Evidence Games: Truth and Commitment*

Sergiu Hart† Ilan Kremer‡ Motty Perry§

July 17, 2015

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 to the mechanism-design setup where he can commit in advance to a reward policy. The main result is that under natural conditions on the truth structure of the evidence, the two setups yield the *same* equilibrium outcome.

*First version: February 2014. The authors thank Elchanan Ben-Porath, Peter DeMarzo, Kobi Glazer, Johannes Hörner, Vijay Krishna, Phil Reny, Ariel Rubinstein, Amnon Schreiber, Andy Skrzypacz, Rani Spiegler, Yoram Weiss, and David Wettstein, for useful comments and discussions.

†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

‡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

§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

Contents

1 Introduction

Ask someone if they deserve a pay raise. The invariable reply (with very few, and therefore notable, exceptions) is, "Of course I do." 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?[1]

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.[2] The present pa-

[1]Thus "single-crossing"-type properties do not hold here, which implies that usual separation methods (as in signaling, etc.; see Section 1.2) cannot help.

[2]Think for instance of the advantage that it confers in bargaining, and in oligopolistic competition (Stackelberg vs. Cournot). See also Example 3 in Section 1.1 that is closer to our setup.

per 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).[3]

An *evidence game* is a standard communication game between an "agent" who is informed and sends a message (that does not affect the payoffs) and a "principal" who chooses the action (call it the "reward"). The two distinguishing features of evidence games are, first, that the agent's private information (the "type") consists of certain pieces of verifiable evidence, and the agent can reveal in his message all this evidence (the "whole truth"), or only some part of it (a "partial truth").[4] The second feature is that the agent's preference order on the rewards is the same regardless of his type—he always prefers the reward to be as high as possible[5]—whereas the principal's utility, which does depend on the type, is single-peaked with respect to the agent's order—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, obtain when 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 infinitesimal increases[6] in the agent's utility and in the probability of telling the whole truth. Specifically, (i) when the reward for revealing some partial truth is the same as the reward for revealing the whole truth, the agent prefers to

[3]See Sections 1.2 and 5.3 where we discuss in detail the relations between the work of Glazer and Rubinstein and the present paper.

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

[5]This is why adding (cheap-talk) messages that any type can use does not help here: all types will use those messages that yield the highest rewards.

[6]Formally, by limits as these increases go to zero.

reveal the whole truth; and (ii) there is a small positive probability that the whole truth is revealed.[7] These conditions, which are part of the setup, and are called *truth-leaning*, are most natural. The truth is after all a focal point, and there must be good reasons for *not* telling it.[8] As Mark Twain wrote, "When in doubt, tell the truth," and "If you tell the truth you don't have to remember anything."[9] With truth-leaning, the resulting equilibria turn out to be precisely those used in the voluntary disclosure literature; moreover, they satisfy the various refinement conditions offered in the literature.[10] See the examples and the discussion in Section 1.1.

The interaction between the two players may be carried out in two distinct ways. 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.[11,12] The resulting equilibria will be referred to as *equilibria* without commitment, and *optimal mechanisms* with commitment, respectively.

We can now state the main equivalence result.

> *In evidence games the equilibrium outcome obtained without commitment coincides with the optimal mechanism outcome obtained with commitment.*

Section 1.1 below provides two simple examples that illustrate the result and the intuition behind it.

An important consequence of the equivalence is that, in the basic case (where the reward equals the conditional expected value), the equilibria yield *constrained Pareto efficient* outcomes (i.e., outcomes that are Pareto efficient

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

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

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

[10] Such as the "intuitive criterion" of Cho and Kreps (1987), "divinity" and "universal divinity" of Banks and Sobel (1987), and the "never weak best response" of Kolberg and Mertens (1986).

[11] The latter is a *Stackelberg* setup, with the principal as leader and the agent as follower.

[12] Interestingly, what distinguishes between "signaling" and "screening" (see Section 1.2 below) is precisely these two different timelines of interaction.

under the incentive constraints).[13] In general, the fact that commitment is not needed in order to obtain optimality is a striking feature of evidence games. Moreover, we will show that the "truth structure" of evidence games (which consists of the partial truth relation and truth-leaning) 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 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. What our result says is that the market's behavior in equilibrium is in fact optimal: it yields the optimal separation that may be obtained between "good" and "bad" firms (i.e., even if mechanisms and commitments were possible they could not be separated more).

The second application has to do with the legal doctrine known as "the right to remain silent." In the United States, this right was 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 light on this debate. Indeed, because equilibria entail (Bayesian) inferences, our result implies that the same inferences apply to the optimal mechanism. Therefore adverse inferences should be allowed, and surely not committedly disallowed. In England, 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

[13]These outcomes yield the maximal separation that is worthwhile for the principal—or the market—to get; see the examples in Section 1.1 and the rest of the paper.

which you later rely on in court," which may be viewed, on the one hand, as allowing adverse inference, and, on the other hand, 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).

The third application concerns medical overtreatment, which is one of the more serious problems in many health systems in the developed world; see Brownlee (2008).[14] A reason that doctors and hospitals overtreat may be fear of malpractice suits; but the more powerful reason is that they are paid more for doing so. One suggestion for overcoming this is to reward doctors for providing evidence. The present paper takes a small step towards a better understanding of an optimal incentive scheme designed to reward revelation of evidence in these and other applications.

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 equilibrium without commitment and optimal mechanism with commitment in evidence games (which, in the basic case of quadratic loss, implies that the equilibria are constrained Pareto efficient); and third, we show that the conditions of evidence games—most importantly, the truth structure—are the *indispensable* conditions beyond which this equivalence no longer holds.

The paper is organized as follows. After the Introduction (which continues below with some examples and a survey of relevant literature), we describe the model and the assumptions in Section 2. The main equivalence result is then stated in Section 3, and proved in Section 4. We conclude with discussions on various extensions and connections in Section 5. The Appendix shows that our conditions are indispensable for the result (Section A.1), and provides a useful alternative proof of one direction of the equivalence result (Section A.2).

[14] Between one-fifth and one-third of U.S. health-care expenditures do nothing to improve health.

1.1 Examples

We provide here two 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,[15] while the professor would naturally like his salary to be as high as possible. The dean, knowing that similar professors' salaries range between, say, 0 and 120, 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 into 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. See Figure 1.

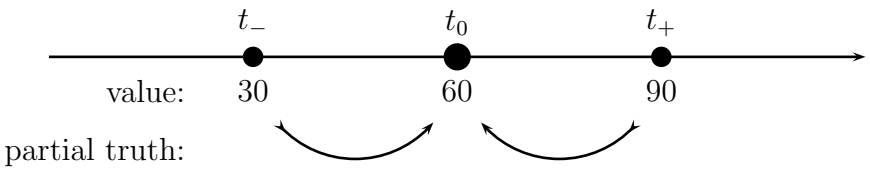

Figure 1: Values and possible partial truth messages in Example 1

Consider first the game setup (without commitment): the professor decides whether to reveal his evidence, if he has any, and then the dean chooses

[15] 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 conditional 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."

the salary. It is easy to verify that there is a unique sequential equilibrium,[16] 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, which is the expected value of the two types that provide no evidence: the no-evidence type together with the negative-evidence type.[17] See Figure 2.

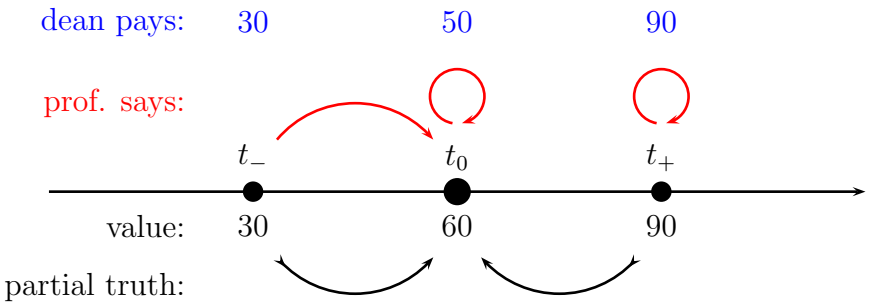

Figure 2: Equilibrium in Example 1

Next, consider the mechanism setup (with commitment): the dean commits to a salary policy (specifically: three salaries, denoted 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 pro-

[16] Indeed, in a 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 someone providing negative evidence must be 30. This shows that the so-called "babbling 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 at the average value of 60—is not a sequential equilibrium here. Finally, we note that truth-leaning yields sequential equilibria.

[17] The conditional expectation is $(50\% \cdot 60 + 25\% \cdot 30)/(50\% + 25\%) = 50$.

vide 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,[18] 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[19] 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. See Figure 3.

[18] 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 3 for a formal statement of this property, which is at the heart of our argument.

[19] There is nothing special about the specific numbers that we use.

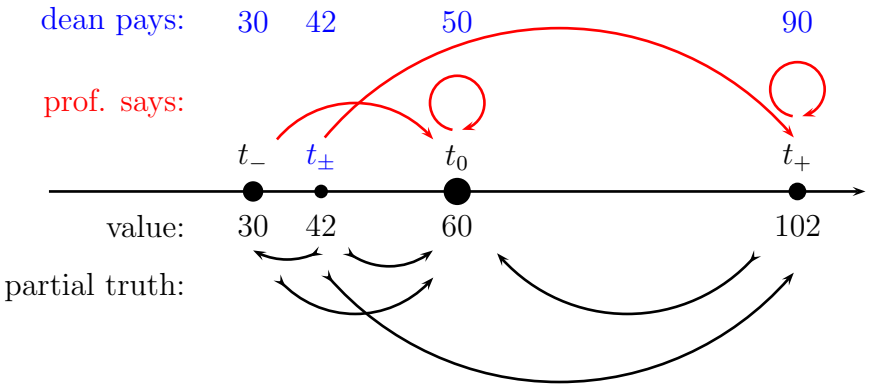

Figure 3: Equilibrium in Example 2

Now, however, the "babbling equilibrium" (in which the professor, regardless of his type, never provides any evidence, and the dean ignores any evidence that might be provided and sets the salary at the average value of 60—clearly, this is worse for the dean as it yields no separation between the types) is a sequential equilibrium. This 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.[20] The babbling 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 at 102, the value of[21] 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 char-

[20] In fact, this babbling equilibrium satisfies all the standard refinements in the literature (intuitive criterion, D1, divinity, never weak best response); see Section 5.4.

