

Truth Elicitation in a Nonstrategic Environment[†]

By JACOB GLAZER AND ARIEL RUBINSTEIN*

A principal decides which party deserves a prize. Only the parties observe the state that determines who is deserving. The principal gives each party a text describing the conditions under which that party deserves the prize and asks each to report the state. The parties are nonstrategic. Each activates a known cheating procedure whose outcome depends on the state and on the text. Implementation is achieved if, whenever the parties disagree, the liar is identified by the “one-way cheating principle”: The truthful party is the unique party who could not have produced his story had the rival’s story been true. (JEL D82, D91)

Two invigilators, B and G, overheard a student receiving a whispered message from another student during an exam. The invigilators have not seen the exam questions but would be able to solve them if they did. The exam includes multiple questions, but only one refers to the variable α and reads as follows: “Solve the equation $\alpha + 1 = 4$.” The student answered the question correctly. Invigilator B claims that the whispered message was “ $\alpha = 3$.” This is a serious allegation; if B is correct, the student’s exam will be disqualified. Invigilator G claims that the whispered message was “Solve the equation $\alpha + 1 = 4$ first.” If G is correct, then the whispered message did not reveal the answer, and there will be no serious consequences. Who should be believed, B or G?

Although there is no definitive proof for either claim, we would choose to believe G. The reason is that if the message were “Solve the equation $\alpha + 1 = 4$ first,” then B could solve the equation himself and claim that the message was “ $\alpha = 3$.” By contrast, if the message were “ $\alpha = 3$,” it is very unlikely that G, who had not seen the exam questions, could guess that the equation solved by $\alpha = 3$ was $\alpha + 1 = 4$ rather than some other equation with the same solution. Hence, there is an asymmetry between the two conflicting claims, which makes it reasonable to conclude that G’s claim is the truthful one.

Asymmetries between statements are often used in practice to decide which of two conflicting statements is true. For example, scholars of old manuscripts who

* Glazer: University of Warwick and Tel Aviv University (email: glazer@tauex.tau.ac.il); Rubinstein: Tel Aviv University and New York University (email: rariel@tauex.tau.ac.il). Authors are listed alphabetically in this paper’s byline. Qingmin Liu was the coeditor for this article. We thank Shlomo Naeh (Department of Jewish Thought, The Hebrew University of Jerusalem) for alerting us to the principle of *lectio difficilior potior*. We also thank Kfir Eliaz, Avner Shaked, Rani Spiegler, Kemal Yildiz, referees of this paper, and especially Áron Tóbiás for their comments. The authors received no financial support for this research and declare that they have no relevant or material financial interests related to the research described in this paper.

[†]Go to <https://doi.org/10.1257/aer.20251879> to visit the article page for additional materials and author disclosure statement(s).

have before them two versions of the same text, only one of which can be genuine, use such asymmetries as a tool to decide which version is the original. A principle known as *lectio difficilior potior* instructs scholars to prefer text A over text B if text B can be seen as a simplification of text A, but not vice versa.

Another natural asymmetry involves word associations; see Michelbacher, Evert, and Schütze (2007). For example, if two parties disagree about which university a certain professor graduated from, one claiming that it is MIT and the other that it is NIT, then we would tend to believe that the professor graduated from NIT.

In the cases above, the asymmetry between the two claims is built in rather than engineered by someone seeking to uncover the truth. In other cases, one might design a mechanism that creates an asymmetry between a truth-teller and a cheater in order to identify the truth-teller with reasonable certainty. The design of such a mechanism is at the core of our analysis.

The approach taken in this paper is related to the biblical Judgment of Solomon. In that story, two women claim to be the mother of the same baby, and the king must decide which woman is telling the truth. There is an asymmetry in the women's preferences regarding the potential consequences of the ruling: The true mother prefers giving up the baby to the other woman over having the baby cut in two, while the fake mother has the opposite preference. King Solomon designs a mechanism that exploits this asymmetry in order to identify the true mother by observing the mothers' simultaneous responses to his order to cut the baby.

Assuming that the true mother is willing to pay more for the baby than the fake mother, Glazer and Ma (1989), in a result later generalized by Perry and Reny (1999), construct a game form with the feature that, regardless of which woman is the true mother, the induced extensive game has a unique subgame perfect equilibrium in which the true mother receives the baby without making any payment.

However, King Solomon's mechanism is different: It relies on the mothers not reasoning strategically. Had the false mother known the king's mechanism and reasoned strategically, she would probably have done better by imitating the true mother's strategy. King Solomon—"the wisest man who ever lived"—thus gained his reputation by being able to predict how the two mothers would respond to the mechanism he had designed. In the wake of King Solomon, this paper proposes a framework for constructing mechanisms that elicit the truth from two parties with opposing interests by exploiting the principal's knowledge of the procedure they use when they cheat.

We consider situations of the following kind: Two parties claim a prize offered by a principal. The principal's view is that only one of them truly deserves the prize, and that party's identity is determined unequivocally by facts, the "state," known to the parties but not to the principal. The parties do not know the principal's view; nonetheless, both insist that they deserve the prize. There are no other asymmetries between them.

The approach taken in this paper is that, unlike telling the truth, effective cheating is often not a simple task and requires the use of some "cheating procedure." We crucially assume that the principal is aware of the parties' cheating procedures and can therefore sometimes design a mechanism in which the parties are asked, "What is the state?" The principal can then identify the liar solely on the basis of their responses without observing an actual "smoking gun."

We propose the following scheme: The principal provides each party with a text describing the circumstances, the states, under which that party deserves the prize. Each party then submits a description of a state that he claims is the true state. The model is enriched with (i) a *language* that the principal can use to compose such texts and (ii) a *cheating procedure* that determines which state a party announces, given the text he has received and the realized state he has observed. The cheating procedure is common to both parties and is known to the principal. Both the language and the cheating procedure are situation specific. For now, we leave their details vague; nonetheless, each situation analyzed below will include a formal specification.

We refer to our concept of implementation as *magical implementation*. A magical implementation mechanism has three stages.

Stage 1.—The principal provides each party with a complete and truthful description of the set of states in which that party deserves the prize. Constrained by a language, many texts may describe this set, and the particular text presented to each party is chosen at the principal's discretion.

