Costly Evidence and the Value of Commitment[*]

Justus Preusser[†]

This version: October 31, 2022

Abstract

A principal has to accept or reject a proposal. The optimal decision depends on the verifiable type of an agent. The agent always wants the proposal to be accepted, and can influence the distribution of the type at a cost. If the principal does not have commitment power, the principal is typically no better off than when acting uninformedly. The principal can be strictly better off by committing to a mechanism. Optimally, the principal commits to sometimes rejecting the proposal when it is optimal to accept, and commit to sometimes accepting the proposal when it is optimal to reject.

Keywords: Information Acquisition, Evidence, Commitment, Mechanism design without transfers

[*] I thank Simon Block, Francesc Dilmé, Deniz Kattwinkel, Stephan Lauermann, Benny Moldovanu, and participants of the Yale Theory Breakfast for their comments. All remaining errors are my own. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project B04) is gratefully acknowledged. I am grateful for the hospitality of Yale University, where parts of this paper were written.

[†] Department of Economics, University of Bonn, *justus.preusser@uni-bonn.de*.

1 Introduction

Consider a principal who decides whether to accept or reject a proposal. The optimal decision depends on the private type of an agent who wants the proposal to be accepted, regardless of the type. Monetary transfers cannot be used to elicit this type, but the type is verifiable; that is, the agent can provide evidence to conclusively prove the type realization.

The existing literature on optimal mechanisms for these kind of problems focuses on situations where the distribution of the agent's type is exogenously given. An important finding is that commitment has no value to the principal. That is, in a game where the principal responds optimally to the disclosed evidence—the *evidence-disclosure* game—the principal is just as well off as when the principal can commit to a mechanism (Ben-Porath et al., 2019; Glazer and Rubinstein, 2004, 2006; Hart et al., 2017; Sher, 2011).

In the present paper, the type distribution is endogenous: the agent covertly acquires a distribution only after the principal announces a mechanism. The main result characterizes an optimal mechanisms and shows that commitment is valuable for the principal.

In the model, the agent's type is commonly-known to lie in $[0, 1]$ and have some fixed mean. We consider an agent who can flexibly acquire any type distribution on $[0, 1]$ with the given mean. The agent incurs some acquisitions costs that are linear in the distribution and decreasing with respect to mean-preserving contractions. Whatever the type realization, the agent chooses whether to provide conclusive proof about the realization or to provide completely uninformative evidence; this is a special case of Dye (1985) evidence.[1] The principal wants to accept the proposal if and only if the type is above some threshold; the agent always prefers the proposal to be accepted.

If the principal has commitment power, the timing is as follows: First, the principal announces a mechanism; then, the agent covertly picks a type distribution; lastly, the type realizes, and the agent decides what evidence and cheap-talk reports to send to the mechanism. In the evidence-disclosure game, the only difference is that

[1] In this interpretation of the model, the agent chooses the distribution of a payoff-relevant state (the type). Another interpretation is that the agent acquires information about a binary state. The agent's type is then interpreted as the agent's posterior belief about the state.

principal best responds to the disclosed evidence.

The main result of the paper characterizes an optimal mechanism. It is parametrized by two type cutoffs θ_0 and θ_1, and by a probability α. If the agent discloses a type below θ_0, the principal accepts with probability α; for types in (θ_0, θ_1), with probability 0; for types above θ_1, with probability 1. In particular, the cutoff θ_1 is strictly above the threshold where the principal would prefer accepting to rejecting; the cutoff θ_0 is strictly below this threshold.

The form of this optimal mechanism suggests that two forms of commitment are valuable: First, the principal commits to rejecting some proposals that are efficiently approved. Intuitively, if the principal did not commit to rejecting these proposals, the agent would pick a distribution whose realizations are high enough to persuade the principal to accept, but not sufficeintly high to strictly benefit the principal. Second, if $\alpha > 0$, then the principal commits to accepting some proposals that are efficiently rejected. Accepting these proposals may be optimal in order to compensate the agent for the costs of acquiring a distribution. Indeed, an example shows that all optimal mechanisms may require the principal to accept some undesirable proposals.

This characterization of an optimal mechanism exploits the assumptions that acquisition costs are linear and that the agent chooses among all distributions on $[0, 1]$ with a fixed mean. These assumptions are restrictive. The upside is that the principal's problem is tractable and that there is a clear sense of how exactly the principal benefits from commitment.[2]

We next study equilibria of the evidence-disclosure game. Under permissive assumptions on the environment, *all* equilibria in a natural class of equilibria give the principal the utility that would obtain if the principal did not consult the agent at all.[3] One such assumption is that the agent's acquisition costs are *strictly* decreasing with respect to mean-preserving contractions (but the magnitude of these costs do not matter). Although this exact result does not seem to have appeared in the literature, it is intuitive from known results on information design.[4] The result is

[2] By contrast, it is not so important that the principal takes a binary decision. In the supplementary appendix, the characterization is extended to a problem with multiple alternatives.

[3] More precisely, we consider perfect Bayesian equilibria with the following additional property: If the agent provides hard evidence about some type realization, then the principal's belief is degenerate on this type even if this happens off the equilibrium path.

[4] For example, the receiver in the courtroom-example from Kamenica and Gentzkow (2011) is as well off as when acting uninformedly. More recently, Titova (2022, Theorem 2) shows that, in a nearby evidence-disclosure game, the receiver is as well off as when acting uninformedly when a

also intuitive from the characterization of optimal mechanisms: as suggested above, the principal commits to rejecting some propsals that are efficiently accepted, and commits to accepting some proposals that are efficiently rejected. We remark that the result can be shown under assumptions on the agent's acquisition costs and the set of available distributions that are far less stringent than those used to characterize optimal mechanisms (see Remark 1 in Section 5.2).

Let us briefly interpret these results in the context of two applications. First, consider a funding agency that has to approve a project proposed by a regional government. The agency prefers to approve if and only if the value of the project exceeds the cost. Suppose that costs are known, but that the value depends on the outcome of a risky investment by the regional government. Now, the funding agency may be committed through regulations that specify certain quality standards. However, these regulations may not be set up to suit a specific one-shot interaction, and hence their design should account for the implied investment incentives. The analysis suggests that if the regional government can flexibly tailor its investments to a given set of regulations, then these regulations are valuable. In particular, regulations optimally set the bar for the project's value higher than the cost. Moreover, to ensure that the regional government invests in the first place, it may be optimal to approve some projects that turn out to be inefficient.

Second, consider a prosecutor who wants to persuade a judge that a defendant is guilty. To convince the judge to accept, the prosecutor acquires information about the defendant's guilt. The judge wants to convict the defendant if and only the probability of guilt exceeds some threshold of doubt. Now, the judge may be commited to certain actions via their rulings in precdental cases. For example, the judge may have built a reputation for being tough on prosecutors who are eager to convict. The analysis suggests that this reputation is valuable—the judge optimally acquits the defendant if the probability of guilt is larger than but too close to the judge's threshold of doubt. Optimal mechanisms may also require the judge to convict defendants that are more likely to be innocent than guilty. This level of commitment power is unrealistic in practice, and hence the results advocate the use of additional instruments.

sender-preferred equilibrium is played.

2 Related literature

The paper aims to contribute to the literature on mechanisms with hard evidence. The aforementioned papers of Ben-Porath et al. (2019), Glazer and Rubinstein (2004, 2006), Hart et al. (2017), and Sher (2011) provide sufficient conditions for commitment to be without value when the type distribution is exogenous. Silva (2020) shows that commitment is valuable if (the type distribution is exogenous but) the agent's evidence is imperfect. Most relevant for us are the recent papers of Ben-Porath et al. (2021) and Migrow and Severinov (2022).

Among other things, Ben-Porath et al. (2021) find sufficient conditions for commitment to be without value in a model with endogenous evidence. In their model, the distribution of the payoff-relevant type is fixed; the agent's actions affect what evidence the agent can present about the type realization. By contrast, in the present model, the agent affects the distribution of the payoff-relevant type; whatever type realizes, it is commonly-known that the agent has evidence about this realization.

Migrow and Severinov (2022) consider a model where a principal decides whether to implement a project. The agent chooses between two (mostly costless) investments: one increases the quality of the project, the other generates a signal about the quality. This contrasts our model in which the agent must be incentivized to undertake a costly investment but is sure to have evidence about its quality. Like us, they find that equilibria are inefficient and commitment has value for the principal. A qualitative difference in the results is that in Migrow and Severinov (2022) the principal optimally accepts when the agent provides weak evidence in the project's favour; this is not so in optimal mechanisms in our model. Moreover, in our model the equilibria of the evidence-disclosure are not just inefficient—the principal is no better off than when taking an uninformed action.

This paper is also related to the literature on mechanism design without transfers but costly verification. See Bayrak et al. (2017), Ben-Porath et al. (2014), Epitropou and Vohra (2019), Erlanson and Kleiner (2019, 2020), Halac and Yared (2020), Kattwinkel and Knoepfle (2019), and Li (2020). All of these consider settings with exogenous types. We show that our results extend in a natural way to a model with costly verification. Ben-Porath et al. (2019) previously pointed out such a connection for models with exogenous types.

Shishkin (2021) and Whitmeyer and Zhang (2022) consider nearby evidence-disclosure games where the agent's information is verifiable and flexibly chosen by the agent.[5] Among other things, they compare outcomes for overt and covert information acquisition strategies. Neither paper considers principal-optimal mechanisms, which are our focus. There are other differences between the respective models and our own which lead to different results. For example, Whitmeyer and Zhang (2022) find that if the agent can commit to what information to acquire, then the agent acquires no information—this is not so in our model.

When the agent but not the principal can commit, the problem is one of Bayesian perusasion. Tsakas et al. (2021) show that the receiver (the principal) may choose to set up a *resistance strategy*. Such a strategy entails a cost for the receiver cost whenever the receiver takes the preferred action of the sender (the agent). Resistance strategies cannot capture the full scope of what one can do with commitment. For example, the principal may optimally commit to taking the agent's preferred action at types where doing so is already suboptimal.

3 Model

We first introduce all relevant objects of the model, and then collect additional important assumptions at the end.

A principal decides whether to accept or reject a proposal. The principal's payoffs depend on the type θ of an agent who wants the proposal to be accepted. It is commonly known that the agent has hard evidence about the type, that the type lies in $[0, 1]$, and that the mean of the type distribution is a value $\mu \in (0, 1)$. However, the type distribution itself is endogenously chosen by the agent.

Let $u_p(\theta)$ and $u_a(\theta)$, respectively, denote the principal's and agent's payoffs, respectively, from accepting the proposal when the agent's type is θ. The payoffs of rejecting in the same situation are 0 for both the principal and the agent. The payoff $u_a(\theta)$ is strictly positive; that is, the agent strictly prefers that the proposal be accepted. The functions u_p and u_a are continuous.

A *type distribution* means a cummulative distribution function (cdf) on $[0, 1]$ whose mean is μ. Let $\mathcal{F}$ denote the set of type distributions. The agent can acquire an arbitrary type distribution, possibly at cost. We represent these costs by a function $K\colon \mathcal{F} \to \mathbb{R}_+$ that is continuous with respect to the L^1-norm on $\mathcal{F}$.

[5] See also Escudé (2020), Titova (2022), and Zhang (2022) for more distantly related work.

Whatever the type distribution, and whatever the realization θ of the type, the agent can provide hard evidence from the set $\{\{\theta\}, [0,1]\}$. Evidence $\{\theta\}$ is interpreted as a conclusive proof that the type is θ; the agent is unable to provide this evidence at other type realizations. Evidence $[0,1]$ is interpreted as the agent proving the trivial event that the type is in $[0,1]$. Let $\mathcal{E} = \{[0,1]\} \cup (\bigcup_{\theta\in[0,1]}\{\{\theta\}\}$ denote all possible pieces of evidence.

To elicit the agent's type, the principal commits to a mechanism. A mechanism consists of a set M of cheap-talk messages and a function $x\colon M \times \mathcal{E} \to [0,1]$. Here, we interpret $x(m,e)$ as the probability that the principal accept the proposal when the agent sends message m and provides evidence e.

The timing is as follows:

(1) The principal commits to a mechanism (M,x).
(2) The agent, knowing the mechanism, picks a type distribution F.
(3) Nature draws the agent's type θ according to F.
(4) The agent, knowing the type θ, picks a message m and evidence e in $\{\{\theta\}, [0,1]\}$.
(5) The mechanism accepts the proposal with probability $x(m,e)$.

The agent acts in steps (2) and (4) to maximize expected utility, breaking ties in favor of the principal.

We can simplify this model. Recall that all types of the agent strictly prefer that the proposal be accepted. Therefore, if the principal commits to rejecting the proposal whenever the agent does not fully disclose the realized type, the agent has a best-response of always disclosing the type. It follows that the principal can implement all functions $x\colon [0,1] \to [0,1]$; here $x(\theta)$ is the probability that the principal accepts when the agent discloses θ. For technical reasons, we only require that x be upper-semicontinuous (usc). That is, henceforth a mechanism simply means a usc function $x\colon [0,1] \to [0,1]$.

The agent's and principal's utilities, respectively, from a mechanism x and a type distribution F are given by

$$U_a(x,F) = \mathbb{E}_F\left[x(\theta)u_a(\theta)\right] - K(F) \quad \text{and} \quad U_p(x,F) = \mathbb{E}_F\left[x(\theta)u_p(\theta)\right],$$

respectively.[6] Let $\mathcal{F}^*(x) = \arg\max_{\hat{F}\in\mathcal{F}} U_a(x,\hat{F})$; we refer $\mathcal{F}^*(x)$ as the set of *agent-*

[6] Throughout the paper, when F is a cdf and $h\colon [0,1] \to \mathbb{R}$ is an F-integrable function, we denote $\mathbb{E}_F[h(\theta)] = \int h dF$.

optimal distributions on x.[7]

Since the agent breaks ties favorably, the principal evaluates a mechanism x via the best-possible utility $\bar{U}_p(x)$ that can arise via an agent-optimal distribution. Formally, let

$$\bar{U}_p(x) = \sup_{F \in \mathcal{F}^*(x)} U_p(x, F).$$

We now introduce additional assumptions that are maintained throughout the paper. The first concerns the principal's payoffs.

Assumption 1. The principal's payoff u_p is convex, crosses 0 exactly once, and does so from below at a point weakly above μ.

We let $u_p^{-1}(0)$ denote the unique point where u_p equals 0.

The second additional assumption concerns the agent's payoffs.

Assumption 2. The agent's payoff u_a is concave. There is a continuous convex function $k\colon [0,1] \to \mathbb{R}_+$ such that for all $F \in \mathcal{F}$ we have $K(F) = \mathbb{E}_F[k(\theta)]$.

In words, the principal prefers accepting if and only if the type is above a threshold value $u_p^{-1}(0)$, where $u_p^{-1}(0)$ is itself higher than the commonly-known mean μ of the type. The other assumptions on u_p, u_a and K hint at a conflict of interest between the principal and the agent with respect to the choice of a type distribution: Given two distributions F and $\hat{F}$, where $\hat{F}$ is an mean-preserving contraction (MPC) of F, the agent prefers $\hat{F}$, and the principal prefers F.[8]

Lastly, we maintain the following non-triviality assumption.

Assumption 3. There exists a mechanism x such that $\bar{U}_p(x) > 0$.

Note that 0 is what the principal can guarantee by always rejecting the proposal. Assumption 3 is a joint assumption on the utilities of the agent and principal, and the prior mean μ. One sufficient condition for Assumption 3 is that K be sufficiently

[7] Since x is usc, since u_a is strictly positive and continuous, and since K is L^1-continuous on $\mathcal{F}$, the agent's utility is L^1-usc on $\mathcal{F}$. The set $\mathcal{F}$ is L^1-compact, and so the existence of an agent-optimal distribution follows from the Extreme Value theorem.

[8] Recall that if F is a cdf on $[0,1]$, a cdf $\hat{F}$ is a *mean-preserving contraction (MPC)* of F if all $t \in [0,1]$ satisfy $\int_0^t (F(s) - \hat{F}(s))\, ds \geq 0$, with equality for $t = 1$. In the same situation, the cdf F is a mean-preserving spread (MPS) of $\hat{F}$.

close to 0. Another sufficient condition is $\mu = u_p^{-1}(0)$, meaning that the principal is indifferent between accepting and rejecting at the mean type μ. See Proposition A.7 in Appendix A.1.1 for a proof of these claims.

4 Optimal mechanisms

4.1 Binary distributions suffice

Fixing a mechanism x, the agent solves an information design problem. The optimal choice of a mechanism takes into account how the agent's solution set $F^*(x)$ changes with the mechanism x. Since the principal can implement all usc functions $x\colon [0,1] \to [0,1]$, there is an enormous set of changes to x that we could consider as improvements.

We begin with a technical result that greatly simplifies the analysis: it is without loss to assume the agent acquires a binary distribution. A distribution is *binary* if its support contains at most two elements.

Lemma 4.1. *Let Assumptions 1 and 2 hold. For all mechanisms x, there exists a binary distribution $F \in \mathcal{F}^*(x)$ such that $U_p(x, F) = \bar{U}_p(x)$; that is, there is an agent-optimal distribution which is binary and maximizes the principal's utility across all agent-optimal distributions.*

See Appendix A.1.2 for a proof.

We clarify in the next subsection how exactly Lemma 4.1 is useful. The idea of the proof is as follows: The extreme points of $\mathcal{F}$ are binary distributions (Winkler, 1988). The agent's utility is linear and usc in the distribution. Hence Bauer's Maximum Principle (Border and Aliprantis, 2006, Theorem 7.69) implies that there is an agent-optimal distribution $F \in \mathcal{F}^*(x)$ which is binary.[9] It requires a little more work to show that there is a binary agent-optimal distribution F solving $U_p(x, F) = \bar{U}_p(x)$. This claim follows from Choquet's theorem (Phelps, 2001, p. 14) if we can show that the principal's utility is continuous in the distribution when restricted to agent-optimal distributions. The principal's utility is *not* generally continuous (or even upper-semicontinuous) on the entire set of distributions; the reason is that $\theta \mapsto x(\theta)u_p(\theta)$ admits a downward jump if x is discontinuous at a point where u_p is strictly negative.