[21] 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 $t_\pm$), this would not suffice in general; see Example 8 in Section A.1.4.

acteristic 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 the general setup of communication games, 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 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).[22]

There can be no separation between the two types in equilibrium, since that would imply that t_0 gets a salary of 60 and t_- gets a salary of 30—but then t_- would prefer not to reveal his evidence and get 60 too. Thus the babbling equilibrium where no evidence is provided and the salary is set at 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).[23] 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. □

[22]Formally, the utility of t_- when he gets salary x is $-(x-50)^2$. The arguments below and the conclusion would not 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. See also Example 9 in the Appendix.

[23]The *optimal* mechanism pays a salary of 70 for no evidence.

To put it formally, *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—see Section A.1.5 in the Appendix; Section A.1.6 there provides another example where commitment yields strictly better outcomes).

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 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 they have different utilities and 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).[24]

When the agent's utility does not however depend on his information, in order to get any separation the agent's utility would need to depend on his communication. A simple and standard way of doing this is for different types to have different sets of possible messages.[25]

First, 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

[24]These setups differ in whether the agent's utility depends on his actions and/or messages—it does in signaling and screening models, but not in cheap talk—and in who plays first—the agent in signaling, the principal in screening (which translates into the distinction between game equilibrium and optimal mechanism). The key condition for separation in these setups is "single-crossing."

[25]Which is the same as taking the cost of the message to be zero when it is feasible, and infinite when it isn't.

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.[26] Since such models typically possess multiple equilibria, these papers focus on what they view as the more natural equilibrium. The selection criteria that they employ are model-specific. However, it may be 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.

Second, 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[27] 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) characterize the social choice functions that can be implemented when agents can also supply hard proofs about their types.

[26]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).

[27]They allow an arbitrary dependence of the agent's utility on his type.

The approach we are taking of comparing equilibria with 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) has provided conditions (namely, concavity) under which the result holds when the principal's decision is no longer binary. See Section 5.3 for a detailed discussion of the Glazer–Rubinstein setup.

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

The model is a communication game in which the set of messages is the set of types and the set of actions is the real line $\mathbb{R}$. The voluntary disclosure models (Grossman 1981, Milgrom 1981, Dye 1985, 1988, Shin 2003, 2006) are all special cases of this 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 distribution $p \in \Delta(T)$ with full support.[28] The principal knows the distribution p but does not know the realized type (which the agent does know).

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 consists of a real number x in $\mathbb{R}$. The message is costless: it does not affect the payoffs of the agent and the principal. The next sections will provide the

[28] $\Delta(T) := \{p = (p_t)_{t \in T} \in \mathbb{R}_+^T : \sum_{t \in T} p_t = 1\}$ is the set of probability distributions on the finite set T. Full support means that $p_t > 0$ for every $t \in T$.

details, including in particular the timeline of the interaction.

The interpretation to keep in mind is that the type is the (verifiable) evidence that the agent possesses, and the message is the evidence that he reveals.

2.1 Payoffs and Single-Peakedness

A basic 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). Since only the ordinal preference of the agent matters,[29] we assume without loss of generality that the agent's payoff is x itself,[30] and refer to x also 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[31] 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 the space of types—the expected utility of the principal is given, for every reward $x \in \mathbb{R}$, by

$$h_q(x) := \sum_{t \in T} q_t \, h_t(x).$$

The functions h_t are assumed to be *differentiable* (in the relevant domain, which will turn out to be a compact interval; see Remark (a) below), and to satisfy a single-peakedness condition. 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 single peak at v, and it strictly increases when $x < v$ and strictly decreases when $x > v$. The assumption on the principal's payoffs is:

[29] See Section 5.2 for randomized rewards.

[30] Formally, let the agent's utility be $U^A(x, s; t)$, where $x \in \mathbb{R}$ is the action, $s \in T$ is the message, and $t \in T$ is the type. Then $U^A(x, s; t)$ is, for every $t \in T$, a strictly increasing function $g_t(x)$ of x; without loss of generality, $g_t(x) = x$ for all t and x.

[31] Formally, writing $U^P(x, s; t)$ for the principal's utility when the action is $x \in \mathbb{R}$, the message is $s \in T$, and the type is $t \in T$, we have $U^P(x, s; t) = h_t(x)$ for all x, s, and t.

(SP) *Single-Peakedness.* The expected utility of the principal $h_q(x)$ is a single-peaked function of the reward x for every probability distribution $q \in \Delta(T)$ on the set of types T.

Thus, (SP) requires each function h_t to be single-peaked, and also each weighted average of such functions to be single-peaked. Let $v(t)$ and $v(q)$ denote the single peaks of h_t and h_q. Thus, $v(t)$ is the reward that the principal views as most fitting, or "ideal," for type[32] t; similarly, $v(q)$ is the ideal reward when the types are distributed according to q. When the distribution of types is given by the prior p, we will at times write $v(T)$ instead of $v(p)$, and, more generally, $v(S)$ instead of $v(p|S)$ for $S \subseteq T$ (where $p|S \in \Delta(T)$ denotes the conditional of p given[33] S).

Basic Example: Quadratic Loss. A particular case, common in much of the literature, uses for each type the quadratic distance from the ideal point: $h_t(x) = -(x - v(t))^2$ for each $t \in T$. In this case, for each distribution $q \in \Delta(T)$, the function h_q has its single peak at the expectation with respect to q of the peaks $v(t)$; i.e., $v(q) = \sum_{t \in T} q_t \, v(t)$.

More generally, for each type $t \in T$ let h_t be a strictly concave function that attains its (unique) maximum at a finite point,[34] denoted by $v(t)$. Since any weighted average of such functions clearly satisfies the same properties, the single-peakedness condition (SP) holds.[35] For instance, take h_t to be the negative of some distance from the ideal point $v(t)$. Even more broadly, the (SP) condition allows one to treat types differently, such as making different h_t more or less sensitive to the distance from the corresponding ideal point

[32] We will at times refer to $v(t)$ also as the *value* of type t (recall the examples in Section 1.1 in the Introduction).

[33] $(p|S)_t = p_t/p(S)$ for $t \in S$ and $(p|S)_t = 0$ for $t \notin S$, where $p(S) = \sum_{t \in S} p_t$ is the probability of S. Thus $v(S)$ is the single peak of $\sum_{s \in S} p_t \, h_t = p(S) \, h_{p|S}$.

[34] Functions that are everywhere increasing or everywhere decreasing are thus not allowed.

[35] Indeed, let h_1 and h_2 be strictly concave, with peaks at $v(1)$ and $v(2)$, respectively. For each $0 \leq \alpha \leq 1$ the function $h := \alpha h_1 + (1 - \alpha) h_2$ is also strictly concave, and it increases for $x < \underline{v} := \min\{v(1), v(2)\}$ and decreases for $x > \bar{v} := \max\{v(1), v(2)\}$ (because both h_1 and h_2 do so), and so attains its (unique) maximum in the interval $[\underline{v}, \bar{v}]$ (see Lemma 1 for the general statement of this "in-betweenness" property).

$v(t)$: e.g., take $h_t(x) = -c_t(x - v(t))^{\gamma_t}$ for appropriate constants c_t and γ_t. Also, the penalties for underestimating vs. overestimating the desired ideal point may be different: take the function h_t to be asymmetric around $v(t)$.

Remarks. *(a)* Let $X = [x_0, x_1]$ be a compact interval that contains the peaks $v(t)$ for all $t \in T$ in its interior (i.e., $x_0 < \min_{t \in T} v(t) \leq \max_{t \in T} v(t) < x_1$); without loss of generality we can then restrict the principal to actions x in X rather than in the whole real line $\mathbb{R}$ (because the functions h_t for all $t \in T$ are strictly increasing for $x \leq x_0$ and strictly decreasing for $x \geq x_1$, and so any x outside the interval X is strictly dominated for the principal by either x_0 or[36] x_1).

(b) The condition that the functions h_t for all $t \in T$ are single-peaked does *not suffice* to get (SP); and (SP) is more general than concavity of the functions h_t; see Section A.1.10 in the Appendix.

2.2 Information and Truth Structure

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." The possible messages of type t, i.e., the types s that t can pretend to be, are therefore those types s that possess less information than t ("less" is taken in the weak sense). This entails two conditions. First, revealing the whole truth is always possible: t can always say t. And second, "less information" is nested: if s has less information than t and r has less information than s, then r has less information than t; that is, if t can say s and s can say r then t can also say r.

This is formalized by a weak order[37] "$\rightarrowtail$" on the set of types T, with

[36] The strict inequalities $x_0 < \min_t v(t)$ and $x_1 > \max_t v(t)$ allow dominated actions to be played (for example, when the principal wants to make the reward for some message the worst, or the best).

[37] A weak order is a reflexive and transitive binary relation, i.e., $t \rightarrowtail t$ for all t, and $t \rightarrowtail s \rightarrowtail r$ implies $t \rightarrowtail r$ for all r, s, t. The order need *not be complete*: there may be t, s for which neither $t \rightarrowtail s$ nor $s \rightarrowtail t$ hold.

"$t \rightarrowtail s$" being interpreted as type t having (weakly) more information (i.e., evidence) than type s; we will say that "s is a partial truth at t" or "s is less informative than t." The set of possible messages of the agent when the type is t, which we denote by $L(t)$, consists of all types that are less informative 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 weak order "$\rightarrowtail$" translate into the following two conditions:

(L1) $t \in L(t)$; and

(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; (L2) says that if s is a partial truth at t and r is a partial truth at s, then r is a partial truth at t; thus, if t can pretend to be s and s can pretend to be r, then t can also pretend to be r.

Some natural models for the "less informative" relation are as follows.

(i) Evidence: Let E be a set of possible "pieces of evidence." A type is identified with a subset of E, namely, the set of pieces of evidence that the agent can provide (e.g., prove in court); thus, $T \subseteq 2^E$ (where 2^E denotes the set of subsets of E). Then $t \rightarrowtail s$ if and only if $t \supseteq s$; that is, s is less informative than t if t has every piece of evidence that s has. It is immediate that $\rightarrowtail$ is a weak order, i.e., reflexive and transitive. The possible messages at t are then either to provide all the evidence he has, i.e., t itself, or to pretend to be a less informative type s and provide only the pieces of evidence in the subset s of t (a partial truth).[38] See Examples 1 and 2 in the Introduction.