Stage 2.—Each party must submit a statement, in the form of a state, which may be true or false, after being informed that (i) if his statement does not justify his claim for the prize, he will certainly not receive it and (ii) if both statements justify his claim, he will receive the prize. The party is not informed about the outcome if both his statement and the other party's statement justify their respective claims for the prize.

Stage 3.—The principal considers the statements submitted by the two parties and makes his decision as follows:

- If both reported states satisfy the same party's text, the principal awards the prize to that party.
- If both parties submit statements justifying their own claims for the prize, then the principal checks whether there is a *unique* party such that, if his statement is true, then the cheating procedure might have enabled the other party to make the statement he made. In this case, the prize is awarded to that party.
- In all other cases, the prize is awarded to neither party.

In other words, the principal provides each party with an accurate description of the conditions under which he deserves the prize. The principal awards the prize if the two parties agree on who deserves it or if he can apply what we call the "*one-way cheating principle*," which is that the truth lies with the party whose statement has the following property: If the statement is true, then the other party's statement could have been obtained from it by applying the cheating procedure, but not vice versa.

Before we formalize the concept of magical implementation, we wish to emphasize several points. Our core idea is the use of the one-way cheating principle, which makes it possible to identify who is the truth-teller and who is the liar when two parties provide conflicting descriptions of the facts. The mechanism relies on the principal being aware of the procedure used by the parties to cheat when they

need to do so. Its success also relies on the parties' ignorance of how the principal will resolve a dispute and on their lack of strategic reasoning. These assumptions might not be reasonable if the parties use advisers or are repeatedly involved in the same situation. However, we find them natural and appropriate in many other common situations.

I. The Formal Framework

Parties 1 and 2 are in a dispute over a single, indivisible prize. Let S be the set of all possible states of the world. The set S is partitioned into two subsets, W^1 and W^2 , where W^i is the set of states in which party i should win the prize. In every state, both parties are informed of the realized state but do not know the partition that determines who deserves the prize. We refer to the tuple $\langle S, W^1, W^2 \rangle$ as an *implementation problem*.

A principal who is not informed of the realized state must rely on the parties in order to award the prize correctly. He constructs a pair of texts, T^1 and T^2 , where T^i is the text provided to party i . We interpret a text as a description of the circumstances in which the party receiving the text deserves the prize. In choosing the texts, the principal uses a *language*, $\mathfrak{L} = \langle \mathfrak{T}, \text{Int} \rangle$, where $\mathfrak{T}$ is a set of *texts* and Int is an interpretation function that assigns to each $T \in \mathfrak{T}$ a set $\text{Int}(T)$ of states in which T is true. Note that different texts may have the same interpretation.

The principal provides the text T^i to party i and asks him to declare the realized state. He informs party i that he will receive the prize if both parties' declared states satisfy T^i and that he will certainly not receive the prize if his declared state does not satisfy T^i . However, the principal does not disclose what will happen in other circumstances, in particular if both parties declare states that satisfy their respective texts.

After receiving a text T , a party sends the principal a message in the form of a state in S . It is assumed that if the realized state s satisfies the text T (i.e., $s \in \text{Int}(T)$), then the party sends the message s . Otherwise, he applies a *cheating procedure*, formalized as a binary relation $\rightarrow_T$ on S . The statement $s \rightarrow_T t$ is interpreted as follows: "If the realized state is s and s does not satisfy T , then the cheating procedure *may* lead the party to claim t , which does satisfy T ." A given state s may be associated with multiple states t for which $s \rightarrow_T t$. We assume that both parties apply the same cheating procedure that is fixed in their minds and is not subject to their choice.

A pair of texts (T^1, T^2) magically implements the problem $\langle S, W^1, W^2 \rangle$ if, assuming that the parties use the cheating procedure, whenever a dispute arises, the principal can apply the one-way cheating principle and correctly infer which party is telling the truth.

Formally, (T^1, T^2) *magically implements* $\langle S, W^1, W^2 \rangle$ if the following two conditions hold:

- $\text{Int}(T^i) = W^i$ for both i . That is, the principal provides each party with a *correct* description of the circumstances under which that party deserves the prize.
- For any two states s and t , if $s \rightarrow_{T^i} t$ (and hence $s \in W^j$ and $t \in W^i$), then $t \not\rightarrow_{T^j} s$.

Discussion.—The Term “Magical Implementation.”—Magicians possess skills that enable them to discern subtle patterns and behavioral cues in human actions, which they can exploit to create the illusion of a miracle. Analogously, our principal exploits his understanding of human imperfections in order to identify which of two rival parties is telling the truth.

Nonstrategic Reasoning: Our analysis does not involve a game. First, the principal provides each party with a text that truthfully describes the circumstances in which that party deserves the prize and commits not to award the prize to that party if his claim does not meet the conditions described in the text. However, the principal does not inform the parties of what will happen if their claims conflict. Unlike standard implementation, the complete outcome rule is not announced to the parties. Our use of the term “implementation” means that, under the stipulated information structure and procedural behavior, the principal always awards the prize correctly.

Second, the parties do not think strategically; that is, they do not take the other party’s actions into consideration. Each party acts as a “problem solver”: He is aware that he will certainly not receive the prize if he does not solve the problem, but he is not told that, even if he successfully solves it, he may still fail to receive the prize. This assumption is particularly appealing when the parties are tightly limited in the time they can spend on deliberation.

Note that if the parties were aware of the principal’s inference method and reasoned strategically, then following the prescribed reporting behavior would not necessarily constitute an *ex post* Nash equilibrium of the induced reporting game. More precisely, when $s \in W^i$, we assume that party j declares a state t such that $s \rightarrow_{T^j} t$, even though $t \not\rightarrow_{T^i} s$. But in that case, party j could do better if he found a state r such that $r \rightarrow_{T^i} s$ and $s \not\rightarrow_{T^j} r$ and, by falsely reporting r , persuaded the principal that party i was the cheater. In some states, party j may not be able to find such a state. However, assume that magical implementation exists and that the cheating procedure is effective in the sense that, for every text T with $\text{Int}(T) \neq \emptyset$ and every state $s \notin \text{Int}(T)$, there exists $t \in \text{Int}(T)$ such that $s \rightarrow_T t$. Under these assumptions, there exists a state in which the cheating party would prefer not to follow the cheating procedure and instead report another state that causes the principal to award him the prize. (To see the logic, consider a particular mechanism that magically implements a problem $\langle S, W^1, W^2 \rangle$. Let $s \in W^1$. By effectiveness, there exists $t \in W^2$ such that $s \rightarrow_{T^2} t$, and there exists $r \in W^1$ such that $t \rightarrow_{T^1} r$. Then, at t , party 1 would do better by reporting s rather than following the cheating procedure and reporting r .)

