Evidence Games: Truth and Commitment[1]

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

March 28, 2016

[1] Previous versions: February 2014; May 2015 (Center for Rationality DP-684). The authors thank Maya Bar-Hillel, Elchanan Ben-Porath, Peter DeMarzo, Kobi Glazer, Ehud Guttel, Johannes Hörner, Vijay Krishna, Rosemarie Nagel, Phil Reny, David Pérez-Castrillo, Uriel Procaccia, Tomás Rodríguez-Barraquer, Ariel Rubinstein, Amnon Schreiber, Andy Skrzypacz, Rani Spiegler, 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.

[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 where the principal chooses the reward after the evidence is disclosed and 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.

1 Introduction

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 relevant literature was on equilibrium and equilibrium prices. This approach was initiated by Akerlof (1970), and followed by the large body of work on voluntary disclosure, starting with Grossman and Hart (1980), Grossman (1981), Milgrom (1981), and Dye (1985). In a different line, the same problem was considered by Green and Laffont (1986) from a general mechanism-design viewpoint, in which 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 Glazer and Rubinstein (2004, 2006).

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

[1] 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 below).

agent can reveal in his message all this evidence (the "whole truth"), 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." Voluntary disclosure games, in which the right reward is the conditional expected value, corresponds to the case where the principal (who may well stand for the "market") has quadratic-loss payoff functions (we refer to this as the "basic case"). See the end of the Introduction for more on this and further applications.

The possibility of revealing the whole truth, an essential feature of evidence games, allows one to take into account the natural property that the whole truth has a slight inherent advantage. This is expressed by slight increases in the agent's utility when telling the whole truth, and in his probability of doing so; the equilibria selected by this approach are called *truth-leaning*. Formally, truth-leaning 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.[4] These simple conditions, which may be viewed as part of the setup or as equilibrium selection criteria, are most natural. The truth is after all a focal point, and there must be good reasons for *not* telling it.[5] As Mark Twain wrote, "When in doubt, tell the truth," and "If you tell the truth you don't have to remember anything."[6] Truth-leaning turns out to be consistent with the various refinement conditions offered in the literature, and equivalent to some of them (such as the equilibria used in the voluntary disclosure literature); see Appendix C.4.

[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.

[4] For example, the agent may be nonstrategic with small but positive probability; cf. Kreps, Milgrom, Roberts, and Wilson (1982).

[5] Psychologists refer to the "sense of well-being" associated with telling the truth.

[6] *Notebook* (1894). When he writes "truth" it means "the whole truth," since any partial truth requires remembering what was revealed and what wasn't.

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; this is the mechanism-design setup).

Our equivalence result can be stated as follows (Section 1.1 below provides simple examples that illustrate the result and the intuition behind it):

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

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.[7]

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 for 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 three such applications. The first one deals with voluntary disclosure in financial markets. Public firms

[7] Moreover, in the basic case where the optimal reward equals the expected value, the truth-leaning equilibria turn out to yield the *constrained Pareto efficient* outcomes; see Remark (c) in Section 3.

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 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 (see Proposition 8 in Appendix C.4), 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 may not, however, go beyond selecting among equilibria the optimal ones, namely, 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 made, 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.[8] Second, truth-leaning may

[8] There are of course other reasons and motivations for the right to remain silent.

well replace commitment: rather than committing to rules such as the right to remain silent and its offshoots, one may strengthen and reinforce the advantages of truth-telling.[9] 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).

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., 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.

To summarize the main contribution of the present paper: first, the class of *evidence games* that we consider models very common and important setups in information economics, setups that lie outside the standard signaling and cheap-talk literature; second, we prove the *equivalence* between truth-leaning equilibria without commitment and optimal mechanisms with commitment in evidence games; and third, 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. The Introduction continues below with some examples and a survey of relevant literature. Section 2 describes the model and the assumptions. The main equivalence result is stated in Section 3, and proved in Section 4 (with one of the proofs relegated to Appendix

[9] Or, at the very least, strengthen and reinforce the *perception* that truth-telling has an advantage.

