# Bilateral communication and Hard evidence[*]

Mehdi AYOUNI[†]

February 21, 2017

**Abstract**

In a persuasion problem, an informed agent who has some certifiable information communicates with the principal who chooses an outcome. In unilateral communication, only the agent sends a message to the principal. In bilateral communication, both exchange messages sequentially. We study and compare these two types of mechanisms under the constraint that the agent can present the same amount of certifiable information in both cases. In the canonical bilateral communication mechanism, after receiving a claim from the agent, the principal asks him to certify a certain event and bases her decision on his ability to do so. The main result of this paper essentially states that if information certification is limited and the limitation prevents the principal from achieving her first-best in unilateral communication then she strictly benefits from bilateral communication.

## 1 Introduction

In a persuasion problem, an agent wishes to influence a principal who has to implement an outcome. The agent privately knows the state of the world, also called his *type*, and has *hard evidence* about it. Any certified message that proves a non trivial statement is considered hard evidence. Formally, a piece of evidence is a message certifying a certain *event*. Namely, that the agent belongs to a certain subset of types. Not all events are necessarily certifiable and the set of certifiable events depends on the problem at hand.

[*] I thank my doctoral advisors Frédéric Koessler (Paris School of Economics – CNRS) and Régis Renault (Université Paris Dauphine) for their valuable guidance and continual help. I also thank Joel Watson (University of California, San Diego), the members of THEMA's Theory Group and the participants of the International Conference on Game Theory at Stony Brook for their suggestions and comments. This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

[†] Thema, Université de Cergy-Pontoise, e-mail: mohamed.ayouni@u-cergy.fr

The principal ignores the state of the world which is relevant to her decision, but she can interact with the agent before implementing an action. The standard setting to model such an interaction is the sender-receiver game: the agent (sender) presents information to the principal (receiver) by sending a message containing certifiable information before the principal chooses an outcome. We call this setting the *unilateral communication* framework. In contrast, the *bilateral communication* setting is one where both the agent and the principal are active in the communication phase, exchanging messages sequentially.

As an illustration, consider the example of a hiring process. The agent is the applicant who knows his skills and abilities which define his type. The principal is the employer who does not observe that information so she interviews him before making a decision. If the employer can learn all information during the interview, she would not gain from being active in the communication phase. However, in some cases only some information can be certified: for example it might not be possible to test all the skills the applicant claims to master due to cost or time constraints. In such cases, bilateral communication might allow the employer to improve the outcome by choosing what the applicant has to certify based on his claims instead of letting him choose the information he presents as in unilateral communication.

In the unilateral communication framework, the principal has to choose an *implementation rule* that assigns an outcome to every possible message the agent can send. In the bilateral communication framework, the principal has to design the *communication mechanism* in addition to the implementation rule. The communication mechanism specifies the active player and the set of available messages at each node. The implementation rule, in this case, specifies the outcome possibly as a function of the history of exchanged messages.

Our goal is to study and compare both frameworks. To that end, we impose the restriction that the same amount of information can be certified in both settings. This guarantees that any difference of implementable outcomes is only the result of the bilateral exchange of non-certifiable information between the principal and the agent.

We show that the *canonical mechanism* in bilateral communication has the following simple structure: a three-stage communication mechanism where (*i*) the agent announces his type, (*ii*) the principal asks him to certify a specific event, (*iii*) he certifies an event of his choice, and an *implementation rule* that selects the outcome based on the announced type and whether the requested event was certified. Namely, for the implementation of an outcome function $f$, if the agent announces type $t$ and certifies the requested event then $f(t)$ is implemented, otherwise a punishment action is implemented. By applying Theorem 6 of Bull and Watson [2007], who study the introduction of hard evidence to mechanism design, we obtain a partial characterization of the canonical mechanism. In order to explicitly determine the principal's message in step (*ii*) and the implementation rule, we use the fact that there is only one agent with type-independent preferences.

Having identified the canonical mechanism, we establish the necessary and sufficient conditions for the implementation of any outcome function in both settings and we show that the sets of implementable outcome functions coincide if and only if the *normality* condition is satisfied. This condition states that every type can certify a maximal evidence event, i.e. an event that is equivalent to certifying all information about that type. In other words, unilateral and bilateral communication are outcome equivalent only when there are no effective limitations on the amount of information that can be certified. Bilateral communication is potentially beneficial to the principal only in settings where normality is not satisfied,. The hiring process is an example of such a setting if it is not possible to certify all (available) events (at least for some types).

Our main result gives sufficient conditions for bilateral communication to improve the outcome for the principal in comparison with unilateral communication. It is essentially shown that if the principal's first-best is well defined (not necessarily by a unique outcome function) and is not achievable in unilateral communication but would be achieved if any amount of information can be certified then bilateral communication strictly increases the principal's expected payoff. In other words, the principal gains from being active in the communication phase if she is unable to achieve her first-best in unilateral communication because of the cost or time constraints that limit information certification.

As an extension, we examine whether the results in the literature about commitment and outcome randomization hold in our framework. Sher [2011] studies the unilateral communication setting and shows that under a concavity assumption, namely that the principal's utility function is a type-dependent concave transformation of the agent's utility function, the principal needs neither commitment over the implementation rule nor randomization of outcome. These results are in fact generalizations of the findings of Glazer and Rubinstein [2006] who considered only binary action spaces, for which the concavity assumption is always satisfied. Hart et al. [2016] show that commitment is unnecessary for a class of certifiability structures (which satisfy normality) and strongly single-peaked preferences. We show that, in bilateral communication under the conditions of our main result and the concavity assumption stated above, randomization is not necessary if the action space is continuous but we give an example with a discrete action space where it is needed. We also give an example where commitment is necessary under the same conditions.

**Related Literature** Certifiable information has been extensively studied in both sender-reciever games and mechanism design by authors such as Green and Laffont [1986], Glazer and Rubinstein [2001, 2004], Forges and Koessler [2005], Bull [2008], Deneckere and Severinov [2008], Ben-Porath and Lipman [2012], Kartik and Tercieux [2012], Koessler and Perez-Richet [2014], Sher [2011], Sher and Vohra [2015], Singh and Wittman [2001], Strausz [2016]. These papers, among others, give rise to important results about implementable allocation rules and some of them establish a revelation principle for settings with certifiable information.

Bull and Watson [2007] study hard evidence in a general mechanism design setting (with multiple agents) and characterize a three-stage communication mechanism in Theorem 6 which we use to determine the canonical bilateral communication mechanism as explained above. We also apply the revelation principle in unilateral communication given in their Theorem 1. Moreover, we show that unilateral and bilateral communication are outcome equivalent if and only if normality is satisfied. The if part holds in general and is established in their Theorem 5. The only if part relies on the agent's type-independent preferences.

Sher [2014] is closely related to our work but focuses on a framework where the decision space of the principal is binary. It is shown for instance, that unilateral communication is optimal under *foresight* which is a condition related to, but weaker than normality. We note that our main result does not apply in that framework because it requires punishment to be non optimal which is impossible with a binary action space.

# 2 The model

## 2.1 The environment

Consider a setting where a principal faces an agent who is privately informed about his type $t$ in a finite set of agent types $T$. The principal ignores $t$ but knows the probability distribution of types. We assume the existence of a *certifiability structure* $\mathcal{C} \subseteq 2^T$, where for every $t \in T$ there exists $C \in \mathcal{C}$ such that $t \in C$. We denote by $\mathcal{C}(t) = \{C \in \mathcal{C} : t \in C\}$ the set of events the agent can certify when his type is $t$.

The principal has to implement an action $a$ in $A$. Prior to her decision, she can communicate with the agent. The principal's goal is to design the *communication mechanism* along with an *implementation rule.* There are two types of communication mechanisms:

- Unilateral Communication: Only the agent is active in the communication. He can certify an event $C$ in $\mathcal{C}$ (which has to be in $\mathcal{C}(t)$ if his type is $t$) and (possibly) send a message $m$ in some predetermined set (independent of the true type).

- Bilateral Communication: Both the agent and the principal partake in sequential communication. The mechanism has to specify the active player at each node and the set of available messages at that node. The only constraint is that, along every possible path, the agent must have exactly one node at which he can certify an event $C$ in $\mathcal{C}$. At every node, the active player chooses a message from a predetermined set of messages.

With these definitions, we can analyze the benefit of active communication for the principal insofar as the same amount of information is certified in both mechanisms: if bilateral communication increases the principal's expected payoff in comparison with unilateral communication then the difference is due only to the non-certifiable information that is exchanged back and forth between the agent and the principal.

The requirement that the agent does not certify more than one event in $\mathcal{C}$ corresponds to given constraints on the amount of information that can be verified during an interaction between the agent and the principal[1]. For example, such constraints apply if the agent has limited time to present this information or the principal has limited time to check it. Hiring processes generally fall in this category when it is impossible to verify whether the applicant fits all the requirements of the job. Recruiters have to decide which aspects to verify and which aspects to ignore.

The implementation rule specifies the principal's action for every possible history in the communication mechanism. In the case of unilateral communication, the history contains exactly one node so that the principal's action is simply a function of the information that the agent presents.