Comparison to Other Implementation Concepts: The one-way cheating principle differs from the basic idea behind standard Nash implementation à la Maskin (1999) in environments with at least three agents. There, the principal draws an inference from a situation in which one agent is an outlier in his report, given a consensus among the other agents about the state, only if that agent’s move leads to an outcome that would be worse for him were the consensus correct.

Standard Nash implementation cannot achieve our objective. Since the parties’ preferences are independent of the state, any fixed game form induces the same strategic game in every state. Its set of Nash equilibrium outcomes is therefore state independent and cannot coincide with the state-dependent allocation specified by $\langle S,$

W^1, W^2). The trick of magical implementation is the design of two texts that, given the cheating procedure, create an asymmetry between a cheater and a truth-teller.

Notice the difference between our approach and two other branches of the implementation literature. In papers such as Green and Laffont (1986); Lipman and Seppi (1995); Glazer and Rubinstein (2006); and Ben-Porath, Dekel, and Lipman (2019), agents are constrained in the lies they can tell about the state. In Kartik and Tercieux (2012), cheating is assumed to involve a cost. In both literatures, the constraints on cheating are given and are not affected by the mechanism. In contrast, in our setting, the constraint is determined by the text given to the party and by the cheating procedure.

Extensions: Our framework can be extended in several directions. For example, consider the case with n parties rather than two, each of whom claims that he deserves the prize although only one does. In this case, magical implementation is carried out by designing a text T^i for each party i such that $\text{Int}(T^i) = W^i$. Magical implementation requires that for every state $t^i \in W^i$ and for each array of statements of the other parties $(t^j)_{j \neq i}$ such that $t^i \rightarrow_{T^i} t^j$ for all $j \neq i$, it should not be possible to explain the entire profile of statements as arising from some t^l where $l \neq i$; that is, for every $l \neq i$, there is a party $k \neq l$ such that $t^l \not\rightarrow_{T^k} t^k$.

In addition, our definition of implementation does not require the parties to have the same cheating procedure, as long as the principal knows their procedures. All we need is to define the binary relation representing the cheating procedure with a party-specific index. Let $s \rightarrow_{T^i}^i t$ mean that party i 's cheating procedure, given text T , may lead him from a realized state s to report t . The definition of magical implementation would then require two texts T^1 and T^2 , satisfying $\text{Int}(T^i) = W^i$, such that for every two states s and t , if $s \rightarrow_{T^i}^i t$ (and thus $s \in W^j$ and $t \in W^i$), then $t \not\rightarrow_{T^j}^j s$.

II. Setting a Trap through Associations

A. A Motivating Story

A village in DrSeussLand is populated by Yooks and Zooks. A person was seen planting a bomb. The police arrest eight suspects in the neighborhood: five Yooks (their names are denoted by y^1, y^2, y^3, y^4, y^5) and three Zooks (their names are denoted by z^1, z^2, z^3). It is certain that one of the suspects is the terrorist. There are two witnesses to the event, one a Yook and the other a Zook. The witnesses are able to recognize the terrorist. They do not know whether he, or any other suspect, is a Yook or a Zook, because Yooks and Zooks cannot be distinguished by appearance or name.

The witnesses are reluctant to identify the terrorist because he might belong to their own group, and such an action would be seen as a betrayal. For a Yook, "getting the prize" means that a Zook is charged with the crime, and for a Zook, "getting the prize" means that a Yook is charged.

The authorities are determined to identify the terrorist. They know which group each of the suspects belongs to and are aware of the witnesses' reluctance to turn in a member of their own group. They construct two websites, one for the Yook witness

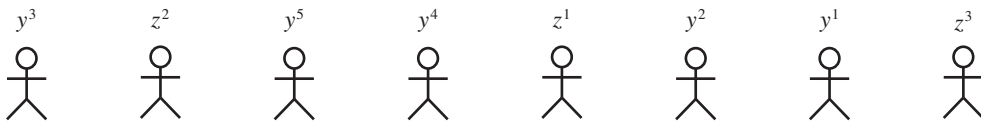


FIGURE 1. THE FIRST PAGE OF BOTH WEBSITES

and the other for the Zook witness. On the first page of each site, they insert the pictures and names of the eight suspects. Recall that a suspect's name does not reveal whether he is a Yook or a Zook. (See Figure 1.)

Each website contains numerous pages, each displaying the names and pictures of two suspects. We consider each page as a means to trigger an association between the two suspects featured on it. In practice, this trigger could be created by emphasizing a prominent characteristic shared by the two suspects. To ensure that a party associates B after considering A, one might include in the description of A a prominent detail that is shared exclusively with B.

Panel A of Figure 2 describes the Yook's website, while panel B describes the Zook's. An edge in a graph means that the two connected suspects appear on one of the pages of the corresponding website. Both witnesses are obliged to name a suspect. The activity on both websites is confidential, and the authorities cannot monitor it.

The Yook witness is informed that a suspect is a Zook if and only if he appears on at least two pages of his website. Similarly, the Zook witness is informed that a suspect is a Yook if and only if he appears on at least two pages of his website. Both witnesses know that the information is accurate, since the authorities are prohibited from cheating.

The authorities are aware of the witnesses' cheating procedure. Each witness naively searches for a suspect who belongs to the other group (i.e., his name appears on at least two pages of his website). He can use a search engine in the process. The witness starts by checking the group identity of the actual terrorist, and if the terrorist does not belong to the opposite group (i.e., his name appears only once on the website), the witness repeats the process starting with the other name that appears with the terrorist on the only page on which the terrorist's name and picture are displayed.