A). In Appendix B it is shown that our conditions are indispensable for the equivalence result, and Appendix C provides additional notes, discussions, and extensions. Appendix D deals with mixed rewards (when there is no concavity) and the connections to the work of Glazer and Rubinstein (2004, 2006) and Sher (2011). Finally, Appendix E presents the construction of equilibria from optimal mechanisms using an extension of Hall's marriage result.

1.1 Examples

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

Example 1 (A simple version of the model introduced by Dye 1985.) A professor negotiates his salary with the dean. The dean would like to set the salary as close as possible to the professor's expected market value,[10] 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 an expected 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

[10] 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."

were t_0); see Figure 1.

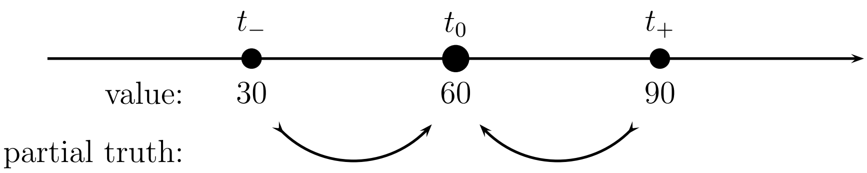


Figure 1: Values and possible partial truth messages in Example 1

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 a professor with positive evidence reveals it and is given a salary of 90 (equal to his value), whereas one with negative evidence conceals it and pretends that he has no evidence. When no evidence is presented the dean's optimal response is to set the salary at $50 = (50\% \cdot 60 + 25\% \cdot 30)/(50\% + 25\%)$, the expected value of the two types that provide no evidence: the no-evidence type together with the negative-evidence type. See Figure 2.

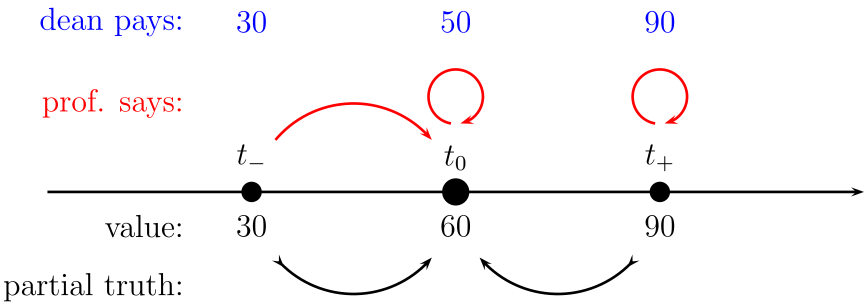


Figure 2: Equilibrium in Example 1

Next, consider the mechanism setup (with commitment): the dean commits to a salary policy (specifically, 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, namely, $x_+ = 90$ and $x_- = x_0 = 50$. Can the dean do better by committing? Can he provide incentives to the negative-evidence type to reveal his information? In order to separate between the negative-evidence type and the no-evidence type, he must give them distinct salaries, i.e., $x_- \neq x_0$. But then the salary for those who provide negative evidence must be higher than the salary for those who provide no evidence (i.e., $x_- > x_0$), because otherwise (i.e., when $x_- < x_0$) the negative-evidence type will pretend that he has no evidence and we are back to the no-separation case. Since the value 30 of the negative-evidence type is lower than the value 60 of the no-evidence type, setting a higher salary for the former than for the latter cannot be optimal (indeed, increasing x_- and/or decreasing 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* the negative-evidence type from the no-evidence type,[11] and so the unique optimal policy is identical to the equilibrium outcome, which is obtained without commitment. □

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 $x_+ = 90$ (their conditional expectation), and types t_0 and t_- provide no evidence, and get the salary $x_0 = 50$ (their conditional expectation). It is not difficult to see that this is also the optimal mechanism outcome.

