Magical Implementation

Jacob Glazer

Department of Economics, the University of Warwick
and Coller School of Management, Tel Aviv University

and

Ariel Rubinstein

School of Economics, Tel Aviv University
and Department of Economics, New York University

July 1, 2025

Abstract: A principal needs to decide which of two parties deserves a prize. Each party privately observes the state of nature that determines which of them deserves the prize. The principal presents each party with a text that truthfully describes the conditions for deserving the prize and asks each of them what the state of nature is. The parties do not behave strategically. Each party can lie by activating a cheating procedure which relates to the state and the text given to him. The principal "magically implements" his goal if he can come up with a pair of texts satisfying that in any dispute, he will recognize the cheater by applying the following rule: the truth is with the party satisfying that if his statement is true, then the other party (using the cheating procedure) could have cheated and made the statement he is making, but not the other way around. Several examples are presented to illustrate the concept.

KEYWORDS: Magical implementation JEL classification: D82

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 and especially Áron Tóbiás for their comments.

1. Introduction

Two invigilators, B and G, have overheard a student receiving a whispered message from another student during an exam. The invigilators have not seen the questions on the exam but would be able to solve them. It is known that B is hostile to the student who received the message while G is sympathetic towards him. 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 answers the question correctly. Invigilator B claims that the whispered message was: "$\alpha = 3$." This is a serious allegation and if correct, the student's exam will be disqualified. Invigilator G claims that the whispered message was: "Solve the equation $\alpha+1=4$ first." If he is right, then the student's answer genuinely reflects his knowledge of the material and there will not be any 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 reasoning would be 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 "$\alpha = 3$". On the other hand, if the message were "$\alpha = 3$", it is very unlikely that G (who, as mentioned, has not seen the exam questions) could guess that the equation solved by the answer $\alpha = 3$ is $\alpha + 1 = 4$ rather than any other equation with the same solution. Hence, there is asymmetry between the two conflicting claims, which makes it possible to reasonably conclude that G's claim is the truthful one.

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

We consider situations of the following nature: Two parties claim a prize being offered by a principal. The principal's view is that only one of them truly deserves the prize, and his identity is determined *unequivocally* by facts known to the two parties but not to the principal. The parties do not know the principal's view; nonetheless both insist that they deserve the prize.

The situations we have in mind are related to the biblical Judgement of Solomon, where two women claim to be the mother of the same baby and the king must decide who is telling the truth. In that story, unlike ours, the women know the circumstances under which King Solomon wishes to deliver the baby to each of them. There is also an asymmetry in the women's preferences with regard to the potential consequences of the ruling: the true mother likes the baby "more" than the fake mother in the sense that the true mother – whichever woman she is – is willing to pay more for the baby than the fake mother. This asymmetry allows Glazer and Ma (1989) (later generalized by Perry and Reny (1999)) to construct a game form with the feature that regardless of who the true mother is, the induced extensive game has a unique subgame perfect equilibrium with the outcome that the true mother gets the baby without making any payment. In contrast, we consider similar disputes without assuming any asymmetry in preferences or information. The only asymmetry between the two parties is that one deserves the prize and can claim it by telling the truth about the state while the other can claim the prize only by telling a lie.

The approach taken in this paper is that, unlike telling the truth, cheating is often not a simple task, and it may require the liar to operationalize some "cheating procedure". We show that a principal who is aware of the agent's cheating procedure can sometimes design a mechanism that will enable him to identify the liar without observing an actual "smoking gun".

We study mechanisms where the principal provides each party with a text describing the circumstances (the set of states) under which he deserves the prize. Each party then submits a description of a state which he claims to be the true state. The model is enriched by (i) a *language* that the principal can use to compose a text, and (ii) a *cheating procedure* that determines the state a party will announce, given the text he received and the realized state he observed. The cheating procedure is common to the two parties and is known to the principal. Both the language and the cheating procedure are situation-specific. For now, their details are left vague; nonetheless, each situation we analyze below will entail a formal specification.

Returning to the biblical "King Solomon Judgement" story, notice that if the mothers were reasoning strategically, then Solomon would not have been able to implement his strategy. In that case, the false mother would have known the king's strategy and would have imitated 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.