An outcome function $f : T \to \Delta(A)$ is a mapping from types to lotteries over actions. The agent has a utility function $u : A \to \mathbb{R}$ which is independent of his type. Let $a_0$ denote an action

[1] This is essentially without loss of generality because if we want to model a limitation to $N$ events instead of one, we would have to replace $\mathcal{C}$ with the set of events that combine up to $N$ elements of $\mathcal{C}$, i.e. $\{\cap_{i=1}^{n} C_i$ s.t. for all $i, C_i \in \mathcal{C}$ and $n \leq N\}$.

such that $u(a_0) = \min_{a \in A} u(a)$ whenever the minimum exists[2]. Throughout the paper, $a_0$ will be called the punishment action and the value of $u(a_0)$ will be set to 0 w.l.o.g. We also assume that $u$ is not constant over $A$ (otherwise all outcome functions would be implementable). The principal has a utility function $v : T \times A \to \mathbb{R}$ which not only depends on the action she chooses to implement, but also on the type of the agent.

## 2.2 The canonical form

Consider the following communication mechanism:

**Definition 1.** A three-stage communication mechanism is a bilateral communication mechanism with the following timing:

- Stage 1 : The agent reports a type.
- Stage 2 : The principal asks the agent to certify a particular event.
- Stage 3 : The agent certifies an event of his choice.

In stage 1, the agent makes a claim by reporting a type $t \in T$. Then the principal asks him to certify a particular event in $\mathcal{C}$. Her choice at stage 2, is given by $\sigma : T \to \Delta(\mathcal{C})$ with $\sigma(t; C)$ denoting the probability of asking the agent to certify $C$ given that he announced type $t$. In stage 3, the agent certifies $C'$ (either the requested $C$ or a different event).

**Definition 2.** For given $f : T \to \Delta(A)$ and $\sigma : T \to \Delta(\mathcal{C})$, the $(\sigma, f)$-mechanism is a three-stage communication mechanism along with an implementation rule such that:

- $\sigma$ is used in stage 2.

[2] It is implicitly assumed that the minimum exists in all the results stated in this paper. But essentially, the results still hold with minor modifications if this assumption is not satisfied. See Appendix for a study of the case where the infimum is not reached.

- If the agent certifies the requested event the outcome $f(t)$ is implemented, otherwise the punishment action $a_0$ is implemented.

As we show in the last part of this section, $(\sigma, f)$-mechanisms are *canonical* in the sense that we can restrict attention to such mechanisms when studying the implementation of a given outcome function $f$. Furthermore, this implementation is achieved with truthful reporting in the first stage: if $\sigma$ is such that for every type $t$, an agent of type $t$ has no incentive to report a different type and is able to certify any $C$ that is requested with positive probability $\sigma(t; C)$ (i.e. the support of $\sigma(t)$ is in $\mathcal{C}(t)$) then $f$ is implementable in the $(\sigma, f)$-mechanism.

**Proposition 1.** *If $f$ is implemented using a general bilateral communication mechanism and a general implementation rule then there exists $\sigma : T \to \Delta(\mathcal{C})$ such that it is also implemented in the $(\sigma, f)$-mechanism with truthful reporting in stage 1.*

**Proof.** See Appendix. □

The argument of this proof is split in two steps. First, note that a bilateral communication mechanism is an extensive form game with three types of nodes :

- Message nodes : one player (principal or agent) sends a message to the other.
- Certification nodes : the agent certifies an event.
- Terminal nodes : the principal implements an outcome.

such that along every path in the game tree, there is exactly one certification node. Theorem 6 of Bull and Watson [2007] guarantees that if an outcome function is implementable using such a general mechanism then it is also implementable using a three-stage mechanism with truthful reporting in stage 1:

- Stage 1 : the agent reports his type to the principal.
- Stage 2 : the principal sends a message to the agent.
- Stage 3 : the agent certifies an event.

This mechanism is similar to our three-stage communication mechanism except that instead of directly asking for evidence, the message of stage 2 identifies an information set for the agent in the original extensive form game, more specifically, the one where he has to present evidence.

In the second step, we use the fact that, in our framework, there is only one agent whose preferences are the same across types in order to show that we can restrict attention even further and focus only on $(\sigma, f)$-mechanisms.

## 3 Implementable outcome functions

Proposition 1 asserts that an outcome function $f$ is implementable in bilateral communication if and only if there exists $\sigma$ that implements it (in the $(\sigma, f)$-mechanism). In this section, we determine the necessary and sufficient conditions for an outcome function to be implementable in bilateral communication. Then, we characterize implementable outcome functions in unilateral communication and identify the link between the two types of mechanisms.

**Lemma 1.** *$\sigma$ implements $f$ with truthful reporting if and only if* [3]

$$\begin{aligned} \forall t, \quad & \sigma_{tt} = 1 \\ \forall t, t' \quad & \sigma_{t't} \leq \frac{u(f(t'))}{u(f(t))} \end{aligned}$$

*where $\sigma_{t't} = \sum_{C \in \mathcal{C}(t')} \sigma(t; C)$ is the probability for an agent of type $t'$ to successfully persuade the principal that he is of type $t$.*

[3] $u(f(t))$ denotes in general the expectation of agent's utility given the lottery over actions $f(t)$.

**Proof.** See Appendix. □

The first set of conditions say that the principal asks only for events that an agent of the announced type can certify. This guarantees that if the agent reports truthfully then the principal will certainly implement the right outcome. The second set of conditions are in fact the incentive compatibility constraints of the agent, which ensure that he reports his type truthfully in the first stage. Truthful reporting in stage 1 is generically necessary to implement the outcome function, and these conditions make sure that the agent has incentive to tell the truth and that the principal does not make the mistake of asking an agent who reported his true type for evidence he cannot present, which in turn, would induce punishment erroneously.

Using Lemma 1 we can determine the necessary and sufficient conditions for an outcome function $f$ to be implementable. We focus on the strategies satisfying the first set of conditions, i.e. strategies such that the support of $\sigma(t)$ is contained in $\mathcal{C}(t)$ for all types $t$, and we study the existence of an incentive compatible strategy among them. Consider an indexing of types in $T$ from 1 to $n$ : $T = \{t_1, \ldots, t_n\}$. Let $q_k$ be the number of events that are certifiable by type $t_k$: $q_k = \text{card}(\mathcal{C}(t_k))$. $\mathcal{C}(t_k)$ may then be written as $\mathcal{C}(t_k) = \{C_k^1, \ldots, C_k^{q_j}\}$. The vector $\sigma(t_k; C)|_{C \in \mathcal{C}(t_k)}$ describes a point $M_k$ in $\mathbb{R}^{q_k}$. Using this definition, the second set of conditions of Lemma 1 can be interpreted as linear inequalities satisfied by the coordinates of the $M_k$'s for $k \in \{1, \ldots, n\}$. From this formulation, we can derive the following result about the implementability of an outcome function $f$:

**Proposition 2.** *An outcome function $f$ is implementable if and only if for all $k \in \{1, \ldots, n\}$, the following linear program $P_k$ has a value greater than or equal to* 1*:*

$$\begin{aligned} \text{Max} \quad & c \cdot x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0, \end{aligned}$$

*where* $x, c \in \mathbb{R}^{q_k}$*,* $b \in \mathbb{R}^{n-1}$ *and* $A$ *a matrix* $(n-1) \times q_k$*.* $\forall l \in \{1, \dots, q_k\}, \forall j \in \{1, \dots, k-1, k+1, \dots, n\}$*,* $c_l = 1$*,* $b_j = \frac{u(f(t_j))}{u(f(t_k))}$ *and* $A_{jl} = \mathbb{1}_{\{t_j \in C_k^l\}}$*.*

**Proof.** From Lemma 1, we know that $f$ is implementable if and only if there exists a strategy $\sigma$ such that

$$\begin{aligned} \forall k, \quad & \sigma_{kk} = 1 \\ \forall k, \forall j, \quad & \sigma_{jk} \leq \frac{u(f(t_j))}{u(f(t_k))} \end{aligned}$$

For a given $k \in \{1, \dots, n\}$, let $x \in \mathbb{R}^{q_k}$ denote the vector $\sigma(t_k; C)|_{C \in \mathcal{C}(t_k)}$, i.e. $x_l = \sigma(t_k; C_k^l)$. The condition $\sigma_{kk} = 1$ is then equivalent to the condition $\sum_{l \in \{1, \dots, q_k\}} x_l = c \cdot x = 1$, where $c \in \mathbb{R}^{q_k}$ and $\forall l, c_l = 1$. Consider the matrix $A$ such that, $\forall l \in \{1, \dots, q_k\}, \forall j \in \{1, \dots, k-1, k+1, \dots, n\}$, $A_{jl} = \mathbb{1}_{\{t_j \in C_k^l\}}$. We can then write $\sigma_{jk} = (Ax)_j$ for every $j$. By defining the vector $b \in \mathbb{R}^{n-1}$ such that $b_j = \frac{u(f(t_j))}{u(f(t_k))}$, we conclude that the set of conditions on $\sigma_{jk}$ for $j \in \{1, \dots, k-1, k+1, \dots, n\}$ is equivalent to $Ax \leq b$. So far, we have shown that $f$ is implementable if and only if for every $k$ there exists a vector $x \in \mathbb{R}^{q_k}$, with positive coordinates, such that

$$\begin{aligned} c \cdot x &= 1 \\ Ax &\leq b \end{aligned}$$

If such a vector exists, then the value of $P_k$ is at least 1. Conversely, if $x^*$ is the solution of $P_k$, with $v = c \cdot x^*$ greater than 1, then the vector $x = \frac{1}{v} x^*$ satisfies the conditions above.

