their outcomes, which is useful when dealing with specific applications.

A *partition* of T consists of disjoint sets $T_1, T_2, ..., T_m$ whose union is T. We will say that the *ordered* partition $(T_1, T_2, ..., T_m)$ is *consistent with* L (more precisely, consistent with the "having more evidence" order on types induced by L; see Appendix C.4) if $s \in L(t)$ for $t \in T_i$ and $s \in T_j$ implies $i \geq j$. Thus, types in T_1 have the least evidence, and those in T_m, the most; and, for any $t \in T_i$, we have $L(t) \subseteq \cup_{j \leq i} T_j$: type t can only pretend to be a type s in the same set or lower.

Proposition 12 *Let π be an optimal mechanism outcome. Then there exists an ordered partition $(T_1, T_2, ..., T_m)$ of T that is consistent with (the order induced by) L such that $v(T_1) < v(T_2) < ... < v(T_m)$ and $\pi_t = v(T_i)$ for every $t \in T_i$.*

necessary for π^* to be an optimal mechanism outcome—i.e., for "no separation" to be optimal. Indeed, if $v(t) > v(T)$ for some $t \neq s$ then put $\pi_t = v(t) > v(T) = \pi_t^*$, and if $v(R) < v(T)$ for some R containing s then put $\pi_r = v(R) < v(T) = \pi_r^*$ for all $r \in R$; in each case the new π satisfies all the constraints and $H(\pi) > H(\pi^*)$.

Proof. Let $\alpha_1 < \alpha_2 < ... < \alpha_m$ be the distinct values of the coordinates of π, and put $T_i := \{t \in T : \pi_t = \alpha_i\}$. This yields a partition that is consistent with L because $s \in L(t)$ implies $\pi_t \geq \pi_s$, and so $t \in T_i$ and $s \in T_j$ imply $i \geq j$. Changing the common value of π_t for all $t \in T_i$ to any other α_i' close enough to α_i so that all (IC) inequalities are preserved (specifically, $\alpha_{i-1} \leq \alpha_i' \leq \alpha_{i+1}$) implies by the optimality of π that α_i must maximize $\sum_{t \in T_i} p_t h_t(x) = p(T_i) h_{T_i}(x)$, and so $\alpha_i = v(T_i)$. ■

Remark. To find the optimal mechanism outcome, one thus needs to check only finitely many outcomes (each one determined by some partition of T).

A converse to Proposition 12 is as follows.

Proposition 13 *Let $(T_1, T_2, ..., T_m)$ be an ordered partition of T that is consistent with (the order induced by) L such that $v(T_1) \leq v(T_2) \leq ... \leq v(T_m)$ and for every $i = 1, 2, ..., m$, the unique optimal mechanism of the problem restricted to T_i is constant (i.e., $\pi_t = \pi_{t'}$ for all $t, t' \in T_i$). Then the unique optimal mechanism outcome is π^* with $\pi_t^* = v(T_i)$ for every $t \in T_i$ and $i = 1, 2, ..., m$.*

Proof. Let (IC') be the set of (IC) constraints $\pi_t \geq \pi_s$ with s, t in the same T_i. The outcome π^* satisfies all (IC') constraints as equalities; moreover, it satisfies the (IC) constraints (because $s \in L(t)$ with $t \in T_i$ and $s \in T_j$ implies $i \geq j$ and so $\pi_t^* = v(T_i) \geq v(T_j) = \pi_s^*$). Therefore, once we show that π^* is the unique maximizer of $H(\pi)$ subject to (IC'), then it is also the unique maximizer subject to (IC).

Now (IC') allows us to consider each T_i separately, and so if π is optimal then $\pi_t = \alpha_i$ for all $t \in T_i$, and so we must have $\alpha_i = v(T_i)$ (otherwise α_i could be slightly modified so that H will increase), which implies that $\pi = \pi^*$. ■

To use Proposition 13 one combines instances where the optimal mechanism outcome is unique. One such instance, where there is a type with minimal amount of evidence, is given by Proposition 5 in Section V.B (see also its generalization, Proposition 10 in Appendix C.9). Another instance, where the value decreases as one has more evidence, is given below.