We also adopt a non-game-theoretic approach (a point that will be discussed later) that is based on a novel concept we refer to as "*magical implementation*". A magical implementation mechanism consists of the following stages (which occur after the parties have been informed about the state):

Stage 1: The principal provides each party with a true and full description of the set of states in which the party deserves the prize. Being constrained by a **language**, there are numerous texts that can describe this set and the particular text presented to a party is chosen at the principal's discretion.

Stage 2: Each party must present a factual statement (true or false) to the principal after he is informed that:
(i) If his statement does not justify his claim for the prize, he will not receive it for certain.
(ii) If his statement justifies his claim while the other agent's does not, then he will receive the prize.
The party is not informed about the outcome if both his and the other agent's statements justify their respective claims.

Stage 3: The principal considers the statements, s_1 and s_2, made by party 1 and party 2, respectively, and makes his decision as follows:

- If both statements imply that the same party deserves the prize, then he awards it to that party.
- If both parties make a statement justifying their own claim to the prize, then the principal checks whether there is a party i such that if s_i is true, then the cheating procedure might have enabled party j to make the statement s_j, whereas if s_j is true, the procedure could not have enabled party i to make the statement s_i. In this scenario, the prize is awarded to party i.
- In all other cases, neither party is awarded the prize.

In other words, the principal provides each party with an accurate description of the conditions under which he deserves the prize. The principal will grant the prize if the two parties agree on who deserves the prize, or if he can apply what we refer to as the "*one-way cheating principle*", according to which the truth is with the party satisfying that if his statement is true, then the other party (using the cheating procedure) could have cheated and made the statement he is making, but not the other way around.

Asymmetries between statements are often used in practice as a tool to decide which of two conflicting statements is true. For example, scholars of old manuscripts who have before them two versions of the same text, but only one of which can be genuine, use such asymmetries as a tool to decide which one is the original. A principle known as "Lectio difficilior potior" instructs scholars to prefer text A over text B if text A can be seen as a simplification of text B, but not vice versa.

Another such 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, where one claims it is MIT and the other that it is NIT, then we would tend to believe that the professor graduated from NIT.

In what follows, we formalize the concept of magical implementation and apply it in three examples.

2. 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 disjoint subsets, W^1 and W^2, where W^i denotes the set of states in which party i should win the prize. In every state both parties are informed about the 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 about the state needs to 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 circumstances in which the party that receives the text deserves the prize. In choosing the texts, the principal uses a *language* $\mathfrak{L} = \langle \mathfrak{T}, Int\rangle$, where $\mathfrak{T}$ is a set of *texts* and Int is an interpretation function that assigns to each $T \in \mathfrak{T}$ a subset $Int(T)$ of states in which T is true. Notice that we allow different texts to have the same interpretation.

The principal gives the text T^i to agent i and asks him to declare the true state. He informs agent i that if the state he decares does not satisfy T^i he will definitely not get the prize. However, the principal does not disclose what would happen if both agents declare states that satisfy their respective texts.

Once a party has received a text T, he sends a message to the principal in the form of a state in S. It is assumed that if the state s satisfies the text T (that is, $s \in Int(T)$) then the party will send the message s. Otherwise, the party applies a *cheating procedure* formalized as a binary relation $\rightarrow_T$ on S. The statement $s \rightarrow_T t$ is interpreted as: "If the true state is s and the text T is not satisfied by s, then the cheating procedure *may* lead the party to claim t which does satisfy T." The same state s may have multiple states t for which $s \rightarrow_T t$. We assume that both parties apply the same cheating procedure. We will consider a different language and a different cheating procedure in each of the sections below.

We say that the pair of texts (T^1, T^2) *magically implements* $\langle S, W^1, W^2 \rangle$ if:
(1) $Int(T^i) = W^i$ for both i. That is, the principal provides each party with a *correct* description of the circumstances under which he deserves the prize.
(2) For every two states s and t, if $s \rightarrow_{T^i} t$ (and thus $s \in W^j$ and $t \in W^i$) then $t \nrightarrow_{T^j} s$. That is, given any state, if the undeserving party cheats, then the principal can apply the *one-way cheating principle* (given the cheating technology) and infer correctly which party is telling the truth.

We use the term *magical implementation* because magicians possess skills to discern subtle patterns and behavioral cues in human actions, which they are able to exploit in order to create the illusion of a miracle. The principal, in his role as magician, exploits his understanding of human imperfections in order to achieve his goal. Whereas a magician wishes to entertain his audience, the principal wishes to identify which of two rival parties is telling the truth.