To demonstrate the authorities' scheme, assume that the terrorist is z^1 . The Yook starts by searching for the name z^1 and finds that it appears on four pages, meaning that the terrorist is a Zook, and thus he can happily announce that he has identified z^1 as the terrorist. The Zook also starts by searching for the name z^1 and discovers that on his website the name appears on only one page, together with y^2 . He concludes that he has to cheat. Hoping that y^2 appears on at least one more page, he continues by searching for y^2 and finds that y^2 appears on his website more than twice and is therefore a Yook. He then announces that the terrorist is y^2 . Thus, both witnesses identify a suspect from the other group. The DrSeussLand authorities now face a dilemma. Nevertheless, they conclude that z^1 is the terrorist. Their logic is based on an understanding of the witnesses' cheating procedure. They know that if the terrorist is z^1 (as the Yook claims), the Yook would claim that the terrorist is z^1 , and

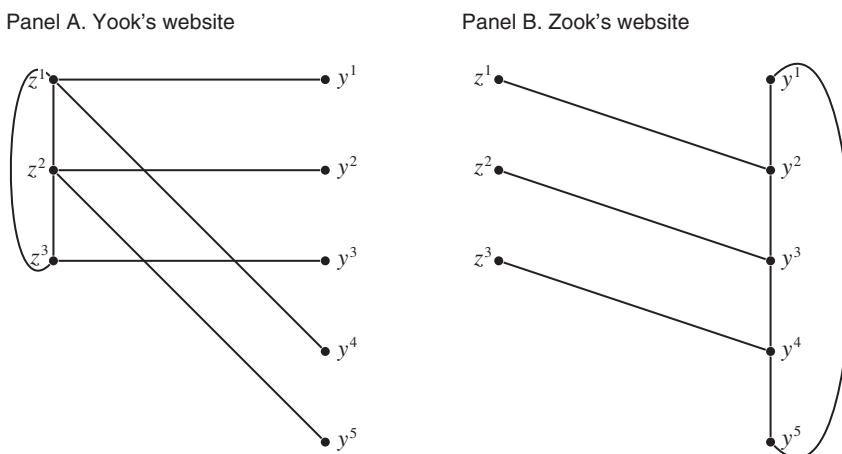


FIGURE 2. THE WITNESSES' WEBSITES

Note: Each edge represents a web page displaying the pictures and names of two paired suspects.

the Zook would claim that the terrorist is y^2 . They also know that if the terrorist is y^2 (as the Zook claims), then the Yook witness would announce z^2 rather than z^1 . This is because the Yook witness would start from y^2 , and after discovering that y^2 is not a Zook, would continue by searching for the name z^2 (suspect y^2 's partner on the only page of the Yook's website on which y^2 appears). The asymmetry leads the authorities to apply the one-way cheating principle and conclude that the terrorist is z^1 and not y^2 . It is easy to verify that this magic works regardless of the identity of the terrorist.

B. The General Case

The Implementation Problem.—The set of states S is finite and partitioned into W^1 and W^2 , each containing at least two states.

The Language.—A text $T(D)$ is characterized by a set D of doubletons of states in S and has the following form: “You deserve the prize if the state of nature is a member of at least two doubletons in D .”

The Cheating Procedure.—A party is endowed with a technology that provides him with answers to questions of the type $Q(D, s)$: “Which sets in D contain s ?” Denote by $A(D, s)$ the answer to the question $Q(D, s)$, which could be either

- none,
- one doubleton $\{s, t\}$ (which contains s), or
- a set of at least two doubletons (each containing s).

We assume that a party who receives a text $T(D)$ and observes the state s activates the following procedure:

- (i) Start by asking the question $Q(D, s)$.
- (ii) If $|A(D, s)| > 1$, declare s .
- (iii) If $A(D, s)$ consists of a single doubleton $\{s, t\}$, ask the question $Q(D, t)$.
 - (a) If $|A(D, t)| > 1$, declare t .
 - (b) Otherwise—that is, if $A(D, s) = \emptyset$, or $A(D, s) = A(D, t) = \{\{s, t\}\}$ —repeat the process starting with an arbitrary state x for which the question $Q(D, x)$ has not previously been asked.
- (iv) If all states in S have been exhausted without finding an element that belongs to two doubletons in D , give up and declare s .

Notice that unless no state satisfies the text $T(D)$, this procedure will always end with the party finding a state that satisfies it. This is the reason why in the present framework, the principal cannot achieve his goal with a single party with the present language and cheating procedure.

The formal description of the cheating procedure is the relation defined by $s \rightarrow_{T(D)} t$ if either

- $A(D, s) = \{\{s, t\}\}$ and $|A(D, t)| > 1$, or
- $|A(D, t)| > 1$ and either $A(D, s) = \emptyset$ or both $A(D, s) = \{\{s, x\}\}$ and $|A(D, x)| = 1$.

In the first option, the party asks the question $Q(D, s)$ and discovers that s appears only in $\{s, t\}$. He is then nudged to ask $Q(D, t)$ and discovers that t satisfies the text.

In the second option, the party starts with $Q(D, s)$ and then is stuck, either because s does not appear in any doubleton in D or because it appears only once, with another state that also appears only once. In either case, the party moves on to explore an arbitrary state not previously explored and ends up with some other state t that appears in at least two doubletons.

CLAIM 1: *Let $\langle S, W^1, W^2 \rangle$ be a problem such that each W^i contains at least two states. If the principal is equipped with the above language and both parties use the above cheating procedure, then the problem is magically implementable.*

PROOF:

Enumerate the states of W^1 and W^2 as $z^1, \dots, z^K$ and $y^1, \dots, y^L$, respectively. Without loss of generality, assume that $K \leq L$. Form a sequence $x^1, \dots, x^{2L}$, starting with y^1 and alternating between states in W^2 and states in W^1 , and let $x^{2L+1} = x^1$. The states of W^2 appear in their given order. The states of W^1 appear cyclically, if necessary, in their given order. For example, if $K = 3$ and $L = 5$, then $(x^1, \dots, x^{10}) = (y^1, z^1, y^2, z^2, y^3, z^3, y^4, z^1, y^5, z^2)$.