Proposition 14 *If $L(t) = \{s : v(s) \geq v(t)\}$ for all t then the outcome π^* with $\pi^*_t = v(T)$ for all t is the unique truth-leaning equilibrium outcome and optimal mechanism outcome.*

Proof. Without loss of generality assume that $T = \{1, 2, ..., n\}$ and v is monotonic: if $t \leq s$ then $v(t) \leq v(s)$. Because $L(t) \supseteq \{t, t+1, ..., n\}$ by the assumption on L, (IC) implies that $\pi_1 \geq \pi_2 \geq ... \geq \pi_n$. Let π be an optimal mechanism outcome. If π is constant (i.e., $\pi_1 = ... = \pi_n$), then optimality implies that $\pi = \pi^*$. If π is not constant, let $1 \leq r < n$ be such that $\alpha := \pi_1 = ... = \pi_r > \pi_{r+1} \geq .. \geq \pi_n$. Because we can slightly modify the common value α of $\pi_1, ..., \pi_r$ without affecting (IC), optimality implies that $\alpha = v(\{1, ..., r\})$, and so $\alpha \leq v(r)$ by in-betweenness. Therefore for every $t \geq r+1$ we have $\pi_t < \alpha \leq v(r) \leq v(t)$, and so $h_t(\pi_t) < h_t(\alpha)$ (the function h_t strictly increases before its peak $v(t)$), implying that

$$H(\pi) = \sum_{t=1}^{r} p_t h_t(\alpha) + \sum_{t=r+1}^{n} p_t h_t(\pi_t) < \sum_{t=1}^{r} p_t h_t(\alpha) + \sum_{t=r+1}^{n} p_t h_t(\alpha) = H(\pi^{(\alpha)})$$

where $\pi^{(\alpha)} := (\alpha, ..., \alpha)$, contradicting the optimality of π. ■

As an application, combining Propositions 14 and 13 provides an alternative proof that the outcome of the sanitizing equilibrium of Shin (2003) is the optimal mechanism outcome (cf. Appendix C.5 (c)); the ordered partition is $(T_0, T_1, ..., T_N)$ with $T_s = \{(s, f) : 0 \leq f \leq N - s\}$).

C.11 Equivalence without Differentiability

Assuming that the functions h_t are differentiable has enabled us to work with the simpler conditions (A0) and (P0) rather than with the limit-of-perturbations approach. However, this was just for convenience: we will show here that the equivalence result holds also in the nondifferentiable case.

We start with a simple example where one of the functions h_t is not differentiable and there is no equilibrium satisfying (A0) and (P0).

Example 12 The type space is $T = \{1, 2\}$ with the uniform distribution, $p_t = 1/2$ for $t = 1, 2$. The principal's payoff functions are $h_1(x) = -(x-2)^2$ for $x \leq 1$ and $h_1(x) = -x^2$ for $x \geq 1$ (and so h_1 is nondifferentiable at its single peak $v(1) = 1$), and $h_2(x) = -(x-2)^2$ (and so h_2 has a single peak at $v(2) = 2$). Both functions are strictly concave, and so h_q has a single peak: $v(q) = 1$ when $q_1 \geq q_2$ and $v(q) = 2q_2$ when $q_1 \leq q_2$ (and thus[27] $v(T) = 1$). Type 1 has more evidence than type 2, i.e., $L(1) = \{1, 2\}$ and $L(2) = \{2\}$.

Let (σ, ρ) be a Nash equilibrium that satisfies (A0) and (P0). If type 1 sends message 1 then $\rho(1) = v(1) = 1$ and $\rho(2) = v(2) = 2$ (both by (P)), contradicting (A): message 1 is not a best reply for type 1. If type 1 sends message 2 then $\rho(1) = v(1) = 1$ (by (P0)) and $\rho(2) = v(T) = 1$ (by (P)), contradicting (A0): message 1 is a best reply for type 1 but he does not use it. Thus there is no truth-leaning equilibrium. □

It may be easily checked that in this example (σ, ρ) is a Nash equilibrium if and only if $\sigma(2|1) = 1$ and $\rho(2) = 1 \geq \rho(1)$, and so the outcome is $\pi = (1, 1)$, the same as the optimal mechanism outcome; truth-leaning yields that $\rho(1) = v(1) = 1$ (by (P0)).

In all our proofs, the differentiability of the functions h_t was used in *only* one place: to get (A0) in the last step of the Proof of Proposition 1 (ii) in Appendix A. All other proofs throughout the paper use only the non-differentiable version of single-peakedness, namely,

($\mathbf{SP}_0$) *Continuous Single-Peakedness.* For every $q \in \Delta(T)$ the principal's utility $h_q(x)$ is a continuous single-peaked function of the reward x.

Thus all the functions h_t are continuous (rather than differentiable), and for every $q \in \Delta(T)$ there is $v(q)$ such that the function $h_q(x)$ is strictly increasing for $x \leq v(q)$ and strictly decreasing for $x \geq v(q)$.