Discussion: The notion of magical implementation is fundamentally different from the classical notion of Nash implementation, in that it does not involve a game. First, the principal provides each party only with a text that truthfully describes the circumstances in which he deserves the prize and commits that the party will not receive the prize 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 contradict each other. Second, the parties do not think strategically, i.e. they do not take into consideration the other party's actions. Each party acts as a "problem solver" *being aware* that he will certainly not get the prize if he does not solve the problem and *without being aware* that even if he successfully solves the problem he still may not get the prize.

Note that if the parties were aware of the principal's inference method and would think strategically, then following the cheating procedure would not constitute a Nash equilibrium. More precisely, when $s \in W^i$, we assume that party j will declare a state t such that $s \rightarrow_{T^j} t$ even if $t \not\rightarrow_{T^i} s$. But in that case, party j would do better if he finds a state r such that $r \rightarrow_{T^i} s$ and $s \not\rightarrow_{T^j} r$ and based on this lie persuades the principal that party i is the cheater.

The one-way cheating principle differs from the basic idea behind the standard Nash implementation mechanism à la Maskin (1999) (in an environment of at least three agents). There, the mechanism accepts an appeal by an agent if and only if the appeal is against the agent's own interests (according to the consensus among the other agents about his preferences). Obviously, Nash implementation is not feasible in our environment. Magical implementation is based on the asymmetry created by the use of the cheating procedure rather than the exploitation of differing interests.

Notice also the difference between our approach and that of the literature on implementation with hard evidence (see, for example, Green and Laffont (1986), Lipman and Seppi (1995), Glazer and Rubinstein (2006)). In that literature, an agent is limited as to the lies he can tell about the state. This limit is given and is not affected by the mechanism designed by the principal. In contrast, in our setting the limit is determined as a function of the text given to the agent and the cheating procedure.

3. Setting a trap through associations

3.1 A motivating story

A village in DrSeussLand is populated by Yooks and Zooks. A person was seen planting a bomb. The police arrests 8 suspects in the neighborhood: 3 Yooks (their names are denoted by y_1, y_2, y_3) and 5 Zooks (their name are denoted by z_1, z_2, z_3, z_4, z_5). It is certain that one of the suspects is the terrorist. There are two witnesses to the event – one of them 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 you can't distinguish Yooks and Zooks 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 while 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 belong 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 and the other for the Zook witness. On the first page of each site, they insert the pictures and names of the 8 suspects. Recall that the name of a suspect does not reveal whether he is a Yook or a Zook.

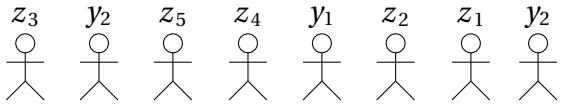

Figure 1. The first page of both websites

Each website comprises 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 emphasizing a prominent characteristic that is shared between the two. To ensure that an agent associates B after considering A, we might include a prominent detail in the description of A that is shared exclusively with B.

The left graph on Figure 2 describes the Yook's website while the right graph de-

scribes the Zook's. An edge in a graph means that the two connected suspects appear on one of the pages in 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.

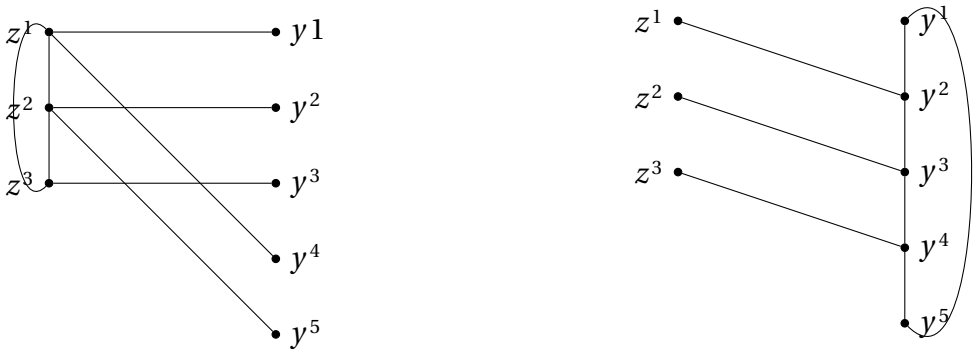

Figure 2. The graph on the left describes the Yook's website while the graph on the right describes the Zook's. Each edge represents one webpage with the pictures and names of two paired suspects.