□

The implementability of an outcome function $f$ is therefore equivalent to conditions on the values of $n$ linear programs. Moreover, if these conditions are satisfied then we obtain a $\sigma$ that implements $f$: $\sigma$ such that $\sigma(t_k; C)|_{C \in \mathcal{C}(t_k)}$ is a solution of $P_k$ divided by its value.

In the second part of this section, we focus on implementation in unilateral communication.

The standard revelation principle (see Theorem 1 of Bull and Watson [2007] or Proposition 2 of Myerson [1982]) applies in this context: if an outcome function $f$ is implementable in unilateral communication then it is implementable in a unilateral communication mechanism where the agent reports a type and certifies an event in $\mathcal{C}$ with truthful type reporting. Using this fact, we can characterize implementable outcome functions in unilateral communication.

In bilateral communication, $\sigma$ is called *deterministic* if for every type $t$, there exists an event $C$ that is requested with certainty if type $t$ is announced in stage 1, i.e. $\sigma(t;C) = 1$.

**Definition 3.** An outcome function $f$ is *implementable in deterministic bilateral communication* if there exists a deterministic $\sigma$ such that $f$ is implemented in the $(\sigma, f)$-mechanism.

The fact that a deterministic $\sigma$ maps every type to one event with certainty makes it possible to reduce the communication phase to a single stage as in the models of Glazer and Rubinstein [2006] and Sher [2011]. Consider an outcome function $f$ and a deterministic $\sigma$ that implements it. In the $(\sigma, f)$-mechanism, if the agent announces a type $t$ then the principal asks him for some $C$ with certainty (which can be denoted $\sigma(t)$), and if he certifies it the outcome $f(t)$ is implemented, otherwise the outcome $a_0$ is implemented. It becomes clear that $f$ is implementable in unilateral communication as follows: if the agent reports $t$ and certifies $\sigma(t)$ for some $t$ then $f(t)$ is implemented, otherwise $a_0$ is implemented.

Notice that if an agent wants to get the outcome $f(t)$ for some type $t$, he just has to be able to certify $\sigma(t)$. Therefore if the agent strictly prefers $f(t)$ to $f(t')$, then the incentive compatibility constraint implies that $t'$ is not in $\sigma(t)$. This property is formalized in the following definition:

**Definition 4.** An outcome function $f$ is *evidence compatible* if for every type $t$ there exists $C$ in $\mathcal{C}(t)$ such that:

$$\forall t', \text{ if } u(f(t')) < u(f(t)) \text{ then } t' \notin C.$$

The evidence compatibility of an outcome function $f$ means that every type $t$ can certify an event that no type with a worse outcome than $f(t)$ can certify. The previous analysis shows that if an outcome function is implementable in deterministic bilateral communication then it is evidence compatible.

We conclude this analysis with the following equivalence result:

**Proposition 3.** *Let $f$ be an outcome function. The three following statements are equivalent:*

*(i) $f$ is implementable in unilateral communication.*

*(ii) $f$ is evidence compatible.*

*(iii) $f$ is implementable in deterministic bilateral communication.*

**Proof.** See Appendix. □

Propositions 2 and 3 characterize the sets of implementable outcome functions in bilateral and unilateral communication. In the remainder of this section, we identify the necessary and sufficient condition for these sets to coincide. This analysis is interesting insofar as it allows us to determine when the principal can potentially benefit from being active in the communication phase. Let $c^*(t)$ denote the intersection of all events that type $t$ can certify:

$$c^*(t) = \bigcap_{C \in \mathcal{C}(t)} C,$$

**Definition 5.** The certifiability structure $\mathcal{C}$ is called *normal*[4] if for every type $t$ there exists a certifiable event providing maximal evidence about $t$, that is:

$$\forall t \in T, c^*(t) \in \mathcal{C}$$

**Proposition 4.** *The sets of implementable outcome functions in unilateral and bilateral communication coincide if and only if the certifiability structure is normal.*

[4] This condition is called normality by Bull and Watson [2007]. It has also been called the full reports condition by Lipman and Seppi [1995] and the minimal closure condition by Forges and Koessler [2005].

**Proof.** First, note that outcome functions that are implementable in unilateral communication are also implementable in bilateral communication (see Proposition 3).

If the certifiability structure is normal, Theorem 5 of Bull and Watson [2007] implies that outcome functions that are implementable in bilateral communication are also implementable in unilateral communication. More specifically, assume that the certifiability structure is normal and consider an outcome function $f$ that is implementable in bilateral communication. From normality and Lemma 1 we get that $f$ is evidence compatible and therefore implementable in unilateral communication (by Proposition 3). Thus the two sets of implementable outcome functions coincide under normality.

To prove the converse, we assume that the certifiability structure is not normal and we construct an outcome function that is implementable in bilateral communication but not implementable in unilateral communication. Under non-normality there exists a type $\bar{t}$ such that $c^*(\bar{t}) \notin \mathcal{C}$. $c^*(\bar{t})$ is not empty (it contains at least $\bar{t}$) and does not contain all types: if we had $c^*(\bar{t}) = T$ then $\mathcal{C}(\bar{t}) = \{T\}$ and as a consequence $c^*(\bar{t})$ would be in $\mathcal{C}$.

Consider an action $a$ such that $u(a) > u(a_0) = 0$ and the outcome function $f_\lambda$ defined by:

$$f_\lambda(t) = \begin{cases} a & \text{if } t \in c^*(\bar{t}) \\ (a_0; a) \text{ with proba. } (\lambda; 1-\lambda) & \text{if } t \notin c^*(\bar{t}) \end{cases}$$

Types in $c^*(\bar{t})$ can certify any event that $\bar{t}$ can certify. Any type that is not in $c^*(\bar{t})$ is unable to certify at least one event in $\mathcal{C}(\bar{t})$. The outcome function $f_\lambda$ separates types in two sets and gives a higher payoff to the set of types that can certify any event in $\mathcal{C}(\bar{t})$. Let $\overline{\lambda} = \frac{1}{\text{card}(\mathcal{C}(\bar{t}))}$, and consider $\sigma$ defined as follows:

$$\sigma(t;C) = \begin{cases} \overline{\lambda} & \text{if } t \in c^*(\overline{t}) \text{ and } C \in \mathcal{C}(\overline{t}) \\ 1 & \text{if } t \notin c^*(\overline{t}) \text{ and } C = T \\ 0 & \text{otherwise} \end{cases}$$

In the $(\sigma, f_\lambda)$-mechanism, if the agent reports a type $t$ in $c^*(\overline{t})$ (i.e., he wants to get the payoff $u(a)$ with certainty), the principal selects an element in $\mathcal{C}(\overline{t})$ randomly (with equal probability) and asks him to certify it. If the agent reports any other type, he is not required to certify any event and he gets $u(a)$ with probability $1-\lambda$ and 0 with probability $\lambda$. It is readily verifiable that this mechanism implements $f_\lambda$ for any $\lambda$ smaller than $\overline{\lambda}$.

We now show that $f_\lambda$ is not implementable in unilateral communication by proving that it is not evidence compatible (see Proposition 3). Indeed, the evidence compatibility condition of $f_\lambda$ would imply that there exists $C$ in $\mathcal{C}(\overline{t})$ that does not contain any type $t$ outside of $c^*(\overline{t})$. Such an event can only be $c^*(\overline{t})$ which is not in $\mathcal{C}$. Thus, $f_\lambda$ is not evidence compatible.

□

**Example 1.** Let $T = \{t_1, t_2, t_3\}$ and $\mathcal{C} = \{\{t_1, t_3\}, \{t_2, t_3\}\}$. $\mathcal{C}$ does not satisfy normality: $c^*(t_3) = \{t_3\}$ is not certifiable. Implementable outcome functions in unilateral communication are the evidence compatible outcome functions. If $f$ is evidence compatible, it follows that $u(f(t_3)) \geq \max\{u(f(t_1)), u(f(t_2))\}$. The reason is that $t_1$ (respectively, $t_2$) cannot certify an event that does not contain $t_3$. Moreover, $u(f(t_3))$ cannot be strictly greater than $\max\{u(f(t_1)), u(f(t_2))\}$, otherwise $t_3$ would have to certify an event that contains neither $t_1$ nor $t_2$, i.e. the event $\{t_3\}$ which is not certifiable. If $u(f(t_3)) = \max\{u(f(t_1)), u(f(t_2))\}$, it is easy to check that $f$ is evidence compatible. In conclusion, $f$ is implementable in unilateral communication if and only if:

$$u(f(t_3)) = \max\{u(f(t_1)), u(f(t_2))\}.$$

Implementable outcome functions in bilateral communication are those that satisfy the conditions of Proposition 2. For type $t_1$ (respectively, $t_2$), we only need to have $u(f(t_1))$ (respectively, $u(f(t_2))$) smaller or equal to $u(f(t_3))$. For type $t_3$, the value of the following linear program has to be greater or equal to 1:

$$\begin{aligned} \text{Max} \quad & x_1 + x_2 \\ \text{s.t.} \quad & x_1 \leq \frac{u(f(t_1))}{u(f(t_3))} \\ & x_2 \leq \frac{u(f(t_2))}{u(f(t_3))} \\ & x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

That is equivalent to the following condition: $u(f(t_1)) + u(f(t_2)) \geq u(f(t_3))$. In conclusion, $f$ is implementable in bilateral communication if and only if:

$$\max\{u(f(t_1)), u(f(t_2))\} \leq u(f(t_3)) \leq u(f(t_1)) + u(f(t_2)).$$

Because $\mathcal{C}$ does not satisfy normality, bilateral communication allows the implementation of more outcome functions than unilateral communication. If the certification structure is *normalized*, i.e. if the event $c^*(t_3) = \{t_3\}$ is added to $\mathcal{C}$, it is easy to check that $f$ is implementable in unilateral (respectively, bilateral) communication if and only if $u(f(t_3)) \geq \max\{u(f(t_1)), u(f(t_2))\}$.