For every state $s \in S$, define $h(s)$ to be the state that follows the first appearance of s in the sequence. Notice that some states in W^1 appear more than once when $K < L$. The key feature of this construction is that $h(s) = s'$ implies $h(s') \neq s$. Indeed, the state following the first appearance of s' belongs to the same winning set as s , but is different from s , by the construction of the sequence and the assumption that each W^i contains at least two states.

We now construct two texts, $T(D^1)$, assigned to party 1, and $T(D^2)$, assigned to party 2.

The set D^1 consists of the doubletons $\{z^1, z^2\}, \{z^2, z^3\}, \dots, \{z^K, z^1\}$ and the doubletons $\{y, h(y)\} | y \in W^2\}$.

The set D^2 consists of the doubletons $\{y^1, y^2\}, \{y^2, y^3\}, \dots, \{y^L, y^1\}$ and the doubletons $\{z, h(z)\} | z \in W^1\}$.

Figure 2 illustrates the construction of the texts $T(D^1)$, on the left, and $T(D^2)$, on the right, when $W^1 = \{z^1, z^2, z^3\}$ and $W^2 = \{y^1, y^2, y^3, y^4, y^5\}$. For any $y \in W^2$, the set $A(D^1, y)$ consists of the single doubleton $\{y, h(y)\}$. Therefore, y does not satisfy $T(D^1)$. The state $h(y)$ belongs to W^1 and satisfies $|A(D^1, h(y))| > 1$. Thus, $y \rightarrow_{T(D^1)} h(y)$. On the other hand, $|A(D^2, y)| \geq 2$, and therefore y satisfies $T(D^2)$. Similarly, every $z \in W^1$ satisfies $T(D^1)$, whereas $A(D^2, z)$ consists of the single doubleton $\{z, h(z)\}$. Therefore, z does not satisfy $T(D^2)$. Since $h(z) \in W^2$ and $|A(D^2, h(z))| > 1$, we have $z \rightarrow_{T(D^2)} h(z)$.

By construction, if $y \rightarrow_{T(D^1)} z$, then $z = h(y)$. Since $h(h(y)) \neq y$, we have $h(z) \neq y$, and hence $z \not\rightarrow_{T(D^2)} y$. Therefore, there are no $z \in W^1$ and $y \in W^2$ such that both $y \rightarrow_{T(D^1)} z$ and $z \rightarrow_{T(D^2)} y$.

It follows that $(T(D^1), T(D^2))$ magically implements the problem. ■

The intuition underlying the construction of the two texts is as follows. If $s \in W^i$, then party i will find that at least two doubletons in D^i contain s , and he will declare s . Party j starts with s and finds that only one doubleton $\{s, t\}$ is in D^j . He then asks about t , ascertains that it belongs to two doubletons in D^j , and declares t . But party j has fallen into a trap! The principal can apply the one-way cheating principle and conclude that j is the cheater: Given the state s , party j could cheat by declaring t , whereas if the state were t , party i could not cheat by declaring s .

III. Setting a Trap through a Logical Riddle

This section investigates the concept of magical implementation when the language used by the principal and the parties' cheating procedure are similar to those discussed in Glazer and Rubinstein (2012) in the context of a *single-agent* implementation problem.

A. A Motivating Example

The rector of a university, a magician in his spare time and known for his eccentric academic opinions, is consulting with two professors, Pro and Con, about whether to appoint professor G to a prestigious university chair. Pro and Con have each interviewed G. The rector knows that even though they agree on G's academic merits, Pro firmly supports the appointment, while Con is vehemently against it.

TABLE 1—THE PROPOSITIONS IN THE TEXTS GIVEN TO PRO AND CON

Professor Pro	Professor Con
$\neg E \wedge L \rightarrow \neg P$	$L \wedge \neg P \rightarrow E$
$E \wedge L \rightarrow P$	$\neg E \wedge P \rightarrow L$
$E \wedge P \rightarrow \neg L$	$\neg L \wedge \neg P \rightarrow \neg E$
$E \wedge \neg L \rightarrow \neg P$	
$\neg E \wedge \neg L \rightarrow P$	

The rector forms his opinion on the basis of the truth or falsity of three statements: E = “G is a genuine scholar of economics,” L = “G is a genuine scholar of law,” and P = “G is a genuine scholar of psychology.” The professors do not know which combinations of the facts will induce the eccentric rector to approve the appointment.

The rector could simply ask the professors to state the facts (on which they agree, as mentioned), but he fears that they may misrepresent the facts because of their strong personal bias for or against G. He suspects that they might not be above using deception in order to “save the university from what they consider to be a catastrophic decision.”

The rector meets Pro and Con separately and asks them whether each of the three statements— E , L , and P —is true or false. Coding “true” as 1 and “false” as 0, there are eight possible configurations of answers (states): 000, 100, . . . , 111 (e.g., 101 stands for E and P being true and L being false).

The rector is not a fan of multidisciplinary experts, and he strongly believes that the chair holder should have only one specialization. Thus, he considers the combinations 100, 010, and 001 as necessary and sufficient for G to be appointed.

Before asking the three questions, the rector (who, as you will recall, is a magician in his spare time) provides each of the professors with a text consisting of a list of propositions that truthfully characterizes the states in which the rector’s view is aligned with the view of the professor who receives the text. In other words, the text that Pro (Con) receives describes the states in which the rector is in favor of (against) the nomination. The rector can do this in a number of ways, and his design of the texts takes into account his knowledge of how each professor will respond to a particular text.

The propositions in each of the texts take the form $A \wedge B \rightarrow C$, interpreted as “if the antecedent is true—that is, if A and B are true—then the consequent (i.e., C) must also be true.” For instance, the proposition $\neg E \wedge L \rightarrow \neg P$ requires that “if G is not a scholar in economics and is a scholar in law, then he should not be a scholar in psychology.” Each of the texts starts with the sentence “You are right if your report satisfies the following propositions.” Table 1 presents the list of propositions given to Pro and Con by the rector.

As mentioned, the rector knows the cheating procedure each professor applies if he finds that the true state does not satisfy all the propositions in his text. In that case, the professor searches for a state that satisfies the text by advancing along a path. He starts with the true state. He finds a proposition (there could be more than one) that the state violates and switches the truth value of its consequent to obtain a new state. If the modified state satisfies the text, then he reports it. Otherwise, he continues the process with the new state until he reaches a state that appeared previously along the

path. He then returns to the true state and looks for another path until he runs out of possibilities, in which case he gives up and announces the true state.