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. A witness starts by checking the group identity of the actual terrorist and if he does not belong to the opposite group (i.e. his name appears only once in the website), he repeats the process starting with the other name that appears with the terrorist on the only page that the terrorist's name and picture are displayed.

In order 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, namely that the terrorist is a Zook and thus he can happily announce that he has identified z_1 to be the terrorist. The Zook also starts by searching for the name z_1 and discovers that on his website the name appears only on 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 he appears on his website more than twice and thus y_2 is 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 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 he would start from y_2, and after discovering that he is not a Zook, would continue by searching for the name z_2 (suspect y_2's partner on the only page in the Yook's website where 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 the above magic would work whichever suspect is the terrorist.

3.2 The general case

The implementation problem: The set of states S is finite and partitioned into W^1 and W^2, each of which has at least two states.

The language: A text is characterized by a set D of doubletons of states in S and has the following form:

$T(D)$: 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 question $Q(D,s)$ which could be either:

(i) none;

(ii) one doubleton $\{s,t\}$ (which contains s); or

(iii) a set of at least two doubletons (each containing s).

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

Start by asking the question $Q(D,s)$.
If $|A(D,s)|>1$, then declare s.
If $A(D,s)$ contains only $\{s,t\}$, then ask the question $Q(D,t)$. If $|A(t)|>1$, then declare t. 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 $Q(D,x)$ was not asked previously.
If you have exhausted all states in S without finding an element that belongs to two sets with labels in D, then give up and declare s.

Notice that unless no state satisfies the text $T(D)$ this procedure will always end up with the party finding a state that satisfies it. This is the reason that given the aforementioned framework, the principal cannot achieve his goal with a single party.

The formal description of the cheating procedure is the relation defined by $s \rightarrow_{T(D)} t$ if either:
(i) $A(D,s)=\{\{s,t\}\}$ and $|A(D,t)|>1$; or
(ii) $|A(D,t)|>1$ and either "$A(D,s)=\emptyset$" or "$A(D,s)=\{\{s,x\}\}$ and $|A(D,x)|=1$".
In option (i), 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 option (ii), the party starts with $Q(D,s)$ and then is stuck, either because s does not appear in any doubleton in D, or it appears only once with another state that also appears only once. In either of these cases, the party picks an arbitrary state not explored before and that state may be t.

Claim A *Let $\langle S, W^1, W^2\rangle$ be a problem satisfying that every 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 $z^1,\ldots,z^K$ and $y^1,\ldots,y^L$, respectively (without loss of generality assume that $K\leq L$). Form a sequence $x^1,\ldots,x^{2L}$ starting with y^1 and alternating between states in W^1 and states in W^2 (for the the case $K=L$, we denote $x^{2L+1}=x^1$). The states of W^2 appear in order. The states of W^1 appear cyclically (if necessary) in their order. Thus, for example, if $K=3$ and $L=5$ the sequence will be: $(x^1,\ldots,x^{10})=(y^1,z^1,y^2,z^2,y^3,z^3,y^4,z^1,y^5,z^2)$. The key feature of this construction is

that there is no pair of states $z \in W^1$ and $y \in W^2$ such that z appears right after y somewhere in the sequence and appears right before y somewhere else in the sequence.

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:
(i) K doubletons $\{z^1, z^2\}, \{z^2, z^3\}, \ldots, \{z^K, z^1\}$; and
(ii) $|W^2|$ doubletons $\{x^k, x^{k+1}\}$, one for each $x^k \in W^2$.

The set D^2 is constructed similarly. Figure 2 above illustrates the construction of the texts $T(D^1)$ (on the left) and $T(D^2)$ (on the right) in the case that $W^1 = \{z_1, z_2, z_3\}$ and $W^2 = \{y_1, y_2, y_3, y_4, y_5\}$.

For any $y \in W^2$, the set $A(D^1, y)$ contains only one doubleton $\{y, z\}$ where z appears right after y in the sequence $(x^1, \ldots, x^{2L})$, and therefore y does not satisfy $T(D^1)$. This z is in W^1 and $|A(D^1, z)| > 1$. Thus, there is a unique $z \in W^1$ (the one which comes after y in the sequence) such that $y \rightarrow_{T(D^1)} z$. On the other hand $|A(D^2, y)| \geq 2$ and therefore y satisfies the text $T(D^2)$. Similarly, any $z \in W^1$ satisfies the text $T(D^1)$ and there is a unique $y \in W^2$ (the one which comes after z in the sequence) such that $z \rightarrow_{T(D^2)} y$.