# 4 The value of bilateral communication

We know that bilateral communication enlarges the set of implementable outcome functions if and only if the certifiability structure $\mathcal{C}$ does not satisfy normality (see Proposition 4). In this section, we establish sufficient conditions for bilateral communication to (strictly) increase the principal's expected payoff. Assume the action space $A$ is a subset of $\mathbb{R}$ (with $a_0 = \min A$ and $A \neq \{a_0\}$), the agent's utility function $u$ is increasing, and both $u$ and $v$ are continuous (on any interval $I \subseteq A$). Frameworks where the principal chooses a reward, a salary or a budget allocation for the agent fit this description.

**Definition 6.** An outcome function $f$ is *weakly evidence compatible* if

$$\forall t, \forall t', \text{ if } u(f(t')) < u(f(t)) \text{ then } t' \notin c^*(t).$$

Recall that an outcome function $f$ is evidence compatible if every type $t$ can certify an event that no type with an outcome worse than $f(t)$ can certify. Weak evidence compatibility only requires that for every type $t$, no type with an outcome worse than $f(t)$ can certify all events in $\mathcal{C}(t)$. Note that if $\mathcal{C}$ satisfies normality, both notions are equivalent.

**Remark 1.** If an outcome function $f$ is implementable in bilateral communication then it is weakly evidence compatible. Indeed, if $f$ is not weakly evidence compatible then there exist two types $t$ and $t'$ such that $t'$ is in $c^*(t)$ and $u(f(t')) < u(f(t))$ and therefore the incentive compatibility constraint of $t'$ is violated because he can perfectly mimic $t$. However, weak evidence compatibility is not sufficient for implementation (see Example 3).

As a consequence of this observation and Proposition 4, there exist outcome functions that are weakly evidence compatible but not evidence compatible if $\mathcal{C}$ does not satisfy normality.

**Definition 7.** The principal's utility function $v$ is *single-plateau* if for every type $t$ there exists $\underline{a}_t$ and $\overline{a}_t$ such that $v(t, \cdot)$ is strictly increasing before $\underline{a}_t$, constant between $\underline{a}_t$ and $\overline{a}_t$, and strictly decreasing after $\overline{a}_t$, i.e. for any action $a$ in $[\underline{a}_t, \overline{a}_t]$, $v(t,a) = v(t, \underline{a}_t) = v(t, \overline{a}_t)$ and for all actions $a'$ and $a''$ in $A$:

$$\text{if } a'' < a' \leq \underline{a}_t \text{ or } \overline{a}_t \leq a' < a'' \text{ then } v(t, a'') < v(t, a').$$

If in addition $\underline{a}_t = \overline{a}_t = a^*_t$ for every type $t$, then $v$ is *single-peaked* at $a^*$.

**Example 2.** In the hiring process example, assume the agent has (regardless of $t$) a quadratic disutility of work: if he works $h$ hours, his disutility is $\frac{h^2}{2}$. If the wage he obtains is $a$ then his surplus is $h(a - \frac{h}{2})$. Therefore the optimal number of hours for the agent is $h = a$. Let $s_t$ be the gross hourly surplus that an agent of type $t$ generates. The principal's payoff if she hires type $t$ at an hourly wage $a$ is therefore $v(t, a) = a(s_t - a)$. It is single-peaked at $\frac{s_t}{2}$.

Let $A^*_t = \arg\max_{a \in A} v(t, a)$ and if it is nonempty for all $t$, i.e. if $v(t, \cdot)$ reaches its maximum in $A$ for all $t$, let $F^*(v)$ denote the set of first-best outcome functions:

$$F^*(v) = \{f : T \to \Delta(A) | f(t) \in \Delta(A^*_t)\},$$

If $v$ is single-plateau, $A^*_t = [\underline{a}_t, \overline{a}_t]$ and the principal wants his action to be in $A^*_t$ if the agent's type is $t$, or as close as possible to this interval.

**Proposition 5.** *Bilateral communication strictly increases the principal's expected payoff if*

*(i) For all $t$, $A^*_t$ nonempty and $a_0 \notin A^*_t$.*

*(ii) No first-best outcome $f^*$ in $F^*(v)$ is evidence compatible.*

*(iii) There exists a weakly evidence compatible first-best outcome $f^*$ in $F^*(v)$.*

*If randomization over actions is not allowed the result holds if in addition, $v$ is single-plateau and $A$ is an interval.*

**Proof.** See Appendix. □

Proposition 5 gives sufficient conditions for bilateral communication to improve the outcome for the principal in comparison with unilateral communication. Such an improvement results only from the principal being active in the communication phase given that the same constraints on information certification apply.

Condition (*i*) guarantees the existence of at least one first-best outcome function, which would be implemented if the principal can observe the agent's type. It also states that, regardless of the agent's type, punishment is not optimal. Under this condition, if $f$ is an optimal outcome function in unilateral communication then $u(f(t)) > 0$ for all $t$. As a consequence, the threat of punishment can be used to increase the principal's expected payoff through bilateral communication (regardless of the specifics of utility functions and type distribution).

Condition (*ii*) guarantees that no first-best outcome function $f^*$ is implementable in unilateral communication (by Proposition 3). This condition is necessary for bilateral communication to improve the outcome for the principal.

Condition (*iii*) states that there exists a weakly evidence compatible first-best outcome function $f^*$. This implies that $f^*$ would be implementable in unilateral communication if the certification structure is *normalized*, i.e. if the maximal evidence events $\{c^*(t)\}_{t \in T}$ are added to $\mathcal{C}$. In other words, $f^*$ can be implemented if the principal can ask the agent to certify all events in $\mathcal{C}(t)$ when he reports type $t$. Under this condition, the constraint on the amount of evidence that can be certified is the reason that the principal is unable to achieve her first-best in unilateral communication. Note that under conditions (*ii*) and (*iii*), $\mathcal{C}$ does not satisfy normality (see Definition 6).

In general, the result depends on the possibility of randomization over actions. But if $v$ is single-plateau and $A$ is an interval, it holds even if randomization is not allowed: instead of improving the outcome by finding an implementable function with a higher probability of choosing an optimal action, we can simply choose a closer action to the interval of optimal actions.

**Example 3.** Consider an employer (principal) who wants to design a hiring process for a job at her firm. There are different profiles (types in the set $T$) of applicants that fit the description of this job. However, these profiles are not equally valued by the employer due to differences in productivity. The action space is $\mathbb{R}_+$: she chooses the wage at which she is willing to hire an applicant (agent). The punishment action is to reject the application, i.e. to choose a wage equal to 0. The applicant wants the highest possible wage. The hiring process is subject to a time limit which implies that a limited amount of information (about the applicant's skills) can be verified. Therefore, a certifiability structure $\mathcal{C}$ can be defined.

If the principal wants to implement an allocation $f$, where $f(t)$ is the wage for type $t$, she can use canonical form bilateral communication mechanism. The hiring process starts when the agent reports a type $t$ by sending his curriculum vitae (which describes his profile). The principal then asks him to certify an element of $\mathcal{C}$ by testing his abilities in certain tasks and/or by asking for third party certifications (such as diplomas). If the applicant passes the test and/or provides the required certificates, he is hired at wage $f(t)$. Otherwise, he is not hired (punishment action). For simplicity, we choose to preclude randomization over actions because they represent wages. A similar analysis can be conducted if randomization is allowed.

Let $T = \{t_1, t_2, t_3\}$ and $\mathcal{C} = \{\{t_1, t_3\}, \{t_2, t_3\}\}$: there are two skills and three possible types with the possibility to verify only one skill during the hiring process. The first type masters the first skill, the second masters the other, while the third masters both skills. Note that $\mathcal{C}$ does not satisfy normality: $c^*(t_3) = \{t_3\}$ is not certifiable.

Let $u(a) = a$: the agent's utility is equal to his wage. Assume that the employer's utility $v$ single-peaked at $a^*$. We give an instance where this condition is satisfied in Example 2 (a change of variable would allow us to have $u(a) = a$ and keep $v$ single-peaked).