[9] An alternate proof of this part of the claim views the agent's optimal choice of a type distribution on a given mechanism as a Bayesian persuasion problem, and then invokes Proposition 4 from the online appendix of Kamenica and Gentzkow (2011).

It turns out, however, that these jumps cannot occur along a sequence of agent-optimal distributions since such a jump would imply an upward jump in the agent's utility, contradicting agent-optimality of the members of the sequence.

4.2 Two-sided cutoff mechanisms

The following is a natural first guess for a good mechanism: fixing some threshold t above $u_p^{-1}(0)$, the principal accepts if and only if the agent discloses a type weakly above t. Let us call this a *one-sided cutoff mechanism.* A potential problem with this mechanism is that the agent is only rewarded for realizations above t. If acquisitions costs are too high (relative to t), the agent may wish to economize on acquisition costs by acquiring the degenerate distribution on μ. Indeed, we later exhibit an example of an environment (that meets the non-triviality Assumption 3 but) where, for all values of t above $u_p^{-1}(0)$, no agent-optimal distribution on the described mechanism places mass above t. In this environment, all one-sided cutoff mechanisms leave the principal with a utility of 0.

How can the principal then obtain a strictly positive utility when acquisition costs are non-negligible? The idea is simply to offer the agent additional compensation by also rewarding sufficiently *low* realizations. This motivates the next definition.

Definition 1. A mechanism x is a **two-sided cutoff mechanism** if there exist probabilities α, β and γ, and points θ_0 and θ_1 in $[0,1]$ such that $\beta \leq \alpha$, $\beta \leq \gamma$, and $\theta_0 \leq \theta_1$, and such that

$$\forall_{\theta \in [0,1]}, \quad x(\theta) = \begin{cases} \alpha, & \text{if } \theta \leq \theta_0 \\ \beta, & \text{if } \theta_0 < \theta < \theta_1 \\ \gamma, & \text{if } \theta_1 \leq \theta. \end{cases}$$

In this case, we refer to $(\theta_0, \theta_1, \alpha, \beta, \gamma)$ as the **parameters** of x.

Interpreted as an indirect mechanism, the agent discloses whether the realized type is "low", "intermediate", or "high". Intermediate types enjoy the lowest winning probabilities. In all two-sided cutoff mechanisms considered below, the intermediate interval contains both the mean μ and the point $u_p^{-1}(0)$ where the principal's payoff from accepting crosses 0.

4.3 Two-sided cutoff mechanisms are optimal

We next show that two-sided cutoff mechanisms are optimal. Let x be a mechanism. Let F^* be a binary distribution, and let its support be $\{\theta_0^*, \theta_1^*\}$. Consider the two-sided cutoff mechanism x^* defined as follows (see Figure 1):

$$\forall_{\theta\in[0,1]}, \quad x^*(\theta) = \begin{cases} x(\theta_0^*), & \text{if } \theta \leq \theta_0^* \\ 0, & \text{if } \theta_0^* < \theta < \theta_1^* \\ x(\theta_1^*), & \text{if } \theta_1^* \leq \theta. \end{cases} \tag{4.1}$$

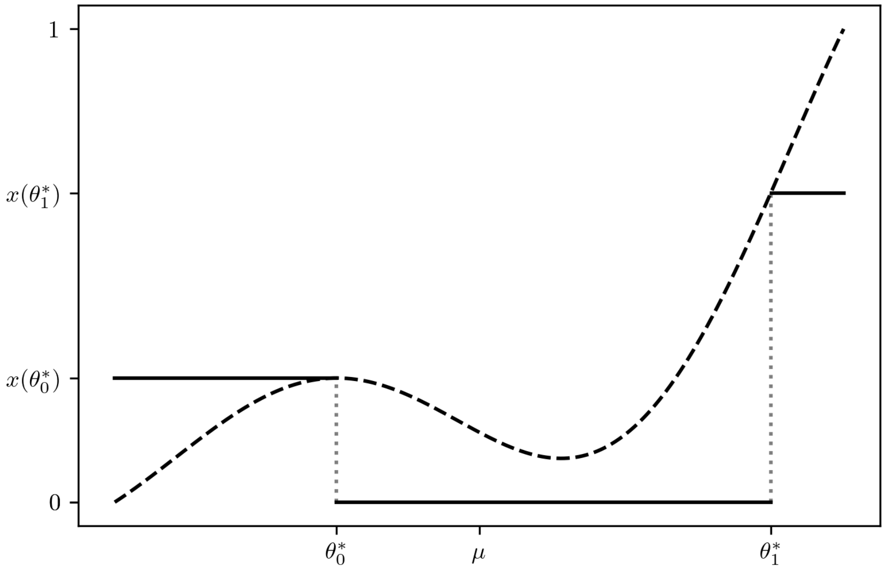

Figure 1: The two-sided cutoff mechanism x^* (solid line) derived from x (dashed line).

Lemma 4.2. *Let Assumptions 1 and 2 hold. Let x be a mechanism such that $\bar{U}_p(x) > 0$. Let $F^* \in \mathcal{F}^*(x)$ be a binary distribution satisfying $U_p(x, F^*) = \bar{U}_p(x)$. Let x^* be the two-sided cutoff mechanism defined in (4.1). Then $\bar{U}_p(x^*) \geq \bar{U}_p(x)$.*

See Appendix A.1.3 for a proof.

Lemma 4.1 asserts that there always exists a binary distribution as in the hypothesis of Lemma 4.2. Hence, whenever $\bar{U}_p(x) > 0$, we can improve on x by passing to a two-sided cutoff mechanism.

To better understand the lemma, let us decompose the passage from x to x^* into two steps. In a first step, we decrease $x(\theta)$ to 0 at all types θ except for the two types θ_0^* and θ_1^* in the support of F^*. This change weakly decreases the agent's utility from all distributions except F^*. Hence F^* remains agent-optimal on the new mechanims, implying that the new mechanism also gives the prinicipal a utility of $\bar{U}_p(x)$.

In a second step, we obtain the mechanism x^* by raising the acceptance probabilities on the subintervals $[0, \theta_0^*)$ and $(\theta_1^*, 1]$, respectively, from 0 to $x(\theta_0^*)$ and $x(\theta_1^*)$, respectively. The content of the lemma is that raising these probabilities weakly increases the principal's utility. Intuitively, raising the probabilities encourages the agent to acquire a distribution that assigns non-zero probability to each of the outer intervals, and no probability to the middle interval (θ_0^*, θ_1^*). Such a distribution is mean-preserving spread of F^*. Since the principal's payoffs u_p are convex, we intuit that raising the probabilities encourages the agent to acquire a distribution that the principal weakly prefers to F^*. (It is in this step that we use the hypothesis $\bar{U}_p(x) > 0$.) Since F^* attains $\bar{U}_p(x)$ on x, we conclude that x^* improves on x.

Lemma 4.2 simplifies the principal's problem from maximizing over all usc functions $x \to [0,1] \to [0,1]$ to maximizing over the parameters $(\theta_0, \theta_1, \alpha, \beta, \gamma)$ of two-sided cutoff mechanisms. We next use this fact to prove the existence of an optimal mechanism. Further, we sharpen the characterization of optimal parameters.

4.4 Optimal parameters

The main result asserts that it is optimal to set the acceptance probability γ^* on the right-most subinterval to 1. Moreover, the principal commits to rejecting at some types where it would be optimal to accept. Lastly, an optimal mechanism exists.

Theorem 4.3. *Let Assumptions 1 to 3 hold. There exists a probability α^*, and θ_0^* and θ_1^* in $[0,1]$ such that $\theta_0^* < \mu \leq u_p^{-1}(0) < \theta_1^*$, and such that the two-sided cutoff mechanism with parameters $(\theta_0^*, \theta_1^*, \alpha^*, 0, 1)$ maximizes $\bar{U}_p$ over the set of mechanisms. Moreover, the binary distribution F^* with support $\{\theta_0^*, \theta_1^*\}$ and mean μ is agent-optimal on x^* and satisfies $\bar{U}_p(x^*) = U_p(x^*, F^*)$.*

See Appendix A.1.4 for a proof.

Recall that $u_p^{-1}(0)$ is the type where the principal's payoffs from accepting crosses 0 from below. The inequality $u_p^{-1}(0) < \theta_1^*$ thus implies that the principal commits to rejecting at some types where accepting is strictly preferred. Moreover, if $\alpha^* > 0$, the inequality $\theta_0^* < \mu$ implies that the principal commits to accepting at some types where rejecting is strictly better.

The intuition for why it is optimal to set $u_p^{-1}(0) < \theta_1^*$ is straightforward. The agent will pick a distribution whose largest realization is θ_1^* (as per the theorem). Hence, if $\theta_1^* \leq u_p^{-1}(0)$, the principal would always end up accepting at a point where the principal is, at best, indifferent between accepting and rejecting.

It is also intuitive why the principal gains from setting γ^*, the acceptance probability on the right-most subinterval, to 1. Since θ_1^* is above $u_p^{-1}(0)$, increasing the acceptance probability on the right-most subinterval contributes positively to the principal's utility. Setting γ^* to 1 turns out to have another benefit: It incentivizes the agent to pick a more variable distribution (in the MPS-ordering), and we already intuited in the proof sketch for Lemma 4.2 that this benefits the principal.

Lastly, consider the value of α^*, the acceptance probability on the left-most subinterval. Accepting the proposal for realizations in this subinterval contributes negatively to the principal's utility. Nevertheless, the principal may find it optimal to pick a non-zero value of α^* in order to reimburse the agent for costs of acquiring F^*. We later present an example where no optimal mechanism can avoid accepting the proposal at types where it is strictly better to reject. See Proposition 5.3 in Section 5.3. (The example serves to illustrate some other points, and hence it is convenient to delay the result.)

5 The evidence-disclosure game

In this section we consider a nearby game where the principal does not have commitment power. We find that, under weak assumptions, the principal's utility is 0 in all equilibria of a natural class of equilibria.

5.1 Setup

The timing of the game is as follows:

(1) The agent covertly picks F in $\mathcal{F}$.

(2) The agent draws θ from F, and discloses evidence e in $\{\{\theta\}, [0, 1]\}$.

(3) The principal observes e, and accepts or rejects the proposal.

The agent's strategy specifies a type distribution F, and a Borel-measurable function $\sigma_A\colon [0,1] \to [0,1]$; here $\sigma_A(\theta)$ is the probability that the agent provides evidence $e = \{\theta\}$ when the realized type is θ; with complementary probability $1 - \sigma(\theta)$ the agent provides the trivial piece of evidence $e = [0,1]$.[10]

A strategy σ_P of the principal specifies a probability $\sigma_P([0,1])$ of accepting when the agent discloses the trivial piece of evidence, and a Borel-measurable function $\sigma_P\colon [0,1] \to [0,1]$; here $\sigma_P(\theta)$ is the acceptance probability when the agent provides evidence that the type is θ.

We consider (perfect Bayesian) equilibria with an additional property.[11] An equilibrium is *truth-leaning* if for *all* $\theta \in [0,1]$, when the agent provides hard evidence that the type is θ, then the principal's beliefs are degenerate on θ. This property has bite for types of the agent that do not provide evidence in equilibrium, or that are not in the support of the equilibrium type distribution. We later discuss its role in the next proposition as well as its relation to the truth-leaning equilibria of Hart et al. (2017).

5.2 The principal's equilibrium utility

The next result characterizes the possible values of the principal's equilibrium utility. Note that 0 is a lower bound on the principal's equilibrium utility since the principal can always reject.

Proposition 5.1. *Let Assumptions 1 and 2 hold. There exists a truth-leaning equilibrium where the principal's equilibrium utility is* 0*. In all truth-leaning equilibria exactly one of the following is true:*

(1) The principal's equilibrium utility is 0*.*

(2) The agent picks a distribution F satisfying $\mathbb{E}_F[u_p(\theta)] > 0$. If the agent does not disclose, the principal accepts with probability 1*. The proposal is accepted with F-probability* 1*.*

[10] We could allow the agent to additionally send costless messages from some message set but this would not alter out conclusions. To elaborate: Suppose there are two messages which all types of the agent can send. If both are played in equilibrium, they must lead to the same acceptance probability since all types of the agent strictly prefer that the proposal be accepted. Hence both messages lead to the same payoff of the principal, too. (If the principal's decision was between three or more alternatives, this argument would not go through.)

[11] See Appendix A.2.1 for a formal definition.

See Appendix A.2.2 for a proof.

The proof is easy to explain. Consider an equilibrium and let F denote the agent's type distribution. Suppose the principal's utility is strictly positive; this implies $\max \operatorname{supp} F > u_p^{-1}(0)$.

In an intermediate step, assume towards a contradiction that the set of types where the proposal is not accepted with probability 1 has non-zero F-probability. These must be types weakly below $u_p^{-1}(0)$ since, given the truth-leaning beliefs of the principal and the principal's lack of commitment, the principal accepts whenever the agent discloses a type strictly greater than $u_p^{-1}(0)$. Given $\max \operatorname{supp} F > u_p^{-1}(0)$, we can consider a perturbation of F where, informally speaking, the maximum of the support is slightly decreased, and where probability mass is shifted to points above $u_p^{-1}(0)$. The resulting distribution is an MPC of F that assigns strictly higher probability to points above $u_p^{-1}(0)$. Thus it yields a strictly higher acceptance probability. Since it is an MPC of F, it is also cheaper to acquire. Thus the perturbation constitutes a profitable deviation for the agent; contradiction.

At this point we know that on the equilibrium path the principal accepts F-almost surely. To complete the argument, we notice that types strictly below $u_p^{-1}(0)$ (which must arise in equilibrium, as else F is degenerate, and then the prinicipal's utility could not be strictly positive), can only get the principal to accept by not disclosing the type. Thus the principal must be accepting whenever the agent does not disclose. In summary, all equilibria where the principal's utility is strictly positive must fall into case (2) of Proposition 5.1. Recalling that the principal's utility is always weakly positive, all other equilibria fall into case (1).

Remark 1. Assumption 2 is stronger than what we really need for this argument. It suffices if the agent can choose all distributions in a non-empty subset of $\mathcal{F}$ that is closed with respect to MPCs, K is decreasing with respect to MPCs (but not necessarily linear), and u_a is strictly positive and concave.[12]

Which of the equilibria in Proposition 5.1 actually arise? Consider an equilibrium that falls into case (2). One possible deviation of the agent is to acquire the degenerate distribution and disclose nothing. Since the degenerate distribution is always the cheapest one to acquire, the given equilibrium can only be sustained if the agent's type

[12] To ensure the existence of a best response, we would also assume that the set of avilable type distributions is L^1-compact, that K is L^1-lower semicontinuous, and that u_a is continuous (at the boundary).

distribution is as cheap as the degenerate distribution. The next result is immediate from these observations.

Corollary 5.2. *Let Assumptions 1 and 2 hold. The principal's utility is* 0 *in all truth-leaning equilibria if at least one of the following holds:*

(1) The function u_a is strictly concave.

(2) The function K is strictly decreasing with respect to MPCs.[13]

(3) All $F \in \mathcal{F}$ satisfy $\mathbb{E}_F[u_p(\theta)] \leq 0$.

Corollary 5.2 speaks to the principal's value of commitment under weak conditions on the environment. Consider condition (2). Suppose that for all F the costs are given by $K(F) = \lambda \bar{K}(F)$, where λ is a strictly positive parameter and $\bar{K}$ is strictly decreasing with respect to MPCs (and satisfies Assumption 2). As λ vanishes, distributions become arbitrarily cheap to acquire. By Corollary 5.2, the principal's utility is equal to 0 in all truth-leaning equilibria of the evidence-disclosure game. Nevertheless, Proposition A.7 in Appendix A.1.1 implies that as $\lambda \to 0$ the principal's utility approaches the utility that would obtain if the principal could freely choose the type distribution.[14] One can make a similar argument involving condition (1).

Under condition (3), there is a conflict of interest. When the type distribution is F and the principal makes a decision without consulting the agent, the principal's utility is $\mathbb{E}_F[u_p(\theta)]$. Hence condition (3) asserts that, no matter F, the principal would never accept the proposal without consulting the agent. Note that if the principal's utility u_p is affine in the type, then condition (3) is implied by the assumption $\mu \leq u_p^{-1}(0)$ (Assumption 1). In particular, if u_p is affine and $u_p^{-1}(0) = \mu$, then all truth-leaning equilibria give the principal a utility of 0 (Corollary 5.2), whereas the principal can obtain a strictly positive utility by committing to a mechanism (recall the paragraph after Assumption 3).

[13] That is, if F is an MPC of F' and $F \neq F'$, then $K(F) < K(F')$.

[14] This recalls the recent result of Ravid et al. (2022). In a model of bilateral trade with costly endogenous information, they show that equilibria approach the Pareto-worst equilibrium with free learning as information costs vanish. In our model, the equilibria of Proposition 5.1 can be Pareto-ranked if u_a is affine and $K = 0$, and in this case the equilibria from case (2) Pareto-dominate the equilibria from case (1). A slight perturbation of the costs K eliminates all equilibria from case (2).

5.3 Truth-leaning beliefs

We now discuss the restriction to truth-leaning beliefs in Proposition 5.1. Hart et al. (2017), who study the value of commitment in a model with an exogenous type distribution, propose truth-leaning equilibria as ones where the agent enjoys some inherent bonus from revealing the truth.[15] They show that in terms of outcomes truth-leaning equilibria are equivalent to perfect Bayesian equilibria in which (i.) the agent discloses the type if doing so is weakly optimal, and (ii.) the principal's beliefs are degenerate on the type when the agent discloses the type. Here we have only imposed an analogue of (ii.).

It is worth emphasizing that Proposition 5.1 does *not* extend to non-truth-leaning equilibria. To see this, let $K = 0$. Consider the one-sided cutoff mechanism x with cutoff 1; that is, the mechanism that accepts if and only if the agent discloses $\theta = 1$. The unique agent-optimal distribution on this mechanism is the binary distribution $\bar{F}$ with support $\{0, 1\}$ and mean μ. In particular, the principal's utility from committing to x is strictly positive. Now consider the following strategy profile and beliefs: The agent acquires $\bar{F}$ and always discloses the type. If the agent discloses a type of 1, the principal's belief is degenerate on 1 and the principal accepts; in *all* other cases, the principal's belief is degenerate on 0 and the principal rejects. Since types in $(0, 1)$ are not in the support of $\bar{F}$, these strategies and beliefs define an equilibrium. It clearly induces the same outcome as the mechanism x. It also does not fall into case (1) of Proposition 5.1 (as the principal's utility is strictly positive) nor case (2) (as the agent always discloses the type).