[11] By contrast, the positive-evidence type is separated from the no-evidence type, because the former has a *higher* value. In general, separation of types with more evidence from types with 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.

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 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.[12] 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[13] t_{+}. □

Communication games, which include evidence games, are notorious for their multiplicity of equilibria. Requiring the equilibria to be sequential may eliminate some of them, but in general this is not enough (cf. Shin 2003). Truth-leaning, which we view as part of the "truth structure" that is characteristic of evidence games, thus provides a natural equilibrium refinement criterion. See Section 2.3.1.

Finally, lest some readers think that commitment is not useful in our general setup, we provide a simple variant of our examples—one that does *not* belong to the class of evidence games—where commitment yields outcomes that are strictly better than anything that can be achieved without it.

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

[12] It may be checked that this uninformative equilibrium satisfies all the standard refinements in the literature; cf. Appendix C.4.

[13] While taking the posterior belief at unused messages to be the conditional prior would suffice to eliminate the babbling equilibrium here (because the belief at message t_{+} would be $80\% - 20\%$ on t_{+} and $t_{\pm}$), it would not suffice in general; see Appendix B.4.

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).[14]

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).[15] □

Note that the 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 paper will show 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).

1.2 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

[14] Formally, take the utility of t_- when he gets salary x to be $-(x-50)^2$. Nothing in the example would be affected if we were to take the utility of t_0 to be $-(x-80)^2$ and to allow both types to send any message—the standard Crawford and Sobel (1982) cheap-talk setup. The fact that commitment may well be advantageous in cheap-talk games is known; see Krishna and Morgan (2007) and Goltsman *et al.* (2009).

[15] In the optimal mechanism the salary for no evidence is set to 70, and t_- (who is now indifferent between revealing and concealing his evidence) reveals his evidence.

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: signaling (Spence 1973 in economics and Zahavi 1975 in biology), screening (Rothschild and Stiglitz 1976), cheap talk (Crawford and Sobel 1982, Krishna and Morgan 2007).

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. In the game setup where the agent moves first, Grossman and O. Hart (1980), Grossman (1981), and Milgrom (1981) initiated the *voluntary disclosure* literature. These papers consider a salesperson who has private information about a product, which he may, if he so chooses, report to a potential buyer. The report is verifiable, that is, the salesperson cannot misreport the information that he reveals; he can, however, conceal it and not report it. These papers show that in every sequential equilibrium the salesperson employs a strategy of full disclosure: this is referred to as "unraveling." The key assumption here that yields this unraveling is that it is commonly known that the agent is fully informed. This assumption was later relaxed, as described below.

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 Jung and 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 the equilibrium is based on a threshold: only types who are informed and whose information is above a certain threshold disclose their information. Shin (2003, 2006), Guttman, Kremer, and Skrypacsz (2014), and Pae (2005) consider an evidence structure in which information is multi-dimensional.[16] Since such models typically possess multiple

[16] While the present paper studies a static model, there is also a literature on dynamic models. See, for example, Acharya, DeMarzo, and Kremer (2011) and Dye and Sridhar (1995).

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 easily verified that all these selected equilibria are in fact "truth-leaning" equilibria; thus truth-leaning turns out to be a natural way to unify all these criteria.

In the mechanism-design setup 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 utility functions is that the message structure be transitive and reflexive—which is satisfied by the voluntary disclosure models, as well as by our more general evidence games. Ben-Porath and Lipman (2012), Kartik and Tercieux (2012), and Koessler and 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 approach we are taking of comparing equilibria and optimal 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. In their work they show that the resulting optimal mechanism can be supported as an equilibrium outcome. More recently, Sher (2011) proved that the result continues to hold when the principal's decision is no longer binary provided that the principal's payoff is concave. See Appendix D.2 for a discussion of the Glazer–Rubinstein setup and the appropriate condition for equivalence.