As established in Example 1, $f$ is implementable in unilateral communication if and only if $f(t_3) = \max\{f(t_1), f(t_2)\}$ and implementable in bilateral communication if and only if $\max\{f(t_1), f(t_2)\} \leq f(t_3) \leq f(t_1) + f(t_2)$. We can easily check that $f$ is weakly evidence compatible if and only if $f(t_3) \geq \max\{f(t_1), f(t_2)\}$.

Let $a_3^*$ be strictly larger than $a_1^*$ and $a_2^*$: the first-best wage for $t_3$ is strictly higher than the first-best wages for $t_1$ and $t_2$. That means $a^*$ is weakly evidence compatible but not evidence compatible. If in addition, $a_k^* > 0$ for all $k$ then all conditions of Proposition 5 are satisfied.

Therefore, bilateral communication strictly increases the principal's payoff in comparison with unilateral communication. In the remainder, we examine how the payoff increase is achieved. Assume w.l.o.g that $a_1^* = \min_k a_k^*$.

If $a^*$ is such that $a_3^* \leq a_1^* + a_2^*$ then it is implementable in bilateral communication. Otherwise, consider $f$ optimal in unilateral communication. We have $f(t_3) = \max\{f(t_1), f(t_2)\}$. Moreover, $f(t_k)$ is in $[a_1^*, a_3^*]$ for all $k$ because $f$ is optimal (in unilateral communication) and $v$ is single-peaked. We show how to construct a function $\hat{f}$ that gives the principal a higher payoff than $f$.

If $f$ is such that $f(t_1) \leq f(t_2) = f(t_3) < a_3^*$, define $\hat{f}$ such that $\hat{f}(t_k) = f(t_k)$ for $k$ in $\{1, 2\}$ and

$$\hat{f}(t_3) = \min\{f(t_1) + f(t_2), a_3^*\}.$$

$\hat{f}$ is implementable in bilateral communication. Given that $f(t_1) \geq a_1^* > 0$, it follows that $f(t_3) < \hat{f}(t_3) \leq a_3^*$. Consequently, $\hat{f}$ gives a strictly higher payoff to the principal than $f$ (because $v$ is single-peaked).

If $f$ is such that $f(t_1) \leq f(t_2) = f(t_3) = a_3^*$, define $\hat{f}$ such that $\hat{f}(t_k) = f(t_k)$ for $k$ in $\{1, 3\}$ and

$$\hat{f}(t_2) = \max\{f(t_3) - f(t_1), a_2^*\}.$$

$\hat{f}$ is implementable in bilateral communication. We have $a_2^* \leq \hat{f}(t_2) < f(t_3)$. Therefore, $\hat{f}$ gives a strictly larger payoff to the principal than $f$ (because $v$ is single-peaked). Similar arguments apply if $f(t_2) < f(t_1) = f(t_3)$.

# 5 Extensions

## 5.1 Tightness of Proposition 5

In this section, we focus on the conditions of the main result. Note that if $A^*_t$ is empty for a given type $t$, these conditions are not well defined. For that reason, we assume that $A^*_t$ is nonempty for all $t$ and we show that if any of the other conditions of Proposition 5 is dropped we can construct an example where the result does not hold, i.e. where the outcome is not improved by bilateral communication. In all examples we have $u(a) = a$, $T = \{t_1, t_2, t_3\}$ and $\mathcal{C} = \{\{t_1, t_3\}, \{t_2, t_3\}\}$ as in Example 3. In the general case, i.e. when randomization is allowed, we choose $A = \mathbb{R}_+$ and single-peaked piecewise linear $v$ which ensures that randomization over actions does not improve the outcome for the principal.

### 5.1.1 Non-optimal punishment condition

Let $a^*$ be such that $0 = a^*_1 < a^*_2 < a^*_3$. $a^*$ is weakly evidence compatible but not evidence compatible. Let $v(t_1, a) = -2a$ and $v(t_k, a) = -|a - a^*_k|$ for $k$ in $\{2, 3\}$ and assume types are uniformly distributed.

Let $f$ be an optimal outcome function in bilateral communication. Implementation conditions (see Example 3) ensure that $\max\{f(t_1), f(t_2)\} \leq f(t_3) \leq f(t_1) + f(t_2)$. If $f(t_1) = 0$, then we necessarily have $f(t_2) = f(t_3)$ and it is in $[a^*_2, a^*_3]$ (because $v$ is single-peaked at $a^*$). Such a function is also implementable in unilateral communication. If $f(t_1) > 0$ then $f(t_3)$ must be equal to $f(t_1) + f(t_2)$ and below $a^*_3$: otherwise we would increases the principal's payoff by lowering $f(t_1)$. It follows that the principal's optimization problem is

$$\begin{aligned} \text{Max} \quad & v(t_1, a_1) + v(t_2, a_2) + v(t_3, a_1 + a_2) \\ \text{s.t.} \quad & a_1 \geq 0, a_2 \geq 0, \\ & a_1 + a_2 \leq a_3^*. \end{aligned}$$

Under these constraints, we have

$$v(t_1, a_1) + v(t_2, a_2) + v(t_3, a_1 + a_2) = -a_1 + a_2 + v(t_2, a_2) - a_3^*.$$

It follows that $a_1 = 0$ at the optimum. Therefore, $f(t_1)$ cannot be strictly positive. In conclusion, the optimal $f$ is implementable in unilateral communication and the outcome is not improved by bilateral communication.

### 5.1.2 Evidence compatibility conditions

It is obvious that if a first-best outcome function is evidence compatible then the result does not hold because it is implementable in unilateral communication. Condition (*ii*) is necessary for the result to hold.

Let $v(t_k, a) = -|a - a_k^*|$ for all $k$ with $0 < a_1^* < a_3^* < a_2^*$. Condition (*iii*) is not satisfied in this case because $a^*$ is the unique first-best and is not weakly evidence compatible. If $f$ is optimal in bilateral communication then we necessarily have $f(t_1) = a_1^*$ and $f(t_2) = f(t_3) \in [a_3^*, a_2^*]$, thus it is implementable in unilateral communication. Bilateral communication cannot improve the outcome.

### 5.1.3 Without randomization over actions

Now we assume randomization over actions is not allowed and we show that if $A$ is not an interval or $v$ is not single-plateau, we can construct an example where the result does not hold.

**The action space is not an interval:** Let $v(t_k, a) = -|a - a_k^*|$ for all $k$ with $0 < a_1^* < a_2^* < a_3^*$ and $a_1^* + a_2^* < a_3^*$. If $A = [0, a_2^*] \cup [a_3^*, +\infty)$, it is easy to see that if $f$ is optimal in bilateral communication then $f(t_1) = a_1^*$ and $f(t_2) = f(t_3) \in \{a_2^*, a_3^*\}$. Therefore, it is also implementable in unilateral communication.

**The principal's utility is not single-plateau:** Assume types are uniformly distributed. Let $A = \mathbb{R}_+$ and $v(t_k, a) = -2|a - a_k^*|$ for $k$ in $\{1, 2\}$ with $0 < a_1^* < a_2^*$. However, let $v(t_k, \cdot)$ be increasing strictly increasing. Let there be $a_3^*$ strictly higher than $a_1^* + a_2^*$. Let $v(t_3, \cdot)$ be strictly increasing for $a < a_2^*$, constant between $a_2^*$ and $a_1^* + a_2^*$, strictly increasing between $a_1^* + a_2^*$ and $a_3^*$, and strictly decreasing for $a > a_3^*$. Assume in addition that $v(t_3, \cdot)$ is piecewise linear with a slope equal to 1 (in absolute value) outside the interval $[a_2^*, a_1^* + a_2^*]$. $a^*$ is the unique first-best and it satisfies the conditions of Proposition 5, i.e. it is weakly evidence compatible but not evidence compatible and $a_k^* > 0$ for all $k$. But $v$ is not single-plateau.

The optimal outcome function in unilateral communication is $f$ such that $f(t_1) = a_1^*$ and $f(t_2) = f(t_3) = a_2^*$. Any outcome function $\hat{f}$ such that $\hat{f}(t_1) = a_1^*$, $\hat{f}(t_2) = a_2^*$ and $\hat{f}(t_3) \in [a_2^*, a_1^* + a_2^*]$ is implementable in bilateral communication and gives the same expected payoff to the principal as $f$. To see that such $\hat{f}$ is optimal, we need to observe that in order to have $v(t_3, \tilde{f}(t_3)) > v(t_3, \hat{f}(t_3))$, we need to have $\tilde{f}(t_3) > a_1^* + a_2^*$. If $\tilde{f}$ is implementable, $\tilde{f}(t_3) \leq \tilde{f}(t_1) + \tilde{f}(t_2)$ and therefore:

$$\tilde{f}(t_3) - (a_1^* + a_2^*) \leq (\tilde{f}(t_1) - a_1^*) + (\tilde{f}(t_2) - a_2^*).$$

Given the chosen function $v$, namely the fact that $v(t_1, \cdot)$ and $v(t_2, \cdot)$ have a twice larger slope than $v(t_3, \cdot)$, it follows that $\tilde{f}$ decreases the principal's payoff in comparison with $f$.