The construction in the previous paragraph relies on $K = 0$. For environments where K is non-zero, it is conceivable that to incentivize the agent's choosing a non-degenerate distribution the principal may also have to accept at types that are *on-path* and where accepting is not a best response. No choice of beliefs can rationalize this behavior of the principal. We already intuited in Section 4.4 that the principal may optmally commit to such behavior. Thus commitment may remain valuable if costs are non-negligible. The next proposition verifies that such environments exist.

[15] In their definition, an equilibrium is truth-leaning if it is obtained as the limit of equilibria of nearby games; in the latter games, the agent enjoys a bonus from revealing the truth and must reveal the truth with a strictly positive. The limit lets the bonus and the lower bound on probability of revealing the truth converge to 0.

Proposition 5.3. *There exist μ, u_a, u_p and K satisfying Assumptions 1 to 3, and such that in this environment all of the following are true:*

(1) In the evidence-disclosure game, the principal's utility is 0 in all equilibria (truth-leaning or not).

(2) Let x be a mechanism. Let $F \in \mathcal{F}$ be agent-optimal on x. If $U_p(x, F) > 0$, then the set $\{\theta \in [0, u_p^{-1}(0)) \colon x(\theta) \in (0,1)\}$ has non-zero F-probability.

See Appendix A.2.3 for a proof.

In words, the principal must accept probabilistically at types that are in the support of type distribution but where it is strictly better to reject.

An implication of (1) is that commitment is valuable (the principal's utility is 0 in all equilibria, but the non-triviality Assumption 3 holds).

An implication of (2) is that deterministic mechanisms do not suffice for solving the principal's problem, the following sense. Suppose the principal attempted to implement a given mechanism x by first randomly selecting a *deterministic* mechanism, and then revealing this deterministic mechanism to the agent. The result implies that if the agent picks the distribution after the deterministic mechanism is revealed, then the agent picks a distribution for which the principal's utility is at most 0. The fact that deterministic mechanisms do not suffice contrasts the case where the distribution is exogenously given; for in that case, the principal's utility from a convex combination of mechanisms is equal to the convex combination of their individual utilities (see Section 6.1).[16]

6 Conclusion

We studied a mechanism design problem where the agent endogenously acquires a type distribution and possesses hard evidence about the type realization. The comparison with a nearby game reveals that the principal benefits from commitment power. In particular, the principal may optimally commit to picking the agent's least favourite alternative when their preferences are aligned, and to picking the agent's favourite alternative when their preferences are misaligned. To wrap up, we discuss some variations on the model.

[16] The existing literature has found general conditions under which deterministic mechanism do suffice in models with exogenous type distribution. See in particular Ben-Porath et al. (2019) and Sher (2011, 2014).

6.1 Exogenous type distributions

Suppose the agent's type distribution is fixed to some distribution $F \in \mathcal{F}$. As before, it is without loss to view the principal as choosing a usc function $x\colon [0,1] \to [0,1]$, giving the principal a utility of $\mathbb{E}_F[x(\theta)u_p(\theta)]$. Clearly, this utility is no greater than $\mathbb{E}_F[\max(0, u_p(\theta))]$. This upper bound is attainable if the principal commits to the one-sided cutoff mechanism with cutoff $u_p^{-1}(0)$.[17] We claim that this one-sided cutoff mechanism is not optimal when the agent's distribution is endogenous. The reason is that the agent will acquire a distribution that is supported on points weakly below $u_p^{-1}(0)$. The principal therefore ends up rejecting with probability 1, or being indifferent between accepting and rejecting. In particular, the principal enjoys a utility of 0. This is precisely the reasoning that also established Proposition 5.1.

6.2 Costly verification

In Appendix B.1 we consider a model where the agent cannot provide costless evidence about the type, but where the principal can audit the agent at a cost. If the agent's type is θ, auditing reveals the type and the principal incurs a cost $c(\theta)$, where $c\colon [0,1] \to \mathbb{R}_+$ is some function. With simple modifications to the current proofs and by strengthening Assumption 1, one can show that two-sided cutoff mechanisms are optimal in this model. The stronger assumption demands that u_p and $u_p - c$ be convex, and that $u_p - c$ crosses 0 exactly once.[18]

The costly verification model admits a new trade-off. Fixing some type distribution F, it turns out that the principal can save on auditing costs by promising to accept the proposal at types *outside* the support of F. This leads to a trade-off since the acceptance-probability at such types of course affects the agent's incentives for actually acquiring F. An implication is that, on a two-sided cutoff mechanism, the acceptance-probability on the middle interval may or may not be 0. A sufficient

[17] In fact, all other optimal mechanisms must agree with this one except on a set having F-probability 0. This result naturally extends if we drop Assumption 1. For general payoffs u_p of the principal, the mechanism that accepts if and only if the agent discloses a type θ such that $u_p(\theta) \geq 0$ is optimal. All other optimal mechanisms agree with this one except on a set having F-probability 0.

[18] This roughly says that the principal prefers a distribution to its mean-preserving contractions, even when the principal is committed to always auditing the agent. The evidence model with Assumption 1 is the special case where $c = 0$. The standing assumption in the literature is that c is constant (Ben-Porath et al., 2014; Erlanson and Kleiner, 2020).

condition for it to be 0 optimally is that all $F \in \mathcal{F}$ satisfy $\mathbb{E}_F[u_p(\theta)] \leq 0$. By contrast, in the evidence model, it is always without loss to reject at types outside the support of F.

6.3 Multiple alternatives

Appendix B.2 considers a more general model with an arbitrary finite set of alternatives. The assumptions on the preferences are that the principal's payoffs (the agent's payoffs) from each alternative are convex (concave) in the type, and that the agent has a least preferred alternative that is independent of the type.[19] We maintain the assumptions on the agent's costs K and the set of available distributions.

The main finding is that the characterization of optimal mechanisms as two-sided cutoff mechanisms extends to multiple alternatives, in the following sense: The agent is asked to disclose whether the type falls into a "low", "intermediate", or "high" interval (and will always disclose truthfully). Disclosing an intermediate type leads to the agent's least-preferred alternative. Disclosing a low type leads to some lottery over alternatives; disclosing a high type leads to a possibly different lottery. In this sense, the result that optimal mechanisms require no more than three messages is unrelated to the number of alternatives, but rather an implication of the assumptions on the set of available distributions (its extreme points are binary) and the preferences (utilities are linear in distributions, the principal's payoffs are convex in the type, the agent's payoffs are concave in the type).

6.4 Learning about an underlying state

The interpretation of the model has so far been that the agent can affect the distribution of a payoff-relevant state—the type. In another interpretation, there is an underlying payoff-relevant state ω that realizes either as 0 or 1, and that has mean μ—in particular, the state is binary. The type distribution represents the distribution of posterior means of ω that obtains when the agent learns about the state through some information structure. The payoffs of $u_p(\theta)$ and $u_a(\theta)$, respectively, are then the principal's and agent's payoffs, respectively, from accepting the proposal when

[19] For binary alternatives, the existence of a type-independent least-preferred alternative implies that the agent's ordinal preferences are type-independent. In Appendix B.2, we do not assume type-independent ordinal preferences.

the *agent*'s posterior mean of the state is θ.[20]

The agent can provide hard evidence about the realization of the posterior mean. The assumption that this realization is verifiable is in line with the existing literature, see e.g. Rappoport and Somma (2017) and Yoder (2022). We also assume that the principal's mechanism cannot condition on the experimental process that generated the posterior.[21] Linear acquisition costs K correspond to the class of posterior-separable cost functions that are assumed in various contributions to the literature.[22]

Appendices

Appendix A Omitted proofs

A.1 Omitted proofs for Section 4

A.1.1 Auxiliary results

Given a mechanism x, recall that $\mathcal{F}^*(x)$ denotes the set of agent-optimal distributions on x. We noted in Section 3 that $\mathcal{F}^*(x)$ is non-empty.

Lemma A.1. *Let Assumptions 1 and 2 hold. For all mechanisms x, the function $F \mapsto U_p(x, F)$ is continuous on $\mathcal{F}^*(x)$.*

Proof of Lemma A.1. Let $\{F_n\}_n$ be a sequence in $\mathcal{F}^*(x)$ converging to $F \in \mathcal{F}^*(x)$. Let $\lambda = \int x u_a \, dF$ and, for all n, let $\lambda_n = \int x u_a \, dF_n$. By agent-optimality, we have

[20] The distinction between whether θ is the agent's or the principal's posterior mean matters if u_p is non-affine. Suppose the agent does not provide evidence. If θ is the agent's posterior mean, the principal's forms a belief about the agent's posterior mean. Denoting this belief by F, the principal's payoff from accepting is $\mathbb{E}_F[u_p(\theta)]$. If θ is the principal's posterior mean, the principal forms a belief F about the state, yielding a payoff from accepting of $u_p(\mathbb{E}_F[\theta])$.

[21] To briefly elaborate, suppose the agent covertly picks a mapping from the state space to some abstract set of messages (an "experiment"). Each message m is associated with a posterior mean $\hat{\theta}(m)$ of the state. The agent's choice of the experiment affects the distribution over messages, but not the associated values $\hat{\theta}(\cdot)$ of the posterior mean. As long as payoffs and costs only depend on posterior means, it is then a normalization to assume that each message is itself a posterior mean. See also the discussion by Shishkin (2021) on the interpretation of evidence in the disclosure game. In Shishkin's model, the principal learns the experiment (which is modelled slightly differently) when the agent discloses the realized message.

[22] See Mensch and Ravid (2022), Thereze (2022), and Yoder (2022) for some recent mechanism design problems with transfers where information is endogenous and costs are posterior-separable.

$\int (xu_a - k)\, dF_n = \int (xu_a - k)\, dF$. Since k is continuous, we have $\int k\, dF_n \to \int k\, dF$, and hence $\lambda_n \to \lambda$.

We now distinguish two cases, depending on whether λ is strictly positive. First, let $\lambda > 0$. In what follows, we understand n to be large enough such that $\lambda_n > 0$. For all Borel sets B of reals, let

$$G(B) = \frac{1}{\lambda} \int_B xu_a\, dF \quad \text{and} \quad G_n(B) = \frac{1}{\lambda_n} \int_B xu_a\, dF_n.$$

Since $u_a > 0$, it follows that G and G_n are probability measures on the Borel subsets of reals.

In an intermediate step, we claim that $\{G_n\}_n$ weak-$*$ converges to G. Let B be a closed set of reals. Since x is usc, we infer from Theorem 15.5 of Border and Aliprantis (2006) that $\limsup_n \int_B xu_a\, dF_n \leq \int_B xu_a\, dF$ holds. Now $\lambda_n \to \lambda$ implies $\limsup_n G_n(B) \leq G(B)$. Since B was an arbitrary closed set, Theorem 15.3 of Border and Aliprantis (2006) implies that $\{G_n\}_n$ weak-$*$ converges to G, as promised.

The principal's utilities from F and F_n, respectively, are given by $\int xu_p\, dF$ and $\int xu_p\, dF_n$, respectively. Note that u_p/u_a is (well-defined and) continuous. Since G and G_n, respectively, are absolutely continuous with respect to F and F_n, respectively, Theorem 13.23 of Border and Aliprantis (2006) implies that we can write the principal's utility from F and F_n, respectively, as $\lambda \int \frac{u_p}{u_a}\, dG$ and $\lambda_n \int \frac{u_p}{u_a}\, dG_n$, respectively. Since $\{G_n\}_n$ weak-$*$ converges to G, and since $\lambda_n \to \lambda$, we conclude that the principal's utility from F_n converges to the utility from F, as promised.

Second, let $\lambda = 0$. Now define G and G_n, respectively, via $G(B) = \int_B xu_a\, dF$ and $G_n(B) = \int_B xu_a\, dF_n$, respectively. Note that $\lim_n \int dG_n = 0$ holds. Since G is absolutely continuous with respect to F, and since u_p/u_a is continuous and bounded, we can again invoke Theorem 13.23 of Border and Aliprantis (2006) to write $\int xu_p\, dF = \int \frac{u_p}{u_a}\, dG = 0$. Similarly,

$$\left| \int xu_p\, dF_n \right| \leq \int \left| \frac{u_p}{u_a} \right| xu_a\, dF_n = \int \left| \frac{u_p}{u_a} \right| dG_n \leq \sup \left| \frac{u_p}{u_a} \right| \int dG_n \to 0.$$

Thus the principal's utility from F_n converges to the utility from F, as desired. □

Given a mechanism x, let $\mathcal{F}^*_B(x)$ denote the set of binary agent-optimal distributions on x.

Lemma A.2. *Let Assumptions 1 and 2 hold. For all mechanisms x, both of the following are true:*

(1) The sets $\mathcal{F}^*(x)$ and $\mathcal{F}_B^*(x)$ are non-empty and compact.

(2) If $F \in \mathcal{F}^*(x)$, then there exists a probability measure ν supported on a subset of $\mathcal{F}_B^*(x)$ such that all continuous linear functions $\Gamma\colon \mathcal{F}^*(x) \to \mathbb{R}$ satisfy

$$\Gamma(F) = \int_{\tilde{F} \in \mathcal{F}_B^*} \Gamma(\tilde{F})\, d\nu(\tilde{F}).$$

Proof of Lemma A.2. Let $\mathcal{F}_B$ denote the set of binary distributions in $\mathcal{F}$. We know from Winkler (1988) that $\mathcal{F}_B$ is the set of extreme points of $\mathcal{F}$. The agent's utility is linear and usc on the compact set $\mathcal{F}$. Hence Bauer's Maximum Principle (Border and Aliprantis, 2006, Theorem 7.69) implies that $\mathcal{F}_B^*(x)$ is non-empty. Hence $\mathcal{F}^*(x)$ is non-empty. Compactness of $\mathcal{F}^*(x)$ and $\mathcal{F}_B^*(x)$, respectively, follows from upper-semicontinuity of the agent's utility and compactness of $\mathcal{F}$ and $\mathcal{F}_B$, respectively.

Claim (2) follows from Choquet's theorem (Phelps, 2001, p.14) if we can show that the set of extreme points of $\mathcal{F}^*(x)$ is a subset of $\mathcal{F}_B^*(x)$. Let F be an extreme point of $\mathcal{F}^*(x)$. Choquet's theorem (Phelps, 2001, p. 14) implies that there is a probability measure ν supported on $\mathcal{F}_B$ and such that all continuous linear functions $\Gamma\colon \mathcal{F} \to \mathbb{R}$ satisfy

$$\Gamma(F) = \int_{\tilde{F} \in \mathcal{F}_B} \Gamma(\tilde{F})\, d\nu(\tilde{F}). \tag{A.1}$$

Now, we recall that k is continuous, that x is usc, and that u_a is strictly positive and continuous. Hence $xu_a - k$ is usc. Hence Theorem 3.13 of Border and Aliprantis (2006) lets us find a sequence $\{h_n\}_n$ of continuous functions from $[0,1]$ to $\mathbb{R}$ converging pointwise to $xu_a - k$. Since h_n is continuous, the mapping $\tilde{F} \mapsto \int_{\theta\in[0,1]} h_n(\theta)\, d\tilde{F}(\theta)$ is continuous (and obviously linear). Hence (A.1) implies

$$\int_{\theta\in[0,1]} h_n(\theta)\, dF(\theta) = \int_{\tilde{F} \in \mathcal{F}_B} \left(\int_{\theta\in[0,1]} h_n(\theta)\, d\tilde{F}(\theta) \right) d\nu(\tilde{F}).$$

The Dominated Convergence theorem implies[23]

$$\int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, dF = \int_{\tilde{F}\in\mathcal{F}_B} \left(\int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, d\tilde{F}(\theta) \right) d\nu(\tilde{F}).$$

Since F is agent-optimal on x, the previous display implies ν-almost all $\tilde{F}$ satisfy $\int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, dF = \int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, d\tilde{F}$. Hence ν is supported on a subset of binary agent-optimal distributions, meaning a subset of $\mathcal{F}_B^*(x)$. We know infer from (A.1) that all continuous linear functions $\Gamma\colon \mathcal{F} \to \mathbb{R}$ satisfy

$$\Gamma(F) = \int_{\tilde{F}\in\mathcal{F}_B^*(x)} \Gamma(\tilde{F})\, d\nu(\tilde{F}). \tag{A.2}$$

Now recall that F is assumed to be an extreme point of $\mathcal{F}^*(x)$. Hence (A.2) and Proposition 1.4 of Phelps (2001) together imply that ν is supported on F. Since ν is supported on a subset of $\mathcal{F}_B^*(x)$, we conclude $F \in \mathcal{F}_B^*(x)$, as promised. □

We next provide a lemma that collects some useful properties of mechanism-distribution pairs that leave the principal with a strictly positive utility.

Lemma A.3. *Let Assumptions 1 and 2 hold. Let x be a mechanism. Let F with support $\{\theta_0, \theta_1\}$ be a binary distribution in $\mathcal{F}$. If $U_p(x, F) > 0$, then all of the following hold:*

(1) $u_p^{-1}(0) \in (\theta_0, \theta_1)$.

(2) F *is non-degenerate.*

(3) $\frac{-x(\theta_0)u_p(\theta_0)}{u_p^{-1}(0)-\theta_0} \leq \frac{x(\theta_1^*)u_p(\theta_1)}{\theta_1-u_p^{-1}(0)}$.

Proof of Lemma A.3. We will show the contrapositive: If one of the conditions in the claim fails to hold, then $U_p(x, F) \leq 0$. Recall that $U_p(x, F) = \mathbb{E}_F\left[x(\theta)u_p(\theta)\right]$ holds.

Assumption 1 implies $\mu \leq u_p^{-1}(0)$. Hence, if $\theta_1 \leq u_p^{-1}(0)$ or if F is degenerate, then F is supported on points for which u_p is weakly negative, implying $\mathbb{E}_F\left[x(\theta)u_p(\theta)\right] \leq 0$.

In what follows, we may thus assume $\theta_1 > u_p^{-1}(0)$ and that F is non-degenerate. Recalling $u_p^{-1}(0) \geq \mu$, we also find $u_p^{-1}(0) \geq \mu > \theta_0$ as else the mean of F would not be μ.