Our paper shows that, in the framework of agents with identical utilities, the addition of the natural truth structure of evidence games—i.e., the partial truth relation and the inherent advantage of the whole truth—yields a stronger result, namely, the equivalence between equilibria and optimal mechanisms.

2 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 $p = (p_t)_{t \in T} \in \Delta(T)$ (where $\Delta(T)$ denotes 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 given by a real number x in $\mathbb{R}$. The message is costless: it does not affect the payoffs of the agent and the principal. As for the action, we assume that there are no further randomizations on the continuous variable[17] x. 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. The voluntary disclosure models (see Section 1.2) are all special cases of our model.

2.1 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 *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—

[17] Randomized rewards are indeed not needed when the principal's utility is concave (i.e., when the functions h_t defined below are all concave, which includes in particular the standard quadratic-loss case). In other cases mixed rewards may be useful; we analyze this in Appendix D.

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[18] 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, it strictly increases for $x \leq v$, and strictly decreases 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 the Introduction). 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, in increasing order of generality:[19]

• *Basic example: Quadratic loss.* For each type t let h_t be the quadratic distance from the ideal point: $h_t(x) = -(x - v(t))^2$. In this case, which is 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.* For each type t let h_t be a strictly concave function that attains its (unique) maximum at a finite point (which then holds for any weighted averages of such functions). For instance, take h_t to be 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., take $h_t(x) =$

[18] Single-peakedness is taken with respect to the order on rewards that is induced by the agent's preference.

[19] In Appendix C.2 we show that concavity is not necessary for (SP), and all the functions h_t being single-peaked is not sufficient for (SP).

$-c_t|x - v(t)|^{\gamma_t}$ (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 now state a useful observation.

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 $x \geq x_1$, all the peaks $v(q)$ for $q \in \Delta(T)$ satisfy $x_0 \leq v(q) \leq x_1$. More generally, if q is a weighted average of probability vectors $q_1, q_2, ..., q_n$ 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 sets $T_1, T_2, ..., T_n$ 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 $v(p|T_i)$ (recall that p is the prior; we write $p|T_i$ for the conditional of p given[20] 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 $x > x_1$).

2.2 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

[20] p is an average of the conditionals $p|T_i$; namely, $p = \sum_i p(T_i)(p|T_i)$, where $p(T_i) = \sum_{t \in T_i} p_t$ is the total probability of T_i.

then either to provide all the evidence E_t that he has ("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").[21] 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 (in the weak sense) evidence than t, i.e., $L(t) := \{s \in T : E_s \subseteq E_t\} \subseteq T$. 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) and Bull and Watson (2007). Appendix C.3 provides additional natural setups where they hold.[22] From now on we abstract away from any specific setup and just assume (L1) and (L2).

Remark. A type t has thus 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 between value and evidence*; in particular, having more evidence need not be associated with having a higher (or lower) value.

2.3 Game and Equilibria

We first consider 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

[21] 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., pretending to be s).

[22] In particular, we show there that one may add messages outside T (for example, "type t_1 *or* type t_2") and the equivalence result continues to hold.

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 to 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

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

for every $t \in T$. Thus π_t is the reward when the type is t, and so the payoffs are π_t for the agent and $h_t(\pi_t)$ for the principal.

2.3.1 Truth-Leaning Equilibria

As discussed in the Introduction, evidence games may have many equilibria; we are interested in those where 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—à la Kreps, Milgrom, Roberts, and Wilson 1982; 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 $0 < \varepsilon_{t|t} < 1$ for all $t \in T$ (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 there are sequences $\varepsilon_t^n \rightarrow_{n\rightarrow\infty} 0$, $\varepsilon_{t|t}^n \rightarrow_{n\rightarrow\infty} 0$, and $(\sigma^n, \rho^n) \rightarrow_{n\rightarrow\infty} (\sigma, \rho)$ 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 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.[23]

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[24] 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 $\Gamma^{\boldsymbol{\varepsilon}}$ for small $\boldsymbol{\varepsilon}$). In Appendix C.4 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.

