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.

Contents

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.

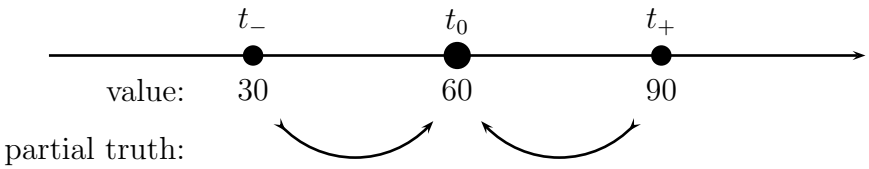


Figure 1: Values and possible partial truth messages in Example 1

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.

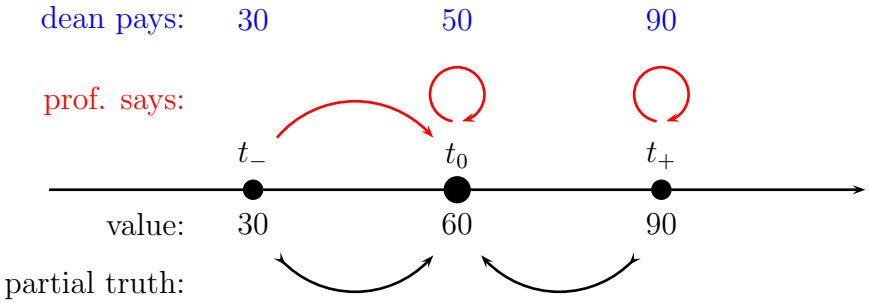


Figure 2: Equilibrium in Example 1

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 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

[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.

of type t can provide (e.g., prove in court). The possible messages of t are 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.

[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.

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 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 $\Gamma^{\boldsymbol{\epsilon}}$ 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 $\Gamma^{\boldsymbol{\epsilon}}$ as all the ε-s converge to 0; i.e., if there 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 $\Gamma^{\boldsymbol{\epsilon}^n}$ 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 Γ^ε for small ε). 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

[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.

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 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 re-

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

wards, 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 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., $\pi_t > v(t)$).

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

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 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 $\pi_t > v(t)$

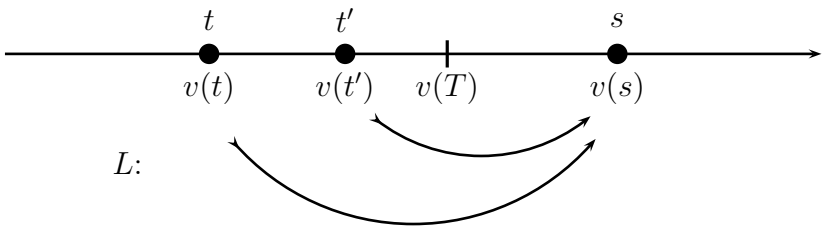
or $\pi_s < v(s)$, and so decreasing π_t or increasing π_s brings it closer to the corresponding peak, and hence increases the value of H. Thus $\pi_t = \pi_s = x$ for some x, and then the maximum is attained when x equals the peak of $h_p(x) = p_s h_s(x) + p_t h_t(x)$, i.e., when $x = v(T)$.

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



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

Figure 3: Proposition 5

$L(s) \subseteq L(t)$ 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 $\Gamma^{\boldsymbol{\epsilon}}$ 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 the final comment in Section 2.1), 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., (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 $\Gamma^{\boldsymbol{\epsilon}^n}$ 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 $\Gamma^{\boldsymbol{\epsilon}^n}$), 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 $\sigma^n(t|t) \geq \varepsilon^n_{t|t} > 0$ 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). ■

[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.

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.

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.

[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 $v(t) = t$).

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 $L(0) = \{0, 4\}$, $L(2) = \{2\}$, 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 $\Gamma^{\boldsymbol{\epsilon}}$ with

[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.

$\varepsilon_t = \varepsilon_{t|t} = \varepsilon$ 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 $L(4) = \{0, 4\}$.

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

[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.

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

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 $\pi = (7, 3, 7, 7)$.

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 $\Gamma^{\boldsymbol{\epsilon}}$ 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 $\Gamma^{\boldsymbol{\epsilon}}$ 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.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 $L(2) = \{2\}$.

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

[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).

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 $\pi = (1,0)$. $\square$

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 $\pi = (5/3, 5/3, 5/3)$. □

C Appendix: Comments and Extensions

C.1 Introduction (Section 1)

(a) Example 1. 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) Interaction timeline. Interestingly, what distinguishes between "signaling" and "screening" (see Section 1.2 for references) is precisely the two different timelines of interaction that we consider: in signaling the agent moves first and the principal responds, in screening the principal moves first and the agent responds.

C.2 Payoffs and Single-Peakedness (Section 2.1)

(a) Averages of single-peaked functions. To get (SP) it does *not suffice* that the functions h_t for $t \in T$ be 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.

(b) 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).[35]

(c) 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 13 in Appendix C.9 shows that without differentianbility 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 $v(q') = v(q)$.

C.3 Evidence and Truth Structure (Section 2.2)

(a) Partial order on types. A general approach to the truth and evidence structure starts from a weak partial order[36] "$\rightarrowtail$" 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

[35] 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.

[36] 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 $t \notin L(s)$).

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[37] to conditions (L1) and (L2).

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

(i) Pieces of evidence: As in Section 2.2, 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 $\Lambda_1, \Lambda_2, ..., \Lambda_n$ 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

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

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).)

(b) 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[38] $(\Omega, \mathcal{F}, \mathbb{P})$). 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, $p_t = \mathbb{P}[\tau(\omega) = t]$ and $U^i(x; t) = \mathbb{E}[U^i(x; \omega) | \tau(\omega) = t]$ 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).

(c) Additional messages. The equivalence result continues to hold if we allow additional messages beyond the set of types T (for instance, messages such as "t_1 *or* t_2" with $t_1 \notin L(t_2)$ and $t_2 \notin L(t_1)$).

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 $L(t) \supseteq L(s)$").

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

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 are 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. ■

(d) 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[39] all $t \neq s$, we identify each m_t with t, which leads to the case $M \supseteq T$ discussed in the previous note (with normality yielding (L2)). Thus, again, the Equivalence Theorem applies here too.

C.4 Truth-Leaning Equilibria (Section 2.3.1)

(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

[39] In Bull and Watson (2007) the messages are taken from $M \times T$, and so if $m_t = m_s$ one uses instead (m_t, t) and (m_s, s), which are different for $t \neq s$.

$\Gamma^{\boldsymbol{\varepsilon}}$, 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) 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 (Kohlberg and Mertens 1986, Banks and Sobel 1987, Cho and Kreps 1987). However, these refinements may not eliminate equilibria such as the uninformative equilibrium of Example 2 in Section 1.1 (see also Example 8 in Appendix B.4); only truth-leaning does.[40] The no-incentive-to-separate (NITS) condition (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 2.1).

(c) 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 equilibria (we will show this 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

[40] 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).

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 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 so also the unique optimal mechanism outcome.*

See Appendix C.8 for an alternative proof.

C.5 Mechanisms and Optimal Mechanisms (Section 2.4)

(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 agents's payoff is $\pi_t := \max_{r \in L(t)} \rho(r)$, and the principal's payoff is[41] $h_t(\pi_t)$. 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 2.1). 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.6 Proof: Preliminaries (Section 4.1)

(a) Pretending and values. Corollary 4 (see the Remark following it) implies that truth-leaning equilibria are not affected when replacing $L(t)$ with $L'(t) := \{s \in L(t) : v(s) > v(t)\} \cup \{t\}$ for all t (i.e., dropping from each $L(t)$ all $s \neq t$ that do not have a higher value than t). Together with our Equivalence Theorem, it follows that the same applies to optimal mechanisms. We provide here a direct 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 all the functions h_t are single-peaked (and not necessarily differentiable).

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

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.7 From Equilibrium to Mechanism (Section 4.2)

(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*[42]

(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.[43]

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 $q(s)$. ■

[42](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 $v(T) = \min_{R: s \in R} v(R)$).

[43]When $L(s) = \{s\}$ and $L(t) = \{t, s\}$ for every $t \neq s$, conditions (i) and (ii) are also 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 $H(\pi) > H(\pi^*)$.

These results are useful in the non-differentiable case (see Appendix C.9).

