Evidence Games: Truth and Commitment[1]
Online Appendix C

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

October 30, 2016

[1] *American Economic Review* (forthcoming).

[2] Department of Economics, Institute of Mathematics, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. *E-mail*: hart@huji.ac.il *Web site*: http://www.ma.huji.ac.il/hart

[3] Department of Economics, Business School, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem; Department of Economics, University of Warwick. *E-mail*: kremer@huji.ac.il

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

C Online Appendix: Extensions and Comments

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

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

C.1 Introduction

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

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

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

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

C.2 Examples (Section I)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

Consider next truth-leaning equilibria (note that truth-leaning makes no requirement on $\rho(m)$ for messages $m \notin T$ that are not used). We claim that none of the messages $m \notin T$ are used in a truth-leaning equilibrium (σ, ρ), i.e., $\bar{\sigma}(m) = 0$ for all $m \notin T$. Indeed, let $m \notin T$; for every type $t \in T$ that uses m, i.e., $\sigma(m|t) > 0$, we get $\pi_t = \rho(m) > \rho(t) = v(t)$ (by (A), (A0), and (P0)). Therefore $\rho(m) > v(q(m))$ by in-betweenness (1), which contradicts (P). Finally, every truth-leaning equilibrium for L' is clearly also a truth-leaning equilibrium for L. ■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix C.10 provides an alternative proof.

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

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

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

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

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

C.7 Proof (Section V)

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

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

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

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

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

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

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

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

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

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

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

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

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

(ii) $\pi_t^* \geq v(t)$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This yields the following generalization of Proposition 6:

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

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

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

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

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

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

C.10 The Optimal Outcome

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

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

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

---

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

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

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

A converse to Proposition 12 is as follows.

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

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

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

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

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

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

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

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

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

C.11 Equivalence without Differentiability

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

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

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

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

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

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

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

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

Equivalence holds also under (SP$_0$):

Proposition 15 *Assume that the principal's payoff function $(h_t)_{t \in T}$ satisfies the continuous single-peakedness condition (SP$_0$). Then there is a unique*

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

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

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

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

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

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

References to Appendix C

(in addition to the references in the paper)

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

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

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

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

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

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

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

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