(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 $t \subseteq s$; that

[38] 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. The only undetectable deviations of t are to pretend that he is another possible type s that has fewer pieces of evidence.

is, s is less informative than t if and only if s provides less information than t, as more states ω are possible at s than at t. For example, take $\Omega = \{1, 2, 3, 4\}$ with the partitions $\Lambda_1 = (1234)$, $\Lambda_2 = (12)(34)$, and $\Lambda_3 = (1)(2)(3)(4)$. There are thus seven types: $\{1, 2, 3, 4\}, \{1, 2\}, \{3, 4\}, \{1\}, \{2\}, \{3\}, \{4\}$ (the first one from Λ_1, the next two from Λ_2, and the last four from Λ_3). Thus type $t = \{1, 2, 3, 4\}$ (who knows nothing) is less informative than type $s = \{1, 2\}$ (who knows that the state of nature is either 1 or 2), who in turn is less informative than type $r = \{2\}$ (who knows that the state of nature is 2); the only thing type t can say is t, whereas type s can say either s or t, and type r can say either r, s, or t. The probability p on T is naturally generated by a probability distribution μ on Ω together with a probability distribution λ on the set of partitions: if t is a ken in the partition Λ_i then $p_t = \lambda(\Lambda_i) \cdot \mu(t)$.

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

Remark. We emphasize that there is no relation between the value of a type and his information; i.e., $v(t)$ is an arbitrary function of t, and having more or less evidence says nothing about the value.

The second ingredient in the truth structure of evidence games is *truth-leaning*, which amounts to giving a slight advantage to revealing the whole truth (i.e., to type t sending the message t, which is always possible by (L1)). Formalizing this would require dealing with sequences of games (see Section 5.4 for details). Since only the equilibrium implications of truth-leaning matter, it is simpler to work directly with the resulting equilibria, which we call *truth-leaning equilibria*; see Section 2.3.1 below.

2.3 Game and Equilibria

We now consider the game where the principal moves after the agent (and cannot commit to a policy); in Section 2.4 we will consider the setup where the principal moves first, and commits to a reward policy before the agent makes his moves.

The *game* Γ proceeds as follows. 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 (mixed)[39] 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 (pure)[40] 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) constitute 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 condition for the agent is:

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

As for the principal, the condition is that the reward $\rho(s)$ to message s satisfies $h_{q(s)}(\rho(s)) = \max_{x \in \mathbb{R}} h_{q(s)}(x)$ for every $s \in T$ that is used, i.e., such that $\bar{\sigma}(s) > 0$. By the single-peakedness condition (SP), this is equivalent to $\rho(s)$ being equal to the single peak $v(q(s))$ of $h_{q(s)}$:

[39] The agent, who plays first, may need to randomize in equilibrium.

[40] As we will see shortly, the principal does not randomize in equilibrium.

(P) for every message $s \in T$: if $\bar{\sigma}(s) > 0$ then $\rho(s) = v(q(s))$.

Note that (P) implies in particular that the strategy ρ is pure.

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

$$\pi_t := \max_{s \in L(t)} \rho(s) \tag{1}$$

for every $t \in T$. Thus π_t is the reward when the type is t; the players' payoffs are then π_t for the agent and $h_t(\pi_t)$ for the principal.

Remark. In the basic quadratic-loss case, where, as we have seen, $v(q)$ equals the expectation of the values $v(t)$ with respect to q, condition (P) implies that the ex-ante expectation of the rewards, i.e., $\sum_{t \in T} p_t \, \pi_t = \mathbb{E}[\pi_t]$, 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 s that is used; now take expectation over s). Therefore *all* equilibria yield the same expected reward $\mathbb{E}[\pi_t]$, namely, the expected value $v(T)$; they may differ however in how this amount is split among the types (cf. Remark (c) in Section 3). For the principal (the "market"), the first best is to give to each type t its value $v(t)$, which also yields the same expected reward $v(T)$; however, this first best may not be achievable because the principal does not know the type.

2.3.1 Truth-Leaning Equilibria

As discussed in the Introduction, truth—more precisely, the whole truth—has a certain prominence. In evidence games, 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.[41] Second, there is an infinitesimal probability that the whole truth is revealed (which may happen because the agent is not strategic and instead always reveals his information—à la Kreps, Milgrom, Roberts, and Wilson 1982—or, because there may be "trembles," such as a slip of the tongue, or of the pen, or a document that is by mistake attached, or an unexpected piece of evidence).

[41] 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.

Formally, this yields the following two additional equilibrium conditions:[42]

(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 $\rho(s) = v(s)$.

Condition (A0) says that when the message t is optimal for type t, it is chosen for sure (i.e., the whole truth t is preferred by type t to any other *optimal* message $s \neq t$). 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 unit mass at s (i.e., s has probability one), to which the principal's optimal response is the peak $v(s)$ of $h_{q(s)} \equiv h_s$.

We will refer to a Nash equilibrium of Γ that satisfies (in addition to (A) and (P)) the conditions (A0) and (P0) as a *truth-leaning equilibrium*. See Section 5.4 for the corresponding "limit of perturbations" approach; we will also see there that truth-leaning satisfies the requirements of most, if not all, of the relevant equilibrium refinements that have been proposed in the literature (and coincides with many of them).

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

This is a standard *mechanism-design* framework. The reward scheme ρ is the *mechanism.* Given ρ, the agent chooses his message so as to maximize his reward; thus, the reward when the type is t equals $\max_{s \in L(t)} \rho(s)$. A reward

[42] We call them (A0) and (P0) since they are conditions (in addition to (A) and (P)) on the strategies σ of the agent and ρ of the principal, respectively.

scheme ρ is an *optimal mechanism* if it maximizes the principal's expected payoff, namely, $\sum_{t\in T} p_t\, h_t(\max_{s\in L(t)} \rho(s))$, among all mechanisms ρ.

The assumptions that we have made on the truth structure, i.e., (L1) and (L2), imply that the so-called "Revelation Principle" applies: any mechanism is equivalent to a direct mechanism where it is optimal for each type to be "truthful," i.e., to reveal his type (see Green and Laffont 1986).[43] Indeed, given a mechanism ρ, let $\pi_t := \max_{s\in L(t)} \rho(s)$ denote the reward (payoff) of type t when the reward scheme is ρ. If type t can send the message s, i.e., $s \in L(t)$, then $L(s) \subseteq L(t)$ by the transitivity condition (L2), and so $\pi_s = \max_{s'\in L(s)} \rho(s') \leq \max_{s'\in L(t)} \rho(s') = \pi_t$. These inequalities, namely, $\pi_t \geq \pi_s$ whenever $s \in L(t)$, are the "incentive compatibility" conditions that guarantee that no t can gain by pretending to be another possible type s (i.e., by acting like s). Conversely, any $\pi = (\pi_t)_{t\in T} \in \mathbb{R}^T$ that satisfies all these inequalities is clearly the outcome of the mechanism whose reward scheme is π itself; such a mechanism is called a *direct mechanism*.

The vector $\pi = (\pi_t)_{t\in T} \in \mathbb{R}^T$ is the *outcome* of the mechanism (when the type is t, the payoffs are π_t to the agent, and $h_t(\pi_t)$ to the principal).[44] The expected payoff of the principal, which he maximizes by choosing the mechanism, is

$$H(\pi) = \sum_{t\in T} p_t\, h_t(\pi_t). \tag{2}$$

In summary, an *optimal mechanism* outcome is a vector $\pi \in \mathbb{R}^T$ that satisfies:

(IC) *Incentive Compatibility.* $\pi_t \geq \pi_s$ for every $t \in T$ and $s \in L(t)$.

(OPT) *Optimality.* $H(\pi) \geq H(\pi')$ for every payoff vector $\pi' \in \mathbb{R}^T$ that satisfies (IC).

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

[43] Green and Laffont (1986) show that (L1) and (L2) are necessary and sufficient for the Revelation Principle to hold for any payoff functions.

[44] Thus, mechanism outcomes are the same as direct mechanisms.

scheme is made known to the agent, who then chooses his message s (out of the feasible set $L(t)$ when the type is t), and the game ends with the reward $\rho(s)$.

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

(c) Optimal mechanisms always exist, since H is continuous and the rewards π_t can be restricted to a compact interval X (see Remark (a) in Section 2.1).

(d) Truth-leaning does not affect optimal mechanisms (it is not difficult to show that incentive-compatible mechanisms with and without truth-leaning yield payoffs that are the same in the limit).

3 The Equivalence Theorem

In general, the possibility of commitment of the principal is significant, and equilibria of the game (where the principal moves second and cannot commit in advance) and optimal mechanisms (where the principal moves first and commits) may be quite different. Nevertheless, in our setup we will show that they coincide.

Our main result is:

Equivalence Theorem *Assume that the payoff functions h_t for all $t \in T$ are differentiable and satisfy the single-peaked condition (SP). Then 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 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).[45] 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 by giving 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 (as for the converse, it is immediate since whatever can be done without commitment can clearly also be done with commitment).

Remarks. *(a)* The Equivalence Theorem is stated in terms of outcomes, and thus payoffs, rather than strategies and reward policies. The reason is that there may be multiple truth-leaning equilibria, and multiple optimal mechanisms—but they all have the same outcome. Indeed, truth-leaning equilibria (σ, ρ) with outcome π coincide in their principal's strategy ρ (which is uniquely determined by π; see (6) in Proposition 2 below), but 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 (see Example 12 in Section A.1.8). 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 be other optimal mechanisms (specifically, the reward for a message t may be lowered when there is a message $s \neq t$ in $L(t)$ with $\pi_s = \pi_t$). Again, we emphasize that the resulting payoffs of both the agent and the principal, for all types t, are uniquely determined.

(b) The uniqueness of the outcome does *not* simply follow from single-peakedness, but is a more subtle consequence of our assumptions (see the examples in Section A.1 in the Appendix).[46]

[45] As much as these conditions seem reasonable, they need *not* hold for equilibria that are not truth-leaning.

[46] When the functions h_t are *strictly concave* for every t (as in the basic canonical case) the uniqueness of the optimal mechanism outcome is immediate, because averaging optimal mechanisms yields an optimal mechanism.