By the construction of the sequence $(x^1, \ldots, x^{2L})$, there is no case where y comes right after z and also z comes right after y. Thus, for no $z \in W^1$ and $y \in W^2$ do we have both $y \rightarrow_{T(D^1)} z$ and $z \rightarrow_{T(D^2)} y$. It follows then 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 the state t and 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 (given the state s, party j could cheat using t and if the state were t, then party i could not cheat by declaring s) and concludes that j is the cheater.

4. Setting a trap through a logical riddle

This section investigates the concept of magical implementation when the language of 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.

4.1 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.

The rector will form his opinion on the basis of the truth or falseness of three statements about G: $E =$"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 (about which they agree, as mentioned) but he fears that they may give him a biased opinion due to their strong personal bias for or against G. He suspects that they may not be above 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 statements E, L and P is True or False. Coding "True" as 1 and "False" as 0, there are 8 possible configurations of answers (states): $000, 100, \ldots, 111$ (for example, 101 stands for E and P are true and L is 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 that 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 will take 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 facts in the antecedent, i.e. A and B, are true in the case of G, then the fact in the consequent, i.e. C, must also be". 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." The following table presents the texts given to *Pro* and *Con* by the rector:

You are right if your report satisfies the following propositions:	
Professor *Pro*	Professor *Con*
$\neg E \wedge L \rightarrow \neg P$	$L \wedge \neg P \rightarrow E$
$E \wedge L \rightarrow P$	$\neg E \wedge P \rightarrow L$
$E \wedge P \rightarrow \neg L$	$\neg L \wedge \neg P \rightarrow \neg E$
$E \wedge \neg L \rightarrow \neg P$	
$\neg E \wedge \neg L \rightarrow P$	

As mentioned, the rector knows the cheating procedure applied by a professor if he finds that the truth does not satisfy all the propositions in his text's list. In that case, the professor searches for a state which satisfies the text by advancing along a path. He starts with the true state. He finds a proposition (in principle, 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 in the case that *Pro* considers that state, it will direct him to consider the state 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 did declare (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).

Figure 3. Each node in the lefthand cube represents a state. The red dots denote the states that support *Pro*'s position, while the black dots denote those that support *Con*'s. The propositions on *Pro*'s (*Con*'s) list are represented by red (black) arrows in the righthand cube. For example, the red arrow from 101 to 100 represents the proposition $E \wedge \neg L \rightarrow \neg P$ on *Pro*'s list.

The left cube in Figure 3 presents the implementation problem. Each node represents a state. The red dots denote the states that support *Pro*'s position, while the black dots denote those that support *Con*'s. The right cube in Figure 3 illustrates the rector's magic. The red arrows correspond to the propositions in *Pro*'s list while 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.

4.2 The general case

The implementation problem: Let $S = \{0,1\}^K$ where $K \geq 3$. A state $x_1 x_2 \ldots x_K$ (shorthand for $(x_1, \ldots, x_K)$) is a vector representing the truth values of the propositional variables $v_1, \ldots, v_K$, where $x_k = 1$ indicates the "truth" of the variable v_k and $x_k = 0$ indicates its "falsity". 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 $v_1, \ldots, v_K$ and has the structure $\wedge_{v \in V} \phi_v \rightarrow \phi_z$ where V is a non-empty 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 ($\wedge_{v \in V} \phi_v$), 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 following form:

$T(\Phi)$: You deserve the prize if the state satisfies all the propositions in Φ.

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 $\rhd_{T(\Phi)}$ on S. The statement $s \rhd_{T(\Phi)} t$ indicates that if the party reaches s then he may consider t. It is required that there be a proposition in Φ which s violates, and t is obtained from s by switching only the truth value of the variable in that proposition's consequent. Formally, $s \rhd_{T(\Phi)} t$ if:

(1) $s \notin int(T(\Phi))$ (that is, s does not satisfy at least one of the propositions in Φ),

(2) t differs from s in only one variable, and

(3) 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:

Start with the true state s^*. Advance along a path of the type $s_1 = s^* \rhd_{T(\Phi)} s_2 \rhd_{T(\Phi)} \cdots \rhd_{T(\Phi)} s_m$ (where all the states are distinct). If you reach a state that satisfies the text, then announce it. If you cannot advance any further without returning to a state on the path, look for a new path starting again from s^*. If you do not find a new path, then announce s^*.