[23] Since $xu_a - k$ is usc and bounded, the mapping $\tilde{F} \mapsto \int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, d\tilde{F}$ is usc and bounded. Hence the integral $\int_{\tilde{F}\in\mathcal{F}_B} \left(\int_{\theta\in[0,1]} (x(\theta)u_a(\theta) - k(\theta))\, d\tilde{F}(\theta) \right) d\nu(\tilde{F})$ is well-defined and the Dominated Convergence theorem can be applied.

Let us now turn to the inequality

$$\frac{-x(\theta_0)u_p(\theta_0)}{u_p^{-1}(0)-\theta_0} \leq \frac{x(\theta_1)u_p(\theta_1)}{\theta_1 - u_p^{-1}(0)}. \tag{A.3}$$

Consider the piece-wise affine function $\omega\colon [\theta_0, \theta_1] \to \mathbb{R}$ defined by

$$\forall_{\theta\in[0,1]}, \quad \omega(\theta) = \begin{cases} x(\theta_0)u_p(\theta_0)\frac{u_p^{-1}(0)-\theta}{u_p^{-1}(0)-\theta_0}, & \text{if } \theta \in [\theta_0, u_p^{-1}(0)) \\ x(\theta_1)u_p(\theta_1)\frac{\theta-u_p^{-1}(0)}{\theta_1-u_p^{-1}(0)}, & \text{if } \theta \in [u_p^{-1}(0), \theta_1]. \end{cases} \tag{A.4}$$

If (A.3) fails, then ω is concave. Moreover, if $\theta \in \{\theta_0, \theta_1\}$, then $\omega(\theta) = x(\theta)u_p(\theta)$. Since $\{\theta_0, \theta_1\}$ is the support of F, we conclude

$$\mathbb{E}_F\left[x(\theta)u_p(\theta)\right] = \mathbb{E}_F\left[\omega(\theta)\right] \leq \omega(\mu).$$

We also know that $\theta_0 < \mu \leq u_p^{-1}(0) < \theta_1$ holds. Since ω is weakly increasing, we have $\omega(\mu) \leq \omega(u_p^{-1}(0))$. Inspection of ω shows $\omega(u_p^{-1}(0)) = 0$. Thus, if (A.3) fails, then $\mathbb{E}_F\left[x(\theta)u_p(\theta)\right] \leq 0$. □

The next lemma is an easy corollary of Lemma A.3. It provides a sufficient condition such that, on a given mechanism, the principal prefers mean-preserving spreads to mean-preserving contractions.

Lemma A.4. *Let Assumptions 1 and 2 hold. Let x be a two-sided cutoff mechanism with parameters $(\theta_0, \theta_1, \alpha, \beta, \gamma)$ such that $\mu \in (\theta_0, \theta_1)$. Let F denote the binary distribution in $\mathcal{F}$ whose support is $\{\theta_0, \theta_1\}$. Let F' and F'' be two binary distributions in $\mathcal{F}$ in such that F is an MPC of F', and F' is an MPC of F''. If $U_p(x, F) > 0$, then $U_p(x, F'') \geq U_p(x, F')$.*

Proof of Lemma A.4. Since $U_p(x, F) > 0$, Lemma A.3 implies $u_p^{-1}(0) \in (\theta_0, \theta_1)$ and

$$\frac{-x(\theta_0)u_p(\theta_0)}{u_p^{-1}(0)-\theta_0} \leq \frac{x(\theta_1)u_p(\theta_1)}{\theta_1 - u_p^{-1}(0)}. \tag{A.5}$$

Consider the function $\bar{\omega}\colon [0,1] \to \mathbb{R}$ defined as follows (the function is well-defined since $u_p^{-1}(0) \in (\theta_0, \theta_1)$):

$$\forall_{\theta\in[0,1]}, \quad \bar{\omega}(\theta) = \begin{cases} x(\theta_0)u_p(\theta), & \text{if } \theta < \theta_0 \\ x(\theta_0)u_p(\theta_0)\frac{u_p^{-1}(0)-\theta}{u_p^{-1}(0)-\theta_0}, & \text{if } \theta \in [\theta_0, u_p^{-1}(0)) \\ x(\theta_1)u_p(\theta_1)\frac{\theta-u_p^{-1}(0)}{\theta_1-u_p^{-1}(0)}, & \text{if } \theta \in [u_p^{-1}(0), \theta_1] \\ x(\theta_1)u_p(\theta), & \text{if } \theta > \theta_1. \end{cases}$$

Assumption 1 and (A.5) together imply that $\bar{\omega}$ is weakly convex.

We next claim that $\bar{\omega}(\theta)$ agrees with $x(\theta)u_p(\theta)$ at all points θ in the supports of F' and F''. To see this, observe that since F, F' and F'' are binary the assumptions on the MPC-ordering implies

$$\min\operatorname{supp} F'' \leq \min\operatorname{supp} F' \leq \theta_0 \leq \theta_1 \leq \max\operatorname{supp} F' \leq \max\operatorname{supp} F''.$$

Recall also that x is a two-sided cutoff mechanism with cutoffs θ_0 and θ_1. In particular, we have that x is constant on $[0, \theta_0]$ and $[\theta_1, 1]$. The claim follows from these observations.

By the previous paragraph, we can write

$$U_p(x, F') = \mathbb{E}_{F'}\left[x(\theta)u_p(\theta)\right] = \mathbb{E}_{F'}\left[\bar{\omega}(\theta)\right].$$

The function $\bar{\omega}$ is convex, and F'' is an MPS of F'. Hence the expression in the previous display is at most as great as $\mathbb{E}_{F''}\left[\bar{\omega}(\theta)\right]$. By rewriting this expectation in the same manner as above, we find that it equals $U_p(x, F'')$. Thus we have shown $U_p(x, F'') \geq U_p(x, F')$, as promised. □

Given a two-sided cutoff mechanism x, let $\underline{\theta}_x = \inf\{\theta \in [0,1] \colon x(\theta) = \min x\}$ and $\bar{\theta}_x = \sup\{\theta \in [0,1] \colon x(\theta) = \min x\}$. (The minimum is well-defined since a two-sided cutoff mechanism assumes at most three distinct values.)

Lemma A.5. *Let x be a two-sided cutoff mechanism such that $\mu \in (\underline{\theta}_x, \bar{\theta}_x)$. If F is agent-optimal on x and assigns non-zero probability to $(\underline{\theta}_x, \bar{\theta}_x)$, then the support of F is a subset of $[\underline{\theta}_x, \bar{\theta}_x]$.*

Proof of Lemma A.5. We shall prove $\min\operatorname{supp} F \geq \underline{\theta}_x$; a similar argument establishes $\max\operatorname{supp} F \leq \bar{\theta}_x$. Towards a contradiction, let $\min\operatorname{supp} F < \underline{\theta}_x$. Let $\beta = \min x$. By definition of $\underline{\theta}_x$, it follows that x is constantly equal to some probability $\alpha \in (\beta, 1]$ on $[0, \underline{\theta}_x]$.

Let f_0 and f_1, respectively, denote the probabilities that F assigns to the subintervals $[0, \underline{\theta}_x]$ and $(\underline{\theta}_x, \bar{\theta}_x)$, respectively. We have $f_0 > 0$ and $f_1 > 0$. Let $\theta_0 = \mathbb{E}_F[\theta | \theta \leq \underline{\theta}_x]$ and $\theta_1 = \mathbb{E}_F[\theta | \theta \in (\underline{\theta}_x, \bar{\theta}_x)]$. Let $\tilde{F}$ denote the distribution that assigns f_0 to θ_0, f_1 to θ_1, and with probability $(1 - f_0 - f_1)$ agrees with the conditional distribution of F on $[\bar{\theta}_x, 1]$.[24] Formally, for all θ, let $\tilde{F}(\theta) = f_0 \mathbf{1}_{(\theta \geq \theta_0)} + f_1 \mathbf{1}_{(\theta \geq \theta_1)} + (F(\theta) - f_0 - f_1)\mathbf{1}_{(\theta \geq \bar{\theta}_x)}$.

Since x is a two-sided cutoff mechanism, it is constant on each of the subintervals $[0, \underline{\theta}_x]$, $(\underline{\theta}_x, \bar{\theta}_x)$ and $[\bar{\theta}_x, 1]$. Since u_a is concave and K is decreasing with respect to MPCs, it follows that $U_a(x, \tilde{F}) \geq U_a(x, F)$ holds. We shall find a distribution $\hat{F}$ such that $U_a(x, \hat{F}) > U_a(x, \tilde{F})$; this contradicts the agent-optimality of F and hence completes the argument.

For a number $\varepsilon > 0$ to be chosen in a moment, let $\eta_\varepsilon > 0$ solve $\theta_0 + \eta_\varepsilon = \frac{f_0\theta_0 + \varepsilon\theta_1}{f_0 + \varepsilon}$. Let $\hat{F}$ denote the distribution that assigns $f_0 + \varepsilon$ to $\theta_0 + \eta_\varepsilon$, $f_1 - \varepsilon$ to θ_1, and with probability $(1 - f_0 - f_1)$ agrees with the conditional distribution of F on $[\bar{\theta}_x, 1]$. Formally, for all θ, let $\hat{F}(\theta) = (f_0 + \varepsilon)\mathbf{1}_{(\theta \geq \theta_0 + \eta_\varepsilon)} + (f_1 - \varepsilon)\mathbf{1}_{(\theta \geq \theta_1)} + (F(\theta) - f_0 - f_1)\mathbf{1}_{(\theta \geq \bar{\theta}_x)}$. If ε is sufficiently small, then $\tilde{F}$ is a well-defined distribution, and in fact an MPC of $\hat{F}$. Moreover, for sufficiently small ε we have $\theta_0 + \eta_\varepsilon < \underline{\theta}_x$. Fixing such a value of ε, we complete the argument by showing $U_a(x, \hat{F}) > U_a(x, \tilde{F})$.

Since $\hat{F}$ is an MPC of $\tilde{F}$, and since K decreases with respect to MPCs, it suffices to show $\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] - \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)] > 0$. This difference spells out to

$$\begin{aligned} &(f_0 + \varepsilon)\alpha u_a(\theta_0 + \eta_\varepsilon) + (f_1 - \varepsilon)\beta u_a(\theta_1) - f_0\alpha u_a(\theta_0) - f_1\beta u_a(\theta_1) \\ =&(f_0 + \varepsilon)\alpha u_a(\theta_0 + \eta_\varepsilon) - f_0\alpha u_a(\theta_0) - \varepsilon\beta u_a(\theta_1). \end{aligned}$$

Since $\theta_0 + \eta_\varepsilon = \frac{f_0\theta_0 + \varepsilon\theta_1}{f_0 + \varepsilon}$, and since u_a is concave, a lower bound on the difference in the previous display is

$$\begin{aligned} &(f_0 + \varepsilon)\alpha \left(\frac{f_0 u_a(\theta_0)}{f_0 + \varepsilon} + \frac{\varepsilon u_a(\theta_1)}{f_0 + \varepsilon} \right) - f_0\alpha u_a(\theta_0) - \varepsilon\beta u_a(\theta_1) \\ =&\varepsilon(\alpha - \beta)u_a(\theta_1). \end{aligned}$$

Since $\alpha > \beta$ (and $\varepsilon > 0$ and $u_a > 0$), this lower bound is strictly positive. □

[24] If $(1 - f_0 - f_1) = 0$, we understand $\tilde{F}$ to simply mean the binary distribution that assigns f_0 to θ_0 and f_1 to θ_1.

The following corollary is easily obtained from Lemma 4.1 and Lemma 4.2; we omit the proof.

Corollary A.6. *Let Assumptions 1 and 2 hold. Let x be a mechanism such that $\bar{U}_p(x) > 0$. There exist θ_0^* and θ_1^*, and probabilities α and γ, such that $0 \leq \theta_0 < \mu \leq u_p^{-1}(0) < \theta_1 \leq 1$ and $\gamma > 0$, and such that two-sided cutoff mechanism x^* with parameters $(\theta_0^*, \theta_1^*, \alpha, 0, \gamma)$ satisfies $\bar{U}_p(x^*) \geq \bar{U}_p(x)$. Moreover, the binary distribution F^* with support $\{\theta_0^*, \theta_1^*\}$ and mean μ is agent-optimal on x^* and satisfies $\bar{U}_p(x^*) = U_p(x^*, F^*)$.*

Lastly, we prove the following claim made in the paragraph following Assumption 3.

Proposition A.7. *Let Assumptions 1 and 2 hold.*

(1) For all $\varepsilon > 0$, if $K(\bar{F})$ is sufficiently close to 0, then there is a one-sided cutoff mechanism x such that $\bar{U}_p(x) \geq \mu u_p(1) - \varepsilon$.

(2) If $\mu = u_p^{-1}(0)$, then there is a one-sided cutoff mechanism x such that $\bar{U}_p(x) > 0$.

Proof of Proposition A.7. Beginning with claim (1), consider the one-sided cutoff mechanism x with cutoff 1. On x, the agent picks F to maximize $F(\{1\})u_a(1) - K(F)$, where $F(\{1\})$ denotes the probability assigned to $\{1\}$. If F is agent-optimal, a lower bound on $U_a(x, F)$ is $U_a(x, \bar{F}) = \mu u_a(1) - K(\bar{F})$. Since all type distributions are MPCs of $\bar{F}$, and since K is decreasing with respect to MPCs, we have $0 \leq K(F) \leq K(\bar{F})$ for all F. It follows that if $K(\bar{F})$ is sufficiently small then all agent-optimal F satisfy $F(\{1\}) \geq \mu - \frac{\varepsilon}{u_p(1)}$. Hence $U_p(x, F) \geq F(\{1\})u_p(1) \geq \mu u_p(1) - \varepsilon$.

Now consider (2). For a number $\varepsilon > 0$ to be chosen in a moment, and consider the binary distribution F_ε that assigns probability $1/2$ to each of the points $\mu - \varepsilon$ and $\mu + \varepsilon$. Let δ_μ denote the degenerate distribution on μ. Consider the one-sided cutoff mechanism x_ε with cutoff $\mu + \varepsilon$. Among distributions which assign no probability to points above $\mu + \varepsilon$, on x_ε the distribution δ_μ is optimal for the agent (since it minimizes acquisition costs). If ε is sufficiently small, then continuity of u_a and K imply $U_a(x_\varepsilon, F_\varepsilon) > K(\delta_\mu)$. Fix such a number ε. Since $K \geq 0$, it follows that all agent-optimal distributions on x_ε assign non-zero probability to points above $\mu + \varepsilon$. The principal's payoff from accepting at these points is strictly positive and bounded away from 0. □

A.1.2 Proof of Lemma 4.1

Proof of Lemma 4.1. According to Lemma A.2, the set $\mathcal{F}^*(x)$ is non-empty and compact. Since $U_p(x, \cdot)$ is continuous on $\mathcal{F}^*(x)$ (Lemma A.1), we conclude that $U_p(x, F) = \bar{U}_p(x)$ holds for some $F \in \mathcal{F}^*(x)$. In view of Lemmata A.1 and A.2, there exists $F^* \in \mathcal{F}^*_B(x)$ such that $\bar{U}_p(x) = U_p(x, F) = U_p(x, F^*)$. □

A.1.3 Proof of Lemma 4.2

Proof of Lemma 4.2. Recall that our candidate two-sided cutoff mechanism x^* is defined as follows:

$$\forall_{\theta \in [0,1]}, \quad x^*(\theta) = \begin{cases} x(\theta_0^*), & \text{if } \theta \leq \theta_0^* \\ 0, & \text{if } \theta_0^* < \theta < \theta_1^* \\ x(\theta_1^*), & \text{if } \theta_1^* \leq \theta. \end{cases} \tag{A.6}$$

Lemma A.3 implies $\theta_1^* > u_p^{-1}(0) \geq \mu > \theta_0^*$ and $x(\theta_1^*) > 0$. By assumption of the lemma we trying to prove, we have $\bar{U}_p(x) = U_p(x, F^*)$. Note that x and x^* agree on the support of F^*, implying $U_p(x, F^*) = U_p(x^*, F^*)$. Hence it suffices to show $\bar{U}_p(x^*) \geq U_p(x^*, F^*)$.

Let us invoke Lemma 4.1 to find an agent-optimal distribution F on x^*. We distinguish two cases.

First, let $x(\theta_0^*) > 0$.

In a first subcase, suppose F assigns probability 0 to (θ_0^*, θ_1^*). Since $\{\theta_0^*, \theta_1^*\}$ is the support of F^*, it follows that F is a mean-preserving spread of F^*. Hence $0 < U_p(x, F^*) = U_p(x^*, F^*)$ and Lemma A.4 imply $U_p(x^*, F^*) \leq U_p(x^*, F)$. Since F was agent-optimal on x^*, we conclude $U_p(x^*, F^*) \leq \bar{U}_p(x^*)$ and we are done.

In a second subcase, suppose F assigns non-zero probability to (θ_0^*, θ_1^*). Since $(\theta_0^*, \theta_1^*) = \{\theta \in [0,1] \colon x^*(\theta) = 0\}$, Lemma A.5 implies that F is supported on a subset of $[\theta_0^*, \theta_1^*]$. In particular, we have $x^* \leq x$ on the support of F and F^*. Recall also that F is agent-optimal on x^*, and that F^* is agent-optimal on x. Hence

$$U_a(x, F) \leq U_a(x, F^*) \leq U_a(x^*, F^*) \leq U_a(x^*, F).$$

Since $x^* \leq x$ holds on the support of F and F^*, we conclude that the inequalities in the previous display are equalities. In particular, we conclude that F^* is agent-optimal on x^*. Hence $U_p(x^*, F^*) \leq \bar{U}_p(x^*)$ and we are done.

At this point we have completed the proof for the case $x(\theta_0^*) > 0$. Now let $x(\theta_0^*) = 0$. Since $\mathbb{E}_F[\theta] = \mu < \theta_1^*$, it follows that F assigns non-zero probability to $[0, \theta_1^*)$. Since $[0, \theta_1^*) = \{\theta \in [0,1] \colon x^*(\theta) = 0\}$, we conclude from Lemma A.5 that F is supported on a subset of $[0, \theta_1^*]$. In particular, the mechanism x^* is weakly below x at all points in the support of F. From here the proof can be completed as above. □

A.1.4 Proof of Theorem 4.3

Proof of Theorem 4.3. We first prove abstractly the existence of an optimal two-sided cutoff mechanism.