Notice that each proposition in both of the texts involves all three variables and has two effects: It rules out one state and directs the professor to a different (unique) state. For example, the proposition $E \wedge \neg L \rightarrow \neg P$ on Pro's list rules out state 101, and if Pro considers that state, it directs him to consider 100.

To see how the mechanism works, assume, for example, that G is an expert in economics (E) and law (L) but not in psychology (P). The state $E, L, \neg P$ (i.e., 110) satisfies all three propositions on Con's list, and therefore Con reports the truth. This state violates (only) the proposition $E \wedge L \rightarrow P$ on Pro's list, leading Pro to consider the state E, L, P (111), which violates (only) the proposition $E \wedge P \rightarrow \neg L$ on his list. Pro then considers the state $E, \neg L, P$ (101), which violates (only) the proposition $E \wedge \neg L \rightarrow \neg P$ on his list. He finally reaches the state $E, \neg L, \neg P$ (100), which satisfies all the propositions on his list, and he therefore reports it to the rector.

The rector now faces two opposing statements: Con's statement— $E, L, \neg P$ (110)—and Pro's statement— $E, \neg L, \neg P$ (100). He decides (correctly) against the nomination by applying the one-way cheating principle: Pro's statement is the output of the cheating procedure if Con's statement is true. On the other hand, if Pro's statement is true, then Con would declare $\neg E, \neg L, \neg P$ (000) instead of what he actually declared (the state $E, \neg L, \neg P$ (100) violates only the proposition $\neg L \wedge \neg P \rightarrow \neg E$ on Con's list, and in that case, Con would be directed to the state $\neg E, \neg L, \neg P$ (000), which satisfies his text).

Panel A of Figure 3 presents the implementation problem. Each node represents a state. The red dots denote the states that support Pro's position, and the black dots denote those that support Con's. Panel B of Figure 3 illustrates the rector's magic. The red arrows correspond to the propositions in Pro's list, and the black arrows correspond to the propositions in Con's list. The graph helps us to verify that the pair of texts presented in the table indeed magically implements the rector's goal.

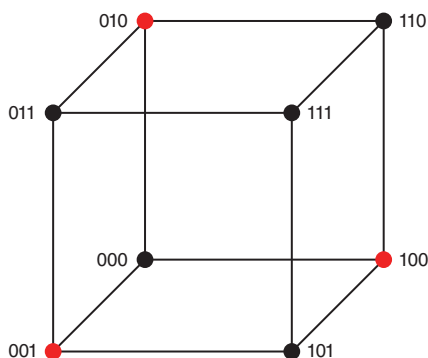
B. The General Case

The Implementation Problem.—Let $S = \{0, 1\}^K$, where $K \geq 3$. A state $x_1 x_2 \dots x_K$ (shorthand for $(x_1, \dots, x_K)$) is a vector representing the truth values of the propositional variables $v_1, \dots, v_K$, where $x_k = 1$ indicates that the variable v_k is true and $x_k = 0$ indicates that it is false. Two states s and t are neighbors, denoted by sNt , if they differ in exactly one component. Assume that each of the sets W^1 and W^2 contains at least two states.

The Language.—A text is characterized by a set of propositions in propositional logic, Φ , each of which uses some of the variables $v_1, \dots, v_K$. Each proposition has the form $\bigwedge_{v \in V} \phi_v \rightarrow \phi_z$, where V is a nonempty subset of variables, z is a variable that is not in V , every ϕ_v is either v (the variable v) or $\neg v$ (the negation of v), and ϕ_z is either z or $\neg z$. Such a proposition should be interpreted as follows: "If the state satisfies the antecedent of the proposition $\bigwedge_{v \in V} \phi_v$, then it should also satisfy its consequent, ϕ_z ." A proposition is *complete* if all K variables appear in it. A complete proposition excludes one state that satisfies the antecedent but not the consequent.

A text $T(\Phi)$ has the form "You deserve the prize if the state satisfies all the propositions in Φ ."

Panel A. The implementation problem



Panel B. Lists of propositions

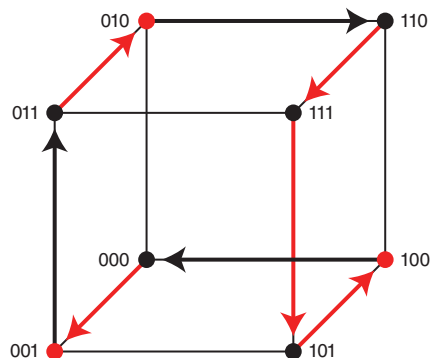


FIGURE 3. THE IMPLEMENTATION PROBLEM AND THE RECTOR'S TEXTS

Notes: In panel A, red dots denote states supporting Pro's position, and black dots denote states supporting Con's position. In panel B, red and black arrows are added and represent propositions on Pro's and Con's lists, respectively. For example, the red arrow from 101 to 100 represents the proposition $E \wedge \neg L \rightarrow \neg P$ on Pro's list.

One additional constraint on a text is the condition of *coherence* described in Glazer and Rubinstein (2012): The text should not include two propositions such that their antecedents do not contradict (i.e., no variable v appears in the antecedents once as v and once as $\neg v$) but their consequents do (i.e., the same variable z appears in the consequents of both propositions—in one case as z and in the other as $\neg z$). For example, a text that includes the two propositions, $v_1 \rightarrow v_3$ and $v_2 \rightarrow \neg v_3$, is not coherent.

The Cheating Procedure.—The fundamental step in the cheating procedure of a party who received a text $T(\Phi)$ is represented by a binary relation $\triangleright_{T(\Phi)}$ on S . The statement $s \triangleright_{T(\Phi)} t$ indicates that if the party reaches s , then he may consider t . It is required that there be a proposition in Φ that s violates, and t is obtained from s by switching only the truth value of the variable in that proposition's consequent.

Formally, $s \triangleright_{T(\Phi)} t$ if

- $s \notin \text{Int}(T(\Phi))$ (i.e., s violates at least one of the propositions in Φ),
- t differs from s in only one variable, and
- there exists a proposition $\phi \in \Phi$ such that s and t satisfy ϕ 's antecedent and s does not satisfy ϕ 's consequent, while t does.