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

Claim B *If the principal is equipped with the above language and both parties use the above cheating procedure, then any implementation problem $\langle S = \{0,1\}^K, W^1, W^2\rangle$ (where $K \geq 3$ and both W^1 and W^2 include at least two states) is magically implementable.*

Proof. Recall that a Hamiltonian cycle of the set S is an enumeration $x_1, \ldots, x_{2^K}$ of all states of S such that $x_k N x_{k+1}$ for all k (and $x_{2^K} N x_1$). The following Lemma is needed:

Lemma: There is an Hamiltonian cycle with more than one block of W^1-states (and thus also more than one block of W^2-states).

Proof: For every $K \geq k \geq 1$ and $\delta \in \{0,1\}$, let $S_{k,\delta}$ be the set of all states s such that $s_k = \delta$. Let k be a dimension for which both $S_{k,0}$ and $S_{k,1}$ include an element of W^2. Either $S_{k,0}$ or $S_{k,1}$ contains two states of W^1, unless W^1 has exactly two states (one in $S_{k,0}$ and one in $S_{k,1}$), in which case W^2 has at least two elements in one of the sets (recall that $K \geq 3$). Therefore, we can assume without loss of generality that $S_{1,1}$ contains at least two states from W^1 and one from W^2 and $S_{1,0}$ contains at least one element from W^2.

Construct an arbitrary Hamiltonian cycle of $S_{1,1}$. If it contains more than one W^1-block we can extend it to an Hamiltonian cycle of S with more than one W^1-block. If not, there must be two W^1-successive states a and b, such that b comes right after a in the cycle. Construct a Hamiltonian cycle of S as follows: start with b, continue in the $S_{1,1}$-cycle to a, move to a's neighbor in $S_{k,0}$ and continue with an Hamiltonian cycle of $S_{k,0}$ that ends with b's neighbor in $S_{k,0}$. This Hamiltonian cycle contains at least two blocks of W^2-states (one in $S_{1,1}$ and one in $S_{1,0}$). Figure 4 demonstrates the construction starting with the Hamiltonian cycle ($a = 110, b = 111, 101, 100$) of $S_{1,1}$:

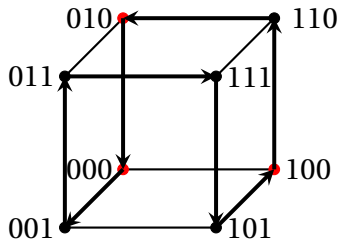

Figure 4. The construction of the Hamiltonian cycle with at least 2 blocks of each color ($a = 110$ and $b = 111$).

We can now prove Claim B. Let $x_1, \ldots, x_{2^K}$ be an 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 $s_k \neq t_k$:
(i) the consequent of $\phi_{s,t}$ is v_k if $t_k = 1$ and $\neg v_k$ if $t_k = 0$; and
(ii) the antecedent of $\phi_{s,t}$ is a conjunction of $K-1$ variables or negations of variables, such that for every $l \neq k$ there is one element in the conjunction: 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 $s \rightarrow^i t$. The set of states that satisfy the text $T(\Phi^i)$ is exactly W^i. Obviously, the texts are coherent. It is left 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 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 $t \in W^2$ in the first W^2-block that follows (in the Hamiltonian cycle of S) the W^1-block that contains s.

The principal can now apply the one-way cheating principle and conclude that 2 is cheating. If the true state were t (as 2 claims), then party 1 would declare the state r, the first element in the W^1-block that follows the W^2-block that contains t. Since there are at least two W^1-blocks, the state r is different from the state s. ■

A question that naturally arises is whether the principal needs to involve both parties or can he make do with involving only one party. With only one party, the implemen-

tation problem is a pair $\langle S, W\rangle$ where W is the set of states in which the party deserves the prize. Implementation would require the existence of a text T with $int(T) = W$ such that there is no $s \notin Int(T)$ and $t \in Int(T)$ 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 complementary set of W contains a cycle of length 4 or more. Returning to this 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 subsection 4.1., this is often not the case and magical implementation requires the involvement of both parties.