Claim A.8. *There exists a two-sided cutoff mechanism x^* such that all mechanisms x satisfy $\bar{U}_p(x) \leq \bar{U}_p(x^*)$*

Proof of Claim A.8. Let P denote the set of possible parameters of two-sided cutoff mechanisms; that is P is the set of vectors $(\theta_0, \theta_1, \alpha, \beta, \gamma)$ such that $0 \leq \theta_0 \leq \theta_1 \leq 1$ and $0 \leq \beta \leq \alpha \leq 1$ and $\beta \leq \gamma \leq 1$. Given $\rho \in \mathrm{P}$, we denote the associated two-sided cutoff mechanisms by x_ρ. Given $\rho = (\theta_0, \theta_1, \alpha, \beta, \gamma)$, let F_ρ denote the binary distribution with support $\{\theta_0, \theta_1\}$ and mean μ, whenever this distribution is well-defined (which is whenever $\mu \in [\theta_0, \theta_1]$).

According to Corollary A.6, for all mechanisms x such that $\bar{U}_p(x) > 0$, there exists $\rho \in \mathrm{P}$ such that $\bar{U}_p(x_\rho) \geq \bar{U}_p(x)$. Since a mechanism x such that $\bar{U}_p(x) > 0$ is assumed to exist, it suffices to show that $\rho \mapsto \bar{U}_p(x_\rho)$ admits a maximizer over P. To that end, let $M = \sup_{\rho \in \mathrm{P}} \bar{U}_p(x_\rho)$. Let $\{\rho_n\}_n$ be a sequence in P such that $\bar{U}_p(x_{\rho_n}) \to M$ as $n \to \infty$. By invoking Corollary A.6, we may assume that for all n the binary distribution F_{ρ_n} is well-defined, agent-optimal on x_{ρ_n}, and such that $U_p(x_{\rho_n}, F_{\rho_n}) = \bar{U}_p(x_{\rho_n})$ holds. By possibly passing to a convergent subsequence, we may assume that $\{\rho_n\}_n$ converges to a point $\rho \in \mathrm{P}$.

A moment's thought reveals that as $n \to \infty$ we have $U_p(x_{\rho_n}, F_{\rho_n}) \to U_p(x_\rho, F_\rho)$. Since $U_p(x_{\rho_n}, F_{\rho_n}) = \bar{U}_p(x_{\rho_n})$ holds for all n, we infer $M = U_p(x_\rho, F_\rho)$. To complete the proof, it thus suffices to argue that F_ρ is agent-optimal on x_ρ.

To that end, let $\tilde{F} \in \mathcal{F}$ be an arbitrary binary distribution. Since there always exists an agent-optimal distribution that is binary, it suffices to show $U_a(x_\rho, \tilde{F}) \leq U_a(x_\rho, F_\rho)$. Let $\{\tilde{\theta}_0, \tilde{\theta}_1\}$ denote the support of $\tilde{F}$. We shall use the following easily established observation: Fixing an arbitrary $\theta \in [0,1]$, the probability $x_{\rho_n}(\theta)$ fails to converge to $x_\rho(\theta)$ only if θ is a discontinuity point of x_ρ and in the interior of $(0,1)$; that is, only if $\theta \in \{\theta_0, \theta_1\} \cap (0,1)$. Now, for a number $\varepsilon > 0$ to be chosen in a moment, let $\tilde{\theta}_{0,\varepsilon} = \max(\tilde{\theta}_0 - \varepsilon, 0)$ and $\tilde{\theta}_{1,\varepsilon} = \min(\tilde{\theta}_1 + \varepsilon, 1)$. Let $\tilde{F}_\varepsilon$ denote the binary distribution supported on $\{\tilde{\theta}_{0,\varepsilon}, \tilde{\theta}_{1,\varepsilon}\}$ and having mean μ. By the previous observation, as $n \to \infty$ we have $U_a(x_{\rho_n}, \tilde{F}_\varepsilon) \to U_a(x_\rho, \tilde{F}_\varepsilon)$. On the other hand, as $\varepsilon \to 0$, we have $U_a(x_\rho, \tilde{F}_\varepsilon) \to U_a(x_\rho, \tilde{F})$. Agent-optimality of F_{ρ_n} on x_{ρ_n} implies $U_a(x_{\rho_n}, \tilde{F}_\varepsilon) \leq U_a(x_{\rho_n}, F_{\rho_n})$. We clearly have $U_a(x_{\rho_n}, F_{\rho_n}) \to U_a(x_\rho, F_\rho)$. In summary, we find $U_a(x_\rho, \tilde{F}) \leq U_a(x_\rho, F_\rho)$, as promised. □

To complete the proof, we show that there exists an optimal two-sided cutoff mechanism x^* whose parameters $(\theta_0^*, \theta_1^*, \alpha^*, \beta^*, \gamma^*)$ are such that $\beta^* = 0$, $\gamma^* = 1$, and $\theta_0^* < \mu \leq u_p^{-1}(0) < \theta_1^*$.

Let x be a two-sided cutoff mechanism as in the conclusion of Claim A.8. By appealing to Corollary A.6 we may assume that the parameters of x are $(\theta_0, \theta_1, \alpha, 0, \gamma)$, where $\theta_0 < \mu \leq u_p^{-1}(0) < \theta_1$ and $\gamma > 0$. Moreover, the binary distribution $F \in \mathcal{F}$ with support $\{\theta_0, \theta_1\}$ is agent-optimal on x and satisfies $U_p(x, F) = \bar{U}_p(x)$.

To complete the proof, it suffices to find a two-sided cutoff mechanism x^* such that $\bar{U}_p(x^*) \geq \bar{U}_p(x)$, and such that the acceptance probability on the right-most interval is 1. There is nothing to show if $\gamma = 1$, so let $\gamma < 1$. Let x^* denote the two-sided cutoff mechanism obtained from x by raising the acceptance probability on the right-most subinterval from γ to 1 (and leaving all other acceptance probabilities unchanged). Note that $u_p^{-1}(0) \leq \theta_1$ implies $U_p(x, F) \leq U_p(x^*, F)$.

Let us fix an arbitrary agent-optimal binary distribution F^* on x^* (existence being guaranteed by Lemma 4.1). In view of $\bar{U}_p(x) = U_p(x, F) \leq U_p(x^*, F)$, we may complete the proof by showing $U_p(x^*, F^*) \geq U_p(x^*, F)$. To show that this inequality holds, it is in turn sufficient to show that F^* is an MPS of F (for then the desired inequality follows from Lemma A.4). To that end, let the support of F^* be denoted $\{\theta_0^*, \theta_1^*\}$, where $\theta_0^* \leq \theta_1^*$. Let f^* denote the probability that F^* assigns to θ_1^*.

Recall that F is agent-optimal on x. Note that the agent's utility from F strictly increases when passing from x to x^* since F assigns non-zero probability to θ_1. Hence, for F^* to be agent-optimal on x^*, we must have $\theta_1^* \geq \theta_1$ and $f^* > 0$. Since $\theta_1^* \geq \theta_1 > \mu$, we thus also have $\theta_0^* < \theta_1$ and $f^* < 1$.

Note that if $\theta_0^* \leq \theta_0$, then $\theta_1^* \geq \theta_1$ implies that F^* is a mean-preserving spread of F, and we are done. We complete the proof by showing that a contradiction obtains if $\theta_0^* > \theta_0$. Let $\theta_0^* > \theta_0$.

Since $\theta_0^* \in (\theta_0, \theta_1)$, it follows from Lemma A.5 that F^* is supported on $\{\theta_0^*, \theta_1\}$. Now, agent-optimality of F^* on x^* and agent-optimality of F on x together imply $U_a(x^*, F^*) + U_a(x, F) \geq U_a(x^*, F) + U_a(x, F^*)$. This inequality rearranges to $\mathbb{E}_{F^*}[(x^*(\theta) - x(\theta))u_a(\theta)] \geq \mathbb{E}_F[(x^*(\theta) - x(\theta))u_a(\theta)]$. The mechanisms x^* and x differ only at points above θ_1. Both F^* and F have θ_1 as their largest realization. Hence the previous inequality simplifies to $f^*(1-\gamma)u_a(\theta_1) \geq f(1-\gamma)u_a(\theta_1)$. Since $\gamma < 1$ and $u_a > 0$, we conclude $f^* \geq f$. Now, since F and F^* both have θ_1 as their largest realization and mean μ, the inequality $f^* \geq f$ implies that the smallest realization of F^*, namely $\theta_0^*\}$, must be weakly than the smallest realization of F, namely θ_0. This contradicts $\theta_0^* > \theta_0$. □

A.2 Omitted proofs for Section 5

A.2.1 Equilibrium definition

We denote the set of Borel probability measures on $[0,1]$ by $\Delta[0,1]$. Recall that $\mathcal{E} = \{[0,1]\} \cup (\bigcup_{\theta \in [0,1]}\{\{\theta\}\}$ denotes the set of all pieces of evidence that the agent can conceivably possess.

A strategy of the agent specifies a type distribution $F \in \mathcal{F}$ and a Borel-measurable function $\sigma_A \colon [0,1] \to [0,1]$. The strategy of the principal specifies a probability $\sigma_P([0,1])$ and a Borel-measurable function $\sigma_P \colon [0,1] \to [0,1]$. The principal's beliefs are given by a function $\beta \colon \mathcal{E} \to \Delta[0,1]$, where for all Borel-subsets B of $[0,1]$ the mapping $\theta \mapsto \beta(B|\theta)$ is Borel-measurable.

Definition 2. Let $(F, \sigma_A, \sigma_P, \beta)$ be a tuple consisting of a strategy of the agent, and a strategy and beliefs of the principal. The tuple $(F, \sigma_A, \sigma_P, \beta)$ is a **(perfect Bayesian) equilibrium** of the evidence-disclosure game if all of the following hold.

(1) The cdf F satisfies

$$F \in \arg\max_{F' \in \mathcal{F}} \left(\mathbb{E}_{F'}[(\sigma_A(F,\theta)\sigma_P(\theta) + (1 - \sigma_A(F,\theta))\sigma_P([0,1]))u_a(\theta)] - K(F')\right).$$

(2) For all $\theta \in [0,1]$ the strategy σ_A satisfies

$$\sigma_A(\theta) \in \arg\max_{y \in [0,1]} \left((y\sigma_P(\theta) + (1-y)\sigma_P([0,1]))u_a(\theta)\right).$$

(3) For all $e \in \mathcal{E}$ the strategy σ_P satisfies $\sigma_P(e) \in \arg\max_{x \in [0,1]} x\mathbb{E}_{\beta(\cdot|e)}[u_p(\theta)]$.

(4) For all Borel-subsets B of $[0,1]$ the beliefs β satisfy

$$\begin{aligned}&\int_{\theta\in[0,1]} \mathbf{1}_{(\theta\in B)}\, dF(\theta)\\ =&\int_{\theta\in[0,1]} \left(\beta(B|\theta)\sigma_A(\theta)+\beta(B|[0,1])(1-\sigma_A(\theta))\right)\, dF(\theta).\end{aligned}$$

(5) For all $\theta\in\operatorname{supp}F$, if $\sigma_A(\theta)>0$, then $\beta(\cdot|\theta)$ is the Dirac measure on $\{\theta\}$.

Conditions (1) and (2) are that the agent plays a best response to the principal's strategy. Condition (3) is that the principal plays a best response to the agent's strategy and given the principal's own beliefs. Condition (4) states that the principal's belief is consistent with Bayes' rule. Condition (5) states that if the agent discloses a type in the support of F that is sometimes disclosed along the equilibrium path, then the principal's belief is degenerate on the type. In a truth-leaning equilibrium, defined next, the principal's belief after disclosure is degenerate on the type even at all types, regardless of whether they are in the support of F or whether these types disclose.

Definition 3. An equilibrium $(F,\sigma_A,\sigma_P,\beta)$ is **truth-leaning** if for all $\theta\in[0,1]$ the probability measure $\beta(\cdot|\theta)$ is the Dirac measure on $\{\theta\}$.

A.2.2 Proof of Proposition 5.1

Proof of Proposition 5.1.

Claim A.9. *In all truth-leaning equilibria exactly one of the following is true:*

(1) The principal's equilibrium utility is 0.

(2) The agent picks a distribution F *satisfying* $\mathbb{E}_F[u_p(\theta)]>0$. *If the agent does not disclose the type, the principal accepts with probability* 1. *Along the path of play, the proposal is accepted with probability* 1.

Proof of Claim A.9. Let $(F,\sigma_A,\sigma_P,\beta)$ be a truth-leaning equilibrium. For all θ, let $x(\theta)=\sigma_A(\theta)\sigma_P(\theta)+(1-\sigma_A(\theta))\sigma_P([0,1])$ denote the induced acceptance probability.

Suppose the principal's utility is strictly positive, meaning $\mathbb{E}_F[x(\theta)u_p(\theta)]>0$. Note that this implies $\max\operatorname{supp}F>u_p^{-1}(0)$, and hence $\min\operatorname{supp}F<\mu\leq u_p^{-1}(0)$. We show that the agent never discloses the type, and that in equilibrium the proposal is always accepted.

Note that if the type realizes strictly above $u_p^{-1}(0)$, then the agent can always disclose it, following which the principal's truth-leaning beliefs and best response imply that the proposal will be accepted with probability 1. Hence we have $x(\theta) = 1$ for all $\theta \in (u_p^{-1}(0), 1]$.

Most of the remaining work shall go towards establishing that

$$\mathbb{E}_F[x(\theta)|\theta < u_p^{-1}(0)] = 1 \tag{A.7}$$

holds; note that the conditional expectation is well-defined since $\min \operatorname{supp} F < u_p^{-1}(0)$ holds. Before tackling the proof of this equality, we argue that it completes the proof of Claim A.9. The equality (A.7) implies that $x(\theta) = 1$ holds for F-almost all $\theta \in [0, u_p^{-1}(0)]$. Hence $\mathbb{E}_F[u_p(\theta)] = \mathbb{E}_F[x(\theta)u_p(\theta)] > 0$. At a type $\theta \in \operatorname{supp} F \cap [0, u_p^{-1}(0))$, the principal's truth-leaning beliefs imply that if θ discloses, then the principal rejects the proposal. Hence F-almost all types in $\operatorname{supp} F \cap [0, u_p^{-1}(0))$ do not disclose. Hence, when the agent does not disclose, the principal accepts with probability 1. It follows that the proposal is accepted with probability 1.

It remains to prove (A.7). Towards a contradiction, let

$$\mathbb{E}_F[x(\theta)|\theta < u_p^{-1}(0)] < 1.$$

Let $\theta_1 = \mathbb{E}_F[\theta|\theta > u_p^{-1}(0)]$, which is well-defined since $\max \operatorname{supp} F > u_p^{-1}(0)$ holds. Let p denote the probability F assigns to the interval $(u_p^{-1}(0), 1]$. Note that $\min \operatorname{supp} F < u_p^{-1}(0)$ implies $p < 1$. Let $\theta_0 = \mathbb{E}_F[\theta|\theta \leq u_p^{-1}(0)]$. Let $\tilde{F}$ denote the distribution that assigns probability p to θ_1, and with probability $1 - p$ agrees with the conditional distribution of F below $u_p^{-1}(0)$. [25] Given $\varepsilon > 0$ to be chosen in a moment, let $\eta_\varepsilon = \theta_1 - \frac{p\theta_1 + \varepsilon\theta_0}{p+\varepsilon}$. Let $\hat{F}$ be the distribution that assigns probability $p + \varepsilon$ to $\theta_1 - \eta_\varepsilon$, and with probability $1 - p - \varepsilon$ agrees with the conditional distribution of $\tilde{F}$ below $u_p^{-1}(0)$. If ε is sufficiently small, then $\theta_1 - \eta_\varepsilon > u_p^{-1}(0)$ and $\hat{F}$ is a well-defined cdf and an MPC of $\tilde{F}$.

Note that $\mathbb{E}_F[x(\theta)u_a(\theta)] - K(F) \leq \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)] - K(\tilde{F})$ holds since x is constant on $(u_p^{-1}(0), 1]$. We shall obtain the desired contradiction by arguing that $\mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)] - K(\tilde{F}) < \mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] - K(\hat{F})$ holds. Since K decreases with respect to MPCs, we have $K(\hat{F}) \leq K(\tilde{F})$. Thus it suffices to show $\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] -$

[25] That is, for all reals θ let $\tilde{F}(\theta) = p\mathbf{1}_{(\theta \geq \theta_1)} + \min(1 - p, F(\theta))$.

$\mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)] > 0$.

For future reference, note that since $u_a > 0$, the inequality $\mathbb{E}_F[x(\theta)|\theta < u_p^{-1}(0)] < 1$ implies

$$\mathbb{E}_F[x(\theta)u_a(\theta)|\theta < u_p^{-1}(0)] < \mathbb{E}_F[u_a(\theta)|\theta < u_p^{-1}(0)].$$

The previous inequality in turn implies

$$\mathbb{E}_F[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)] < \mathbb{E}_F[u_a(\theta)|\theta \leq u_p^{-1}(0)]. \tag{A.8}$$

We now show $\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] - \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)] > 0$. Since x is constantly equal to 1 strictly above $u_p^{-1}(0)$, we have

$$\begin{aligned}&\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] - \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)]\\ =&(p+\varepsilon)u_a(\theta_1 - \eta_\varepsilon) + (1-p-\varepsilon)\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)]\\ -&pu_a(\theta_1) - (1-p)\mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)].\end{aligned}$$

Note that the distributions of $\tilde{F}$ and $\hat{F}$ agree conditional on realizations below $u_p^{-1}(0)$. Hence the above term simplifies to

$$(p+\varepsilon)u_a(\theta_1 - \eta_\varepsilon) - pu_a(\theta_1) - \varepsilon\mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)]. \tag{A.9}$$

By concavity of u_a and the choice of η_ε, a lower bound on this difference is

$$\begin{aligned}&(p+\varepsilon)\left(\frac{p}{p+\varepsilon}u_a(\theta_1) + \frac{\varepsilon}{p+\varepsilon}u_a(\theta_0)\right) - pu_a(\theta_1) - \varepsilon\mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)]\\ =&\varepsilon\left(u_a(\theta_0) - \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)]\right).\end{aligned} \tag{A.10}$$

We next invoke (A.8) to infer

$$u_a(\theta_0) - \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)|\theta \leq u_p^{-1}(0)] > u_a(\theta_0) - \mathbb{E}_{\tilde{F}}[u_a(\theta)|\theta \leq u_p^{-1}(0)]. \tag{A.11}$$

By concavity of u_a and the definition of θ_0, we have $u_a(\theta_0) \geq \mathbb{E}_{\tilde{F}}[u_a(\theta)|\theta \leq u_p^{-1}(0)]$. Collecting our work, we conclude that $\mathbb{E}_{\hat{F}}[x(\theta)u_a(\theta)] > \mathbb{E}_{\tilde{F}}[x(\theta)u_a(\theta)]$ holds, as promised. □

Claim A.10. *There exists a truth-leaning equilibrium in which the principal's utility is* 0*.*

Proof of Claim A.10. Below, we will distinguish between two cases, and define the equilibrium differently depending on the case. In both cases, however, we have the following: If the agent disclose the type realization, the principal accepts if and only if revealed type is weakly above $u_p^{-1}(0)$. If the agent discloses no evidence, the principal rejects. The agent always discloses the type.