C.8 The Optimal Outcome

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

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.3) 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 $t \in T_i$.*

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 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 can be slightly modified such that H increases), which implies that $\pi = \pi^*$. ■

To use Proposition 13 one 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 4.2 (see also its generalization Proposition 10 in Appendix C.7). 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.4 (c)); the ordered partition is $(T_0, T_1, ..., T_N)$ with $T_s = \{(s, f) : 0 \leq f \leq N - s\}$).

C.9 Equivalence without Differentiability

Assuming that the functions h_t are differentiable 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 13 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[44] $v(T) = 1$). 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)),

[44]The *strict* in-betweenness of Appendix C.2 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.

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,

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

Proof. We will use Proposition 11 in Appendix C.7 (which generalizes Proposition 6 in Section 4.2). 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 $\varepsilon^n_t \to_n 0^+$, $\varepsilon^n_{t|t} \to 0^+$, 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), $v(q^n(s)) = \rho^n(s) \geq \rho^n(t) + \varepsilon^n_t > \rho^n(t) = v(t)$ 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 13 shows); however, we can obtain, again without changing the outcome, a weaker version of (A0):

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

here t is required to choose t for sure when it is a best reply for t *only* provided that 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 $\mathbf{1}_t$).

D Appendix: Randomized Rewards

Consider the case where the principal may choose randomized (or mixed) rewards; i.e., the reward $\rho(s)$ to each message s is a probability distribution ξ on $\mathbb{R}$ rather than a pure $x \in \mathbb{R}$. The utility functions of the two players are now von Neumann–Morgenstern utilities on $\mathbb{R}$, and so the utility of a randomization ξ equals its expected utility: $\mathbb{E}[\xi]$ for the agent and $h_t(\xi) := \mathbb{E}_{x\sim\xi}[h_t(x)]$ for the principal, for each type[45] $t \in T$. While we no longer assume single-peakedness, we do assume that all the functions h_t are continuous, and that only a compact interval $X = [x_0, x_1] \subset \mathbb{R}$ of rewards matters (for instance, let all the functions $h_t(x)$ increase for $x < x_0$ and decrease for $x > x_1$ for all $t \in T$).

The following example shows that the equivalence result may not hold for randomized rewards, even when single-peakedness holds.

Example 14 The type space is $T = \{0, 2\}$ with the uniform distribution, i.e., $p_t = 1/2$ for $t = 0, 2$. The principal's payoff functions h_0 and h_2 are both strictly increasing for $x < 0$, strictly decreasing for $x > 2$, and piecewise linear[46] in the interval $[0, 2]$ with values at $x = 0, 1, 2$ as follows: $12, 2, 0$ for h_0, and $-30, 2, 6$ for h_2 (note that (SP) holds: the function $h_q = q_0 h_0 + (1-q_0)h_2$ is single-peaked for every $q_0 \in [0, 1]$, because $x = 1$ is never the worst reward: for $q_0 \leq 3/4$ we have $h_q(0) < 2 = h_q(1)$ and for $q_0 \geq 3/4$ we have $h_q(2) < 2 = h_q(1)$). Assume that type 2 has less evidence than type 0, i.e., $L(0) = \{0, 2\}$ and $L(2) = \{2\}$.

Consider first only pure rewards. We claim that every Nash equilibrium (σ, ρ), whether truth-leaning or not, yields the outcome $\pi = (2, 2)$. Indeed, type 2 can send only message 2, and so the posterior $q(2)$ after message 2 must put at least as much weight on type 2 as on type 0 (i.e., $q_2(2) \geq 1/2 \geq q_0(2)$; recall that the prior is $p_0 = p_2 = 1/2$). Therefore the principal's best reply to message 2 is always 2 (because $h_{q(2)}(0) \leq 0$, $h_{q(2)}(1) = 2$, and $h_{q(2)}(2) \geq 3$).

[45] We have assumed that x is that pure reward that yields utility x to the agent. For the principal, we extend the functions h_t to mixed rewards.

[46] The example is not affected if the two functions h_0, h_2 are made differentiable (by smoothing out the kinks at $x = 0, 1$, and 2).

Therefore type 0 will never send the message 0 with positive probability (because then $q(0) = (1, 0)$ and so $\rho(0) = v(0) = 0 < 2$). Thus both types send only message 2, and we get an equilibrium if and only if $\rho(2) = 2 \geq \rho(0)$ (and, in the unique truth-leaning equilibrium, $\rho(0) = v(0) = 0$ by (P0)), resulting in the outcome $\pi = (2, 2)$.

The optimal mechanism (with pure rewards) cannot separate between the types (because, again, that would require the principal to reward message 0 strictly more than message 2, which is not optimal; cf. Proposition 5); therefore the best for him is to set $\pi_0 = \pi_2 = 2$, yielding an expected utility of $(1/2) \cdot 0 + (1/2) \cdot 6 = 3$.

Now allow randomized rewards. The Nash equilibria do not change (because for every posterior belief $q(s)$ after signal s the principal has a unique best reply, which is pure: the single peak of $h_{q(s)}$). However, there is now a better optimal mechanism: signal 0 is rewarded with the half-half mixture between 0 and 2, signal 2 with the pure reward 1, and the agent reveals the whole truth (i.e., type 0 sends 0 and type 2 sends 2). The outcome is $\pi' = (1, 1)$ (and so incentive compatibility (IC) is satisfied), and the principal's expected utility is now $(1/2) \cdot ((1/2) \cdot 12 + (1/2) \cdot 0) + (1/2) \cdot 2 = 4$, higher than 3, which was the most that he could get when the rewards were pure. □

This example shows that with randomized rewards the principal can separate between agent's types while giving them the same expected rewards (and so incentive compatibility holds).[47] This yields a better outcome for the principal, but requires commitment: in the example, after seeing that the type is 0 the principal prefers to change the reward to 0.

We are again looking for conditions under which commitment yields no advantage, and so we must rule out such situations. Formally, we introduce the following condition:

(PUB) *Principal's Uniform Best.* For every utility level of the agent $x \in X$

[47] As shown by Chakraborti and Harbaugh (2010), cheap talk may yield separation for an agent (the sender) with type-independent utility: the agent is indifferent, whereas the principal (the receiver) is not.

there is a mixed reward ξ_x such that $\mathbb{E}[\xi_x] = x$ and $h_t(\xi_x) \geq h_t(\xi')$ for all types $t \in T$ and every mixed reward ξ' with $\mathbb{E}[\xi'] = x$.

Thus, for every utility level for the agent x in X there is a mixed reward ξ_x that yields utility x to the agent and is preferred by the principal, for every type t, to any other mixed reward that yields the same utility x to the agent ("uniform" refers here to "for every t").

A condition that is equivalent to (PUB) (and is more practical) uses the concept of "concavification." The *concavification* $\text{cav}f$ of a function $f : X \to \mathbb{R}$ is the smallest concave function that is everywhere no less than f; i.e., $\text{cav}f : X \to \mathbb{R}$ is a concave function, $f(x) \leq (\text{cav}f)(x)$ for every $x \in X$, and $(\text{cav}f)(x) \leq g(x)$ for any concave function g that satisfies $f(x) \leq g(x)$ for all x. Equivalently, the hypograph[48] of $\text{cav}f$ equals the convex hull of the hypograph of f; or,

$$\begin{aligned}(\text{cav}f)(x) &= \max\{\mathbb{E}[f(\xi)] : \mathbb{E}[\xi] = x\} \qquad (11)\\ &= \max\{\lambda f(y_1) + (1-\lambda)f(y_2) : \lambda y_1 + (1-\lambda)y_2 = x,\ \lambda \in [0,1]\}\end{aligned}$$

for every x. Consider the following condition:

(CAV) $(\text{cav}(\sum_{t\in T} q_t h_t))(x) = \sum_{t\in T} q_t(\text{cav}h_t)(x)$ for all $x \in X$ and *all* $q \in \Delta(T)$

What (CAV) says is that the concavification of each average of the functions h_t equals the average of the concavifications.

Remark. Some cases where (CAV) holds are:

(i) all the functions h_t are concave (and then $\text{cav}h_t = h_t$),

(ii) all the functions h_t are convex (and then all concavifications use only the two extreme rewards x_0 and x_1; i.e., $(\text{cav}h_t)(\lambda x_0 + (1-\lambda)x_1) = \lambda h_t(x_0) + (1-\lambda)h_t(x_1)$ for every t and $\lambda \in [0,1]$).