## 5.2 Optimal solutions: Randomization over actions and Commitment

An outcome function $f$ is optimal if it maximizes the principal's expected payoff among the set of implementable outcome functions. In unilateral communication, Sher [2011] shows that there exists an optimal function that does not involve randomization over actions and can be implemented without the principal's commitment if the following concavity assumption is satisfied: the principal's utility function is a type-dependent concave transformation of the agent's utility function. Hart et al. [2016] show that commitment is unnecessary for a class of certifiability structures satisfying a stronger condition than normality if the principal's preferences are strongly single-peaked, that is every convex combination of elements of $\{v(t, \cdot)\}_{t \in T}$ is single-peaked.

In this section, we show that under the conditions of Proposition 5 and the concavity assumption stated above, randomization over actions is not needed if the action space is an interval and we give an example with a discrete action space where it is necessary. We also give an example where commitment is necessary at the optimum under the same conditions.

We focus on settings where the conditions of Proposition 5 are satisfied because it guarantees that bilateral communication is beneficial and it is then interesting to study the properties of optimal solutions in comparison with unilateral communication.

### 5.2.1 Randomization over actions

The result of Sher [2011] about randomization holds whether the principal's actions space $A$ is continuous or discrete. We show that it holds in bilateral communication if $A$ is an interval and we give an example that illustrates the need for randomization for discrete $A$.

**Proposition 6.** *Assume the conditions of Proposition 5, $A$ is an interval and the following concavity assumption are satisfied: for all $t$, there exists a concave function $c_t$ such that $v(t,\cdot) = c_t(u(\cdot))$. Then there exists an optimal outcome function in bilateral communication $f$ such that for all $t$, $f(t) \in A$.*

**Proof.** Consider an outcome function $f$. For a given type $t$, the outcome $f(t)$ is a distribution over actions. Let $\mathbb{E}_{f(t)}(u)$ be the expected utility of an agent under the lottery $f(t)$. Because $u$ is continuous over $A$, there exists an action $\widehat{f}(t) \in A$ such that.

$$u(\widehat{f}(t)) = \mathbb{E}_{f(t)}(u)$$

This defines a deterministic outcome function $\widehat{f}$. If $f$ is implementable then $\widehat{f}$ is also implementable (because the agent's expected utilities are identical for both outcome functions). Now we compare the principal's utilities under $f$ and $\widehat{f}$ when she faces an agent of type $t$.

$$\begin{aligned}
\mathbb{E}_{f(t)}(v(t,\cdot)) &= \mathbb{E}_{f(t)}(c_t(u)) \\
&\leq c_t(\mathbb{E}_{f(t)}(u)) \quad \text{(concavity of } c_t) \\
&\leq c_t(u(\widehat{f}(t))) = v(t, \widehat{f}(t)).
\end{aligned}$$

The principal is therefore (weakly) better off not randomizing over actions. The conclusion follows. □

As we can see in the proof above, the fact that $A$ is an interval plays an essential role in the argument. In the following example, we consider a discrete actions space and we find that randomization is necessary at the optimum.

**Example 4.** Consider a setting where $u(a) = a$, $T = \{t_1, t_2, t_3\}$ and $\mathcal{C} = \{\{t_1, t_3\}, \{t_2, t_3\}\}$ as in Example 3. Assume types are uniformly distributed, $A = \{0, 1, 3\}$ and let the principal's utility function be given by the following table:

| $v(t,a)$ | 0 | 1 | 3 |
|---|---|---|---|
| $t_1$ | 0 | 1 | 0 |
| $t_2$ | 0 | 1 | 0 |
| $t_3$ | 0 | 1 | 3 |

Note that the concavity assumption is satisfied. The best deterministic outcome function in this case is one such that type $t_3$ receives the action 3 along with one of the other two types, while the remaining type receives the outcome 1. The utility of the principal for such a function is $V = 4$. This function is not optimal though. The optimal solution assigns action 3 to type $t_3$ and the same randomized outcome to types $t_1$ and $t_2$ such that action 1 has a probability $\frac{3}{4}$ and action 3 has a probability $\frac{1}{4}$. The optimal payoff of the principal is $V = \frac{9}{2}$.

### 5.2.2 Commitment

We now give an example where commitment is necessary at the optimum in bilateral communication under the conditions of Proposition 5 and the concavity assumption.

**Example 5.** Consider the same framework as Example 1 with the utility functions of Example 2. Namely, $T = \{t_1, t_2, t_3\}$, $\mathcal{C} = \{\{t_1, t_3\}, \{t_2, t_3\}\}$, $A = [0, +\infty)$, $u(a) = \frac{a^2}{2}$ and $v(t, a) = a(s_t - a)$ with $s_t$ the surplus that type $t$ generates. Assume in addition that $s_{t_1} = s_{t_2} = 1$ and $s_{t_3} = 2$. $v$ is single peaked at $a^* = (\frac{1}{2}, \frac{1}{2}, 1)$ which is weakly evidence compatible but not evidence compatible. Also, punishment is not optimal regardless of the agent type. All conditions of Proposition 5 are satisfied. In addition the concavity assumption is satisfied: for every type $t$, $v(t, \cdot) = c_t(u(\cdot))$, where $c_t(x) = s_t\sqrt{2x} - 2x$. In fact, the strict concavity ensures that there can be no randomization over actions at the optimum.

It follows from the characterization given in Example 1 that $f$, such that $f(t_k) = a_k$ for all $k$, is implementable if and only if

$$\max\{a_1^2, a_2^2\} \leq a_3^2 \leq a_1^2 + a_2^2.$$

The optimal solution is such that $a_1 = a_2 = \frac{1+\sqrt{2}}{4}$ and $a_3 = \frac{2+\sqrt{2}}{4}$. The principal needs commitment in order to implement this outcome function because these actions are not rational given her beliefs at the time of implementation: for example, when choosing action $a_3 = \frac{2+\sqrt{2}}{4}$, she knows that the agent's type is $t_3$ and her rational decision would be $a = 1$.

# 6 Appendix

**Proof of Proposition 1.** Bull and Watson [2007] show that if $f$ is implemented in a general mechanism, then it is also implemented in a special three-stage mechanism characterized by $g : T \times M \times \mathcal{C} \to \Delta(A)$ and $\sigma : T \to \Delta(M)$ with truthful reporting at stage 1. $g(t, m, C)$ is the outcome when the agent reports $t$, the principal sends message $m$ and the agent certifies $C$. $\sigma(t, m)$ is the probability that the principal sends the message $m$ if the agent reports $t$. Therefore, for every type $t$ and every message $m$, there must exist an event $C_{t,m}$ in $\mathcal{C}(t)$ such that the outcome $f(t)$ is implemented whenever the agent announces $t$, the principal sends $m$ and the agent shows $C$. Formally:

$$\forall t, \forall m, \exists C_{t,m} \in \mathcal{C}(t) \text{ such that } g(t, m, C_{t,m}) = f(t)$$

For every type $t$, let $\phi_t$ be a mapping from messages $m$ to events $C_{t,m}$ :

$$\forall m, \phi_t(m) \in \mathcal{C}(t) \text{ and } g(t, m, \phi_t(m)) = f(t)$$

Incentive compatibility constraints are given by:

$$\forall t, \forall t', \sum_m \sigma(t', m) \max_{C \in \mathcal{C}(t)} u(g(t', m, C)) \leq u(f(t)).$$

Consider the mechanism $\widehat{\sigma}$ and $\widehat{g}$ defined by:

- $\forall t, \forall C$, $\widehat{\sigma}(t,C) = \sum_{m \in \phi_t^{-1}(C)} \sigma(t,m)$.
- $\forall t, \forall C$, $\widehat{g}(t,C,C) = f(t)$.
- $\forall t, \forall C' \neq C$, $\widehat{g}(t,C,C') = a_0$.

We can easily check that $\forall t, \sum_{C \in \mathcal{C}(t)} \widehat{\sigma}(t,C) = 1$. Note that this is a description of the three-stage $(\widehat{\sigma}, f)$-mechanism. In order to prove that $\widehat{\sigma}$ and $\widehat{g}$ implement $f$, we check that incentive compatibility constraints are satisfied. First, using the definition of $\widehat{\sigma}$, we have:

$$\forall t, \forall t', \sum_C \widehat{\sigma}(t',C) \max_{C' \in \mathcal{C}(t)} u(\widehat{g}(t',C,C')) = \sum_C \sum_{m \in \phi_{t'}^{-1}(C)} \sigma(t',m) \max_{C' \in \mathcal{C}(t)} u(\widehat{g}(t',C,C'))$$

By definition, if $m \in \phi_{t'}^{-1}(C)$ then $g(t',m,C) = g(t') = \widehat{g}(t',C,C)$, and for $C' \neq C$, we have $u(g(t',m,C')) \geq u(a_0) = u(\widehat{g}(t,C,C'))$. Therefore

$$\forall t, \forall t', \sum_C \widehat{\sigma}(t',C) \max_{C' \in \mathcal{C}(t)} u(\widehat{g}(t',C,C')) \leq \sum_C \sum_{m \in \phi_{t'}^{-1}(C)} \sigma(t',m) \max_{C' \in \mathcal{C}(t)} u(g(t',m,C'))$$

The r.h.s term is equal to $\sum_m \sigma(t',m) \max_{C \in \mathcal{C}(t)} u(g(t',m,C))$ and we finally get:

$$\forall t, \forall t', \sum_C \widehat{\sigma}(t',C) \max_{C' \in \mathcal{C}(t)} u(\widehat{g}(t',C,C')) \leq u(f(t)).$$

Which proves that $\widehat{\sigma}$ implements $f$. □

**Proof of Lemma 1.** Given the structure of the $(\sigma, f)$-mechanism, $\sigma$ implements $f$ with truthful reporting if and only if $(i)$ the support of $\sigma(t)$ is in $\mathcal{C}(t)$ and $(ii)$ no type has an incentive to misreport.