(c) In the basic quadratic-loss case, where $h_t(x) = -(x - v(t))^2$, the agent is indifferent among all equilibria (because his expected payoff equals the expected value $v(T)$; see the Remark in Section 2.3). As for the principal, the Equivalence Theorem implies that the truth-leaning equilibria are the ones that maximize his payoff. Therefore the truth-leaning equilibria are precisely the constrained Pareto efficient equilibria.[47]

The Equivalence Theorem is proved in the next section. After some preliminaries in Section 4.1—which in particular provide useful properties of truth-leaning equilibria and optimal mechanisms—we prove, first, that the outcome of any truth-leaning equilibrium equals the unique optimal mechanism outcome (Proposition 5 in Section 4.2), and second, that truth-leaning equilibria always exist (Proposition 6 in Section 4.3). An alternative proof showing that an optimal mechanism outcome can be obtained by a truth-leaning equilibrium is provided in the Appendix (Proposition 7 in Section A.2).

In Section A.1 in the Appendix we show the tightness of the Equivalence Theorem: dropping any single condition allows examples where the conclusion does not hold. Specifically, these indispensable conditions are:

- truth structure: reflexivity (L1) of the "partial truth" relation;
- truth structure: transitivity (L2) of the "partial truth" relation;
- truth-leaning: condition (A0) that revealing the whole truth is slightly better;
- truth-leaning: condition (P0) that revealing the whole truth is slightly possible;
- the agent's utility: independent of type;
- the principal's utility: single-peakedness (SP); and
- the principal's utility: differentiability.

[47] That is, the equilibria that are ex-ante Pareto efficient among all equilibria.

4 Proof of the Equivalence Theorem

Throughout this section we maintain the assumptions of the Equivalence Theorem: the functions h_t are differentiable and satisfy the single-peakedness condition (SP).

4.1 Preliminaries

We start with a simple implication of single-peakedness: an *in-betweenness* property of the peaks.

Lemma 1 (In-Betweenness) *Let the probability vector $q \in \Delta(T)$ be a convex combination of probability vectors $q_1, q_2, ..., q_n \in \Delta(T)$ (i.e., $q = \sum_{i=1}^n \lambda_i q_i$ for some $\lambda_i > 0$ with $\sum_{i=1}^n \lambda_i = 1$). Then*

$$\min_{1\leq i\leq n} v(q_i) \leq v(q) \leq \max_{1\leq i\leq n} v(q_i). \tag{3}$$

Moreover, both inequalities are strict unless the $v(q_i)$ are all identical.

Proof. Single-peakedness (SP) implies that the functions h_{q_i} all increase when $x < \min_i v(q_i)$, and all decrease when $x > \max_i v(q_i)$; the same therefore holds for h_q, since $q = \sum_i \lambda_i q_i$ implies $h_q = \sum_i \lambda_i h_{q_i}$, and so the single-peak of h_q must lie between $\min_i v(q_i)$ and $\max_i v(q_i)$. Moreover, if $\min_i v(q_i) < \max_i v(q_i)$ then $h'_q(x) = \sum_i \lambda_i h'_{q_i}(x)$ is positive at $x = \min_i v(q_i)$ and negative at $x = \max_i v(q_i)$. ■

Remark. If T is partitioned into disjoint nonempty sets $T_1, T_2, ..., T_n$ then (3) yields $\min_{1\leq i\leq n} v(T_i) \leq v(T) \leq \max_{1\leq i\leq n} v(T_i)$, because p is an average of the conditionals $p|T_i$, namely, $p = \sum_i p(T_i)\,(p|T_i)$.

Next we provide useful properties of truth-leaning equilibria and their outcomes.

Proposition 2 *Let (σ, ρ) be a truth-leaning equilibrium, 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, for every $t \in T$,

$$\rho(t) = \min\{\pi_t, v(t)\}. \tag{6}$$

Thus, in truth-leaning equilibria, the reward $\rho(t)$ assigned to message t never exceeds the peak $v(t)$ of type t. Moreover, each type t that reveals the whole truth gets an outcome that is at most his value (i.e., $\pi_t \leq v(t)$), whereas each type t that does not reveal the whole truth gets an outcome that exceeds his value (i.e., $\pi_t > v(t)$). This may perhaps sound strange at first. The explanation is that the lower-value types are the ones that have the incentive to pretend to be a higher-value type, and so each message t that is used is sent by t as well as by "pretenders" of lower value. In equilibrium, this effect is taken into account by the principal—or, the market—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 (Lemma 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 3 *Let (σ, ρ) be a truth-leaning equilibrium. If $\sigma(s|t) > 0$ and $s \neq t$ 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; see for instance Examples 1 and 2 in Section 1.1 in the Introduction). As a consequence, replacing the set of possible messages $L(t)$ with its subset $L'(t) := \{s \in L(t) : v(s) > v(t)\} \cup \{t\}$ for every type t affects neither truth-leaning equilibria nor, by our Equivalence Theorem, optimal mechanisms; note that L' also satisfies (L1) and (L2), and as L' is smaller it is simpler to handle.

In the case where evidence has always positive value—i.e., if t has more information than s then the value of t is at least as high as the value of s (formally, $s \in L(t)$ implies $v(t) \geq v(s)$)—Corollary 3 implies that truth-leaning equilibria are fully revealing (i.e., $\sigma(t|t) = 1$ for every type t).[48]

4.2 From Equilibrium to Mechanism

We now prove that the outcome of a truth-leaning equilibrium is the unique optimal mechanism outcome.

We consider a special case first, which will turn out to provide the core of the argument in the general case.

Proposition 4 *Assume that there is a type $s \in T$ such that: (i) $s \in L(t)$ for every t, and (ii) $v(t) < v(T)$ for every $t \neq s$. Then the outcome π^* with*

[48] See Proposition 4 below for the the case where evidence has negative value and truth-leaning equilibria are completely nonrevealing.

$\pi_t^* = v(T)$ *for all* $t \in T$ *is the unique optimal mechanism outcome; i.e.,*

$$\sum_{t \in T} p_t \, h_t(\pi_t) \leq \sum_{t \in T} p_t \, h_t(\pi_t^*) = h_p(v(T))$$

for every π *that is incentive-compatible (i.e., satisfies (IC)), with equality only if* $\pi_t = \pi_t^*$ *for all* $t \in T$.

Condition (i) says that type s is a "least informative" type (in most examples, this is the type that has no evidence at all); condition (ii) implies, by Lemma 1, that we cannot have $v(s) < v(T)$, and so $v(s)$ is the highest peak: $v(t) < v(T) \leq v(s)$ for all $t \neq s$. To get some intuition, consider the simplest case where there are only two types, say, $T = \{s, t\}$. The two peaks $v(t)$ and $v(s)$ satisfy $v(t) < v(s)$, whereas the (IC) constraint $\pi_s \leq \pi_t$ (which corresponds to $s \in L(t)$) goes in the opposite direction. This implies that the maximum of $H(\pi) = p_s h_s(\pi_s) + p_t h_t(\pi_t)$ subject to $\pi_s \leq \pi_t$ is attained when π_s and π_t are taken to be equal (if $\pi_s < \pi_t$ then increasing π_s and/or decreasing π_t would bring at least one of them closer to the corresponding peak, and hence would increase the value of H). Thus $\pi_s = \pi_t = x$ 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. Put $\alpha := v(T)$. We will show that even if we were to consider *only* the (IC) constraints $\pi_t \geq \pi_s$ for all $t \neq s$ and ignore the other (IC) constraints—which can only increase the value of the objective function $H(\pi)$—the maximum of $H(\pi) = \sum_{t \in T} p_t \, h_t(\pi_t)$ is attained when all the π_t are equal, and thus $\pi_t^* = \alpha$ for all $t \in T$.

Thus, consider an optimal mechanism outcome π^0 for this relaxed problem, and put $\beta := \pi_s^0$. Since the only constraint on π_t for $t \neq s$ is $\pi_t \geq \beta$, the fact that h_t has its single peak at $v(t)$ implies the following: if β lies before the peak, i.e., $\beta \leq v(t)$, then we must have $\pi_t^0 = v(t)$, and if β is after the peak, i.e., $\beta \geq v(t)$, then we must have $\pi_t^0 = \beta$. Thus,

$$\pi_t^0 = \max\{\beta, v(t)\} \quad \text{for every } t \neq s. \tag{7}$$

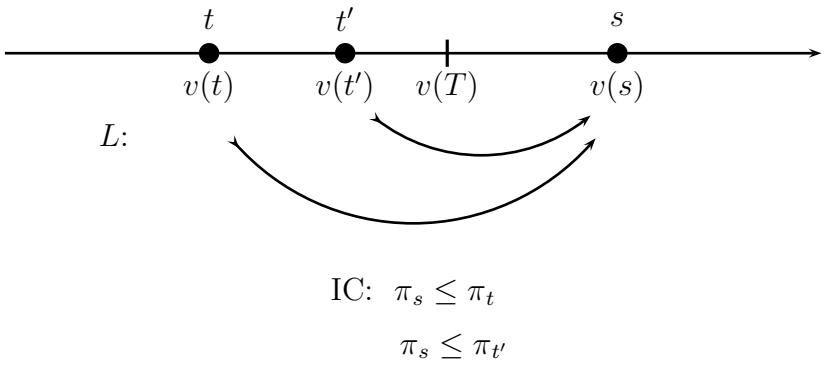

Figure 4: Proposition 4

Put $T^0 := \{t \in T : \pi_t^0 = \beta\}$. We claim that

$$v(T^0) \geq \alpha. \tag{8}$$

Indeed, otherwise $v(T^0) < \alpha$ together with $v(t) < \alpha$ for every $t \notin T^0$ (which holds by assumption (ii) because $s \in T^0$ and so $t \neq s$) would have yielded $v(T) < \alpha$ by Lemma 1, a contradiction. Now the optimality of π^0 implies that β, the common value of π_t^0 for all $t \in T^0$, must be a (local) maximand of $\sum_{t \in T^0} p_t h_t(x)$ (since we can slightly increase or decrease β without affecting the other constraints, namely, $\pi_t \geq \pi_s$ for all $t \notin T^0$, which π^0 satisfies as strict inequalities); therefore β equals the single peak of T^0, i.e., $\beta = v(T^0)$. Hence $\beta > v(t)$ for every $t \neq s$ (by (8) and assumption (ii)), which yields $\pi_t^0 = \beta$ (by (7)). This shows that T^0 (recall its definition) contains all $t \neq s$, as well as s, and so $T^0 = T$ and $\beta = v(T^0) = v(T) = \alpha$, completing the proof that $\pi_t^0 = \alpha$ for every t, i.e., $\pi^0 = \pi^*$. ■

Remark. It is not difficult to show directly (that is, without appealing to our Equivalence Theorem) that under the assumptions of Proposition 4 there is a unique truth-leaning equilibrium outcome, namely, the same π^* with $\pi_t^* = v(T)$ for all $t \in T$ (specifically, the babbling equilibrium (σ, ρ) with $\sigma(t) = s$ for every t, and $\rho(s) = v(T)$ and $\rho(t) = v(t)$ for every $t \neq s$, is a truth-leaning equilibrium here). The conditions of Proposition 4 essentially

(up to replacing some strict inequalities with equalities) identify the case where one *cannot separate* between the types, whether the principal commits or not.