The cheating procedure executes the following algorithm:

- (i) Start with the true state s^* .
- (ii) Advance along a path of distinct states $s_1 = s^* \triangleright_{T(\Phi)} s_2 \triangleright_{T(\Phi)} \cdots \triangleright_{T(\Phi)} s_m$.
 - (a) If you reach a state that satisfies the text, announce that state.
 - (b) If you cannot advance any further without returning to a state already on the path, look for another path starting from s^* .
- (iii) If no such path exists, announce s^* .

The procedure induces a binary relation $\rightarrow_{T(\Phi)}$ on S defined by $s \rightarrow_{T(\Phi)} t$ if there exists a sequence of at least two states $s_1 = s, s_2, \dots, s_m = t$ such that $s_1, \dots, s_{m-1}$ do not satisfy $T(\Phi)$, t does, and $s_l \triangleright_{T(\Phi)} s_{l+1}$ for $l = 1, \dots, m-1$.

CLAIM 2: *Let $\langle S, W^1, W^2 \rangle$ be an implementation problem, where $K \geq 3$, $S = \{0, 1\}^K$, and both W^1 and W^2 contain at least two states. If the principal is equipped with the above language and both parties use the above cheating procedure, then the implementation problem is magically implementable.*

PROOF:

Recall that a Hamiltonian cycle of the hypercube graph on S is an enumeration $x_1, \dots, x_{2^K}$ of all states of S such that $x_k N x_{k+1}$ for all k (and $x_{2^K} N x_1$). We first show that there is a Hamiltonian cycle with more than one block of W^1 states (equivalently, with more than one block of W^2 states).

For every $k = 1, \dots, K$ and $\delta \in \{0, 1\}$, let $S_{k,\delta}$ be the set of all states s such that $s_k = \delta$. Since W^2 contains at least two states, there is a dimension k such that both $S_{k,0}$ and $S_{k,1}$ contain a state from W^2 . If one of these two sets contains at least two states from W^1 , we may assume without loss of generality that $S_{1,1}$ contains at least two states from W^1 and one from W^2 , while $S_{1,0}$ contains at least one state from W^2 .

Otherwise, each of $S_{k,0}$ and $S_{k,1}$ contains at most one state from W^1 . Since W^1 contains at least two states, it then contains exactly two states, one in each set. In this case, we interchange the roles of W^1 and W^2 . Since $K \geq 3$, each of $S_{k,0}$ and $S_{k,1}$ contains at least two states from the new W^1 and one state from the new W^2 . Thus, after relabeling, we may assume again that $S_{1,1}$ contains at least two states from W^1 and one from W^2 , while $S_{1,0}$ contains at least one state from W^2 .

Construct a Hamiltonian cycle C_1 of $S_{1,1}$. Suppose first that C_1 contains more than one W^1 block. Choose two consecutive states a and b on C_1 , and let a' and b' be their respective neighbors in $S_{1,0}$. Since every edge of a hypercube belongs to a Hamiltonian cycle, there is a Hamiltonian cycle C_0 of $S_{1,0}$ containing the edge $\{a', b'\}$. Delete the edges $\{a, b\}$ and $\{a', b'\}$ from C_1 and C_0 , respectively, and add the edges $\{a, a'\}$ and $\{b, b'\}$. The resulting graph is a Hamiltonian cycle of S . Moreover, replacing the edge $\{a, b\}$ by a path whose internal vertices lie in $S_{1,0}$ cannot merge two distinct W^1 blocks of C_1 . Hence, the resulting Hamiltonian cycle still contains more than one W^1 block.

Suppose now that C_1 contains only one W^1 block. Since $S_{1,1}$ contains at least two states from W^1 , there are two consecutive W^1 states a and b on C_1 . Let a' and b' be their respective neighbors in $S_{1,0}$, and choose a Hamiltonian cycle C_0 of $S_{1,0}$ containing the edge $\{a', b'\}$. Delete the edges $\{a, b\}$ and $\{a', b'\}$ from C_1 and C_0 , respectively, and add the edges $\{a, a'\}$ and $\{b, b'\}$. Equivalently, start with b , follow C_1 until reaching a , move to a' , follow the Hamiltonian path obtained from C_0 by deleting the edge $\{a', b'\}$, and finally move from b' to b .

The resulting Hamiltonian cycle contains a W^2 block in $S_{1,1}$ and a distinct W^2 block in $S_{1,0}$. The two connecting states a and b belong to W^1 , so the W^2 states in the two half cubes cannot belong to the same block. Hence, the cycle contains more than one W^2 block and, equivalently, more than one W^1 block. This establishes the existence of the required Hamiltonian cycle.

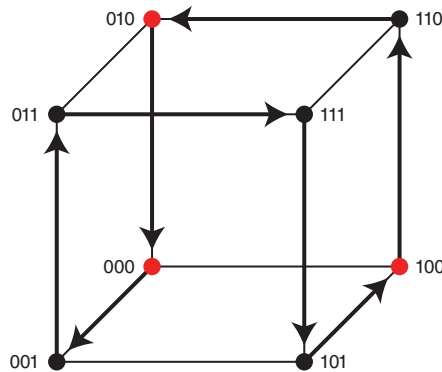


FIGURE 4. A HAMILTONIAN CYCLE WITH AT LEAST TWO BLOCKS OF EACH COLOR ($A = 110$ AND $B = 111$)

Figure 4 demonstrates the construction starting with the Hamiltonian cycle ($a = 110, b = 111, 101, 100$) of $S_{1,1}$.

We can now prove the claim. Let $x_1, \dots, x_{2^k}$ be a Hamiltonian cycle of S that has at least two blocks of W^1 states (and thus also of W^2 states). Let $\rightarrow$ be the binary relation on S defined by $s \rightarrow t$ if t appears right after s in the cycle. Define $s \rightarrow^1 t$ if $s \rightarrow t$ and $s \in W^2$ and $s \rightarrow^2 t$ if $s \rightarrow t$ and $s \in W^1$.