2.4 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).

[23] We could well have started directly with the natural conditions (A0) and (P0); however, we find the limit-of-small-perturbations approach to be more basic.

[24] The proof of (ii) turns out to be somewhat more delicate than the arguments above suggest; in particular, it needs the differentiability of the functions h_t. As for existence (i), it follows from a standard fixed-point argument and compactness.

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 where it is optimal for each type to be "truthful" and reveal his type (see Green and Laffont 1986, or Appendix C.5). The incentive compatibility constraints are:

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

Indeed, type t can pretend to be type s only if he can send message s, i.e., $s \in L(t)$; then $L(t) \supseteq L(s)$ 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).

Remarks. *(a)* An optimal mechanism is just a Nash equilibrium of the game where the principal moves first and chooses the reward scheme.

(b) The outcome π of any Nash equilibrium (σ, ρ) of the game Γ of the previous section satisfies (IC) (by transitivity (L2)), and so an optimal mechanism can yield only a higher payoff to the principal: commitment can only help the principal.

3 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 the whole truth, i.e., 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).[25] Thus t and s are *not separated* in equilibrium, and we claim that in this case 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 between types more than truth-leaning equilibria do (as for the converse, it 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.8).

(b) Tightness of the result. All the assumptions except differentiability are indispensable for 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.9).

(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., $\mathbb{E}\left[\pi_t\right] = \sum_{t \in T} p_t \, \pi_t$, equals the ex-ante expectation of the values $\mathbb{E}\left[v(t)\right] = \sum_{t \in T} p_t \, v(t) = v(T)$ (because $\mathbb{E}\left[\pi_t | s\right] = v(q(s)) = \mathbb{E}\left[v(t) | s\right]$ 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 $\mathbb{E}\left[\pi_t\right] = v(T)$

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

(they differ ex-post, however, in the way this amount is split among the 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).

4 Proof of the Equivalence Theorem

The proof proceeds as follows. We start with some useful and interesting properties of truth-leaning (Section 4.1), and then prove that the outcome of any truth-leaning equilibrium outcome is an optimal mechanism outcome, which is moreover unique (Proposition 6 in Section 4.2). Together with the existence of truth-leaning equilibria (Proposition 1 (i) in Section 2.3.1) this yields the result.[26]

4.1 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 perhaps 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

[26] An alternative proof, which also shows how to obtain a truth-leaning equilibrium from an optimal mechanism, is provided in Appendix E.

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 the Introduction). In particular, in cases where evidence has always positive value—i.e., if t has more 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).

Remark. One may thus drop from $L(t)$ every $s \neq t$ with $v(s) \leq v(t)$; this affects neither the truth-leaning equilibrium outcomes 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)).

4.2 From Equilibrium to Mechanism

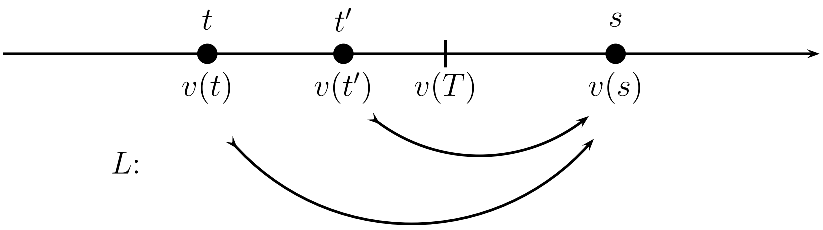
This section proves that any truth-leaning equilibrium outcome is an optimal mechanism outcome and, moreover, that the latter is unique. We first deal with a special case where there is no separation, and then show how a truth-leaning equilibrium yields a decomposition into instances of this special 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 3. To get some intuition, consider the simplest case of only 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 $\pi_t > v(t)$



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

Figure 3: Proposition 5

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) (if $s \in L(t)$ then $L(s) \subseteq L(t)$ by (L2), and so $\pi_s^* = \max_{r \in L(s)} \rho(r) \leq \max_{r \in L(t)} \rho(r) = \pi_t^*$), 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 such s, let $T_s := \{t \in T : \sigma(s|t) > 0\}$ be the set of types that play s. For each $t \neq s$ 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 split 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, G. A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84, 488–500.