(iii) X is split into a number of intervals and all the functions h_t are concave in some intervals, convex in the rest, and the slopes decrease from one interval to the next.

[48]The *hypograph* of a real function $F : X \to \mathbb{R}$ consists of all points below the graph, i.e., $\{(x,\alpha) \in X \times \mathbb{R} : \alpha \leq F(x)\}$.

Proposition 16 *Conditions (PUB) and (CAV) are equivalent.*

Proof. (PUB) implies (CAV). For every $q \in \Delta(T)$ we have $(\text{cav} \sum_{t\in T} q_t h_t)(x) = \max\{\sum_{t\in T} q_t h_t(\xi) : \mathbb{E}[\xi] = x\} = \sum_{t\in T} q_t h_t(\xi_x)$ by (11) and then (PUB); taking in particular $q = \mathbf{1}_t$ yields $(\text{cav}h_t)(x) = h_t(\xi_x)$ for every t, and so we have (CAV).

(CAV) implies (PUB). Take $q \in \Delta(T)$ such that $q_t > 0$ for all t (for instance, $q = p$), and let ξ_x be such that $\text{cav}(\sum_{t\in T} q_t h_t)(x)$ is attained at ξ_x (see (11)), i.e., $\mathbb{E}[\xi_x] = x$ and $\text{cav}(\sum_{t\in T} q_t h_t)(x) = \sum_{t\in T} q_t h_t(\xi_x)$. The right-hand side is $\leq \sum_{t\in T} q_t \, \text{cav}h_t(x)$ (by (11)); because (CAV) says that we have equality, and all the q_t are positive, it follows that $h_t(\xi_x) = \text{cav}h_t(x)$ for every t, and hence $h_t(\xi_x) \geq h_t(\xi')$ for every ξ' with $\mathbb{E}[\xi'] = x$ (again by (11)). ■

Remarks. *(a)* The proof above shows that we can replace (CAV) with the weaker condition that there is *some* q in $\Delta(T)$ with strictly positive coordinates (i.e., $q_t > 0$ for all t) such that $(\text{cav}(\sum_{t\in T} q_t h_t))(x) = \sum_{t\in T} q_t(\text{cav}h_t)(x)$ for all $x \in X$.

(b) Call a mixed reward ξ *undominated* if there is no other mixed reward ξ' with the same expectation, i.e., $\mathbb{E}[\xi'] = \mathbb{E}[\xi]$, such that $h_t(\xi') \geq h_t(\xi)$ for all types $t \in T$, with strict inequality for at least one t. Then it can be shown that (PUB) is equivalent to (cf. (SP)):

(SP-M) *Single-Peakedness for Mixed Rewards.* For every probability distribution $q \in \Delta(T)$ on the set of types T, the expected utility of the principal is a single-peaked function of the agent's utility on the class of undominated mixed rewards; i.e., there exists a weakly[49] single-peaked function $g_q : X \to \mathbb{R}$ such that $h_q(\xi) = g_q(\mathbb{E}[\xi])$ for every undominated ξ.

If (SP-M) holds then, in particular, for undominated mixed rewards the utility of the principal is a *function* of the agent's utility: $\mathbb{E}[\xi] = \mathbb{E}[\xi']$

[49] A real function φ is *weakly single-peaked* if there exist $a \leq b$ such that φ increases for $x < a$, is constant for $a \leq x \leq b$, and decreases for $x > b$ (thus the interval $[a, b]$ is now the single flat top of φ; note that concave functions are weakly single-peaked). This weakening is needed since we want to allow piecewise linear functions, as we will see below.

implies[50] $h_q(\xi) = h_q(\xi')$. Moreover, $g_q(x) = \text{cav}h_q(x)$ by (11) (because the maximum there is always attained at some undominated ξ), and so weak single-peakedness follows because the functions g_q are concave. Thus (SP-M) is in fact equivalent to the simpler condition:

(SP-M1) For every type $t \in T$ there exists a function $g_t : X \to \mathbb{R}$ such that $h_t(\xi) = g_t(\mathbb{E}[\xi])$ for every undominated mixed reward ξ.

(It is enough to require this for every t, because $h_q = \sum q_t h_t$ gives $g_q = \sum q_t g_t$.)

These conditions are equivalent because they all amount to the requirement that for each utility level x of the agent, the set of feasible vector payoffs of the principal $\{(h_t(\xi))_{t\in T} : \mathbb{E}[\xi] = x\}$, which is a convex subset of $\mathbb{R}^T$, has a unique Pareto point (namely, $(h_t(\xi_x))_{t\in T} = (\text{cav}h_t(x))_{t\in T}$); in this case, "undominated" and "dominating everything else" are equivalent.

Because we no longer have single-peakedness (concavity yields a single flat plateau rather than a single peak), optimal mechanisms and truth-leaning equilibria need no longer yield a unique outcome. The simplest example has only one type t, with $h_t(x)$ concave and having a single flat for, say, $x \in [0, 1]$; then every $\pi_t \in [0, 1]$ is an optimal mechanism outcome, as well as a truth-leaning equilibrium outcome. However, the truth-leaning equilibrium outcomes are still optimal mechanism outcomes:

Theorem 17 *Assume that the payoff functions h_t satisfy the Principal-Uniform-Best condition (PUB). Then truth-leaning limit-equilibria exist, and their outcomes are optimal mechanism outcomes.*

Proof. Using for every utility level x for the agent the mixed reward ξ_x given by (PUB) is equivalent to replacing the principal's payoff functions h_t with their concavifications $g_t = \text{cav}h_t$. The functions g_t are concave, and so the result follows from Proposition 18 of the next section. ■

[50] If we were to require this for *all* mixed rewards it would follow that $h_q(x)$ are all linear (more precisely, affine) functions of x—which is much too restrictive a condition.

D.1 Weakly Single-Peaked Payoffs

In this section we prove the following generalization of our result, in the case of *pure* rewards. A real function f is *weakly-single-peaked* if there exist $x_0 \leq x_1$ such that f is strictly increasing for $x \leq x_0$, constant for $x_0 \leq x \leq x_1$, and strictly decreasing for $x \geq x_1$ (i.e., $x < x' \leq x_0$ implies $f(x) < f(x')$; $x_0 \leq x < x' \leq x_1$ implies $f(x) = f(x')$; and $x_1 \leq x < x'$ implies $f(x) > f(x')$). Single-peakedness (SP) is now weakened to

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

(WSP) holds in particular when all the h_t functions are concave (because then h_q is concave, and so weakly single-peaked).

Proposition 18 *Assume (WSP). Then truth-leaning equilibria exist, and their outcomes are optimal mechanism outcomes.*

Proof. For every $q \in \Delta(T)$ let $V(q) \equiv [v_0(q), v_1(q)]$ be the interval where h_q is maximal; then $X := [\min_t v_0(t), \max_t v_1(t)]$ is a compact interval that contains all the $V(q)$.

First, we claim that each Γ^ε with $0 < \varepsilon < 1$ has a Nash equilibrium. Indeed, the correspondence that assigns to each pair of strategies (σ, ρ) all (σ', ρ') with σ' a best reply to ρ and ρ' a best reply to σ (with values in X) is upper hemicontinuous and has closed and convex values, and so satisfies all the conditions of Kakutani's fixed-point theorem. Second, because (σ, ρ) belongs to $\Delta(T)^T \times X^T$, which is compact, limit points as $\varepsilon \to 0^+$ exist.

Let (σ, ρ) be a truth-leaning limit equilibrium: (σ_n, ρ_n) is a Nash equilibrium of Γ^{ε_n} with $\varepsilon_n \to 0^+$ and $(\sigma_n, \rho_n) \to (\sigma, \rho)$. Let $t \neq s$ be such that $\sigma(s|t) > 0$; then $\sigma_n(t|s) > 0$ for all large enough n. Then $\rho_n(t) \in V(t)$ and $\rho_n(s) \in V(q(s))$, and $\rho_n(s) \geq \rho_n(t) + \varepsilon_n > \rho_n(t)$, and thus $\rho_n(s) > v_0(t)$ for every $t \neq s$. Therefore, for every R that contains s we get $\rho_n(s) \leq v_1(q_n(s)|R)$; indeed, $h_{q_n(s)}(x)$, which is maximal at $\rho_n(s)$, is the average of $h_{q_n(s)|R}(x)$ and $h_t(x)$ for all $t \notin R$ with $\sigma_n(s|t) > 0$, all of which are nonincreasing at $\rho_n(s)$ (because, as we have seen, $\rho_n(s) > v_0(t)$), and so $h_{q_n(s)|R}(x)$

cannot be strictly decreasing there—hence the desired inequality. Because $q_n(s) \to q(s)$ and $q_n(s)|R \to q(s)|R$ (the denominators of these posteriors are bounded away from 0 by $p_s\sigma^n(s|s) = p_s > 0$) we get in the limit $v_0(t) \leq \rho(s) \leq \limsup_n v_1(q_n(s)|R) \leq v_1(q(s)|R)$ (the last inequality because a limit of maximizers is a maximizer in the limit).