Let $x\colon [0,1] \to [0,1]$ denote the acceptance probability induced by these strategies and when the agent always disclses the type (this is nothing but the one-sided cutoff mechanism with cutoff $u_p^{-1}(0)$). Let $F \in \arg\max_{\hat{F}\in\mathcal{F}} U(x, \hat{F})$. We distinguish two cases.

First, suppose $\max\operatorname{supp} F < u_p^{-1}(0)$. In this case, using the fact that K decreases with respect to MPCs, the degenerate distribution on μ must also be in $\arg\max_{\hat{F}\in\mathcal{F}} U_a(x, \hat{F})$. We now complete the description of equilibrium as follows: The agent acquires the degenerate distribution on μ, and if the agent does not disclose the type the principal's belief is degenerate on μ. One may verify that this yields a truth-leaning equilibrium in which the principal's utility is 0.

Second, suppose $\max\operatorname{supp} F \geq u_p^{-1}(0)$. By retracting the arguments that established Claim A.9, one may verify that there is a cdf $\tilde{F}$ in $\arg\max_{\hat{F}\in\mathcal{F}} U_a(x, \hat{F})$ such that $\max\operatorname{supp} \tilde{F} = u_p^{-1}(0)$ (but these arguments cannot be used to infer $\max\operatorname{supp} F = u_p^{-1}(0)$). We now complete the description of equilibrium as follows: The agent acquires the distribution $\tilde{F}$, and if the agent does not disclose the type the principal's belief is that the type is distributed according to $\tilde{F}$. Since $\max\operatorname{supp} \tilde{F} = u_p^{-1}(0)$, we necessarily have $\mathbb{E}_{\tilde{F}}[u_p(\theta)] \leq 0$, and hence it is rational for the principal to reject if the agent does not disclose. Hence the agent has a best response of always disclosing. Lastly, since $\max\operatorname{supp} \tilde{F} = u_p^{-1}(0)$, it is also clear that the principal's utility is 0. □

□

A.2.3 Proof of Proposition 5.3

Proof of Proposition 5.3. Let $h\colon [0,1] \to \mathbb{R}$ be a function that is convex, continuous (at the boundary), strictly increasing, and such that $h(0) \leq \frac{4}{3}$ and $h(\frac{1}{2}) = 0$ and $h(1) = \frac{4}{3}$. Fixing such a function h, we next define our candidate environment (in case the reader prefer a specific example, let h be defined for all θ by $h(\theta) = \frac{16}{9}\left(\theta^2 - \frac{1}{4}\right)$). Let t_0 solve $h(t_0) = 1$.

Now let $\mu = 1/2$. For all θ, let

$$u_p(\theta) = \max\left(h(\theta) - 1, \frac{1}{3}\left(\frac{\theta}{t_0} - 1\right)\right)$$

and $k(\theta) = \max(h(\theta), 0)$ and $u_a(\theta) = 1$. For all F, let $K(F) = \mathbb{E}_F[k(\theta)]$.

We omit the straightforward verification that Assumptions 1 and 2 hold. For later reference, note that all $F \in \mathcal{F}$ satisfy

$$\mathbb{E}_F[u_p(\theta)] \leq \mathbb{E}_{\bar{F}}[u_p(\theta)] = \frac{h(1) - 1}{2} - \frac{1}{2}\frac{1}{3} = 0. \tag{A.12}$$

We also note that $u_p^{-1}(0) = t_0$ holds, and that all $\theta \in [u_p^{-1}(0), 1]$ satisfy $u_p(\theta) = -(u_a(\theta) - k(\theta))$. Lastly, for all mechanisms x the degenerate distribution δ_μ on μ satisfies $U_a(x, \delta_\mu) \geq 0$.

The next claim shows that Assumption 3 holds.

Claim A.11. *The two-sided mechanism x with parameters $(\theta_0, \theta_1, \alpha, \beta, \gamma) = (0, 1, \frac{1}{2}, 0, 1)$ satisfies $\bar{U}_p(x) > 0$.*

Proof of Claim A.11. Direct computation shows $U_p(x, \bar{F}) = \frac{1}{2}\left(h(1) - 1 - \frac{1}{2}\frac{1}{3}\right) = \frac{1}{4}\frac{1}{3}$. Hence we can show $\bar{U}_p(x) > 0$ by showing that $\bar{F}$ is uniquely agent-optimal on x. Let $v(\theta) = \frac{1}{2}\mathbf{1}_{(\theta=0)} + \mathbf{1}_{(\theta=1)} - \max(h(\theta), 0)$. The agent's utility from a type distribution F is $\mathbb{E}_F[x(\theta)u_a(\theta) - k(\theta)] = \mathbb{E}_F[v(\theta)]$. One may verify that v is weakly below the function $\hat{v}$ defined by $\hat{v}(\theta) = \frac{1}{2} - \left(\frac{1}{3} - \frac{1}{2}\right)\theta$. (The function $\hat{v}$ is the concave closure of v.) In fact, note that v is strictly below $\hat{v}$ at all points in $[0, 1]$ except 0 and 1. Hence $\bar{F}$ is the unique agent-optimal distribution on x. □

We next prove the following auxiliary claim.

Claim A.12. *Let x be a mechanism, let F be binary and agent-optimal on x. If all $\theta \in [0, u_p^{-1}(0))$ satisfy $x(\theta) \in \{0, 1\}$, then $U_p(x, F) \leq 0$.*

Proof of Claim A.12. Let the support of F be $\{\theta_0, \theta_1\}$, where $\theta_0 \leq \theta_1$. If $\theta_1 \leq u_p^{-1}(0)$ or F is degenerate, then clearly $U_p(x, F) \leq 0$. Thus let $\theta_1 > u_p^{-1}(0)$ and let F be non-degenerate. Since $\mu \leq u_p^{-1}(0)$, we have $\theta_0 < u_p^{-1}(0)$, and hence $x(\theta_0) \in \{0, 1\}$. If $x(\theta_0) = 1$, then $\theta_1 \geq u_p^{-1}(0)$ and (A.12) together imply $U_p(x, F) = \mathbb{E}_F[x(\theta)u_p(\theta)] \leq$

$\mathbb{E}_F[u_p(\theta)] \leq 0$. Thus let $x(\theta_0) = 0$. Recall that $u_p = -(u_a - k)$ holds at points above $u_p^{-1}(0)$. Hence the agent's utility satisfies $U_a(x, F) = \mathbb{E}_F[x(\theta)u_a(\theta)] - K(F) \leq -\mathbb{E}_F[x(\theta)u_p(\theta)]$. Hence the fact that the agent weakly prefers F to the degenerate distribution on μ implies $V(x, F) = \mathbb{E}_F[x(\theta)u_p(\theta)] \leq 0$. □

The next claim concerns part (2) of the proposition.

Claim A.13. *Let x be a mechanismm, and let F be agent-optimal on x. Let $T = \{\theta \in [0, u_p^{-1}(0))\colon x(\theta) \in (0,1)\}$. If $U_p(x, F) > 0$, then T has non-zero F-probability.*

Proof of Claim A.13. We show the contrapositive: If T has F-probability 0, then $U_p(x, F) \leq 0$. Consider the mechanism y defined for all θ by

$$y(\theta) = \begin{cases} 0, & \text{if } \theta \in T \\ x(\theta), & \text{else.} \end{cases}$$

(The fact that x is usc implies that y is usc, and hence y is indeed a mechanism.) Since T has F-probability 0, we have $U_a(x, F) = U_a(y, F)$ and $U_p(x, F) = U_p(y, F)$. We clearly have $x \geq y$. It follows that F is agent-optimal on y, and that we may complete the proof by showing $U_p(y, F) \leq 0$.

Let us denote by $\mathcal{F}_B^*(y)$ the set of binary distributions in $\mathcal{F}$ that are agent-optimal on y. Lemma A.1 tells us that $U_p(x, \cdot)$ is continuous on $\mathcal{F}^*(y)$. Hence Lemma A.2 implies that there exists a probability measure ν supported on $\mathcal{F}_B^*(y)$ such that

$$U_p(y, F) = \int_{\tilde{F} \in \mathcal{F}_B^*(y)} U_p(y, \tilde{F})\, d\nu(\tilde{F}).$$

Now let $\tilde{F} \in \mathcal{F}_B^*(y)$. Since all $\theta \in [0, u_p^{-1}(0))$ satisfy $y(\theta) \in \{0, 1\}$, we infer from Claim A.12 that $U_p(y, \tilde{F}) \leq 0$ holds. Hence $U_p(y, F) \leq 0$, as promised. □

Lastly, we consider the evidence-disclosure game (part (1) of the proposition).

Claim A.14. *In all equilibria of the evidence-disclosure game, the principal's utility is* 0.

Proof of Claim A.14. Let (F, σ_A, σ_P) be an equilibrium. For all θ, let

$$x(\theta) = \sigma_A(\theta)\sigma_P(\theta) + (1 - \sigma_A(\theta))\sigma_P([0, 1])$$

denote the induced acceptance probability at type θ. The principal's utility cannot be strictly negative as the principal can always reject for a sure payoff of 0. Towards a contradiction, let the principal's utility be strictly positive; that is, let $\mathbb{E}_F[x(\theta)u_p(\theta)] > 0$. This requires F to be non-degenerate (since $u_p(\mu) \leq 0$).

Let $S = \{\theta \in [0, u_p^{-1}(0)) \cap \operatorname{supp} F \colon 0 < x(\theta)\}$. As an intermediate step, we claim that S has non-zero F-probability. Towards a contradiction, suppose not. Then F-almost all θ in $[0, u_p^{-1}(0))$ satisfy $x(\theta) = 0$. In an intermediate step, we claim that all of the following hold:

$$\begin{aligned} 0 \leq& \mathbb{E}_F[x(\theta)u_a(\theta) - k(\theta)] \\ \leq& \mathbb{E}_F[u_a(\theta)\mathbf{1}_{(u_p^{-1}(0)\leq\theta)} - k(\theta)] = \mathbb{E}_F[-u_p(\theta)\mathbf{1}_{(u_p^{-1}(0)\leq\theta)}]. \end{aligned}$$

The agent's equilibrium utility $\mathbb{E}_F[x(\theta)u_a(\theta) - k(\theta)]$ is at least 0 since the agent can always acquire the degenerate distribution (which has a cost of 0 in this environement); the second inequality follows from $u_a > 0$ and the assumption that F-almost all θ in $[0, u_p^{-1}(0))$ satisfy $x(\theta) = 0$; the third is from inspection of u_a and k. We conclude from the previous display that $\mathbb{E}_F[u_p(\theta)\mathbf{1}_{(u_p^{-1}(0)\leq\theta)}] \leq 0$ holds. Using again that S has F-probability 0, we have $\mathbb{E}_F[x(\theta)u_p(\theta)] = \mathbb{E}_F[u_p(\theta)\mathbf{1}_{(u_p^{-1}(0)\leq\theta)}] \leq 0$. Thus the principal's utility is at most 0; contradiction.

Now consider a type θ in S. If this type discloses with non-zero probability in equilibrium, then, since θ is in the support of F, the principal's beliefs are degenerate on θ. Since the type is less strictly than $u_p^{-1}(0)$, disclosing the type leads to the principal rejecting. But by definition of S we have $x(\theta) > 0$. Hence type θ must strictly prefer not to disclose. Thus all types in S do not disclose.

Now let $\alpha = \sigma_P([0,1])$ denote the probability that the principal accepts if the agent does not disclose. Since no type in S discloses, we have $\alpha = x(\theta) > 0$ for all $\theta \in S$. Note that since all types can choose not to disclose we must have $S = \operatorname{supp} F \cap [0, u_p^{-1}(0))$.

At this point we know $\alpha > 0$. We next argue that $\alpha = 1$ and $\alpha \in (0,1)$ both imply a contradiction, completing the proof.

First, let $\alpha = 1$. Since all types can choose not to disclose, it follows that $x(\theta) = 1$ holds for all F-almost all types. Hence the principal's equilibrium utility satisfies $\mathbb{E}_F[x(\theta)u_p(\theta)] = \mathbb{E}_F[u_p(\theta)] \leq \mathbb{E}_{\bar{F}}[u_p(\theta)] = \frac{1}{2}\left(h(1) - 1\right) - \frac{1}{2}\frac{1}{3} = 0$; this contradicts the assumption that the principal's utility is strictly positive.

Second, let $\alpha \in (0,1)$. Let $G = \beta(\cdot|[0,1])$ denote the principal's belief about the type conditional on the agent not disclosing. Let μ_G denote the mean of G. For $\alpha \in (0,1)$ to be optimal for the principal, we must have $\mathbb{E}_G[u_p(\theta)] = 0$. Note that F would be the principal's belief after non-disclosure if F-almost all types did not disclose. We already know that all types in $S = \operatorname{supp} F \cap [0, u_p^{-1}(0))$ do not disclose. Hence $\mu_G \leq \mu_F$, with equality if and only if F-almost all types do not disclose. Let $\bar{G}$ denote the binary distribution that assigns probability μ_G to 1, and probability $1 - \mu_G$ to 0. We now distinguish two cases. First, let $\mu_G = \mu_F$. Then, since F-almost all types do not disclose, the principal's equilibrium utility is $\alpha\mathbb{E}_F[u_p(\theta)] \leq 0$; contradiction. Second, let $\mu_G < \mu_F$. Then we have $\mathbb{E}_G[u_p(\mu)] \leq \mathbb{E}_{\bar{G}}[u_p(\mu)] < \mathbb{E}_{\bar{F}}[u_p(\theta)] \leq 0$ since u_p is convex and since $u_p(0) < u_p(1)$. But $\mathbb{E}_G[u_p(\mu)] < 0$ contradicts the fact that it is optimal for the principal to accept with probability $\alpha \in (0,1)$ when the agent does not disclose. □

Claims A.11, A.13 and A.14 together complete the proof. □

Appendix B Supplementary material

B.1 Costly verification

This part of the appendix discusses optimal mechanisms for the model with costly verifcation from Section 6.2.

Recall that $c\colon [0,1] \to \mathbb{R}_+$ denotes the principal's cost of verifying the agent. Throughout this appendix, we maintain the following assumption in place of Assumption 1.

Assumption 4. The functions $u_p - c$ and u_p are convex, continuous, and satisfy the following: There is a point s_0 in $[\mu, 1)$ such that $u_p - c$ is strictly negative on $[0, s_0)$, and strictly positive on $(s_0, 1]$.

B.1.1 Model

A mechanism $\mathcal{M}$ is now a tuple (M, a, p, q) consisting of a metric space M of messages, and functions $a\colon M \to [0,1]$, $p\colon M \to [0,1]$ and $q\colon M \times [0,1] \to [0,1]$. If the agent sends message m at type θ, then $a(m)$ is the probability that the principal *a*udits the agent, $p(m)$ is the acceptance probability conditional on not having audited, and $q(m,\theta)$ is the acceptance probability conditional on having audited. Let the acceptance probability $x_{\mathcal{M}}\colon M \times [0,1] \to [0,1]$ be defined for all m and θ by

$$x_{\mathcal{M}}(m,\theta) = a(m)q(m,\theta) + (1 - a(m))p(m).$$

A *strategy* (for reporting to the mechanism) is a Borel-measurable function $\sigma\colon [0,1] \to M$. A strategy is a *best-response (for reporting to the mechanism)* if

$$\forall_{\theta \in [0,1]}, \quad \sigma(\theta) \in \operatorname*{arg\,max}_{m \in M} x_{\mathcal{M}}(m,\theta).$$

We denote the set of best-responses by $\Sigma^*(\mathcal{M})$. For technical reasons, we restrict the set of allowed mechanisms by requiring that M is compact, that $x_{\mathcal{M}}$ is usc on $M \times [0,1]$, that a is Borel-measurable, and that there exists a best-response.[26]

The timing is now:

(1) The principal commits to a mechanism $\mathcal{M}$.

(2) The agent, knowing the mechanism, picks a type distribution F from $\mathcal{F}$.

(3) Nature draws the agent's type according to θ.

(4) The agent, knowing θ, reports to the mechanism.

(5) The mechanism possibly audits the agent, and then accepts or rejects the proposal.

All ties are broken in favour of the principal.