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

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

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

Chakraborty, A. and R. Harbaugh (2010), "Persuasion by Cheap Talk," *American Economic Review* 100, 2361–2382.

Chen, Y., N. Kartik, and J. Sobel (2008), "Selecting Cheap-Talk Equilibria," *Econometrica* 76, 117–136.

Cho, I. K. and D. M. Kreps (1987), "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics* 102, 179–221.

Crawford, V. and J. Sobel (1982), "Strategic Information Transmission," *Econometrica* 50, 1431–1451.

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

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

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

Goltsman M., J. Hörner, G. Pavlov, and F. Squintani (2009), "Mediation, Arbitration and Negotiation," *Journal of Economic Theory* 144, 1397–1420.

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

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

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

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

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

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

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

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

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

Kohlberg E. and J.-F. Mertens (1986), "On the Strategic Stability of Equilibria," *Econometrica* 54, 1003–1037.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory* 27, 245–252.

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

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

Myerson, R. B. (1979), "Incentive-Compatibility and the Bargaining Problem," *Econometrica* 47, 61–73.

Rothschild, M. and J. Stiglitz (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics* 90, 629–649.

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

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

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

Spence, M. (1973), "Job Market Signalling," *The Quarterly Journal of Economics* 87, 355–374.

Zahavi, A. (1975), "Mate Selection—A Selection for a Handicap," *Journal of Theoretical Biology* 53, 205–214.

A Appendix: Proof of Proposition 1

We prove here the existence of truth-leaning equilibria, and their payoff-equivalence to equilibria that satisfy (A0) and (P0).

Proof of Proposition 1. *(i) Existence.* First, a standard fixed-point argument shows that the game $\Gamma^{\boldsymbol{\varepsilon}}$ possesses a Nash equilibrium. Let $\Sigma^{\boldsymbol{\varepsilon}}$ be the set of strategies of the agent in $\Gamma^{\boldsymbol{\varepsilon}}$; then $\Sigma^{\boldsymbol{\varepsilon}}$ 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 $\Gamma^{\boldsymbol{\varepsilon}}$ 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 $\sigma \in \Sigma^{\varepsilon}$ to the set of all $\sigma' \in \Sigma^{\varepsilon}$ that are best replies to ρ^{σ} in Γ^{ε} 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 Γ^{ε}.

Second, the strategy sets of the two players are compact (for the principal, see the final comment in Section 2.1), and so limit points of Nash equilibria of Γ^{ε}—i.e., truth-leaning equilibria of Γ—exist (it is immediate to verify that any limit point of Nash equilibria of Γ^{ε} 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 Γ^{ε^n} for every n (which is easily seen to imply that (σ, ρ) is a Nash equilibrium of Γ, i.e., (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 so $\sigma^n(s|t) > 0$ for all (large enough) n. In Γ^{ε^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 Γ^{ε^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[27] 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 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 condition (except for differentiability, which is assumed for convenience; see Appendix C.9) 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 truth-leaning equilibria need be neither unique nor pure (Sections B.8 and B.9).

B.1 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.

[27] Example 13 in Section C.9 shows that this property need *not* hold without differentiability. The argument below amounts to *strict* in-betweenness; see Section C.2.

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.,[28] $\pi = (\pi_0, \pi_2, \pi_4) = (0, 3, 3)$.