Now apply Proposition 20 below. ■

The following example shows that not all optimal mechanism outcomes are obtained as truth-leaning equilibrium outcomes.

Example 15 $T = \{0, 2\}$, $p = (1/2, 1/2)$, $h_0(x) = -|x|$, $h_2(x) = -|x-2|$ (the functions h_t are concave, and therefore (WSP) holds; however, $h_p(x) = 1$ for $x \in [0, 2]$ and $= -|x-1|$ otherwise, and so h_p is not single-peaked: it has a plateau between 0 and 2). Let $L(0) = \{0, 2\}$ and $L(2) = \{2\}$.

The optimal mechanism outcomes are all $\pi = (\alpha, \alpha)$ for $\alpha \in [0, 2]$.

Only $\pi^* = (2, 2)$ is a truth-leaning equilibrium outcome: if (σ, ρ) is a Nash equilibrium of $\Gamma^{\boldsymbol{\epsilon}}$ then the posterior $q(2)$ after message 2 satisfies $q_2(2) > 1/2$ (because $\sigma(2|2) = 1 > 1 - \varepsilon_{0|0} = \sigma(0|2)$ and $p_0 = p_2$), and so $h_q(x)$ has a single peak at 2, which yields $\rho(2) = 2$ (and $\rho(0) = 0$ because only 0 can send 0), and the limit outcome is indeed $\pi^* = (2, 2)$. □

To prove Proposition 20 we start with a simpler case (cf. Proposition 5 in Section E).

Proposition 19 *Assume (WSP). If there is a type $s \in T$ such that $s \in L(t)$ for every t, and*[51]

$$\max_{t \neq s} v_0(t) \leq \min_{R: s \in R} v_1(R) \tag{12}$$

then for every $z \in V(T)$ the outcome $\pi^{(z)}$ with $\pi_t^{(z)} = z$ for all $t \in T$ is an optimal mechanism outcome.

Proof. Any constant outcome, such as $\pi^{(z)}$, satisfies all (IC) constraints (as equalities).

[51] Condition (12) is easily seen to be equivalent to $\max_{Q: s \notin Q} v_0(Q) \leq \min_{R: s \in R} v_1(R)$.

Let π be an optimal mechanism outcome for which $\min_{t\in T}\pi_t$ is maximal; i.e., π maximizes $H(\pi)$ subject to (IC), and every other maximizer π' satisfies $\min_t \pi'_t \le \min_t \pi_t$. Put $\alpha := \min_t \pi_t$ and $R := \{t \in T : \pi_t = \alpha\}$; then $s \in R$ (since $\pi_s \le \pi_t$ for every t by (IC)). 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, $\alpha' \le \beta$ 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 $\alpha \in v(R)$; moreover, the choice of π so that $\alpha = \min_t \pi_t$ is maximal implies that in fact $\alpha = v_1(R)$. Assumption (12) implies that $\alpha \ge v_0(t)$ for every $t \notin R$ (because then $t \ne s$), and so $h_t(\pi_t) \le h_t(\alpha)$ (because $\pi_t > \alpha \ge v_0(t)$ and the function h_t is nonincreasing for $x \ge v_0(t)$). Therefore $H(\pi) = \sum_t p_t h_t(\pi_t) = \sum_{t\in R} p_t h_t(\alpha) + \sum_{t\notin R} p_t h_t(\pi_t) \le \sum_t p_t h_t(\alpha) = h_T(\alpha) \le h_T(z) = H(\pi^{(z)})$ for every $z \in V(T)$; again, the optimality of π implies that we have equality, and so $\pi^{(z)}$ is an optimal mechanism outcome. ■

This yields the following generalization of Proposition 11:

Proposition 20 *Assume (WSP). Let (σ, ρ) be a Nash equilibrium that satisfies, for every message s that is used (i.e., $\bar\sigma(s) > 0$),*

(i) $v_0(t) \le \rho(s)$ for every $t \ne s$ that plays s (i.e., $\sigma(s|t) > 0$); and

(ii) $v_1(q(s)|R) \ge \rho(s)$ for every R containing s (i.e., $s \in R$).

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

Proof. Use the decomposition (7) of H given by σ, and for each s with $\bar\sigma(s) > 0$ apply Proposition 19 with $T = T_s$ and $p = q(s)$. ■

D.2 The Glazer–Rubinstein Setup

As stated in the Introduction, the work that is conceptually closest to the present paper is Glazer and Rubinstein (2004, 2006), or **GR** for short. The GR setup is more general than ours in the communication structure—having arbitrary messages rather than our truth structure (where messages are types and the mapping L satisfies (L1) and (L2))—and less general in the payoff structure—having only two pure rewards rather than single-peaked payoffs.

The first difference implies that in the GR setup truth and truth-leaning are not defined. Thus only one direction of the result holds: optimal mechanisms are always obtained by equilibria, but the converse is not true.[52]

As for the second difference, GR showed that their result cannot be extended in general to more than two pure rewards (the example at the end of Section 6 in Glazer and Rubinstein 2006[53]); Sher (2011) later showed that it does hold when the principal's payoff functions are concave.

The discussion on single-peakedness (SP-M) (and its equivalent condition (PUB)) in the previous sections of this Appendix D helps clarify this.[54] First, take the GR setup with only two pure rewards, say, 0 and 1. For every $x \in [0,1]$ there is a unique mixed reward yielding utility x to the agent, namely, getting 1 with probability x and 0 otherwise; therefore $h_t(x) = (1-x)h_t(0) + xh_t(1)$ is an affine function of x, and so is $h_q(x)$ for every $q \in \Delta(T)$, which implies that (SP-M) (or, equivalently, (PUB)) always holds in this case of only two pure rewards. Next, take the GR setup with more than two pure rewards. As Example 14 above showed, single-peakedness (SP-M) now becomes a restrictive condition that no longer holds in general. The Remark before Proposition 16 shows that it holds when all the functions h_t are concave—the assumption of Sher (2011)—as well as in other cases (for instance, when all the functions h_t are convex).

Finally, if one were to add to the GR setup the truth structure with conditions (L1) and (L2), then, under the (SP-M) or (PUB) condition, Theorem 17 would imply that truth-leaning equilibria yield optimal mechanisms.

[52] See the examples in Appendices B.1 and B.2.

[53] While the discussion there considers only pure rewards, it can be checked that the same holds for mixed rewards as well.

[54] It turns out to apply also to the case where there are finitely many rewards and randomizations are not allowed: it can be shown that (SP-M) is equivalent to the concavity of the functions h_t after a suitable increasing transformation is applied to the rewards.

E Appendix: From Mechanism to Equilibrium

We provide here a complete proof that every optimal mechanism yields a truth-leaning equilibrium with the same outcome.

Proposition 21 *Let $\pi^* \in \mathbb{R}^T$ be the outcome of an optimal mechanism; then there exists a truth-leaning equilibrium yielding the outcome π^*.*

To illustrate the idea of the proof, consider first the special case where the optimal mechanism outcome $\pi^* \in \mathbb{R}^T$ gives the same reward, call it α, to all types: $\pi^*_t = \alpha$ for all $t \in T$. Recalling Proposition 3, we define the strategy ρ of the principal by $\rho(t) = \min\{v(t), \pi^*_t\} = \min\{v(t), \alpha\}$ for all $t \in T$. As for the agent, let $S := \{t \in T : v(t) \geq \alpha\}$; the elements of S will be precisely the messages used in equilibrium, and we put $\sigma(t|t) = 1$ for all $t \in S$ and $\sigma(t|t) = 0$ for $t \notin S$. The question is how to define $\sigma(\cdot|t)$ for $t \notin S$.