Given a mechanism $\mathcal{M}$, a type distribution F in $\mathcal{F}$, and a strategy σ, the expected utilities of the agent and the principal, respectively, are

$$U_a(\mathcal{M},F,\sigma) = \mathbb{E}_F\left[x_{\mathcal{M}}(\sigma(\theta),\theta)u_a(\theta)\right] - K(F) \tag{B.1a}$$

$$U_p(\mathcal{M},F,\sigma) = \mathbb{E}_F\left[x_{\mathcal{M}}(\sigma(\theta),\theta)u_p(\theta) - a(\sigma(\theta))c(\theta)\right]. \tag{B.1b}$$

A type distribution $\mathcal{F}$ is *agent-optimal* on $\mathcal{M}$ if for some (arbitrary) best response σ we have $F \in \operatorname{arg\,max}_{\tilde{F} \in \mathcal{F}} U_a(\mathcal{M},\tilde{F},\sigma)$. The set of agent-optimal distributions is

[26] That is, the arg max-correspondence should admit a Borel-measurable selection. One sufficient condition for this is that, in addition to the other restrictions, $x_{\mathcal{M}}$ be continuous in m (see Border and Aliprantis (2006, Theorem 18.19, p. 605)). In particular, a sufficient condition is that the mechanism have finitely many messages.

denoted $\mathcal{F}^*(\mathcal{M})$. The principal's utility $\bar{U}_p(\mathcal{M})$ from a mechanism $\mathcal{M}$ is

$$\bar{U}_p(\mathcal{M}) = \sup_{F \in \mathcal{F}^*(\mathcal{M}), \sigma \in \Sigma^*(\mathcal{M})} U_p(\mathcal{M}, F, \sigma).$$

We understand the non-triviality Assumption 3 to now refer to the newly-defined utility $\bar{U}_p$ of the principal.

Lemma B.1. *Let Assumptions 2 and 4 hold. For all mechanisms, there exists an agent-optimal distribution.*

Proof of Lemma B.1. All best-responses σ of the agent must induce the same acceptance-probability $\theta \mapsto x_{\mathcal{M}}(\sigma(\theta), \theta)$. Since M is compact and $x_{\mathcal{M}}$ is usc, Lemma 17.30 of Border and Aliprantis (2006, p. 569) the induced acceptance-probability usc in the type. Hence the agent's utility is usc in the type distribution. The claim follows from compactness of $\mathcal{F}$ and the Extreme Value theorem. □

B.1.2 Preliminaries

The first step in the analysis is to argue that there are again no limitations as to what the principal can implement, provided the principal is willing to incur sufficiently high costs. This follows earlier observations in the literature, see e.g. Ben-Porath et al. (2014). Formally, let us say a mechanism is *direct* if its message space is $[0, 1]$. A direct mechanism $\mathcal{M}$ is *incentive-compatible (IC)* if the agent cannot increase the acceptance-probability by misreporting; that is, for all θ and θ' in $[0, 1]$ we have $x_{\mathcal{M}}(\theta, \theta) \geq x_{\mathcal{M}}(\theta', \theta)$.

Lemma B.2. *For all mechanisms $\mathcal{M}$, there exists a direct IC mechanism $\tilde{\mathcal{M}} = ([0, 1], \tilde{a}, \tilde{p}, \tilde{q})$ satisfying all of the following:*

(1) We have $\bar{U}_p(\tilde{\mathcal{M}}) \geq \bar{U}_p(\mathcal{M})$.

(2) The function $\theta \mapsto x_{\tilde{\mathcal{M}}}(\theta, \theta)$ is usc.

(3) All $\theta \in [0, 1]$ satisfy $\tilde{a}(\theta) = x_{\tilde{\mathcal{M}}}(\theta, \theta) - \inf_{\theta' \in [0,1]} x_{\tilde{\mathcal{M}}}(\theta', \theta')$.

Conversely, for all usc functions $x\colon [0, 1] \to [0, 1]$, there is a direct IC mechanism $\tilde{\mathcal{M}}$ which induces x under truth-telling, and where $\tilde{a}$ and all $\theta \in [0, 1]$ satisfy $\tilde{a}(\theta) = x(\theta) - \inf_{\theta' \in [0,1]} x(\theta')$.

The proof follows below.

The relationship between $\tilde{a}$ and the acceptance probability $x_{\tilde{\mathcal{M}}}$ is suggested by the following heuristic: To incentivize truth-telling in the direct mechanism, the principal should maximally punish the agent when an audit finds the agent to have lied. Similarly, if auditing reveals that the agents has told the truth, then the principal should maximally reward the agent. This heuristic pins down $\tilde{a}$ via the equation

$$\inf_{\theta' \in [0,1]} x_{\tilde{\mathcal{M}}}(\theta', \theta') = x_{\tilde{\mathcal{M}}}(\theta, \theta) - \tilde{a}(\theta).$$

The right side is what the agent gets when *falsely* reporting a type θ. The audit probability is just high enough to make the worst-off types of the agent, who get the infimum on the left side, indifferent.

It is worth pointing out that the infimum in $\inf_{\theta' \in [0,1]} x_{\tilde{\mathcal{M}}}(\theta', \theta')$ is indeed taken over all θ in $[0,1]$. The agent may ultimately acquire a signal whose support is a strict subset of $[0,1]$. Hence the principal chooses the auditing rule so as to prevent *all* types θ from misreporting, including the types the principal does not expect to arise. The reason is simply that the auditing rule at these types nevertheless affects the agent's incentives to acquire information.

In view of Lemma B.2, we can take a mechanism to simply mean a usc function $x\colon [0,1] \to [0,1]$. Denoting $\underline{x} = \inf_{\theta \in [0,1]} x(\theta)$, the auditing rule is understood to be $a = x - \underline{x}$. The agent's and principal's expected utilities from x and F are given by

$$\begin{aligned} U_a(x, F) &= \mathbb{E}_F \left[x(\theta) u_a(\theta) \right] - K(F) \\ U_p(x, F) &= \mathbb{E}_F \left[x(\theta) u_p(\theta) - (x(\theta) - \underline{x}) c(\theta) \right]. \end{aligned}$$

The agent's utility is the same is in the model with evidence. The only change in the principal's utility are the additional costs.

We can now state an analogue of Lemma 4.1; the proof is analogous and omitted.

Lemma B.3. *Let Assumptions 2 and 4 hold. For all mechanisms x, there exists a binary distribution $F \in \mathcal{F}^*(x)$ such that $U_p(x, F) = \bar{U}_p(x)$.*

B.1.3 Optimality of two-sided cutoff mechanisms

Let x be a mechanism. Let F^* be a binary signal, and let its support be $\{\theta_0^*, \theta_1^*\}$. Our candidate two-sided cutoff mechanism x^* for improving on x is defined as follows.

$$\forall_{\theta\in[0,1]}, \quad x^*(\theta) = \begin{cases} x(\theta_0^*), & \text{if } \theta \leq \theta_0^* \\ \underset{\sim}{x}, & \text{if } \theta_0^* < \theta < \theta_1^* \\ x(\theta_1^*), & \text{if } \theta_1^* \leq \theta. \end{cases} \tag{B.2}$$

Lemma B.4. *Let Assumptions 2 and 4 hold. Let x be a mechanism. Let $F^* \in \mathcal{F}_B^*(x)$ satisfy $U_p(x, F) = \bar{U}_p(x)$. Let x^* be the two-sided cutoff mechanism defined in (B.2). If $U_p(x, F^*) > 0$, then $\bar{U}_p(x^*) \geq \bar{U}_p(x)$.*

Note there is difference to Lemma 4.2. The improving two-sided cutoff mechanism has an acceptance probability of $\underset{\sim}{x}$ on the middle subinterval; in Lemma 4.2, this probability is 0. Having the probability at $\underset{\sim}{x}$ ensures that x and x^* have the same infimum, which aids the comparison of the principal's utility under F^*. Apart from this detail, the proof from the evidence-model requires only minor modifications, and is therefore omitted.

Theorem B.5. *Let Assumptions 2 to 4 hold. There exists probabilities α^* and β^*, and θ_0^* and θ_1^* in $[0, 1]$ such that $\theta_0^* < \mu \leq s_0 < \theta_1^*$, and such that the two-sided mechanism x^* with parameters $(\theta_0^*, \theta_1^*, \alpha^*, \beta^*, 1)$ maximizes $\bar{U}_p$ over the set of mechanisms. Moreover, the binary distribution F^* with support $\{\theta_0^*, \theta_1^*\}$ and mean μ is agent-optimal on x^* and satisfies $\bar{U}_p(x^*) = U_p(x^*, F^*)$.*

The proof is analogous to the proof of Theorem 4.3 and is omitted. Note that we are not claiming that β^*, the acceptance-probability on the middle interval, is optimally set to 0. Intuitively, a high value of β^* lets the principal save on auditing costs. Conversely, if β^* is too high, then the agent will acquire the degenerate distribution on μ.

With a bit of work, one can show that if all $F \in \mathcal{F}$ satisfy $\mathbb{E}_F[u_p(\theta)] \leq 0$, then one can optimally set β^* to 0. To see this intuitively, recall that the principal's utility from a mechanism x and a distribution F is $\mathbb{E}_F\left[x(\theta)u_p(\theta) + (x(\theta) - \underset{\sim}{x})c(\theta)\right]$. Decreasing x everywhere by a constant does not affect the expectation $\mathbb{E}_F\left[(x(\theta) - \underset{\sim}{x})c(\theta)\right]$, but increases $\mathbb{E}_F\left[x(\theta)u_p(\theta)\right]$. This suggests that if x is a two-sided cutoff mechanism, then we can improve on x by passing to $x - \underline{x}$; the acceptance-probability of $x - \underline{x}$ on the middle interval is 0. The only complication in this argument is that changing x also distorts the agent's incentives when picking the type distribution. For the following result, we thus verify that such a distortion cannot make the principal worse off.

Lemma B.6. *Let all $F \in \mathcal{F}$ satisfy $\mathbb{E}_F[u_p(\theta)] \leq 0$ Let x be a two-sided cutoff mechanism with cutoffs $\{\theta_0, \theta_1\}$. Let F with support $\{\theta_0, \theta_1\}$ be agent-optimal on x and such that $\bar{U}_p(x) = U_p(x, F) > 0$ holds. Then $\bar{U}_p(x - \underline{x}) \geq \bar{U}_p(x)$.*

The proof follows below.

Using Lemma B.6, one may verify that if all $F \in \mathcal{F}$ satisfy $\mathbb{E}_F[u_p(\theta)] \leq 0$, then there is an optimal two-sided cutoff mechanism where the acceptance-probability on the middle (right-most) interval is 0 (is 1).

B.1.4 Proof of Lemma B.2

Proof of Lemma B.2. Consider the first part of the claim.

We begin by deriving a lower bound on the costs that the principal has to incur when the agent best responds. Let σ be an arbitrary best-response on $\mathcal{M}$. Let θ and θ' in $[0, 1]$ be arbitrary. We have the following (the first inequality follows from the optimality of $\sigma(\theta')$ at θ'; the equality follows by spelling out $x_{\mathcal{M}}(\sigma(\theta), \theta)$; the last inequality is immediate):

$$\begin{aligned}
x_{\mathcal{M}}(\sigma(\theta'), \theta') &\geq x_{\mathcal{M}}(\sigma(\theta'), \theta') \\
&= x_{\mathcal{M}}(\sigma(\theta), \theta) + a(\sigma(\theta))q(\sigma(\theta), \theta') - a(\sigma(\theta))q(\sigma(\theta), \theta) \\
&\geq x_{\mathcal{M}}(\sigma(\theta), \theta) - a(\sigma(\theta)).
\end{aligned}$$

The previous display implies $a(\sigma(\theta)) \geq x_{\mathcal{M}}(\sigma(\theta), \theta) - x_{\mathcal{M}}(\sigma(\theta'), \theta')$ for arbitrary θ and θ' in $[0, 1]$. Let $\underline{x}_{\mathcal{M}} = \inf_{\theta' \in [0,1]} x_{\mathcal{M}}(\sigma(\theta'), \theta')$ We conclude that for all θ in $[0, 1]$,

$$a(\sigma(\theta)) \geq x_{\mathcal{M}}(\sigma(\theta), \theta) - \underline{x}_{\mathcal{M}}.$$

Note that if σ' is another best-response on $\mathcal{M}$, then $x_{\mathcal{M}}(\sigma(\theta), \theta) = x_{\mathcal{M}}(\sigma'(\theta), \theta)$ holds for all θ. Therefore, repeating the previous calculation, we find that for all θ and all best-responses σ',

$$a(\sigma'(\theta)) \geq x_{\mathcal{M}}(\sigma(\theta), \theta) - \underline{x}_{\mathcal{M}}. \tag{B.3}$$

We are ready to define our candidate direct IC mechanism. Note that $\theta \mapsto x_{\mathcal{M}}(\sigma(\theta), \theta)$ is usc (see, e.g., Lemma 17.30 of Border and Aliprantis (2006, p. 569)), and hence Borel-measurable. Consider the direct mechanism $\tilde{\mathcal{M}} = ([0,1], \tilde{a}, \tilde{q}, \tilde{p})$ defined by the following rules:

$$\begin{aligned}
\tilde{q}(\theta', \theta) &= \begin{cases} 1, & \text{if } \theta = \theta', \\ 0, & \text{else} \end{cases} \\
\tilde{p}(\theta) &= \begin{cases} \frac{\underline{x}_{\mathcal{M}}}{1 - x_{\mathcal{M}}(\sigma(\theta), \theta) + \underline{x}_{\mathcal{M}}}, & \text{if } 1 - x_{\mathcal{M}}(\sigma(\theta), \theta) + \underline{x}_{\mathcal{M}} > 0, \\ 0, & \text{else} \end{cases} \\
\tilde{a}(\theta) &= x_{\mathcal{M}}(\sigma(\theta), \theta) - \underline{x}_{\mathcal{M}}.
\end{aligned}$$

All three of these functions map to $[0,1]$ and are Borel-measurable.

Let us first show that $x_{\tilde{\mathcal{M}}}(\theta, \theta) = x_{\mathcal{M}}(\sigma(\theta), \theta)$ holds for all θ. Indeed, we note that $1 - x_{\mathcal{M}}(\sigma(\theta), \theta) + \underline{x}_{\mathcal{M}} \leq 0$ holds if and only if $x(\sigma(\theta), \theta) = 1$ and $\underline{x}_{\mathcal{M}} = 0$. Therefore, for all θ,

$$\begin{aligned}
x_{\tilde{\mathcal{M}}}(\theta, \theta) &= \tilde{a}(\theta) + (1 - \tilde{a}(\theta))\tilde{p}(\theta) \\
&= x_{\mathcal{M}}(\sigma(\theta), \theta) - \underline{x}_{\mathcal{M}} + (1 - x_{\mathcal{M}}(\sigma(\theta), \theta) + \underline{x}_{\mathcal{M}})\tilde{p}(\theta) \\
&= x_{\mathcal{M}}(\sigma(\theta), \theta).
\end{aligned}$$

Thus $\theta \mapsto x_{\tilde{\mathcal{M}}}(\theta, \theta)$ is usc. Moreover, for all θ we have $\tilde{a}(\theta) = x_{\tilde{\mathcal{M}}}(\theta, \theta) - \inf_{\theta' \in [0,1]} x_{\tilde{\mathcal{M}}}(\theta', \theta')$.

We next show that $\tilde{\mathcal{M}}$ is IC. To that end, let θ' be different from θ. We have

$$x_{\tilde{\mathcal{M}}}(\theta', \theta) = (1 - \tilde{a}(\theta'))\tilde{p}(\theta') = x_{\mathcal{M}}(\sigma(\theta'), \theta') - \tilde{a}(\theta').$$

The choice of $\tilde{a}(\theta')$ as $\underline{x}_{\mathcal{M}}$ clearly implies $x_{\mathcal{M}}(\sigma(\theta), \theta) \geq x_{\mathcal{M}}(\sigma(\theta'), \theta') - \tilde{a}(\theta')$, and hence $x_{\tilde{\mathcal{M}}}(\theta, \theta) \geq x_{\tilde{\mathcal{M}}}(\theta', \theta)$. Thus $\tilde{\mathcal{M}}$ is IC.

It remains to show that $\bar{U}_p(\tilde{\mathcal{M}}) \geq \bar{U}_p(\mathcal{M})$ holds. We have shown that $x_{\tilde{\mathcal{M}}}(\theta, \theta) = x_{\mathcal{M}}(\sigma(\theta), \theta)$ holds for all best-responses σ on $\mathcal{M}$. Hence the two mechanisms admit the same set of agent-optimal distributions. Lastly, when the agent best responds the induced auditing cost are weakly lower under $\tilde{\mathcal{M}}$ than under $\mathcal{M}$, as we infer from (B.3) and the definition of $\tilde{a}$. Hence $\bar{U}_p(\tilde{\mathcal{M}}) \geq \bar{U}_p(\mathcal{M})$.

Lastly, consider the second part of the claim. The construction of a direct IC mechanism with the desired properties is completely analogous to the construction of $\tilde{\mathcal{M}}$ above, if one only replaces the above instance of $x_{\mathcal{M}}$ and $\inf x_{\mathcal{M}}$, respectively, with x and $\inf x$, respectively. We omit the details. □

B.1.5 Proof of Lemma B.6

Proof of Lemma B.6. Let the parameters of x be denoted $(\theta_0, \theta_1, \alpha, \beta, \gamma)$. There is nothing to show if $\beta = 0$; thus let $\beta > 0$.

As discussed in the paragraph preceding Lemma B.6, we have $U_p(x, F) \leq U_p(x - \beta, F)$. It follows that there is nothing to prove if F is agent-optimal on $x - \beta$. Thus assume F fails to be agent-optimal on $x - \beta$.

We invoke Lemma B.3 to find an agent-optimal binary distribution $\tilde{F}$ on $x - \beta$ such that $U_p(x - \beta, \tilde{F}) = \bar{U}_p(x - \beta)$. To prove the claim, it suffices to show $U_p(x - \beta, F) \leq U_p(x - \beta, \tilde{F})$. We will show that $\tilde{F}$ must be an MPS of F. Arguments analogous to those that established Lemma A.4 then imply that $U_p(x - \beta, F) \leq U_p(x - \beta, \tilde{F})$ holds.

To show that $\tilde{F}$ is an MPS of F, we proceed in two steps.

First, we claim that if $\hat{F}$ is agent-optimal on $x - \beta$, then it is not an MPC of F. Since F is agent-optimal on x but not on $x - \beta$, we have $U_a(x - \beta, \hat{F}) + U_a(x, F) > U_a(x - \beta, F) + U_a(x, \hat{F})$. Spelling out this inequality yields $\beta\left(\mathbb{E}_F[u_a(\theta)] - \mathbb{E}_{\hat{F}}[u_a(\theta)]\right) > 0$. Since u_a is concave and $\beta > 0$, we conclude that $\hat{F}$ cannot be an MPC of F.