Truth-leaning entails no restrictions here: types 0 and 2 have a single message each (their type), and type 4 cannot send the 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 $H(\pi) > H(\pi')$. □

The evidence structure in this example is not "normal" (Bull and Watson 2007—see Appendix C.3—because there is no message m_4 for type 4). We therefore provide an additional example where normality holds.

Example 5 The same as above, with $M = \{a, b, c\}$, $L(0) = \{a, c\}$, $L(2) = L(4) = \{b, c\}$; the evidence structure is normal: take $m_0 = a$, $m_2 = b$, and $m_4 = c$. The optimal mechanism yields $\pi = (0, 3, 3)$, and the equilibrium (σ, ρ) with outcome $\pi' = (2, 2, 2)$ where types 0 and 4 send c and type 2 sends b, and $\rho(a) = 0$, $\rho(b) = \rho(c) = 2$, satisfies (A0) and (P0) (revealing the truth for type t means sending the message m_t).

B.2 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 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 all t. The allowed messages are $L(0) = \{0, 4\}$, $L(2) =$

[28] The order on types that is used when writing vectors such as π is increasing in value (thus here $\pi = (\pi_0, \pi_2, \pi_4)$; recall that $v(t) = t$).

$\{2\}$, 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[29] the reward scheme $\rho = (0, 3, 0)$, with outcome $\pi = (0, 3, 3)$; indeed, if 2 and 4 are separated then the best is 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 the best is 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 $\pi_2 = \pi_4 = 3$), 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:[30] take Γ^{ε} with $\varepsilon_t = \varepsilon_{t|t} = \varepsilon$ for all t, then π obtains from the limit of[31] $\sigma^{\varepsilon}(\cdot|0) = (\varepsilon, 0, 1 - \varepsilon)$, $\sigma^{\varepsilon}(\cdot|4) = (0, 1 - \varepsilon, \varepsilon)$, 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.3 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 7 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 $L(4) = \{0, 4\}$.

[29] 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 2.4; 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).

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

[31] $\sigma^{\varepsilon}(\cdot|0) = (\varepsilon, 0, 1 - \varepsilon)$ 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}.

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[32] $\rho = (0, 0, 3)$. 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.4 Without (P0)

Example 2 in the Introduction has an equilibrium (the uninformative equilibrium) that satisfies (A0) but does not satisfy (P0), and its outcome differs from the 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 t, and $\rho(t) = v(q(t)) = L^{-1}(t))$ (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 8 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 the Remark on L' at the end of Section 4.1; note that $v(q(10)) = v(q(11)) = v(\{0, 10, 11\}) = 7$). The unique truth-leaning and optimal mechanism outcome is thus $\pi = (7, 3, 7, 7)$.

[32] By Corollary 4 (see L' in the paragraph following it) we may drop 0 from $L(2)$.

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.5 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 9 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 some negative evidence, type 6 some positive evidence, and type 2 both pieces of evidence; this is the same evidence structure as in Example 2 in the Introduction[33]).

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 Γ^ε 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 Γ^ε with $\sigma^\varepsilon(0|0) = \sigma^\varepsilon(2|2) = \varepsilon$, $\sigma^\varepsilon(4|0) = \sigma^\varepsilon(4|2) = 1 - \varepsilon$,

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

$\sigma^\varepsilon(4|4) = \sigma^\varepsilon(4|6) = 1$, 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.6 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 10 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[34] in the interval $[0, 2]$ with values at $x = 0, 1, 2$ as follows: $-3, 0, -2$ 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

[34] 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).

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.7 Agent's Payoffs Depend on Type

Example 3 in the Introduction—which may be viewed also as a Crawford and Sobel (1982) standard cheap-talk game—shows that the equivalence result fails when the agent's types do not all have the same preference.

B.8 Multiple Truth-Leaning Equilibria

All the 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 $\alpha \in [0, 1/3]$. □

B.9 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 8 above is another such case).

Example 12 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 $\pi = (5/3, 5/3, 5/3)$. □