If S consists of a single element s, then we put $\sigma(s|t) = 1$ for every t (and it is easy to verify that (ρ, σ) is then indeed a truth-leaning equilibrium). In general, however, S is not a singleton, and then we need carefully to assign to each type t those messages $s \in L(t) \cap S$ that t plays (we will see that the optimality of π^* implies that every t has some message to use, i.e., $L(t) \cap S \neq \emptyset$; see Claim 1 in the Proof of Proposition 21 below).

Consider a simple case (such as Example 11 in the Appendix) where $T = \{t, s, s'\}$, $S = \{s, s'\}$, $L(t) = T$, and the principal's payoff is quadratic (the value $v(R)$ of a set R is thus the expected value of its elements). How does type t choose between s and s' ? First, we have $v(t) < \alpha \leq v(s), v(s')$ by the definition of S. Second, again using the optimality of π^*, we get $v(T) = \alpha$ (otherwise, moving α toward $v(T)$ would increase the principal's payoff $H(\pi)$; see Claim 2 in the proof). Third, $v(\{t, s\}) \leq \alpha$ (because $v(\{t, s, s'\}) \equiv v(T) = \alpha$ and $v(s') \geq \alpha$), and similarly $v(\{t, s'\}) \leq \alpha$ (see Claims 3 and 4 in the proof; the argument in the general case is more complicated, and also relies on the optimality of π^*). Thus $v(\{t, s\}) \leq \alpha \leq v(s')$, and so there is some fraction $\lambda \in [0, 1]$ such that $v(\{\lambda * t, s\}) = \alpha$, where $\lambda * t$ denotes the λ-fraction of t (i.e., the value of the set containing s and the fraction λ of t is

exactly[55] α). Therefore $v(\{(1-\lambda)*t, s'\}) = \alpha$ too (because $v(T) = \alpha$), and we define $\sigma(s|t) = \lambda$ and $\sigma(s'|t) = 1-\lambda$.

When S contains more than two elements we get sets R_s for all $s \in S$ whose union is T, such that the value of each R_s, as well as the value of each union of them, is always $\leq \alpha$ (i.e., $v(\cup_{s\in Q} R_s) \leq \alpha$ for every $Q \subset S$; in the three-type example above $R_s = \{t, s\}$ and $R_{s'} = \{t, s'\}$). Using a simple extension of the classical Marriage Theorem of Hall (1935) to continuous measures due to Hart and Kohlberg (1974) (see Section E.1 below)[56] yields a partition of the set of types T into disjoint "fractional" sets F_s such that each F_s is a subset of R_s with value exactly α, i.e., $v(F_s) = \alpha$. This fractional partition gives the strategy σ, as above.

When we go beyond the quadratic case and the value v is not an expectation (and thus corresponds to an additive measure), we use the strict in-betweenness property instead (see Appendix C.2 (c)). Formally, we find it easier to replace conditions such as $v(R) \leq \alpha$ with their derivative counterparts $h'_R(\alpha) \leq 0$ (since being after the peak means being in the region where the function decreases), or, equivalently, $\sum_{t\in R} p_t\, h'_t(\alpha) \leq 0$. These derivative conditions add up over disjoint sets R, and they yield an additive measure to which the Marriage Theorem can be applied.[57]

Finally, the general case (where π^*_t is not the same for all t) is handled by partitioning T into disjoint "layers" $T^\alpha := \{t \in T : \pi^*_t = \alpha\}$ corresponding to the distinct values α of the coordinates of π^*, and then treating each T^α separately as in the special case above. One may verify that there is no

[55] Formally, $(\lambda p_t v(t) + p_{s_1} v(s_1))/(\lambda p_t + p_{s_1})$ is a continuous function of λ, which is $\geq \alpha$ at $\lambda = 0$ and $\leq \alpha$ at $\lambda = 1$.

[56] Hall's (1935) result is the following. There are n boys and n girls, each girl knows a certain set of boys, and we are looking for a one-to-one matching between boys and girls such that each girl is matched with a boy that she knows. Clearly, for such a matching to exist it is necessary that any k girls know together at least k different boys; Hall's result is that this condition is also sufficient.

Glazer and Rubinstein (2006) used a different line of proof (the "bridges problem") for a parallel result: construct an equilibrium (but without the additional requirement of getting it to be truth-leaning) from an optimal mechanism. We find that the very short inductive proof of Halmos and Vaughan (1950), as used in Hart and Kohlberg (1974), provides a simple procedure for constructing the agent's strategy; see below.

[57] One may instead directly apply continuity arguments to the v function, as in Hart and Kohlberg (1974).

interaction between the different layers (because T is finite there is a minimal positive gap $\delta_0 > 0$ between distinct values, and then we take the "slight" changes in the arguments above to be less than δ_0). Moreover, one advantage of the translation to conditions on derivatives, which are additive over sets, is that it allows us to carry out the arguments globally, without having to refer explicitly to the separate layers.

Proof of Proposition 21. Given π^*, define the strategy ρ of the principal by $\rho(t) = \min\{\pi^*_t, v(t)\}$ for all $t \in T$. It remains to construct the strategy σ of the agent so that (σ, ρ) is a truth-leaning equilibrium.

Let $S := \{t \in T : \pi^*_t \leq v(t)\} = \{t \in T : \rho(t) = \pi^*_t\}$; in view of Proposition 3, S contains those messages that will be used in equilibrium (i.e., σ will satisfy $\bar{\sigma}(t) > 0$ if and only if $t \in S$). For each $s \in S$ put $T_s := \{t \in T : s \in L(t) \text{ and } \pi^*_t = \pi^*_s\}$ and $R_s := T_s \cap (T \backslash S) \cup \{s\} \subseteq T_s$. The set R_s contains all the types that may potentially choose the message s in equilibrium: type s itself, together with all types $t \notin S$ such that $s \in L(t)$ and $\pi^*_t = \pi^*_s = \rho(s)$ (thus σ will satisfy $\sigma(s|t) > 0$ only if $t \in R_s$). The reason that we use the set R_s rather than T_s is that we want not only to obtain a Nash equilibrium, but also to satisfy the truth-leaning condition (A0), which will require every $s \in S$ to choose only s itself (the difference between T_s and R_s is that T_s may contain other $s' \in S$ in addition to s).

The strategy σ will correspond to a partition of the set of types T into disjoint subsets F_s (which consists of those types t that will choose s according to σ) such that for every $s \in S$ we have $F_s \subseteq R_s$, and also $v(F_s) = \pi^*_s$ (this is the principal's equilibrium condition (P')). As seen in the discussion preceding the proof, these sets may well be fractional sets, and then $F_s \subseteq R_s$ becomes "if $\sigma(s|t) > 0$ then $t \in R_s$," and $v(F_s) = \pi^*_s$ becomes $v(q(s)) = \pi^*_s$ (recall that $q(s)$ is the posterior given the message s, i.e., the "composition" of F_s). The existence of a fractional partition is obtained using an appropriate "marriage theorem"; the conditions needed to apply this result are provided in the following claims.

The first claim shows that for every type t there is a message in S that he may use to get his reward (i.e., $\pi^*_t = \pi^*_s = \rho(s)$ for some $s \in L(t) \cap S$).

Let $\delta_0 > 0$ be such that the gap between any two distinct values of π^* is at least δ_0; i.e., $\delta_0 := \min\{|\pi^*_t - \pi^*_{t'}| : \pi^*_t \neq \pi^*_{t'}\}$.

Claim 1 *Every $t \in T$ belongs to some R_s; i.e., $\cup_{s\in S} R_s = T$.*

Proof. Since $s \in R_s$ for every $s \in S$, we need to show that for every $t \notin S$ there is $s \in L(t)$ such that $\pi^*_s = \pi^*_t$ and $s \in S$. Let $K(t) := \{s \in L(t) : \pi^*_s = \pi^*_t\}$; the set $K(t)$ is nonempty since $t \in K(t)$. Assume by way of contradiction that $K(t) \cap S = \emptyset$, and so $\pi^*_s > v(s)$ for every $s \in K(t)$. For $0 \leq \delta \leq \delta_0$ let $\pi^\delta_s := \pi^*_s - \delta$ if $s \in K(t)$ and $\pi^\delta_s := \pi^*_s$ if $s \notin K(t)$. Then π^δ satisfies all the (IC) constraints. Indeed, take such a constraint $\pi_s \geq \pi_r$ for $r \in L(s)$. If π^* satisfied it as a strict inequality, then π^δ satisfies it because $\delta \leq \delta_0$ (which is the minimal gap); and if π^* satisfied it as an equality, π^δ satisfies it because π^*_s decreases by δ only when $s \in K(t)$, and then $r \in K(t)$ too (since $r \in L(t)$ by (L2) and $\pi^*_t = \pi^*_s = \pi^*_r$), and so π^*_r too decreases by δ. But $\pi^*_s > v(s)$ for all $s \in K(t)$, and so, for $\delta > 0$ small enough (so that $\pi^\delta_s \geq v(s)$ for all $s \in K(t)$), we get $H(\pi^\delta) - H(\pi^*) = \sum_{s\in K(t)} p_s \left(h_s(\pi^\delta_s) - h_s(\pi^*_s)\right) > 0$ (because π^δ_s is closer to $v(s)$ than π^*_s for all $s \in K(t)$). This contradicts the optimality of π^*. ■

The second claim corresponds to $v(T) = \alpha$ in the discussion at the beginning of the section.