Let $\sigma_{t't} = \sum_{C \in \mathcal{C}(t')} \sigma(t,C)$ denote the probability for an agent of type $t'$ to successfully persuade the principal that he is of type $t$. Using this notation, $(i)$ states that for all $t$,

$\sigma_{tt} = 1$. $(ii)$ describes the incentive compatibility constraints and states that for all $t$ and $t'$, $\sigma_{t't} u(f(t)) \leq u(f(t'))$. Note that for $t$ such that $u(f(t)) = 0$, these conditions are satisfied and do not constrain the choice of $\sigma(t)$. Therefore, we can write $(ii)$ as follows:

$$\forall t, t' \sigma_{t't} \leq \frac{u(f(t'))}{u(f(t))},$$

with the right hand side equal to $+\infty$ if $u(f(t)) = 0$.

□

**Proof of Proposition 3.** Let $f$ be an outcome function. Recall the three statements:

(i) $f$ is implementable in unilateral communication.

(ii) $f$ is evidence compatible.

(iii) $f$ is implementable in deterministic bilateral communication.

In order to prove the equivalence, we will show the following implications: (i)⇒(ii)⇒(iii)⇒(i).

**(i)⇒(ii)** Consider an outcome function $f$ implementable in unilateral communication. Using the revelation principle, we know that there must exist a unilateral communication mechanism that implements it with truthful type reporting. For every type $t$, there must exist at least $C$ in $\mathcal{C}(t)$ such that if the agent reports $t$ and certifies $C$, the principal implements $f(t)$. Consequently, if $t'$ is such that $u(f(t')) < u(f(t))$ then $t' \notin C$: otherwise $t'$ benefits from deviating by reporting $t$ and certifying $C$. Therefore, $f$ is evidence compatible.

**(ii)⇒(iii)** If $f$ is evidence compatible, define the deterministic $\sigma$ such that for every $t$, $\sigma(t)$ is an element of $\mathcal{C}(t)$ that contains no type $t'$ with $u(f(t')) < u(f(t))$. It is guaranteed to exist by evidence compatibility. It is readily verified that $\sigma$ implements $f$.

**(iii)⇒(i)** If there exists a deterministic $\sigma$ that implements $f$, consider the following implementation rule in unilateral communication: if the agent reports type $t$ and certifies $\sigma(t)$ implement $f(t)$, otherwise implement $a_0$. This mechanism implements $f$.

□

**Proof of Proposition 5.** Let $f$ be an optimal outcome function in unilateral communication. Proposition 3 guarantees the existence of deterministic $\sigma$ that implements $f$ in bilateral communication. Our goal is to slightly modify the $(\sigma, f)$-mechanism so that we obtain an implementable outcome function $\hat{f}$ with a strictly higher expected payoff for the principal than $f$.

Let $f^*$ be a weakly evidence compatible element of $F^*(v)$ (it is guaranteed to exist by condition $(iii)$). Consider an indexing of types in $T$ from 1 to $n : T = \{t_1, \ldots, t_n\}$. Given that $A^*_{t_k}$ is nonempty for all $k$, let $\overline{a}_k$ (respectively, $\underline{a}_k$) denote its largest (respectively, smallest) element. Let the indexing be such that $u(f(t_1)) \leq u(f(t_2)) \leq \cdots \leq u(f(t_n))$ and if there exist $k$ and $l$ such that $u(f(t_{k-1})) < u(f(t_k)) = u(f(t_{k+1})) = \cdots = u(f(t_{k+l})) < u(f(t_{k+l+1}))$, rearrange the indexing so that $u(f^*(t_k)) \leq u(f^*(t_{k+1})) \leq \cdots \leq u(f^*(t_{k+l}))$.

For any two lotteries over actions $\mu$, $\mu'$ and any $\alpha \in [0, 1]$ let $L(\mu, \mu', \alpha) = (1-\alpha)\mu + \alpha\mu'$. If $u(f(t_k))$ is in $[u(\underline{a}_k), u(\overline{a}_k)]$ then $v(t_k, f(t_k))$ is necessarily maximal because $f$ is optimal. Otherwise $f(t_k)$ can be replaced with $L(\overline{a}_k, \underline{a}_k, \alpha)$ where $\alpha$ is such that $u(f(t_k)) = \alpha u(\underline{a}_k) + (1-\alpha)u(\overline{a}_k)$. This would make $v(t_k, f(t_k))$ maximal without affecting the evidence compatibility constrains of $f$.

Condition $(ii)$ implies that $f$ is not in $F^*(v)$ (because $f$ is evidence compatible). Therefore, using the previous observation, there must exist $\tilde{k}$ such that $u(f(t_{\tilde{k}}))$ is not in $[u(\underline{a}_{\tilde{k}}), u(\overline{a}_{\tilde{k}})]$, and as a consequence $v(t_{\tilde{k}}, f(t_{\tilde{k}})) < v(t_{\tilde{k}}, f^*(t_{\tilde{k}}))$. Let $J = \{j : u(f(t_j)) = u(f(t_{\tilde{k}}))\}$. It follows that there exist $l$, $l'$, $m$ and $m'$ such that

$$J = \{\tilde{k} - l, \cdots, \tilde{k} - l', \cdots, \tilde{k}, \cdots, \tilde{k} + m', \cdots, \tilde{k} + m\}$$

with:

$$\begin{cases} u(f^*(t_j)) < u(f^*(t_{\tilde{k}})) & \text{if } \tilde{k} - l \leq j < \tilde{k} - l' \\ u(f^*(t_j)) = u(f^*(t_{\tilde{k}})) & \text{if } \tilde{k} - l' \leq j \leq \tilde{k} + m' \\ u(f^*(t_j)) > u(f^*(t_{\tilde{k}})) & \text{if } \tilde{k} + m' < j \leq \tilde{k} + m \end{cases}$$

It follows from condition $(i)$ that $\underline{a}_k > a_0$ for all $k$. Therefore, $u(f(t_1)) > 0$: otherwise $f(t_1) = a_0$ and it would be possible to improve the outcome for the principal (and the agent incidentally) by replacing $f(t_1)$ with $L(a_0, a_1, \alpha)$ where $\alpha > 0$ is small enough for $f$ to remain evidence compatible (the same argument applies if more than one type receive the punishment action). Let $\varepsilon$ be such that $0 < \varepsilon < \min\{\frac{u(f(t_1))}{u(f(t_{\tilde{k}}))}, \frac{1}{2}\}$. We now construct an outcome function $\hat{f}$ that gives the principal a strictly higher expected payoff than $f$ by having $v(t, \hat{f}(t)) \geq v(t, f(t))$ for all $t$ and $v(t_{\tilde{k}}, \hat{f}(t_{\tilde{k}})) > v(t, f(t_{\tilde{k}}))$.

(I) If $u(f(t_{\tilde{k}})) > u(\overline{a}_{\tilde{k}})$ then we necessarily have $m > m'$: if $m$ is equal to $m'$, we can replace $f(t_j)$ with $L(f(t_j), f^*(t_j), \alpha_j)$ for every $j$ in $J$ with $\alpha_j > 0$ such that this lottery gives the agent a payoff equal to $\max\{u(f^*(t_{\tilde{k}})), u(f(t_{\tilde{k}-l-1}))\}$. This change would increase the principal's expected payoff without affecting the evidence compatibility of $f$, which contradicts the fact that $f$ is optimal in unilateral communication.

Let $\underline{J} = \{\tilde{k} - l, \cdots, \tilde{k} + m'\}$ and $\overline{J} = \{\tilde{k} + m' + 1, \cdots, \tilde{k} + m\}$. $\underline{J}$ contains $\tilde{k}$ and $\overline{J}$ is nonempty because $m > m'$. We also have $u(f^*(j)) > u(f^*(j'))$ for all $j \in \overline{J}$ and $j' \in \underline{J}$.

For every $j$ in $\overline{J}$, note that $t_{j'} \notin c^*(t_j)$ for all $j' \in \underline{J}$ because $f^*$ is weakly evidence compatible. Let $C_j$ denote the event $\sigma(t_j)$. $C_j$ may contain types $t_{j'}$ with $j' \in \underline{J}$ but cannot be certified by any type $t_k$ with $k < \tilde{k} - l$ (because $\sigma$ implements $f$).

Let $\hat{\sigma}$ be identical to $\sigma$ except for types $t_j$ with $j$ in $\overline{J}$. For each of these types and every $j'$ such that $j' \in \underline{J}$ and $t_{j'} \in C_j$ choose $C_{jj'}$ in $\mathcal{C}(t_j)$ that does not contain $t_{j'}$. Let $l_j$ be the number of $C_{jj'}$'s. If $l_j = 0$, let $\hat{\sigma}(t_j, C_j) = 1$. Otherwise, let $\hat{\sigma}(t_j, C_j) = 1 - \varepsilon$ and $\hat{\sigma}(t_j, C_{jj'}) = \frac{\varepsilon}{l_j}$ for each $C_{jj'}$. Let $\tilde{l} = \max_{j \in \overline{J}} l_j$.