5. Making cheating too risky

This section is motivated by the two-invigilator scenario presented in the introduction. We will demonstrate that sometimes the principal may be able to distinguish between a truth-teller and a cheater only after he shares some of his information with the parties. First, however, we need to adjust the notion of implementation problem and the definition of magical implementation.

The implementation problem: We expand the notion of an implementation problem to be a tuple $\langle S, W^1, W^2, I, I_p, \mu\rangle$ where the three additional elements are:

I: The parties' common information structure of S (a partition of S). In state $s \in S$, each party is informed about $I(s)$, the cell in the partition that contains s.

I_p: The principal's information structure of S. In state $s \in S$, the principal is informed about $I_P(s)$, the cell in the partition that contains s.

μ: A probability measure on S.

It is required that any information set in I is either a subset of W^1 or of W^2, that is, the information possessed by the parties is sufficient to determine which party deserves the prize. If this condition were also applied to I_p, then the case would be trivial: the principal would arrive at the correct conclusion without having to elicit any information from the parties.

The language: A text has two parameters: A set $Y \subseteq S$, a union of sets in I_p which constitutes the information provided to the parties by the principal (in addition to what the

parties already know according to the information structure I), and a set $W \subseteq S$, a union of the cells in I. A text $T(Y,W)$ has the following form:

$T(Y,W)$: The state is in Y. You deserve the prize if the state is in W.

Once a party has received a text $T(Y,W)$, he needs to send a message to the principal in the form of an information set in I. It is assumed that a party is aware that the information he receives from the principal is truthful (but does not necessarily consist of all the information possessed by the principal).

The cheating procedure: The cheating procedure in this section is more conventional than in the previous two. A party's willingness to cheat by reporting a false information set in I depends on his fear of getting caught. We say that a party is *caught cheating* if the information he reports and the principal's knowledge do not intersect. It is assumed that a party considers cheating (by announcing an untrue information set in I) only if he believes that the probability of getting caught does not exceed some threshold τ. A party bases his belief on the prior μ, the information set in I which he initially received, the principal's announcement Y, and the principal's information structure I_p.

To summarize, the procedure followed by a party after receiving the text $T(Y,W)$ and given that he initially received the information set $K \in I$ is as follows:

If $K \subseteq W$, then declare K. If $K \not\subseteq W$, then search for an $L \in I$ in which you deserve the prize ($L \subseteq W$) and the probability of being caught after declaring L is below τ, that is, $\mu(\{s \mid L \cap I_p(s) = \emptyset\} \mid K \cap Y) \leq \tau$. If you find such an L, then declare it; otherwise declare K.

This cheating procedure generates the binary relation $\rightarrow_{T(Y,W)}$ on the information sets in I, defined by $K \rightarrow_{T(Y,W)} L$ if $K \not\subseteq W$, $L \subseteq W$, and a party that initially receives the information K and cheats by declaring L gets caught with probability (conditional on $K \cap Y$) not exceeding τ.

Finally, we modify the definition of magical implementation. The principal chooses a partition I_p^* that is coarser than I_p, and after being informed about an information set M in I_p he provides each party i with the text $T(Y, W^i)$ where Y is the cell in I_p^* that includes M. Thus, it is assumed that the principal provides both parties with the same

additional information. In order to magically implement his goal, it is required that whenever the parties' claims differ (and assuming that the parties follow the cheating procedure), the principal will be able to activate the one-way cheating principle and correctly identify the deserving party.

Formally, we say that $\langle S, W^1, W^2, I, \mu, I_p \rangle$ is magically implementable if there is a partition I_p^* coarser than I_p such that for any set $Y \in I_p^*$ the texts $T(Y, W^1)$ and $T(Y, W^2)$ satisfy that if $K, L \in I$, $K \cap Y \neq \emptyset$, $L \cap Y \neq \emptyset$ and $L \rightarrow_{T(Y,W^i)} K$, then $K \not\rightarrow_{T(Y,W^j)} L$.

Notice the difference between catching a cheater and inferring that a party is cheating using the one-way cheating principle. The former occurs only when the principal has solid proof that a party is cheating, namely the party's statement contradicts the information possessed by the principal. The latter occurs when there is a dispute between the parties and the principal applies the one-way cheating principle *without any solid evidence* that one of the parties is cheating.