Claim 2 $\sum_{t\in T} p_t\, h'_t(\pi^*_t) = 0$.

Proof. For every δ (positive, zero, and negative) let $\pi^\delta_s := \pi^*_s + \delta$ for all $s \in T$; then clearly π^δ satisfies all the (IC) constraints (since π^* does). The optimality of $\pi^* = \pi^0$ implies that $H(\pi^\delta) \leq H(\pi^0)$ for every δ, and so $H(\pi^\delta) = \sum_{t\in T} p_t\, h_t(\pi^*_t + \delta)$ is maximized at $\delta = 0$. Therefore its derivative with respect to δ vanishes at $\delta = 0$, i.e., $\sum_{t\in T} p_t\, h'_t(\pi^*_t) = 0$. ■

The next two claims correspond to the inequalities $v(\cup_{s\in Q} R_s) \leq \alpha$ for all $Q \subseteq S$ (again, see the discussion at the beginning of the section). We prove this first for the sets T_s in Claim[58] 3, and then for the sets R_s in Claim 4. For every nonempty subset $Q \subseteq S$ put $T_Q := \cup_{s\in Q} T_s$ and $R_Q := \cup_{s\in Q} R_s$.

[58]To get a Nash equilibrium that is not necessarily truth-leaning one works with the sets T_s instead of R_s, and then Claim 3 suffices.

Claim 3 $\sum_{t\in T_Q} p_t\, h_t'(\pi_t^*) \le 0$ *for every* $Q \subseteq S$.

Proof. For every $0 \le \delta \le \delta_0$ let $\pi_t^\delta := \pi_t^* + \delta$ if $t \in T_Q$ and $\pi_t^\delta := \pi_t^*$ if $t \notin T_Q$. Similarly to the argument in the proof of Claim 1, π^δ satisfies every (IC) constraint $\pi_{t'}^\delta \ge \pi_t^\delta$ (for $t \in L(t')$). If π^* satisfied it as a strict inequality, because $\delta \le \delta_0$; if π^* satisfied it as an equality, then if the right-hand side increased by δ then so did the left-hand side: $t \in T_Q$ implies[59] $t' \in T_Q$ (indeed: $t \in T_Q$ implies $t \in T_s$ for some $s \in Q$, and hence $s \in L(t)$ and $\pi_t^* = \pi_s^*$; together with $t \in L(t')$ and $\pi_{t'}^* = \pi_t^*$, as π^* satisfied this constraint as an equality, it follows that $s \in L(t')$ and $\pi_{t'}^* = \pi_s^*$, which means that $t' \in T_s \subseteq T_Q$).

Now $\sum_{t\in T_Q} p_t\ (h_t(\pi_t^* + \delta) - h_t(\pi_t^*)) = H(\pi^\delta) - H(\pi^*) \le 0$ for every $0 \le \delta \le \delta_0$ (by the optimality of π^*), and so the derivative at $\delta = 0$ is ≤ 0, which proves the claim. ■

Claim 4 $\sum_{t\in R_Q} p_t\, h_t'(\pi_t^*) \le 0$ *for every* $Q \subseteq S$.

Proof. We have $\sum_{t\in R_Q} p_t\, h_t'(\pi_t^*) = \sum_{t\in T_Q} p_t\, h_t'(\pi_t^*) - \sum_{t\in T_Q\setminus R_Q} p_t\, h_t'(\pi_t^*)$ (because $R_Q \subseteq T_Q$). Now $t \in T_Q\setminus R_Q$ implies $t \in S\setminus Q \subseteq S$, and so $h_t'(\pi_t^*) \ge 0$ (because $\pi_t^* \le v(t)$), which, together with Claim 3, completes the proof. ■

We can now conclude the proof of Proposition 21.

Proof of Proposition 21 (continued). First, Claim 1 implies that every $t \notin S$ belongs to R_s for some $s \in S$; together with $s \in R_s$ we get $R_S = \cup_{s\in S} R_s = T$. Let $\gamma_t := -p_t\, h_t'(\pi_t^*)$; the collection of sets $(R_s)_{s\in S}$ satisfies $\sum_{t\in R_Q} \gamma_t \ge 0$ for every $Q \subseteq S$ (by Claim 4), with equality for $Q = S$ (by Claim 2 since $R_S = T$). Applying Corollary 25 in Appendix E.1 to the collection $(R_s)_{s\in S}$ together with $\alpha_s = 0$ for every $s \in S$ yields $\sigma : T \to \Delta(S)$ such that, first,

$$\sigma(s|t) > 0 \text{ implies } t \in R_s. \tag{13}$$

And second, $h'_{q(s)}(x) = (1/\bar\sigma(s)) \sum_{t\in T} p_t\, \sigma(s|t)\, h_t'(x)$ vanishes at the point $x = \pi_s^* = \pi_t^*$ for all $t \in T_s$, because $\sum_{t\in T_s} p_t\, \sigma(s|t)\, h_t'(\pi_t^*) = -\sum_{t\in T} \sigma(s|t)\, \gamma_t = 0$.

[59] The reason that, unlike in Claim 2, we cannot take $\delta < 0$ is that there may be (IC) constraints for which we have equality $\pi_{t'}^* = \pi_t^*$, but $t' \in T_Q$ and $t \notin T_Q$.

The single-peakedness condition (SP) then implies that π^*_s is the single peak of $h_{q(s)}$, i.e.,

$$\pi^*_s = v(q(s)). \tag{14}$$

To conclude we verify that (σ, ρ) is indeed a truth-leaning equilibrium with outcome π^*. Recall that $\rho(s) = \pi^*_s \leq v(s)$ iff $s \in S$ and $\rho(t) = v(t) < \pi^*_t$ iff $t \notin S$. Then $\pi^*_t = \max_{r \in L(t)} \pi^*_r \geq \max_{r \in L(t)} \rho(r)$ by (IC), and Claim 1 implies that there is equality; thus the outcome is π^*. The agent's equilibrium condition (A) holds by (13): $\sigma(s|t) > 0$ implies $s \in S$ and $t \in R_s$, and so $s \in L(t)$ and $\pi^*_t = \pi^*_s = \rho(s)$. The truth-leaning condition (A0) holds because $\rho(s) = \pi^*_s$ iff $s \in S$, and then, since the only $R_{s'}$ that contains s is R_s, we have $\sigma(s|s) = 1$ by (13). The principal's equilibrium condition (P) holds because $\bar{\sigma}(s) > 0$ iff $s \in S$ by (13) and (A0), and then $\rho(s) = \pi^*_s = v(q(s))$ by (14). Finally, the truth-leaning condition (P0) holds because $\bar{\sigma}(t) = 0$ iff $t \notin S$, and then $\rho(t) = v(t)$. ■

Remarks. *(a)* For every value α of π^*, let $S^\alpha := \{s \in S : \pi^*_s = \alpha\}$ be the set of messages that yield outcome α. For $Q = S^\alpha$ we get $= 0$ (instead of ≤ 0) in Claims 3 and 4, because in the Proof of Claim 3 we can take also negative δ (with $|\delta| \leq \delta_0$), and $R_{S^\alpha} = T_{S^\alpha} = \{t : \pi^*_t = \alpha\}$ by Claim 1. Therefore the construction of σ can be carried out for each layer α separately.