Proposition 5 *Let $\pi^* \in \mathbb{R}^T$ be the outcome of a truth-leaning equilibrium (σ, ρ); then π^* is the unique optimal mechanism outcome.*

Proof. By Proposition 2, $\pi^*_t = \max_{s \in L(t)} \rho(s)$ and $\rho(t) = \min\{\pi^*_t, v(t)\}$ for every $t \in T$. Thus π^* satisfies (IC): if $t' \in L(t)$ then $L(t') \subseteq L(t)$ and so $\pi^*_{t'} = \max_{t'' \in L(t')} \rho(t'') \leq \max_{t'' \in L(t)} \rho(t'') = \pi^*_t$.

We will show $H(\pi^*) > H(\pi)$ for every $\pi \neq \pi^*$ that satisfies (IC). Let $S := \{s \in T : \pi^*_s = \rho(s)\}$ be the set of messages s in T that are used in the equilibrium (ρ, σ) (cf. (4) in Proposition 2); and, for every such $s \in S$, let $T_s := \{t \in T : \sigma(s|t) > 0\}$ be the set of types that play s. For any $\pi \in \mathbb{R}^T$, split the principal's payoff $H(\pi)$ as follows:

$$H(\pi) = \sum_{t \in T} p_t \, h_t(\pi_t) = \sum_{s \in S} \bar{\sigma}(s) \sum_{t \in T_s} q_t(s) \, h_t(\pi_t) \tag{9}$$

(recall that, given the strategy σ, we write $\bar{\sigma}(s)$ for the probability of s, and $q(s) \in \Delta(T)$ for the posterior on T given that s was chosen).

Take $s \in S$, and let $\alpha := \rho(s) = \pi^*_s$ be the reward there; the principal's equilibrium condition (P) implies that

$$\rho(s) = v(q(s)). \tag{10}$$

For every $t \in T_s$, $t \neq s$ we have $\pi^*_t = \rho(s)$ (since $\sigma(s|t) > 0$), and so t is unused (since $\sigma(t|t) \neq 1$) and $v(t) < \pi^*_t$ (by (5)), and hence

$$v(t) < v(q(s)) \text{ for all } t \in T_s, t \neq s. \tag{11}$$

We can thus apply Proposition 4 to the set of types T_s with the distribution $q(s)$, to get

$$\sum_{t \in T_s} q_t(s) \, h_t(\pi_t) \leq \sum_{t \in T_s} q_t(s) \, h_t(\pi^*_t) \tag{12}$$

for every π that satisfies (IC), with equality only if $\pi_t = \pi_t^*$ for every $t \in T_s$. Multiplying by $\bar{\sigma}(s) > 0$ and summing over $s \in S$ yields $H(\pi) \leq H(\pi^*)$ (use (9) for both π and π^*) for every π that satisfies (IC). Moreover, to get equality we need equality in (12) for each $s \in S$, that is, $\pi_t = \pi_t^*$ for every $t \in \cup_{s \in S} T_s = T$, which completes the proof. ■

4.3 Existence of Truth-Leaning Equilibrium

Here we prove that truth-leaning equilibria exist. The proof uses perturbations of the game Γ where a slight advantage is given to revealing the whole truth, both in payoff and in probability. We show that the limit points of Nash equilibria of the perturbed games (existence follows from standard arguments) are essentially—up to an inessential modification—truth-leaning equilibria of the original game. See also the discussion in Section 5.4.

Proposition 6 *There exists a truth-leaning equilibrium.*

Proof. For every $0 < \varepsilon < 1$ let Γ^ε be the following ε-perturbation of the game Γ. First, the agent's payoff is[49] $x + \varepsilon \mathbf{1}_{s=t}$ when the type is $t \in T$, the message is $s \in T$, and the reward is $x \in \mathbb{R}$; and second, the agent's strategy σ is required to satisfy $\sigma(t|t) \geq \varepsilon$ for every type $t \in T$. Thus, first, the agent gets an ε "bonus" in his payoff if he reveals the whole truth, i.e., his type; and second, he must do so with probability at least ε.

A standard argument shows that the game Γ^ε possesses a Nash equilibrium. Let Σ^ε be the set of strategies of the agent in Γ^ε; then Σ^ε is a compact and convex subset of $\mathbb{R}^{T \times T}$. 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 Γ^ε every message is used: $\bar{\sigma}(s) \geq \varepsilon p_s > 0$). The mapping from σ to ρ^σ is continuous: the posterior $q(s) \in \Delta(T)$ is a continuous function of σ (because $\bar{\sigma}(s)$ is bounded away from 0), and $v(q)$ is a continuous function of q (by the Maximum Theorem together with the single-peakedness condition (SP), which gives the uniqueness of the maximizer). The set-valued function Φ that maps each $\sigma \in \Sigma^\varepsilon$ to the set of all $\sigma' \in \Sigma^\varepsilon$ that are best replies to

[49] $\mathbf{1}_{s=t}$ is the indicator that $s = t$ (i.e., it equals 1 if $s = t$, and 0 otherwise).

ρ^σ in Γ^ε is therefore upper hemicontinuous, and a fixed point of Φ, whose existence is guaranteed by the Kakutani fixed-point theorem, is precisely a Nash equilibrium of Γ^ε.

Let (σ, ρ) be a limit point as $\varepsilon \to 0^+$ of Nash equilibria of Γ^ε (the strategy spaces are compact; for the principal, recall Remark (a) in Section 2.1); thus, there are sequences $\varepsilon_n \to_n 0^+$ and $(\sigma_n, \rho_n) \to_n (\sigma, \rho)$ such that (σ_n, ρ_n) is a Nash equilibrium in Γ^{ε_n} for every n. It is immediate to verify that (σ, ρ) is a Nash equilibrium of Γ, i.e., (A) and (P) hold.

Let s be such that $\sigma(s|s) < 1$. Then there is $r \neq s$ in $L(s)$ such that $\sigma(r|s) > 0$, and so $\sigma_n(r|s) > 0$ for all large enough n. Hence in particular $\rho_n(r) \geq \rho_n(s) + \varepsilon_n > \rho_n(s)$, which implies that s is not optimal in Γ^{ε_n} for any $t \neq s$ (because $s \in L(t)$ implies $r \in L(t)$ by transitivity (L2) of L, and r gives to t a strictly higher payoff than s in Γ^{ε_n}); thus $\sigma_n(s|t) = 0$. Taking the limit yields the following property of the equilibrium (σ, ρ):

$$\text{if } \sigma(s|s) < 1 \text{ then } \sigma(s|t) = 0 \text{ for all } t \neq s; \tag{13}$$

what this says is that if s does not choose s for sure, then no other type chooses s. Moreover, the posterior $q_n(s)$ after message s puts all the mass on s (since $\sigma_n(s|s) \geq \varepsilon_n > 0$ whereas $\sigma_n(s|t) = 0$ for all $t \neq s$), i.e., $q_n(s) = \mathbf{1}_s$, and so $\rho_n(s) = v(q_n(s)) = v(s)$; in the limit,

$$\text{if } \sigma(s|s) < 1 \text{ then } \rho(s) = v(s). \tag{14}$$

This in particular yields (P0): $\sigma(s|s) = 0$ implies $\rho(s) = v(s)$.

To get (A0) we may need to modify σ slightly, as follows. Let $s \in T$ be such that s is a best reply for s (i.e., $\rho(s) = \max_{r \in L(s)} \rho(r)$), but $\sigma(s|s) < 1$. Then $\rho(s) = v(s)$ by (14), and every $r \neq s$ that s uses, i.e., $\sigma(r|s) > 0$, gives the same reward as s, and so $v(q(r)) = \rho(r) = \rho(s) = v(s)$. Therefore we define σ' to be identical to σ except that type s chooses only message s; i.e., $\sigma'(s|s) = 1$ and $\sigma'(r|s) = 0$ for every $r \neq s$. We claim that (σ', ρ) is a Nash equilibrium: the agent is indifferent between s and r, and, for the principal, the new posterior $q'(r)$ satisfies $v(q'(r)) = v(q(r)) = v(s)$ (by

Lemma 1, because $q(r)$ is an average of $q'(r)$ and $\mathbf{1}_s$; note that $\bar{\sigma}'(r)$ since $\sigma'(r|r) = \sigma(r|r) = 1$ by (13)). Clearly (13–14), hence (P0), continue to hold. Proceeding this way for every s as needed yields also (A0). ■

5 Extensions

In this section we discuss various extensions and related setups.

5.1 State Space

A useful setup that reduces to our model is as follows.

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[50] $(\Omega, \mathcal{F}, \mathbb{P})$). 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 $x \in \mathbb{R}$. 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, this setup reduces to our model, where $p_t = \mathbb{P}[\tau(\omega) = t]$ 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 Section 2.2), which is useful in many applications (such as the voluntary disclosure setup).

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

5.2 Randomized Rewards

Assume that the principal may choose randomized (or mixed) rewards; i.e., the reward $\rho(s)$ to each message s is now a probability distribution ξ on $\mathbb{R}$ rather than a pure $x \in \mathbb{R}$. The utility functions of the two players are taken as von Neumann–Morgenstern utilities on $\mathbb{R}$, and so the utility of a randomization ξ is its expected utility: $\mathbb{E}_{x\sim\xi}[g_t(x)]$ for the agent and $\mathbb{E}_{x\sim\xi}[h_t(x)]$ for the principal, for each type $t \in T$; we will denote these by $g_t(\xi)$ and $h_t(\xi)$, respectively.