Next, recall that $\tilde{F}$ is binary, and that the support of F is $\{\theta_0, \theta_1\}$. If both realizations of $\tilde{F}$ are outside the interval (θ_0, θ_1), then $\tilde{F}$ is an MPS of F and we are done. We complete the proof by arguing that a contradiction to the previous paragraph obtains if $\tilde{F}$ has a realization in (θ_0, θ_1). Suppose $\tilde{F}$ has such a realization t. Denote the other realization of $\tilde{F}$ by t'. Note that Lemma A.5 from the model with evidence applies to the model with verification since the lemma is a statement about agent-optimal distributions (which do not depend on the principal's verification costs). If $t' = t$, then $\tilde{F}$ is degenerate, and hence an MPC of F; contradiction. If $t' > \theta_1$, then Lemma A.5 implies that $\tilde{F}$ is supported on a subset of $[\theta_0, \theta_1]$. In particular, $\tilde{F}$ is an MPC of F; contradiction. If $t' \in (\theta_0, \theta_1)$ or $\alpha = \beta$, then the degenerate distribution on μ is another agent-optimal distribution. Since this distribution is an MPC of F, we have a contradiction. Lastly, if $t' \leq \theta_0$ and $\alpha > \beta$, then Lemma A.5 implies that $\tilde{F}$ is supported on a subset of $[\theta_0, \theta_1]$. In particular, $\tilde{F}$ is an MPC of F; contradiction. □

B.2 Multiple Alternatives

In this part of the appendix, we consider optimal mechanisms in an extension with multiple alternatives. We maintain the assumption that the agent chooses between all distributions on $[0,1]$ with some fixed mean, but allow for richer preferences than in the basic model.

B.2.1 Setup

Let us agree to the following notation: When Y is a metric space, then ΔY means the set of Borel probability distributions over Y. When y and u are vectors in Euclidean space, then $y \cdot u$ refers to their inner product.

There is a finite set X of alternatives. The agent's and principal's payoffs, respectively, are given by continuous functions $u_a\colon X \times [0,1] \to \mathbb{R}$ and $u_p\colon X \times [0,1] \to \mathbb{R}$, respectively. We further maintain the following:

Assumption 5. For all $x \in D$, the functions $\theta \mapsto u_p(x, \theta)$ is convex and $\theta \mapsto u_a(x, \theta)$ is concave. There exists $x_0 \in X$ such that, for all $x \in X$, the difference $\theta \mapsto u_a(x, \theta) - u_a(x_0, \theta)$ is weakly positive and concave.

Let $\mu \in (0,1)$. As in the basic model, the agent chooses a distribution F from the set $\mathcal{F}$ of distributions on $[0,1]$ with mean μ. Here we assume that all distributions are costless. This assumption is without loss in the following sense: If the costs are given as the integral of some convex function k, then we can simply redefine $u_a(x, \theta)$ as $u_a(x, \theta) - k(\theta)$.

As in the main text, we assume that, no matter the type distribution, the agent can provide hard evidence that fully reveals the type realization. In view of Assumption 5, we may therefore assume that the agent always discloses the type (else, the principal commits to implementing x_0).

A mechanism is a function[27] $x\colon [0,1] \to \Delta X$ satisfying the following.

[27] There should be no risk of confusion by using the symbol x for alternatives as well as for mechanisms.

(1) The mapping $\theta \mapsto x(\theta) \cdot u_a(\theta)$ is upper semi-continuous (usc).

(2) There exists a sequence $\{h_n\}_{n\in\mathbb{N}}$ of continuous real-valued functions on $[0,1]$ that converges pointwise to $\theta \mapsto x(\theta) \cdot u_p(\theta)$.

Restriction (1) is the natural analogue of the restriction from the basic model. Restriction (2) is that the principal's payoffs are sufficiently well-behaved; it holds if, for example, the mechanism admits finitely-many discontinuities.

The agent's and principal's utility from a mechanism x and a distribution F are $U_a(x,F) = \int x \cdot u_a \, dF$ and $U_p(x,F) = \int x \cdot u_p \, dF$, respectively. The set of x-agent-optimal distributions is $\mathcal{F}^*(x) = \arg\max_{F\in\mathcal{F}} U_a(x,F)$. The principal's utility from x (when the agent best responds and breaks ties favourably) is $\bar{U}_p(x) = \sup_{F\in\mathcal{F}^*(x)} U_p(x,F)$.

B.2.2 Two-sided cutoff mechanisms

Let δ_{x_0} denote the degenerate distribution on x_0.

Definition 4. A mechanism x is a **two-sided cutoff mechanism** if there exist $p_0, p_1 \in \Delta X$ and $\theta_0, \theta_1 \in [0,1]$ such that

$$\forall_{\theta\in[0,1]}, \quad x(\theta) = \begin{cases} p_0, & \text{if } \theta \le \theta_0, \\ \delta_{x_0}, & \text{if } \theta_0 < \theta < \theta_1, \\ p_1, & \text{if } \theta_1 \le \theta. \end{cases}$$

In this case, we refer to $(\theta_0, \theta_1, p_0, p_1)$ as the **parameters** of x.

The main result of this part of the appendix is that two-sided cutoff mechanisms are optimal.

Theorem B.7. *Let Assumption 5 hold. Let x be a mechanism. For all $\varepsilon > 0$, there is a two-sided cutoff mechanism x^* satisfying $\bar{U}_p(x^*) \ge \bar{U}_p(x) - \varepsilon$.*

B.2.3 Proof of Theorem B.7

The proof follows the one from the basic model. Given a mechanism x, let $\mathcal{F}^*_B(x)$ denote the set of binary distributions in $\mathcal{F}^*(x)$ (which at this point we are not claiming to be non-empty). We first establish a sense in which it suffices to consider binary distributions.

Lemma B.8. *Let x be a mechanism, and let $F \in \mathcal{F}^*(x)$. There exists a distribution $F^* \in \mathcal{F}_B^*(x)$ such that $U_p(x, F) \leq U_p(x, F^*)$.*

Proof of Lemma B.8. Recall that $\theta \mapsto x(\theta) \cdot u_a(\theta)$ is usc (by definition of a mechanism). By using this fact and retracing in the proof of Lemma A.2, one may verify that there is a Borel-probability measure ν over $\mathcal{F}_B^*(x)$ such that all continuous linear functions $\Gamma \colon \mathcal{F}^*(x) \to \mathbb{R}$ satisfy

$$\Gamma(F) = \int_{\tilde{F} \in \mathcal{F}_B^*} \Gamma(\tilde{F}) \, d\nu(\tilde{F}). \tag{B.4}$$

Now, let $\{h_n\}_{n \in \mathbb{N}}$ be a sequence of continuous real-valued functions that converges pointwise to $\theta \mapsto x(\theta) \cdot u_p(\theta)$.

For all n, continuity of h_n implies that the function $\tilde{F} \mapsto \int h_n \, d\tilde{F}$ is continuous on $\mathcal{F}^*(x)$. Hence (B.4) and the Dominated Converge theorem imply

$$\int_{\theta \in [0,1]} x(\theta) \cdot u_p(\theta) \, dF = \int_{\tilde{F} \in \mathcal{F}_B^*} \int_{\theta \in [0,1]} x(\theta) \cdot u_p(\theta) \, d\tilde{F} \, d\nu(\tilde{F}).$$

Equivalently,

$$U_p(x, F) = \int_{\tilde{F} \in \mathcal{F}_B^*} U_p(x, \tilde{F}) \, d\nu(\tilde{F}).$$

Thus there exists $F^* \in \mathcal{F}_B^*(x)$ such that $U_p(x, F) \leq U_p(x, F^*)$, as promised. □

We next establish a sufficient condition for the principal's on a two-sided cutoff mechanism to increase with respect to mean-preserving spreads. Let $\bar{u}_p(\mu) = \max_{x \in X} u_p(x, \mu)$.

Lemma B.9. *Let x be a two-sided cutoff mechanism with parameters $(\theta_0, \theta_1, p_0, p_1)$ such that $\mu \in (\theta_0, \theta_1)$. Let F denote the binary distribution in $\mathcal{F}$ whose support is $\{\theta_0, \theta_1\}$. Let F' and F'' be two binary distributions in $\mathcal{F}$ in such that F is an MPC of F', and F' is an MPC of F''. If $U_p(x, F) > \bar{u}_p(\mu)$, then $U_p(x, F'') \geq U_p(x, F')$.*

The proof follows the same ideas as the proof of Lemma A.4.

Proof of Lemma B.9. Let $\omega\colon [0,1] \to \mathbb{R}$ be defined by

$$\forall_{\theta\in[0,1]}, \quad \omega(\theta) = \begin{cases} x(\theta_0)\cdot u_p(\theta), & \text{if } \theta \le \theta_0, \\ x(\theta_0)\cdot u_p(\theta_0) + \frac{\theta-\theta_0}{\mu-\theta_0}\left(\bar{u}_p(\mu) - x(\theta_0)\cdot u_p(\theta_0)\right), & \text{if } \theta_0 < \theta \le \mu, \\ \bar{u}_p(\mu) + \frac{\theta-\mu}{\theta_1-\mu}\left(x(\theta_1)\cdot u_p(\theta_1) - \bar{u}_p(\mu)\right), & \text{if } \mu < \theta \le \theta_1, \\ x(\theta_1)\cdot u_p(\theta), & \text{if } \theta_1 < \theta. \end{cases}$$

The assumption on the MPC-ordering of F, F', and F'' implies $\operatorname{supp} F' \cup \operatorname{supp} F'' \subseteq [0,\theta_0] \cup [\theta_1, 1]$. Note also that x is constant on $[0,\theta_0] \cup [\theta_1,1]$. Hence, if θ is a point in the support of F, F', or F'', then $\omega(\theta)$ and $x(\theta)\cdot u_p(\theta)$ coincide. The claim therefore follows if we can show that ω is convex on $[0,1]$.

We first verify that the restriction of ω to $[\theta_0,\theta_1]$ is convex. Towards a contradiction, suppose this restriction is non-convex. Since this restriction is piecewise affine with at most one change in slope, it is concave. Hence

$$U_p(x,F) = \int x(\theta)\cdot u_p(\theta)\,dF = \int \omega\,dF \le \omega(\mu) = \bar{u}_p(\mu),$$

and we have a contradiction to the assumption $U_p(x,F) > \bar{u}_p(\mu)$.

Now, since $x(\theta_0)\cdot u_p(\mu) \le \bar{u}_p(\mu)$ and since the functions $(u_p(x,\cdot))_{x\in X}$ are all convex, one may verify that the restriction of ω to $[0,\mu]$ is convex. A similar argument shows that the restriction of ω to $[\mu,1]$ is convex. It follows from these facts and the previous paragraph that ω is convex on $[0,1]$. □

We now complete the proof of Theorem B.7.

Proof of Theorem B.7. Recall the definition $\bar{u}_p(\mu) = \max_{x\in X} u_p(x,\mu)$. If $\bar{U}_p(x) \le \bar{u}_p(\mu)$, then the claim follows trivially by taking x^* to be two-sided cutoff mechanism that is constantly δ_{x_0}. (On this mechanism, the degenerate distribution on μ is agent-optimal, and this distribution generates a utility of $\bar{u}_p(\mu)$ for the principal.) Thus let $\bar{U}_p(x) > \bar{u}_p(\mu)$.

Let ε. In view of Lemma B.8, we may find a binary distribution F^* in $\mathcal{F}^*(x)$ such that $U_p(x,F^*)$ is within $\frac{\varepsilon}{2}$ of $\bar{U}_p(x)$ and such that $U_p(x,F^*) > \bar{u}_p(\mu)$. Let $\{\theta_0^*,\theta_1^*\}$ denote the support of F^*. The inequality $U_p(x,F^*) > \bar{u}_p(\mu)$ implies that F^* is non-degenerate, meaning $\mu \in (\theta_0^*,\theta_1^*)$.

Our candidate for x^* is the two-sided cutoff mechanism with parameters $(\theta_0^*, \theta_1^*, x(\theta_0^*), x(\theta_1^*))$. Invoking Lemma B.8, let us find a binary distribution F in $\mathcal{F}^*(x)$ such that $U_p(x^*, F)$ is within $\frac{\varepsilon}{2}$ of $\bar{U}_p(x^*)$. We consider two cases.

If the support of F consists of a point below θ_0^* and a point above θ_1^*, then F is a mean-preserving spread of F^*. Hence Lemma B.9 implies $U_p(x^*, F) = U_p(x^*, F^*)$. Since we also have $U_p(x^*, F^*) = U_p(x, F^*)$, and hence the choice of F^* and F imply $\bar{U}_p(x^*) \geq \bar{U}_p(x) - \varepsilon$, as desired.

Thus suppose the support of F contains a point in (θ_0^*, θ_1^*). By analogizing the argument that established Lemma A.5, one may verify that there is another x^*-agent-optimal distribution $\tilde{F}$ whose support is a subset of $[\theta_0^*, \theta_1^*]$. Agent-optimality of $\tilde{F}$ on x^* and agent-optimality of F^* on x imply $U_a(x, F^*) \geq U_a(x, \tilde{F})$ and $U_a(x^*, F^*) \leq U_a(x^*, \tilde{F})$. The definition of x^* and the inclusion $\operatorname{supp} \tilde{F} \subseteq [\theta_0^*, \theta_1^*]$ moreover imply $U_a(x, \tilde{F}) \geq U_a(x^*, \tilde{F})$. Hence $U_a(x^*, F^*) = U_a(x^*, \tilde{F})$, meaning that F^* is another agent-optimal distribution on x^*. Hence $\bar{U}_p(x^*) \geq U_p(x^*, F^*) = U_p(x, F^*) \geq \bar{U}_P(x) - \frac{\varepsilon}{2}$. In particular, we have $\bar{U}_p(x^*) \geq \bar{U}_p(x) - \varepsilon$, as desired. □

References

Bayrak, Halil I, Kemal Güler, and Mustafa Ç Pınar (2017). "Optimal allocation with costly inspection and discrete types under ambiguity". In: *Optimization Methods and Software* 32.4, pp. 699–718.

Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman (2019). "Mechanisms with evidence: Commitment and robustness". In: *Econometrica* 87.2, pp. 529–566.

— (2021). "Mechanism Design for Acquisition of/Stochastic Evidence".

— (2014). "Optimal allocation with costly verification". In: *American Economic Review* 104.12, pp. 3779–3813.

Border, Kim C. and Charalambos D. Aliprantis (2006). *Infinite Dimensional Analysis : A Hitchhikers Guide*. Springer Berlin Heidelberg. DOI: 10.1007/3-540-29587-9.

Dye, Ronald A (1985). "Disclosure of nonproprietary information". In: *Journal of accounting research*, pp. 123–145.

Epitropou, Markos and Rakesh Vohra (2019). "Optimal On-Line Allocation Rules with Verification". In: *International Symposium on Algorithmic Game Theory*. Springer, pp. 3–17.

Erlanson, Albin and Andreas Kleiner (2019). "A note on optimal allocation with costly verification". In: *Journal of Mathematical Economics* 84. Publisher: Elsevier, pp. 56–62.

— (2020). "Costly verification in collective decisions". In: *Theoretical Economics* 15.3. Publisher: Wiley Online Library, pp. 923–954.

Escudé, Matteo (2020). "Communication with partially verifiable endogenous information".

Glazer, Jacob and Ariel Rubinstein (2004). "On optimal rules of persuasion". In: *Econometrica* 72.6, pp. 1715–1736.

— (2006). "A study in the pragmatics of persuasion: a game theoretical approach". In: *Theoretical Economics* 1.4, pp. 395–410.

Halac, Marina and Pierre Yared (2020). "Commitment versus Flexibility with Costly Verification". In: *Journal of Political Economy* 128.12. Publisher: The University of Chicago Press Chicago, IL, pp. 4523–4573.

Hart, Sergiu, Ilan Kremer, and Motty Perry (2017). "Evidence games: Truth and commitment". In: *American Economic Review* 107.3, pp. 690–713.

Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian persuasion". In: *American Economic Review* 101.6, pp. 2590–2615.

Kattwinkel, Deniz and Jan Knoepfle (2019). "Costless Information and Costly Verification: A Case for Transparency". In: *Available at SSRN 3426817*.

Li, Yunan (2020). "Mechanism design with costly verification and limited punishments". In: *Journal of Economic Theory* 186, p. 105000. ISSN: 0022-0531.

Mensch, Jeffrey and Doron Ravid (2022). "Monopoly, Product Quality, and Flexible Learning".

Migrow, Dimitri and Sergei Severinov (Aug. 2022). "Investment and Information Acquisition". In: *American Economic Journal: Microeconomics* 14.3, pp. 480–529. DOI: 10.1257/mic.20200115. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20200115.

Phelps, Robert R (2001). *Lectures on Choquet's theorem.* Springer Science & Business Media.

Rappoport, Daniel and Valentin Somma (2017). "Incentivizing information design".

Ravid, Doron, Anne-Katrin Roesler, and Balázs Szentes (2022). "Learning before trading: on the inefficiency of ignoring free information". In: *Journal of Political Economy* 130.2, pp. 346–387.

Sher, Itai (2011). “Credibility and determinism in a game of persuasion”. In: *Games and Economic Behavior* 71.2, pp. 409–419.

— (2014). “Persuasion and dynamic communication”. In: *Theoretical Economics* 9.1, pp. 99–136.

Shishkin, Denis (2021). “Evidence Acquisition and Voluntary Disclosure”. In: *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 817–818.

Silva, Francisco (2020). “The importance of commitment power in games with imperfect evidence”. In: *American Economic Journal: Microeconomics* 12.4, pp. 99–113.

Thereze, João (2022). “Screening Costly Information”.

Titova, Maria (2022). “Persuasion with verifiable information”.

Tsakas, Elias, Nikolas Tsakas, and Dimitrios Xefteris (2021). “Resisting persuasion”. In: *Economic Theory* 72.3, pp. 723–742.

Whitmeyer, Mark and Kun Zhang (2022). “Costly Evidence and Discretionary Disclosure”.

Winkler, Gerhard (1988). “Extreme points of moment sets”. In: *Mathematics of Operations Research* 13.4, pp. 581–587.

Yoder, Nathan (2022). “Designing incentives for heterogeneous researchers”. In: *Journal of Political Economy* 130.8, pp. 000–000.

Zhang, Kun (2022). “Withholding Verifiable Information”.