(b) The short inductive proof of Lemma 4 in Hart and Kohlberg (1974) yields the following simple procedure for constructing the strategy σ. If there is a nonempty $Q_0 \subsetneqq S$ for which we have equality in Claim 3, then solve separately the two smaller problems $(R_s)_{s \in Q_0}$ and $(R_s \backslash R_{Q_0})_{s \in S \backslash Q_0}$. If there is strict inequality in Claim 3 for every $Q \neq S, \emptyset$, then take some $s_0 \in S$ and replace R_{s_0} with R'_{s_0} such that $R_{s_0} \backslash R_{S \backslash \{s_0\}} \subseteq R'_{s_0} \subseteq R_{s_0}$ and there is equality in Claim 4 for at least one $Q \neq S, \emptyset$.

Combining this with Remark (a) above implies that one can carry out this construction separately for each value α of π^*.

E.1 Hall's Marriage Theorem and Extensions

This appendix deals with the famous "Marriage Theorem" of Hall (1935) and its extensions that are used in our proofs.

Hall's result is as follows. A necessary and sufficient condition to be able to choose a distinct element from each one of a finite collection of finite sets is that the union of any k of these sets contains at least k distinct elements, for any k. Thus let $(W_m)_{m\in M}$ be a finite collection of finite sets, and let $W := \cup_{m\in M} W_m$ be their union. Then there exists a collection $(w_m)_{m\in M}$ of distinct elements of W (i.e., $w_m \neq w_{m'}$ for $m \neq m'$) such that $w_m \in W_m$ for all $m \in M$ if and only if[60] $|\cup_{m\in M} W_m| = |M|$ and $|\cup_{m\in K} W_m| \geq |K|$ for every $K \subseteq M$. For the connection to "marriage," let W_m be the set of women that man m knows; then Hall's Theorem tells us exactly when every man can be matched to a distinct woman whom he knows. To prepare for our extension, we state this formally as follows.[61]

Theorem 22 (Hall 1935) *Let M be a finite set, and $(W_m)_{m\in M}$ a finite collection of finite sets; put $W := \cup_{m\in M} W_m$. Let μ be the counting measure on M and ν the counting measure on W. If*

$$\nu(\cup_{m\in M} W_m) = \mu(M), \quad \textit{and} \tag{15}$$

$$\nu(\cup_{m\in K} W_m) \geq \mu(K) \quad \textit{for every } K \subseteq M, \tag{16}$$

then there exists a partition of[62] W into disjoint sets $(V_m)_{m\in M}$ satisfying

$$V_m \subseteq W_m \qquad \textit{for every } m \in M, \textit{ and} \tag{17}$$

$$\nu(V_m) = \mu(\{m\}) \quad \textit{for every } m \in M. \tag{18}$$

Can one extend this result to arbitrary measures (a measure λ on a finite set N is given by weights $\lambda_n \equiv \lambda(\{n\})$ for $n \in N$, i.e., $\lambda(I) = \sum_{n\in I} \lambda_n$

[60] For a finite set A, we write $|A|$ for the cardinality of A, i.e., the number of elements of A. We refer to this as the *counting measure* of A.

[61] We state only the nontrivial direction that the conditions are sufficient; see the Remark following Proposition 23.

[62] I.e., the sets V_m are disjoint and their union is W.

for $I \subseteq N$)? Consider the following example: $M = \{1, 2\}$; $W_1 = \{a, b\}$ and $W_2 = \{b, c\}$; μ and ν are the uniform probability measures on M and $W = \{a, b, c\}$, respectively (i.e., $\mu(\{m\}) = 1/2$ for $m = 1, 2$ and $\nu(w) = 1/3$ for $w = a, b, c$). Conditions (15) and (16) clearly hold, but we cannot partition $W = \{a, b, c\}$ into two disjoint sets $V_1 \subseteq W_1$ and $V_2 \subseteq W_2$ with probability $1/2$ each, as that would require us to "split" the element b half-half between V_1 and V_2.

We will show that the extension is indeed possible when such splitting is not needed (namely, when the measure ν is continuous and has no atoms), or when it is allowed (in the form of "fractional" sets).

We start with the nonatomic case where the set W is infinite and the measure ν has no atoms (the finiteness of M is kept throughout).

Proposition 23 (Hart–Kohlberg 1974) *Let M be a finite set and $(W_m)_{m \in M}$ a finite collection of sets; put $W := \cup_{m \in M} W_m$. Let μ be a measure on M and ν a* nonatomic *finite measure on*[63] *W. If (15) and (16) hold, then there exists a partition of W into disjoint sets $(V_m)_{m \in M}$ satisfying (17) and (18).*

Proof. This is the lemma in Section 4 of Hart and Kohlberg (1974),[64] with two minor improvements: first, the measure μ is not required to be nonnegative (a condition that appears in the Hart–Kohlberg statement but is not used in the proof there); and second, the sets V_m that are obtained satisfy in addition $\cup_{m \in M} V_m = W = \cup_{m \in M} W_m$ (which is easily seen to hold by the inductive construction in the proof there). ■

Remark. The converse (i.e., a partition of W exists only if (15) and (16) hold) is no longer true (it is when ν is a nonnegative measure, since then (17) implies $\nu(\cup_{m \in K} V_m) \leq \nu(\cup_{m \in K} W_m)$).

When the measure ν has atoms (as is the case when W is a finite set), we introduce the possibility of splitting atoms between sets. Formally, we

[63] Formally, ν is defined on a σ-field $\mathcal{F}$ of subsets of W, which contains all the relevant sets. The measure ν is *nonatomic* if for every S with $\nu(S) \neq 0$ there is $S_1 \subset S$ such that $\nu(S_1) \neq 0$ and $\nu(S \backslash S_1) \neq 0$. All subsets of W and all functions on W that we use are taken to be measurable.

[64] Whose simple proof is inspired by the simple inductive proof of Halmos and Vaughan (1950) of Hall (1935)'s Marriage Theorem.

identify a subset V of W with its characteristic function $V : W \to \{0,1\}$ (where $w \in V$ if and only if $V(w) = 1$), and we define a *fractional* subset V of W as a function $V : W \to [0,1]$, where $V(w)$ is understood as the fraction of w that belongs to V. A *partition* of W into disjoint *fractional* sets[65] $(V_m)_{m\in M}$ requires that each element $w \in W$ belong in certain proportions to the various sets V_m, and these proportions add up to unity; that is, $V_m : W \to [0,1]$ for each $m \in M$ and $\sum_{m\in M} V_m(w) = 1$ for each $w \in W$. For fractional sets V_m, the inclusion $V_m \subseteq W_m$ says that if $V_m(w) > 0$ then[66] $w \in W_m$, and the measure $v(V_m)$ is given by[67] $\int_W V_m \, d\nu$. We have:[68]

Corollary 24 *Let M be a finite set and $(W_m)_{m\in M}$ a finite collection of sets; put $W := \cup_{m\in M} W_m$. Let μ be a measure on M and ν a finite measure on W. If (15) and (16) hold, then there exists a partition of W into disjoint* fractional *sets $(V_m)_{m\in M}$ satisfying (17) and (18).*

Proof. Replace each atom w of the measure ν with a nonatomic continuum C_w with the same measure and apply Proposition 23; $V_m(w)$ in the original space is then the proportion of C_w that belongs to V_m in the nonatomic space. ■

The partition $(V_m)_{m\in M}$ of W into fractional sets may equivalently be described by a function σ that assigns to each element w in W a probability distribution on M that gives the fractions of w in the various[69] V_m; that is, $\sigma : W \to \Delta(M)$ with[70] $\sigma(m|w) := V_m(w)$ for each $m \in M$ and $w \in W$. When W is a finite set and the measures μ and ν are given by the weights $(\mu_m)_{m\in M}$ and $(\nu_n)_{w\in W}$, Corollary 24 may be restated as follows.

[65] Known also as a "partition of unity"; fractional sets are also referred to as "fuzzy sets" and "ideal sets."

[66] Viewing W_m as $W_m : W \to \{0,1\}$ allows us to write this condition as $V_m \leq W_m$ (i.e., $V_m(w) \leq W_m(w)$ for every $w \in W$).

[67] When W is a finite set, $\nu(V_m) = \sum_{w\in W} V_m(w)\, \nu(\{w\})$.

[68] The extension of Hall's Theorem to fractional sets may thus be called "Hall's Hull," short for "The Convex Hull of Hall's Theorem."

[69] Referred to as a "Markov kernel."

[70] We write $\sigma(m|w)$ for the m-th coordinate of the probability distribution $\sigma(w) \in \Delta(M)$.