Our assumption on payoffs requires that there be an order on rewards such that for every type the agent's utility agrees with this order, and the principal's utility is single-peaked with respect to this order. Applying this to mixed rewards implies, for the agent, that g_t must be the same function for all t; reparametrizing[51] x allows us to take $g_t(x) = x$ for all x. For the principal, it includes in particular the requirement that his utility be *a function* of the agent's utility. This entailed no restriction in the case of pure rewards, where for every utility level of the agent x there is a unique reward yielding him x (namely, the pure reward x itself). It does however become significant in the mixed case, where all ξ with expectation x yield the same utility x to the agent—and they would all need to yield the same utility to the principal too.[52] This is clearly much too strong a requirement, as it amounts to $h_t(x)$ being linear in[53] x, for each t.

It turns out that there is a way to overcome this, namely, to consider only "undominated" rewards. Specifically, let ξ and ξ' be two mixed rewards with the same expectation, i.e., $\mathbb{E}[\xi] = \mathbb{E}[\xi']$ (the agent is thus indifferent between ξ and ξ'); then ξ *dominates* ξ' if $h_t(\xi) \geq h_t(\xi')$ for all types $t \in T$, with strict inequality for at least one t. The single-peakedness condition for mixed rewards is:

[51] Take x to be that reward that yields utility x to the agent.

[52] If ξ and ξ' both yield the same utility to the agent, which one will the principal choose? Think moreover of the case where the same message is used by more than one type, and then there must be a clear way to determine the right ξ. This is what condition (PUB) below does.

[53] This will come up in the discussion on the connection to the work of Glazer and Rubinstein (2004, 2006) in Section 5.3.

(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[54] single-peaked function $f_q : \mathbb{R} \to \mathbb{R}$ such that $h_q(\xi) = f_q(\mathbb{E}[\xi])$ for every undominated ξ.

Let $X \subset \mathbb{R}$ be a compact interval containing all the peaks (cf. Remark (a) in Section 2.1); all x and ξ below will be assumed to lie in X. Let $\Xi^{\mathrm{U}}(x)$ denote the set of all undominated mixed rewards ξ with $\mathbb{E}[\xi] = x$, and $\Xi^{\mathrm{D}}(x)$ the set of dominated mixed rewards with $\mathbb{E}[\xi] = x$. It is immediate that every dominated $\xi' \in \Xi^{\mathrm{D}}(x)$ is in particular dominated by some *undominated* $\xi \in \Xi^{\mathrm{U}}(x)$, and therefore (SP-M) yields[55]

$$f_q(x) = h_q(\xi) \geq h_q(\xi') \tag{15}$$

for every $\xi \in \Xi^{\mathrm{U}}(x)$, $\xi' \in \Xi^{\mathrm{D}}(x)$, and $q \in \Delta(T)$. Therefore

$$f_q(x) = \max\{h_q(\xi) : \mathbb{E}[\xi] = x\} \tag{16}$$

for every x, which implies that f_q equals the *concavification* $\widehat{h_q}$ of h_q, the smallest concave function that is everywhere no less than h_q (its hypograph is the convex hull of the hypograph of h_q). Since for every x we have $f_q(x) = \sum_{t \in T} q_t\, f_t(x)$ (take $\xi \in \Xi^{\mathrm{U}}(x)$ in (15)), it follows that the maximum in (16) must be reached *at the same* $\xi \in \Xi^{\mathrm{U}}(x)$ for *all* $q \in \Delta(T)$; equivalently, *at the same* $\xi \in \Xi^{\mathrm{U}}(x)$ for *all* $t \in T$. We state this condition:

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

[54] 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 a single flat top of φ; note that concave functions are weakly single-peaked). This weakening is needed since, as we will see below, we may get piecewise linear functions.

[55] We do not use here the single-peakedness of f_q, but only the existence of such a function f_q (which maps the utility of the agent to the utility of the principal for undominated mixed rewards).

types $t \in T$ and every ξ' with $\mathbb{E}[\xi'] = x$.

(Clearly, ξ_x is undominated: $\xi_x \in \Xi^{\mathrm{U}}(x)$). Thus what we have shown above is that (SP-M) implies (PUB); surprisingly, perhaps, the converse also holds: (PUB) implies (SP-M). Indeed, (PUB) implies that $\widehat{h_q} = \sum_t q_t \widehat{h_t}$ for every $q \in \Delta(T)$ (since for each x the concavifications are all obtained from the same mixed reward ξ_x); since $f_q = \widehat{h_q}$ is a concave function, it is weakly single-peaked.

For example, take $T = \{1, 2\}$ and $X = [-1, 1]$; the functions $h_1(x) = -x^2$ and $h_2(x) = x^2$ do *not* satisfy (PUB). Indeed, $f_1(x) = \widehat{h_1}(x) = h_1(x)$ (because h_1 is concave); $f_2(x) = \widehat{h_2}(x) = 1$ (because $h_2(-1) = h_2(1) = 0 \geq h_2(x)$ for all $x \in X$); for, say, $p = (1/2, 1/2)$, we have $h_p(x) = 0$ and so $f_p(x) = \widehat{h_p}(x) = 0$, which is different from $p_1 f_1(x) + p_2 f_2(x) = -x^2/2$. Take for instance $x = 0$: the maximum in (16) for $t = 1$ (i.e., $q = (1, 0)$) is attained only at the pure reward 0, whereas for $t = 2$ (i.e., $q = (0, 1)$), only at the half-half mixture of 1 and -1.

To summarize, (SP-M) and (PUB) are equivalent requirements; moreover, results similar to those proved in Sections 4.2 and A.2 may then be obtained.

5.3 The Glazer–Rubinstein Setup

As stated in the Introduction, the work closest to the present paper is Glazer and Rubinstein (2004, 2006), to which we will refer as **GR** for short. The GR setup is more general than ours in the communication structure—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—only two pure rewards rather than single-peaked payoffs. The first difference implies that in the GR setup only one direction of our equivalence holds: optimal mechanisms are always obtained by equilibria,[56] but the converse is not true.[57] As for the second difference, GR show that their result cannot be extended in general to more than two pure rewards (the example at the

[56] In our setup we moreover show that these are truth-leaning equilibria.

[57] Cf. the examples in Sections A.1.1 and A.1.2 in the Appendix.

end of Section 6 in Glazer and Rubinstein 2006[58]); Sher (2011) later showed that it does hold when the principal's payoff functions are concave.

The discussion of Section 5.2 above helps clarify all this.

First, consider the GR setup where there are only two pure rewards, say, 0 and 1; then 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. Moreover, the principal's utility $h_t(x)$, as a von Neumann–Morgenstern utility, is an affine function of x, and so is necessarily single-peaked (types t with $h_t(0) = h_t(1)$, and so with constant h_t, do not affect anything and may be ignored). Thus (SP) always holds in this case of only two pure rewards.[59]

However, when there are more than two pure rewards, the single-peakedness condition (SP-M) becomes restrictive, and no longer holds in general. As seen in Section 5.2, there are now multiple mixed rewards ξ yielding the same payoff x to the agent (i.e., $\mathbb{E}[\xi] = x$); the uniformity condition (PUB) says that among them there is one that is best for the principal no matter what the type is. For example, if the pure rewards are $0, 1, 2$, then the $1/2 - 1/2$ mixture between 0 and 2 is the same for the agent as getting the pure reward 1, and so (PUB) requires that either $h_t(1) \geq (1/2)h_t(0) + (1/2)h_t(2)$ hold *for all* t, or $h_t(1) \leq (1/2)h_t(0) + (1/2)h_t(2)$ hold *for all* t (in the above-mentioned example in Glazer and Rubinstein 2006, this indeed does not hold: for h_1 we get $>$ and for h_2 we get $<$). A particular case where (PUB) holds is therefore the case where all the functions h_t are concave, because then the pure x is uniformly best for the principal among all mixed ξ with $\mathbb{E}[\xi] = x$; this is the assumption of Sher (2011). But it also holds, for instance, when all the functions h_t are convex (because then ξ_x is the appropriate mixture of the two extreme rewards, 0 and 2), as well as in many other cases.

The *single-peakedness condition* (and its equivalent version (PUB)) appears thus as a good way to generalize and unify these assumptions.[60]

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

[59] Moreover (SP-M) is the same as (SP), as there are no dominated mixed rewards (because for every $x \in [0,1]$ there is a single ξ with $\mathbb{E}[\xi] = x$).

[60] 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.

5.4 Truth-Leaning

An alternative definition of truth-leaning equilibria is based on a "limit of small perturbations" approach (a simple version of which was used to prove existence in Proposition 6). We do it here.

Given $0 < \varepsilon_t^1, \varepsilon_t^2 < 1$ for every $t \in T$ (denote this collection by $\boldsymbol{\varepsilon}$), let $\Gamma^{\boldsymbol{\varepsilon}}$ be the perturbation of the game Γ where the agents's payoff is $x + \varepsilon_t^1 \mathbf{1}_{s=t}$, and his strategy is required to satisfy $\sigma(t|t) \geq \varepsilon_t^2$ for every type $t \in T$. A *TL'-equilibrium* is then defined as a limit point of Nash equilibria of $\Gamma^{\boldsymbol{\varepsilon}}$ as $\boldsymbol{\varepsilon} \to 0$. As in the proof of Proposition 6, one can show that a TL'-equilibrium (σ, ρ) satisfies (A), (P), (13), and (14) (and thus (P0)). Condition (13) is a slight weakening of (A0), as it requires that $\sigma(s|s) = 1$ when message s is optimal for type s *and* some other type $t \neq s$ uses s. The difference between (A0) and (13) is insignificant, because the outcomes are identical, and one can easily modify the equilibrium to get (A0) (as we did in the last paragraph of the proof of Proposition 6). We view (A0) as a slightly more natural condition. However, we could have worked with (13) instead, and all the results would have carried through: the Equivalence Theorem holds for TL'-equilibria too.

Finally, we indicate why 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., $\pi_s = \min_{t:s\in L(t)} \pi_t$). 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 that deviated is allowed by all the refinements, specifically, the intuitive criterion, the D1 condition, universal divinity, and never weak best reply (Kohlberg and Mertens 1986, Banks and Sobel 1987, Cho and Kreps 1987). However, these refinements may not eliminate equilibria such as the babbling equilibrium of Example 2 in Section 1.1 (see also Example 8 in Section A.1.4 in the Appendix); only truth-leaning does.[61]

[61] Interestingly, if we consider the perturbed game $\Gamma_1^{\boldsymbol{\varepsilon}}$ where the agent's payoff is $x + \varepsilon_t^1 \mathbf{1}_{s=t}$ (but his strategy is *not* required to satisfy $\sigma(t|t) \geq \varepsilon_t^2$), the refinements D1,

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.