Given any two neighboring states s and t , let $\phi_{s,t}$ be the complete proposition that is satisfied by t but not by s . (That is, given that $s_k \neq t_k$,

- the consequent of $\phi_{s,t}$ is v_k if $t_k = 1$ and $\neg v_k$ if $t_k = 0$, and
- the antecedent of $\phi_{s,t}$ is a conjunction of $K - 1$ variables or negations of variables, such that for every $l \neq k$ the conjunction contains one element: either v_l if $s_l = t_l = 1$ or $\neg v_l$ if $s_l = t_l = 0$.)

Finally, let Φ^i be the set of all propositions $\phi_{s,t}$ for which $s \rightarrow^i t$. For every $s \notin W^i$, the set Φ^i contains exactly one proposition that is violated by s , namely the proposition $\phi_{s,t}$, where t is the state that immediately follows s in the Hamiltonian cycle. This proposition is violated only by s . Every state in W^i , on the other hand, satisfies every proposition in Φ^i . It follows that $\text{Int}(T(\Phi^i)) = W^i$.

Moreover, every state $s \notin W^i$ has exactly one $\triangleright_{T(\Phi^i)}$ successor: the state immediately following s in the Hamiltonian cycle. Hence, starting from any state outside W^i , the cheating procedure follows the Hamiltonian cycle uniquely until it reaches the first state in W^i .

The texts are also coherent. Indeed, suppose that two propositions in Φ^i had non-contradictory antecedents and contradictory consequents. Since both propositions are complete, they would correspond to the two opposite orientations of the same edge of the hypercube. But a directed Hamiltonian cycle cannot contain both orientations of the same edge.

It remains to show that the pair of texts $(T(\Phi^1), T(\Phi^2))$ magically implements $\langle S, W^1, W^2 \rangle$. Consider, for example, $s \in W^1$. The state s satisfies $T(\Phi^1)$, and thus party 1 declares s . Party 2 finds that s does not satisfy $T(\Phi^2)$. The cheating procedure leads him to declare the first state t in the W^2 block that follows the W^1 block containing s .

If the true state were t , as party 2 claims, then party 1 would declare the state r , the first element in the W^1 block that follows the W^2 block containing t . Since there are at least two W^1 blocks, $r \neq s$. By the uniqueness of the cheating path, it follows that $t \not\rightarrow_{T(\Phi^1)} s$. Thus, $s \rightarrow_{T(\Phi^2)} t$, whereas $t \not\rightarrow_{T(\Phi^1)} s$. The principal can therefore apply the one-way cheating principle and conclude that party 2 has cheated.

The argument for $s \in W^2$ is symmetric. Therefore, $(T(\Phi^1), T(\Phi^2))$ magically implements $\langle S, W^1, W^2 \rangle$. ■

A question that naturally arises is whether the principal needs to involve both parties or can make do with involving only one party. With only one party, the implementation problem is a pair $\langle S, W \rangle$, where W is the set of states in which the party deserves the prize. Implementation requires the existence of a text T with $\text{Int}(T) = W$ such that there is no $s \notin \text{Int}(T)$ and $t \in \text{Int}(T)$ such that $s \rightarrow_T t$. It follows from Glazer and Rubinstein (2012) that such an implementation is possible if and only if each connected component (with respect to the neighboring binary relation) of the complement of W contains a cycle length of four or more. Returning to the present paper, if either W^1 or W^2 has this feature, then implementation can be achieved by involving only one of the parties. However, as in the motivating example in this section, this is often not the case and implementation under the present logical language and cheating procedure requires the involvement of both parties.

IV. Takeaway

The paper's takeaway is conceptual. It draws attention to the potential of designing mechanisms that rely on the procedure of cheating. As already mentioned, our notion of implementation is not game theoretic. Although game theory obviously provides an interesting approach to the study of implementation problems, it is not the only approach, and models that ignore parties' strategic reasoning may be interesting in their own right.

We demonstrate that a mechanism designer who understands the cheating procedure used by the parties can sometimes design a mechanism that exploits the inherent asymmetry between a truth-teller and a liar in order to identify the liar without observing an actual smoking gun. There are other ways to exploit procedural properties of decision-making in order to elicit the truth from parties with conflicting interests. For example, a reliable lie detector could also do the job. It is also reasonable to expect that the response time of a cheater will be statistically longer than that of a truth-teller. In this paper, however, the framework is similar to the standard implementation literature in the sense that each party sends a message and the principal acts as a function of the messages alone without observing any additional information. It differs, however, in its embedded inference rule, which is very different from the standard Nash equilibrium approach, and in its nonstrategic environment.

REFERENCES

- Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman. 2019. "Mechanisms with Evidence: Commitment and Robustness." *Econometrica* 87 (2): 529–66.
- Glazer, Jacob, and Ching-To Albert Ma. 1989. "Efficient Allocation of a 'Prize'—King Solomon's Dilemma." *Games and Economic Behavior* 1 (3): 222–33.

- Glazer, Jacob, and Ariel Rubinstein.** 2006. "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach." *Theoretical Economics* 1 (4): 395–410.
- Glazer, Jacob, and Ariel Rubinstein.** 2012. "A Model of Persuasion with Boundedly Rational Agents." *Journal of Political Economy* 120 (6): 1057–82.
- Green, Jerry R., and Jean-Jacques Laffont.** 1986. "Partially Verifiable Information and Mechanism Design." *Review of Economic Studies* 53 (3): 447–56.
- Kartik, Navin, and Olivier Tercieux.** 2012. "Implementation with Evidence." *Theoretical Economics* 7 (2): 323–55.
- Lipman, Barton L., and Duane J. Seppi.** 1995. "Robust Inference in Communication Games with Partial Provability." *Journal of Economic Theory* 66 (2): 370–405.
- Maskin, Eric.** 1999. "Nash Equilibrium and Welfare Optimality." *Review of Economic Studies* 66 (1): 23–38.
- Michelbacher, Lukas, Stefan Evert, and Hinrich Schütze.** 2007. "Asymmetric Association Measures." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, edited by Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, 367–72. INCOMA Ltd.
- Perry, Motty, and Philip J. Reny.** 1999. "A General Solution to King Solomon's Dilemma." *Games and Economic Behavior* 26 (2): 279–85.