Corollary 25 *Let M be a finite set and $(W_m)_{m\in N}$ a finite collection of finite sets; put $W := \cup_{m\in M} W_m$. Let μ_m for each $m \in M$ and ν_w for each $w \in W$ be real numbers such that*

$$\begin{aligned} \sum_{w\in W} \nu_w &= \sum_{m\in M} \mu_m \quad \textit{and} \\ \sum_{w\in \cup_{m\in K} W_m} \nu_w &\geq \sum_{m\in K} \mu_m \quad \textit{for every } K \subseteq M. \end{aligned}$$

Then there exists a function $\sigma : W \to \Delta(M)$ such that for every $m \in M$

$$\begin{aligned} &\sigma(m|w) > 0 \quad \textit{implies} \quad w \in W_m, \quad \textit{and} \\ &\sum_{w\in W} \sigma(m|w)\, \gamma_w = \mu_m \ . \end{aligned}$$

For an application, consider a school where each student registers in one or more clubs (the chess club, the singing club, the writing club, and so on). Assume that the average grade of all the students in the school equals $\bar{g}$, and that the average grade of all the students registered in each club, as well as in each collection of clubs, is at least[71] $\bar{g}$ (for a collection of clubs K, we take all the students that registered in at least one of the clubs in K and average their grades). Corollary 25 implies that there is a way to divide each student's time among the clubs in which he registered, in such a way that the average grade in *each* club is *exactly* $\bar{g}$ (the average is now a weighted average, with each student's weight being his relative time in the club).[72]

[71] This is consistent with the tendency of high-grade students to register in more clubs than low-grade ones.

[72] Let M be the set of clubs, W_m the set of students in club m, and $W := \cup_{m\in M} W_m$ the set of all students. Let g_w be the grade of student w; then $\bar{g} = \sum_{w\in W} g_w/|W|$ is the average grade. Finally, let the measure ν on W be given by the weights $\nu_w = g_w - \bar{g}$, and let $\mu = 0$ be the measure on M.

E.2 An Alternative Proof of the Glazer–Rubinstein–Sher Result

In the Glazer–Rubinstein–Sher setup, where the sets are arbitrary subsets of a message space M (and so neither (L1) nor (L2) can be assumed), we have the following:

Proposition 26 *In the setup with a general message space: if all the functions h_t are concave, then every optimal mechanism outcome is also a Nash equilibrium outcome.*

Proof. We claim that it suffices to prove the result when the functions h_t are strictly concave and differentiable.

Indeed, as in the Proof of Theorem 17 in Section D.1, we first replace the functions h_t with their concavifications $g_t = \mathrm{cav}h_t$. Next, let π^* be an optimal mechanism outcome for g, i.e., $\pi^* \in \mathrm{OM}(g)$. For every $\varepsilon > 0$, put $g_t^\varepsilon(x) := g_t(x) - \varepsilon(x - \pi_t^*)^2$; then π^* is an optimal mechanism outcome for g^ε; because the functions g_t^ε are strictly concave, it is the unique element there, i.e., $\{\pi^*\} = \mathrm{OM}(g^\varepsilon)$. Next smooth out g^ε: for each $\delta > 0$ put $g^{\varepsilon,\delta}(x) := (2\delta)^{-1} \int_{-\delta}^{\delta} g_t^\varepsilon(x+y)\mathtt{d}y$; then the function $g_t^{\varepsilon,\delta}$ is strictly concave and differentiable; let $\pi^{\varepsilon,\delta}$ be the unique optimal mechanism outcome for $g^{\varepsilon,\delta}$, i.e., $\{\pi^{\varepsilon,\delta}\} = \mathrm{OM}(g^{\varepsilon,\delta})$. Assuming that we have shown the result for strictly concave and differentiable functions, it follows that $\pi^{\varepsilon,\delta} \in \mathrm{EQ}(g^{\varepsilon,\delta})$. Letting $\delta \to 0^+$ while keeping ε fixed yields $g^{\varepsilon,\delta} \to_\delta g^\varepsilon$, and so, by the upper hemicontinuity of OM, $\pi^{\varepsilon,\delta} \to_\delta \pi^*$ (which is the unique element of $\mathrm{OM}(g^\varepsilon)$. Now use the upper hemicontinuity of EQ: first, $\pi^{\varepsilon,\delta} \to_\delta \pi^*$ and $g^{\varepsilon,\delta} \to_\delta g^\varepsilon$ imply $\pi^* \in \mathrm{EQ}(g^\varepsilon)$; and second, $g^\varepsilon \to_\varepsilon g$ implies $\pi^* \in \mathrm{EQ}(g)$.

We thus assume that all the functions h_t are strictly concave and differentiable.

Let π be the unique optimal mechanism outcome, and let ρ be the reward scheme. For every $m \in M$ let $T_m := \{t \in T : m \in L(t) \text{ and } \rho(m) = \pi_t\}$ be the set of types for which message m is optimal for the payment scheme ρ. For every set of messages $Q \subset M$, let $T_Q := \cup_{m \in Q} T_m$ be the set of types for which some message in Q is optimal.

Claim 1. $\sum_{t\in T} p_t\, h_t'(\pi_t) = 0$.

Proof. For every δ let $\rho^\delta(m) := \rho(m) + \delta$ for all $m \in M$; then $\pi_t^\delta := \max_{m\in L(t)} \rho^\delta(t) = \pi_t + \delta$ for every $t \in T$. The optimality of ρ as a solution of MD implies that $\sum_{t\in T} p_t\, h_t(\pi_t) \geq \sum_{t\in T} p_t\, h_t(\pi_t^\delta) = \sum_{t\in T} p_t\, h_t(\pi_t + \delta)$ for every δ, and so this expression is maximized at $\delta = 0$. Therefore its derivative with respect to δ vanishes at $\delta = 0$, i.e., $\sum_{t\in T} p_t\, h_t'(\pi_t) = 0$. □

Claim 2. $\sum_{t\in T_Q} p_t\, h_t'(\pi_t) \leq 0$ for every $Q \subseteq M$.

Proof. Let $\delta_0 > 0$ be such that the gap between any two distinct values of ρ is at least δ_0 (i.e., $\delta_0 := \min\{|\rho(m) - \rho(m')| : \rho(m) \neq \rho(m')\}$; recall that M is finite). For every $\delta \geq 0$ such that $\delta < \delta_0$, let $\rho^\delta(m) := \rho(m) + \delta$ if $m \in Q$ and $\rho^\delta(m) := \rho(m)$ if $m \notin Q$; put $\pi_t^\delta := \max_{m\in L(t)} \rho^\delta(t)$ for every $t \in T$. Then $\pi_t^\delta = \pi_t + \delta$ for $t \in T_Q$ and $\pi_t^\delta = \pi_t$ for $t \notin T_Q$, since the maximal payoff increases by δ for those types for which some message in Q was optimal under ρ (because $\delta < \delta_0$, the maximal payoff for all other types is unchanged). The optimality of ρ as a solution of MD implies that

$$\begin{aligned} 0 &\leq \sum_{t\in T} p_t\, h_t(\pi_t) - \sum_{t\in T} p_t\, h_t(\pi_t^\delta) \\ &= \sum_{t\in T_Q} p_t\, h_t(\pi_t) - h_t(\pi_t + \delta)) \end{aligned}$$

for every $0 \leq \delta < \delta_0$. Therefore the derivative at $\delta = 0$ is ≤ 0, which gives our result. □

Let $\gamma_t := -p_t\, h_t'(\pi_t)$; the collection of sets $(T_m)_{m\in M}$ satisfies $\sum_{t\in T_Q} \gamma_t \geq 0$ for every $Q \subseteq M$ (by Claim 2), with equality for $Q = M$ (by Claim 1). Therefore, by Proposition 25 in Appendix E.1 (with $\alpha_m = 0$ for all $m \in M$) there exists a strategy $\sigma : T \to \Delta(M)$ such that, first, $\sigma(m|t) > 0$ only if $t \in T_m$, and so $\sigma(\cdot|t)$ gives positive probability only to messages that are optimal for t under ρ. Second, the derivative of $h_{q(m)}(x) = (1/\bar\sigma(m)) \sum_{t\in T} p_t\, \sigma(m|t)\, h_t(x)$ at $x = \rho(m)$ $(= \pi_t$ for all $t \in T_m)$ equals zero since $\sum_{t\in T_m} p_t\, \delta(m|t)\, h_t'(\pi_t) = -\sum_{t\in T} \sigma(m|t)\, \gamma_t = 0$, and thus $h_{q(m)}(x)$ is maximized at $x = \rho(m)$. ■