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.

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.

universal divinity, and never weak best reply (but not the intuitive criterion) yield the (P0) condition, and thus truth-leaning as $\varepsilon \to 0$ (we thank Phil Reny for this observation).

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.

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. (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

A.1 Tightness of the Equivalence Theorem

We will show here that our Equivalence Theorem is tight. First, we show that dropping any single condition allows examples where the equivalence between optimal mechanisms and truth-leaning equilibria does not hold (Sections A.1.1 to A.1.7). Second, we show that the conclusion cannot be strengthened: truth-leaning equilibria need be neither unique nor pure (Sections A.1.8 and A.1.9).

A.1.1 The Mapping L Does Not Satisfy 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 $T = \{0, 1, 3\}$ 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 1 have less information than type 3, but message 3 is not allowed; i.e., $L(0) = \{0\}$, $L(1) = \{1\}$, and $L(3) = \{0, 1\}$.

The unique optimal mechanism outcome is: $\pi_0 = v(0) = 0$ and $\pi_1 = v(\{1, 3\}) = 2$, i.e., $\pi = (\pi_0, \pi_1, \pi_3) = (0, 2, 2)$.

Truth-leaning entails no restrictions here: types 0 and 1 have a single message each (their type), and type 3 cannot send the message 3. There are three Nash equilibria: (1) $\sigma(1|3) = 1$, $\rho(0) = 0$, $\rho(1) = 2$, with $\pi = (0, 2, 2)$ (which is the optimal mechanism outcome); (2) $\sigma(0|3) = 0$, $\rho(0) = 3/2$, $\rho(1) = 1$, with $\pi' = (3/2, 1, 3/2)$; and (3) $\sigma(0|3) = 4/5$, $\rho(0) = \rho(1) = 4/3$, with $\pi'' = (4/3, 4/3, 4/3)$. Note that $H(\pi) > H(\pi') > H(\pi'')$. □

A.1.2 The Mapping L Does Not Satisfy Transitivity (L2)

We provide an example where (L2) is not satisfied—the "less informative than" relation induced by L is not transitive—and there is no truth-leaning equilibrium.

Example 5 The type space is $T = \{0, 1, 3\}$ 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 $L(0) = \{0, 1\}$, $L(1) = \{1, 3\}$, and $L(3) = \{3\}$. This cannot be represented by a transitive order, since type 0 can send message 1 and type 1 can send message 3, but type 0 cannot send message 3.

Let (σ, ρ) be a truth-leaning equilibrium; then $\rho(0) = 0$ (by (P) if 0 is used, and by (P0) if it isn't); similarly, $0 \leq \rho(1) \leq 1$ and $2 \leq \rho(3) \leq 3$. Therefore type 1 chooses message 3, and so only type 0 may choose message 1. If he does so then $\rho(1) = 0$ (by (P)), but then $\rho(0) = \rho(1)$, which contradicts (A0); and if he doesn't, then $\rho(1) = 1$ by (P0), and then $\rho(0) < \rho(1)$, which contradicts the best-replying condition (A).

The unique optimal mechanism is given by[62] $\rho(0) = 0$ and $\rho(1) = \rho(3) = 2$, with outcome $\pi = (0, 2, 2)$ (indeed, types 1 and 3 cannot be separated, since type 1 can say 3 and $v(1) < v(3)$; cf. Proposition 4). □

While truth-leaning equilibria do not exist in Example 5, the slightly more general TL'-equilibrium of Section 5.4 does exist: types 1 and 3 choose message 3, type 0 chooses message 1, and $\rho(0) = 0, \rho(1) = 0, \rho(3) = 2$. We therefore provide another example where even TL'-equilibria do not exist.

Example 6 The type space is $T = \{0, 5, 8, 10\}$ 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 all t. The allowed messages are $L(0) = \{0, 10\}$, $L(5) = \{5, 8\}$, $L(8) = \{8, 10\}$, and $L(10) = \{10\}$. This cannot be represented by a transitive order, since type 5 can send message 8 and type 8 can send message 10, but type 5 cannot send message 10.

A truth-leaning equilibrium (and thus also a TL'-equilibrium) (σ, ρ) is: types 0 and 10 say 10, and types 5 and 8 say 8; the rewards are $\rho(0) = 0, \rho(5) = 5, \rho(8) = v(\{5, 8\}) = 6.5$, and $\rho(10) = v(\{0, 10\}) = 5$. The

[62] While type 0 can send message 1, he *cannot* fully mimic type 1, because he cannot send message 3, which type 1 can. Therefore the incentive-compatibility constraints are *not* $\pi_t \geq \pi_s$ for $s \in L(t)$ as in Section 2.4, but rather $\pi_t = \max\{\rho(s) : s \in L(t)\}$ where $\rho \in \mathbb{R}^T$ is a reward scheme (cf. Green and Laffont 1986).

resulting outcome $\pi = (5, 6.5, 6.5, 5)$ is the optimal mechanism outcome. There is however another TL'-equilibrium (σ', ρ'): types $0, 8$, and 10 say 10, and type 5 says 8; the rewards are $\rho'(0) = 0, \rho'(5) = 5, \rho'(8) = 5$, and $\rho'(10) = v(\{0, 8, 10\}) = 6$. The outcome is $\pi' = (6, 5, 6, 6)$, which is worse than π, since $H(\pi) = -13.625 > H(\pi') = -14$. □

A.1.3 Equilibrium That Does Not Satisfy (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 $T = \{0, 1, 2\}$ 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 information than type 2 who has less information than type 1; i.e., $L(0) = \{0\}$, $L(1) = \{0, 1, 2\}$, and $L(2) = \{0, 2\}$.

The unique optimal mechanism outcome is $\pi = (0, 3/2, 3/2)$, and the unique truth-leaning equilibrium has types 1 and 2 choosing message 2 (type 0 must choose 0) and[63] $\rho(0) = 0, \rho(1) = 1, \rho(2) = 3/2$. There is however another (sequential) equilibrium: type 1 chooses message 2 and type 2 chooses message 0, and $\rho'(0) = \rho'(1) = \rho'(2) = 1$, with outcome $\pi' = (1, 1, 1)$, which is not optimal ($H(\pi') < H(\pi)$). At this equilibrium (P0) is satisfied (since $\rho'(1) = v(1)$ for the unused message 1), but (A0) is not satisfied (since message 1 is optimal for type 1 but he chooses 2). □

A.1.4 Equilibrium That Does Not Satisfy (P0)

We provide an example of a sequential equilibrium that does not satisfy the (P0) condition of truth-leaning, and whose outcome differs from the unique optimal mechanism outcome.

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 $h_t(x) = -(x -$

[63] By Corollary 3 (see L' in the paragraph following it) we may drop 0 from $L(2)$.

$t)^2$ (and so $v(t) = t$) for each $t \in T$. Types 10 and 11 both have less information than type 0, and more information 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 truth-leaning equilibrium is mixed: $\sigma(10|0) = 3/7$, $\sigma(11|0) = 4/7$, all the other types $t \neq 0$ reveal their type, and $\rho(0) = v(0) = 0$, $\rho(3) = v(3) = 3$, and $\rho(10) = \rho(11) = v(\{0, 10, 11\}) = 7$. The optimal mechanism outcome is thus $\pi = (7, 3, 7, 7)$.

Take the babbling equilibria where every type sends message 3 and $\rho(3) = v(T) = 6$ and $\rho(t) \leq 6$ for $t \neq 3$; they do not satisfy (P0) (for types 10 and 11), and so it is not truth-leaning.

Suppose we were to require instead of (P0) that the belief after an unused message t be that it was sent by some of the types t' that could send it, rather than by t itself (as required by (P0)); specifically, put[64] $\rho(t) = v(L^{-1}(t))$ instead of $\rho(t) = v(t)$ (i.e., use the prior probabilities on those types that can send message t). Then the babbling equilibrium satisfies this requirement, since $\rho(0) = v(0) = 0$, $\rho(10) = v(\{0, 10\}) = 5$, and $\rho(11) = v(\{0, 11\}) = 5.5$, and these rewards are all less than $\rho(3) = v(T) = 6$. □

A.1.5 Agent's Payoffs Depend on Type

We show here that it is crucial that the agent's types all have the same preference (see also Example 3 in the Introduction).

Example 9 Consider the standard cheap-talk games of Crawford and Sobel (1982), where all the agent's types can send the same messages (there is no verifiable evidence), but the types differ in their utilities. Specifically, consider the following example taken from Krishna and Morgan (2007). The type t is uniformly distributed on[65] $[0, 1]$. The utilities are $-(x - t)^2$ for the principal and $-(x - t - b)^2$ for the agent, where b is the "bias" parameter that measures how closely aligned the preferences of the two players are.

It is easy to verify (see Krishna and Morgan 2007) that when, say, $b = 1/4$, no information is revealed in any sequential equilibrium, and so the unique

[64] $L^{-1}(t) := \{t' \in T : t \in L(t')\}$ is the set of types that can send message t.

[65] The fact that the type space is not finite does not matter, as a large finite approximation will yield similar results.

outcome of the game is $\pi_t = \mathbb{E}[t] = 1/2$ for all t, which yields an expected payoff of $-1/12$ to the principal.

By contrast, consider the mechanism with reward function $\rho(s) = s+1/4$. The agent's best response to this policy is to report t truthfully, i.e., $s = t$, and so there is full separation and the principal's expected payoff increases to $-1/16$. Thus commitment definitely helps here. □

A.1.6 Principal's Payoffs Are Not Single-Peaked (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 that 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[66] 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 information than type 1, i.e., $L(1) = \{1, 2\}$ and $L(2) = \{2\}$.

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

Next, we will show that every Nash equilibrium (σ, ρ), whether truth-leaning or not, yields the outcome $\pi' = (2, 2)$. Indeed, type 2 can only send message 2, and so the posterior $q(2)$ after message 2 must put on type 2 at least as much weight 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

[66] 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).