Equivalence holds also under (SP_0):

Proposition 15 *Assume that the principal's payoff function $(h_t)_{t \in T}$ satisfies the continuous single-peakedness condition (SP_0). Then there is a unique*

[27] The *strict* in-betweenness of Appendix C.3 does *not* hold here: the peak of h_1 is strictly less than the peak of h_2, and the peak of their average equals the peak of h_1.

truth-leaning equilibrium outcome, a unique optimal mechanism outcome, and these two outcomes coincide.

Proof. We will use Proposition 11 in Appendix C.9 (which generalizes Proposition 6 in Section V.B). We thus need to show that every truth-leaning limit equilibrium (σ, ρ) satisfies conditions (i) and (ii) of this proposition. We proceed as in the Proof of Proposition 1 (ii). Let $\varepsilon_t^n \to_n 0^+$, $\varepsilon_{t|t}^n \to 0^+$, and $(\sigma^n, \rho^n) \to_n (\sigma, \rho)$ be such that (σ^n, ρ^n) is a Nash equilibrium in $\Gamma^{\boldsymbol{\varepsilon}^n}$ for every n. If $\sigma(s|t) > 0$ for $t \neq s$, then, as in the arguments leading to (8) and (9), $v(q^n(s)) = \rho^n(s) \geq \rho^n(t) + \varepsilon_t^n > \rho^n(t) = v(t)$ for all large enough n. For every $R \subseteq T$ that contains s the posterior $q^n(s)$ is a weighted average of $q^n(s)|R$, the conditional of $q^n(s)$ on R, and $\mathbf{1}_t$ for all $t \notin R$ with $\sigma^n(s|t) > 0$, for all of which $v(q^n(s)) > v(t)$, as we have just seen; therefore in-betweenness (1) implies that $v(q^n(s)) \leq v(q^n(s)|R)$. Thus $v(t) < v(q^n(s)) \leq v(q^n(s)|R)$ for all large enough n; the continuity of v together with $q^n(s) \to q(s)$ and $q^n(s)|R \to q(s)|R$ (because, by (8) and $s \in R$, the limit denominators are bounded away from zero by $p_s \sigma(s|s) = p_s > 0$) yield conditions (i) and (ii) in the limit, as claimed. ■

Remark. As shown in the Proof of Proposition 1 (ii), every truth-leaning equilibrium (σ, ρ) satisfies (P0) and, assuming differentiability, can be modified without changing the outcome so as to satisfy also (A0). Without differentiability the latter is no longer true (as Example 12 shows); however, we can obtain, again without changing the outcome, a weaker version of (A0):

$$\text{(10)} \qquad \text{if } \rho(t) = \max_{r \in L(t)} \rho(r) \text{ } \textit{and} \text{ } \bar{\sigma}(t) > 0 \text{ then } \sigma(t|t) = 1;$$

here the condition that t chooses t for sure when it is a best reply for t is required *only* when message t is used at all). To get (10): if $\sigma(t|t) = 0$ then $\bar{\sigma}(t) = 0$ by (8) and no change is needed; and if $0 < \sigma(t|t) < 1$ then put $\sigma'(t|t) := 0$ and $\sigma'(s|t) := \sigma(s|t) + \sigma(t|t)$ for some $s \neq t$ that is played by t, i.e., $\sigma(s|t) > 0$ (because both t and s are played by t it follows that $v(t) = \rho(t) = \pi_t = \rho(s) = v(q(s))$, and so $v(q'(s)) = \pi_t$ by in-betweenness (1), as $q'(s)$ is a weighted average of $q(s)$ and $\mathbf{1}_t$).

References to Appendix C

(in addition to the references in the paper)

Banks, Jeffrey S. and Sobel, Joel (1987), "Equilibrium Selections in Signaling Games," *Econometrica* 55, 647–661.

Brownlee, Shannon (2007), "*Overtreated: Why Too Much Medicine Is Making Us Sicker and Poorer*," Bloomsbury.

Chen, Ying, Kartik, Navin, and Sobel, Joel (2008), "Selecting Cheap-Talk Equilibria," *Econometrica* 76, 117–136.

Cho, In-Koo and Kreps, David M. (1987), "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics* 102, 179–221.

Hall, Philip (1935), "On Representatives of Subsets," *Journal of the London Mathematical Society* 10, 26–30.

Halmos, Paul R. and Vaughan, Herbert E. (1950), "The Marriage Problem," *American Journal of Mathematics* 72, 214–215.

Hart, Sergiu and Kohlberg, Elon (1974), "Equally Distributed Correspondences," *Journal of Mathematical Economics* 1, 167–174.

Kohlberg, Elon and Mertens, Jean-François (1986), "On the Strategic Stability of Equilibria," *Econometrica* 54, 1003–1037.