Let $\hat{f}$ be identical to $f$ except for types $t_{j'}$ such that $j' \in \underline{J}$. For each of these types, let $\hat{f}(t_{j'}) = L(f(t_{j'}), f^*(t_{j'}), \alpha_{j'})$ with $\alpha_{j'} > 0$ such that this lottery gives the agent an expected payoff equal to $u(f(t_{\tilde{k}})) - \eta$ for some $\eta$ satisfying the following condition:

$$0 < \eta \leq u(f(t_{\tilde{k}})) - \max\{u(f^*(t_{\tilde{k}})), u(f(t_{\tilde{k}-l-1}))\}.$$

This outcome function gives the principal a strictly higher expected payoff than $f$. If $\tilde{l} = 0$, $\hat{f}$ would be evidence compatible and $f$ would not be optimal in unilateral communication. Therefore $\tilde{l} > 0$ and $\hat{\sigma}$ is not deterministic. Moreover, $\hat{\sigma}$ implements $\hat{f}$ if

$$\begin{aligned}(1 - \frac{\varepsilon}{\tilde{l}})u(f(t_{\tilde{k}})) &\leq u(f(t_{\tilde{k}})) - \eta \\ \varepsilon u(f(t_{\tilde{k}})) &\leq u(f(t_1))\end{aligned}$$

The first condition guarantees that types $t_{j'}$ such that $j' \in \underline{J}$ have no incentive to deviate and is satisfied for $\eta > 0$ small enough. The second condition guarantees that types below $t_{\tilde{k}-l}$ have no incentive to deviate and is satisfied by definition of $\varepsilon$.

(II) If $u(f(t_{\tilde{k}})) < u(\underline{a}_{\tilde{k}})$ then $l > l'$: if $l$ is equal to $l'$, we can replace $f(t_j)$ with $L(f(t_j), f^*(t_j), \alpha_j)$ for every $j$ in $J$ with $\alpha_j > 0$ such that this lottery gives the agent a payoff equal to $\min\{u(f^*(t_{\tilde{k}})), u(f(t_{\tilde{k}+m+1}))\}$. This change would increase the principal's expected payoff without affecting the evidence compatibility of $f$, which contradicts the fact that $f$ is optimal in unilateral communication.

Let $\underline{J} = \{\tilde{k}-l, \cdots, \tilde{k}-l'-1\}$ and $\overline{J} = \{\tilde{k}-l', \cdots, \tilde{k}+m\}$. $\underline{J}$ is nonempty because $l > l'$ and $\overline{J}$ contains $\tilde{k}$. We also have $u(f^*(j)) > u(f^*(j'))$ for all $j \in \overline{J}$ and $j' \in \underline{J}$ so that we construct $\hat{\sigma}$ in the same way as in (I).

Let $\hat{f}$ be identical to $f$ except for types $t_j$ with $j$ in $\overline{J}$. For each of these types, let $\hat{f}(t_j) = L(f(t_j), f^*(t_j), \alpha_j)$ with $\alpha_j > 0$ such that this lottery gives the agent an expected payoff equal to $u(f(t_{\tilde{k}})) + \eta$ for some $\eta$ satisfying the following condition:

$$0 < \eta \leq \min\{u(f^*(t_{\tilde{k}})), u(f(t_{\tilde{k}+l+1}))\} - u(f(t_{\tilde{k}})).$$

Similarly to (I), this outcome function gives the principal a strictly higher expected payoff than $f$ and we have $\tilde{l} > 0$ and $\hat{\sigma}$ non-deterministic. Moreover, $\hat{\sigma}$ implements $\hat{f}$ if

$$\begin{aligned}(1 - \frac{\varepsilon}{\tilde{l}})(u(f(t_{\tilde{k}})) + \eta) &\leq u(f(t_{\tilde{k}})) \\ \varepsilon(u(f(t_{\tilde{k}})) + \eta) &\leq u(f(t_1))\end{aligned}$$

These conditions are analogous to those of (I) and are satisfied for $\eta > 0$ small enough. This concludes the proof in the general case.

In order to prove the result when $v$ is single-plateau, $A$ is an interval and randomization over actions is not allowed, we simply have to replace $L(a, a', \alpha)$ with the action $(1-\alpha)a + \alpha a'$ for any actions $a$ and $a'$.

□

**Study of the case where $\mathbf{a_0}$ does not exist.** This happens when $\inf_{a \in A} u(a)$ is not attained.

If $\inf_{a\in A} u(a) = -\infty$ then the punishment can be as big as the principal wants. Formally, $f$ is implementable if and only if there exists $\alpha \in \mathbb{R}$ and $\sigma$ such that

$$\forall t, t' \quad \sigma_{t't} u(f(t)) - (1-\sigma_{t't})\alpha \leq u(f(t'))$$

First, note that if $t' \in c^*(t)$ then $\sigma_{t't}$ is necessarily equal to 1 which implies $u(f(t)) \leq u(f(t'))$. Consider the following mechanism : if the agent reports type $t$, ask for all events in $\mathcal{C}(t)$ with the same probability. Then $\forall t, t'$, if $t' \notin c^*(t)$, the above inequality is satisfied for $\alpha$ large enough. Because we have a finite number of such inequalities, we can take the largest $\alpha$ to satisfy all of them. We conclude that if $\inf_{a\in A} u(a) = -\infty$, $f$ is implementable if and only if

$$\forall t, t', \text{ if } u(f(t')) < u(f(t)) \text{ then } t' \notin c^*(t),$$

that is, $f$ weakly evidence compatible. If the punishment can be as large as we want, all weakly evidence compatible outcome functions are implementable and the limitation on information certification has no effect.

If on the other hand $\inf_{a\in A} u(a)$ is finite, we can set it to 0 w.l.o.g and denote by $a_\epsilon$ an action such that $u(a_\epsilon) = \epsilon$ for all $\epsilon > 0$. By continuity of $u$, such action always exists. In this case, $f$ is implementable if and only if there exists $\sigma$ such that

$$\begin{aligned} \exists \epsilon > 0 \text{ s.t } \forall t, t' \qquad & \sigma_{t't} u(f(t)) + (1-\sigma_{t't})\epsilon \leq u(f(t')) \\ \Leftrightarrow \forall t, t' \qquad & \text{if } u(f(t')) < u(f(t)) \text{ then } \sigma_{t't} u(f(t)) < u(f(t')) \end{aligned}$$

We conclude that $\sigma$ implements $f$ iff

$$\forall t, t', \text{ if } u(f(t')) < u(f(t)) \text{ then } \sigma_{t't} < \frac{u(f(t'))}{u(f(t))}$$

Implementation results follow from Lemma 1, where, in this context, certain inequalities are replaced with strict inequalities. The subsequent results still hold but have to be modified accordingly.

□

# References

Elchanan Ben-Porath and Barton L Lipman. Implementation with partial provability. *Journal of Economic Theory*, 147(5):1689–1724, 2012.

Jesse Bull. Mechanism Design with Moderate Evidence Cost. *The BE Journal of Theoretical Economics (Contributions)*, 8(5), 2008.

Jesse Bull and Joel Watson. Hard evidence and mechanism design. *Games Econ. Behav.*, 58: 75–93, 2007.

Raymond Deneckere and Sergei Severinov. Mechanism design with partial state verifiability. *Games and Economic Behavior*, 64(2):487–513, 2008.

Françoise Forges and Frédéric Koessler. Communication equilibria with partially verifiable types. *Journal of Mathematical Economics*, 41(7):793–811, 2005.

Jacob Glazer and Ariel Rubinstein. Debates and decisions: On a rationale of argumentation rules. *Games and Economic Behavior*, 36(2):158–173, 2001.

Jacob Glazer and Ariel Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6): 1715–1736, 2004.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1:395–410, 2006.

Jerry R. Green and Jean-Jacques Laffont. Partially verifiable information and mechanism design. *Review of Economic Studies*, 53(3):447–456, 1986.

Sergiu Hart, Ilan Kremer, and Motty Perry. Evidence games: Truth and commitment. *American Economic Review*, Forthcoming, 2016.

Navin Kartik and Olivier Tercieux. Implementation with evidence. *Theoretical Economics*, 7 (2):323–355, 2012.

Frederic Koessler and Eduardo Perez-Richet. Evidence based mechanisms. *mimeo*, 2014.

Barton L. Lipman and Duane J. Seppi. Robust inference in communication games with partial provability. *Journal of Economic Theory*, 66:370–405, 1995.

Roger B. Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics*, 10(1):67–81, June 1982.

Itai Sher. Credibility and determinism in a game of persuasion. *Games Econ. Behav.*, 71: 409–419, 2011.

Itai Sher. Persuasion and dynamic communication. *Theoretical Economics*, 9:99–136, 2014.

Itai Sher and Rakesh Vohra. Price discrimination through communication. *Theoretical Economics*, 10:597–648, 2015.

Nirvikar Singh and Donald Wittman. Implementation with partial verification. *Review of Economic Design*, 6(1):63–84, 2001.

Roland Strausz. Mechanism design with partially verifiable information. *mimeo*, 2016.