(because $h_{q(2)}(0) < 0$, $h_{q(2)}(1) = 0$, and $h_{q(2)}(2) > 0$). Therefore type 1 will never send the message 1 with positive probability (because then $q(1) = (1, 0)$ and so $\rho(1) = v(1) = 1 < 2$). Thus both types only send message 2, and we get an equilibrium if and only if $\rho(2) = 2 \geq \rho(1)$ (and, in the unique truth-leaning equilibrium, (P0) implies $\rho(1) = v(1) = 1$), resulting in the outcome $\pi' = (2, 2)$, which is not optimal: the optimal mechanism outcome is $\pi = (1, 0)$. □

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

A.1.7 Principal's Payoffs Are Not Differentiable

We provide an example where one of the functions h_t is not differentiable, and there exists no truth-leaning equilibrium.

Example 11 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 (SP) holds: the peak $v(q)$ for $q \in \Delta(T)$ equals 1 when $q_1 \geq q_2$ and it equals $2q_2$ when[67] $q_1 \leq q_2$ (and thus $v(T) = 1$). Type 2 has less information than type 1, i.e., $L(1) = \{1, 2\}$ and $L(2) = \{2\}$.

Let (σ, ρ) be a truth-leaning Nash equilibrium. If $\sigma(1|1) = 1$ 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 $\sigma(1|1) = 0$ then $\rho(1) = v(1) = 1$ (by (P0)) and $\rho(2) = v(T) = 1$ (by (P)), contradicting (A0): message 1 is a best reply for type 1. Thus there is no truth-leaning equilibrium.

It may be checked that the Nash equilibria are given by $\sigma(2|1) = 1$ and $\rho(1) \leq \rho(2) = 1$, and the optimal mechanism outcome is $\pi = (1, 1)$. □

[67] This shows that the strict in-betweenness property may not hold without differentiability (cf. Remark (b) after Lemma 1).

A.1.8 Nonunique Truth-Leaning Equilibrium

We provide here an example where there are multiple truth-leaning equilibria (all having the same outcome).

Example 12 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 the "strictly less information" relation is $4 \lhd 3 \lhd 1 \lhd 0$. 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 $\alpha \in [0, 1/3]$. □

A.1.9 Mixed Truth-Leaning Equilibrium

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 13 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 information 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 $2/(1 + \alpha) = 3/(2 - \alpha)$, hence $\alpha = 1/5$. This is therefore the unique truth-leaning equilibrium; its outcome is $\pi = (5/3, 5/3, 5/3)$. □

A.1.10 On the Single-Peakedness Assumption (SP)

We conclude with two comments on the single-peakedness condition (SP) (see Section 2.1).

First, 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 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.

Second, 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).

A.2 From Mechanism to Equilibrium (Short Version)

We show here how to construct a truth-leaning equilibrium from an optimal mechanism.

To illustrate the idea, consider first a 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 2, 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 (it can be shown

that the optimality of π^* implies that every t has some message to use, i.e., $L(t) \cap S \neq \emptyset$).

Take a simple case (such as Example 12 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 towards $v(T)$ would increase the principal's payoff $H(\pi)$). Third, $v(\{t, s\}) \leq \alpha$ (because $v(\{t, s, s'\}) \equiv v(T) = \alpha$ and $v(s') \geq \alpha$), and similarly $v(\{t, s'\}) \leq \alpha$. Thus $v(\{t, s\}) \leq \alpha \leq v(s')$, and so there is some fraction $\lambda \in [0, 1]$ so 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[68] 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 define the sets $R_s := \{t \notin S : s \in L(t)\} \cup \{s\}$ for all $s \in S$. Their union is T, and 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 $R_{s'} = \{t, s'\}$); this follows from the optimality of π^* (increasing π_t for $t \in \cup_{s \in Q} R_s$ can only decrease H). Using a simple extension of the classical Marriage Theorem of Hall (1953) to continuous measures due to Hart and Kohlberg (1974, Lemma in Section 4) 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.,[69]

[68] 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 $\lambda = 1$.

[69] 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.

This result is extended to nonatomic measures in Hart and Kohlberg (1974, Lemma in Section 4); replacing each atom by a continuum yields the fractional result that is needed. 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 $\bar{g}$ (for a collection of clubs K, we take all the students that registered in at least one of the clubs

$v(F_s) = \alpha$. This fractional partition gives the strategy σ, as above. (When we go beyond the quadratic case and the value v is not an expectation, we use the in-betweenness property instead.)

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 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 changes in π_t in the arguments above to be less than δ_0).

A.3 From Mechanism to Equilibrium (Long Version)

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

Proposition 7 *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 2, 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 K and average their grades). The result is 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).

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.

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 7 below).

Consider a simple case (such as Example 12 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 towards $v(T)$ 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 $\lambda \in [0, 1]$ so 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[70] 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 (1953) to continuous measures due to Hart and Kohlberg (1974) (see Section A.3.1 below)[71] yields

[70] 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 $\lambda = 1$.

[71] 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.

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 Remark (b) after Lemma 1). 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.[72]

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 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 7. 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 2, 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 $s \in L(t)$

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

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 $\pi_s^\delta := \pi_s^* - \delta$ if $s \in K(t)$ and $\pi_s^\delta := \pi_s^*$ 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^* also decreases by δ. But $\pi_s^* > v(s)$ for all $s \in K(t)$, and so, for $\delta > 0$ small enough (so that $\pi_s^\delta \geq v(s)$ for all $s \in K(t)$), we get $H(\pi^\delta) - H(\pi^*) = \sum_{s \in K(t)} p_s \left(h_s(\pi_s^\delta) - h_s(\pi_s^*)\right) > 0$

(because π_s^δ 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 $\pi_s^\delta := \pi_s^* + \delta$ 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 3,[73] 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 $R_Q := \cup_{s\in Q} R_s$.

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

Proof. For every $0 \leq \delta \leq \delta_0$ 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 $\pi_{t'}^\delta \geq \pi_t^\delta$ (for $t \in L(t')$). If π^* satisfied it as a strict inequality, because $\delta \leq \delta_0$; and 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[74] $t' \in T_Q$ (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 $\sum_{t\in T_Q} p_t\ (h_t(\pi_t^* + \delta) - h_t(\pi_t^*)) = H(\pi^\delta) - H(\pi^*) \leq 0$ for every $0 \leq \delta \leq \delta_0$ (by the optimality of π^*), and so the derivative at $\delta = 0$ is ≤ 0, which proves the claim. ■

[73] 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.

[74] 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 $t \notin T_Q$.

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 7.

Proof of Proposition 7 (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 11 in Appendix A.3.1 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,

$$\sigma(s|t) > 0 \text{ implies } t \in R_s. \tag{17}$$

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 $\sum_{t\in T_s} p_t\, \sigma(s|t)\, h_t'(\pi_t^*) = -\sum_{t\in T} \sigma(s|t)\, \gamma_t = 0$. The single-peakedness condition (SP) then implies that π_s^* is the single peak of $h_{q(s)}$, i.e.,

$$\pi_s^* = v(q(s)). \tag{18}$$

To conclude we verify that (σ, ρ) is indeed a truth-leaning equilibrium with outcome π^*. Recall that $\rho(s) = \pi_s^* \le v(s)$ iff $s \in S$ and $\rho(t) = v(t) < \pi_t^*$ iff $t \notin S$. Then $\pi_t^* = \max_{r\in L(t)} \pi_r^* \ge \max_{r\in L(t)} \rho(r)$ by (IC), and Claim 1 implies that there is equality; thus the outcome is π^*. The agent's equilibrium condition (A) holds by (17): $\sigma(s|t) > 0$ 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 (17). The principal's equilibrium condition (P) holds because $\bar\sigma(s) > 0$ iff $s \in S$ by (17) and (A0), and then $\rho(s) = \pi_s^* = v(q(s))$ by (18). 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^*.

A.3.1 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[75] $|\cup_{m \in M} W_m| = |M|$ 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.[76]

[75] 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 also as the *counting measure* of A.

[76] We state only the nontrivial direction that the conditions are sufficient; see the Remark following Proposition 9.

Theorem 8 (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*

$$\nu(\cup_{m\in M} W_m) = \mu(M), \quad \textit{and} \tag{19}$$

$$\nu(\cup_{m\in K} W_m) \geq \mu(K) \quad \textit{for every } K \subseteq M, \tag{20}$$

then there exists a partition of[77] *W into disjoint sets $(V_m)_{m\in M}$ satisfying*

$$V_m \subseteq W_m \qquad \textit{for every } m \in M, \textit{ and} \tag{21}$$

$$\nu(V_m) = \mu(\{m\}) \quad \textit{for every } m \in M. \tag{22}$$

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., $\lambda(I) = \sum_{n\in I} \lambda_n$ 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 (19) and (20) 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 9 (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*[78] *W. If (19) and (20) hold, then there exists a partition of W into disjoint sets $(V_m)_{m\in M}$ satisfying (21) and (22).*

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

[78] Formally, ν is defined on a σ-field $\mathcal{F}$ of subsets of W, which contains all the relevant sets. ν is *nonatomic* if for every S with $\nu(S) \neq 0$ there is $S_1 \subset S$ such that $\nu(S_1) \neq 0$

Proof. This is the lemma in Section 4 of Hart and Kohlberg (1974),[79] 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 also $\cup_{m\in M} V_m = W = \cup_{m\in M} W_m$ (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 (19) and (20) hold) is no longer true (it is when ν is a nonnegative measure, since then (21) 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 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[80] $(V_m)_{m\in M}$ requires that each element $w \in W$ belongs 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[81] $w \in W_m$, and the measure $v(V_m)$ is given by[82] $\int_W V_m \, d\nu$. We have:[83]

Corollary 10 *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 (19) and (20) hold, then there exists a partition of W into disjoint* fractional *sets $(V_m)_{m\in M}$ satisfying (21) and (22).*

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.

[79] Whose simple proof is inspired by the simple inductive proof provided by Halmos and Vaughan (1950) to Hall's Marriage Theorem.

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

[81] 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 $w \in W$).

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

[83] 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."

Proof. Replace each atom w of the measure ν with a nonatomic continuum C_w with the same measure and apply Proposition 9; $V_m(w)$ 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 associates to each element w in W a probability distribution on M that gives the fractions of w in the various[84] V_m; that is, $\sigma : W \to \Delta(M)$ with[85] $\sigma(m|w) := V_m(w)$ 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 10 may be restated as follows.

Corollary 11 *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[86] $\bar{g}$ (for a collection of clubs K, we

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

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

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

take all the students that registered in at least one of the clubs in K and average their grades). Corollary 11 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).[87]

[87]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.