The exam example: We now formalize the example presented in the introduction. The set S consists of four states which are depicted as cells in the table below: a row indicates the content of the whispered message (the solution or just the equation) while a column specifies the equation (assuming, for simplicity, that there are only two possible equations):

the whispered message	The exam equation	
	$\alpha+1=4$	$\alpha+2=5$
a solution	a	b
an equation	c	d

The winning sets are $W^B = \{a, b\}$ and $W^G = \{c, d\}$. Each party's information partition is $I = \{\{a, b\}, \{c\}, \{d\}\}$ while the principal's is $I_p = \{\{a, c\}, \{b, d\}\}$. Assume that $\mu(a) = \mu(b) > 0$.

When $\tau < 1/2$, the texts $T(S, W^B)$ and $T(S, W^G)$ (the principal does not provide any additional information to the parties) magically implement the problem:
– When the parties are informed that the state is c (d), they know that the principal possesses the information $\{a, c\}$ ($\{b, d\}$). Then, party B is not afraid to cheat and declare $\{a, b\}$ since he will with certainty not be caught cheating.

– When the parties are given the information $\{a,b\}$, they do not know whether the principal received the information $\{a,c\}$ or the information $\{b,d\}$. Party G is afraid to report $\{c\}$ since if he does, he will be caught cheating with probability $\mu(b)/[\mu(a)+\mu(b)] = 1/2 > \tau$. Similarly, G is afraid to cheat by reporting $\{d\}$.

Thus, $\{c\} \rightarrow_{T(X,W^B)} \{a,b\}$ and $\{d\} \rightarrow_{T(X,W^B)} \{a,b\}$ but $\{a,b\} \not\rightarrow_{T(X,W^G)} \{c\}$ and $\{a,b\} \not\rightarrow_{T(X,W^G)} \{d\}$ and the problem is magically implementable.

Note that if $\tau > 1/2$, then also $\{a,b\} \rightarrow_{T(X,W^G)} \{c\}$ and $\{a,b\} \rightarrow_{T(X,W^G)} \{d\}$ and the pair of texts fails to magically implement the problem. This will also be the case if the principal reveals his knowledge. Then, in state a (for example), the principal will announce $\{a,c\}$ and party G (who possesses the information $\{a,b\}$) will not be afraid to cheat by declaring $\{c\}$. At the same time, party B will not be afraid to cheat in state c by declaring $\{a,b\}$. Thus, full information revelation will not enable the principal to perform his magic.

The asymmetry between a cheater and a truth-teller emerged on its own in the above example. The following example demonstrates that providing information is sometimes necessary for magical implementation.

Setting a trap by providing additional information:

$S = \{a,b,c,d,e,f\}$,

$W^1 = \{a,b,c\}$, $W^2 = \{d,e,f\}$,

$I = \{W^1, W^2\}$,

$I_p = \{Z_1 = \{a\}, Z_2 = \{c,d\}, Z_3 = \{b,e\}, Z_4 = \{f\}\}$,

$\mu =$ the uniform probability measure on S.

Assume that τ is above and below

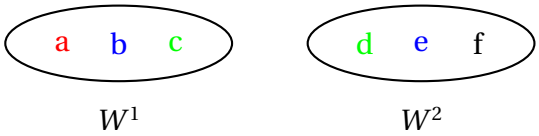

Figure 5. The information partition I is represented by the ellipses while I_p is represented by the colors.

It is easy to verify that as long as $\tau > 1/3 = \mu(a)/\mu(W^1) = \mu(f)/\mu(W^2)$, the problem is not magically implementable without the principal providing additional infor-

mation to the parties. However, if $\tau < \mu(a)/[\mu(a)+\mu(c)] = \mu(f)/[\mu(f)+\mu(d)] = 1/2$, magical implementation is achieved by the principal committing to supply additional information according to the information structure $I_p^* = \{Z_1 \cup Z_3, Z_2 \cup Z_4\}$. Assume, for example, that the principal has announced $Z_1 \cup Z_3 = \{a, b, e\}$. In the case that the parties are initially informed of W^1, each party will conclude that the state of nature is either a or b. Party 2 is deterred from cheating since he will be caught with probability $1/2 = \mu(a)/[\mu(a)+\mu(b)] > \tau$. In the case that the parties are initially informed of W^2, they will conclude that the state of nature is e and therefore the principal possesses the information $\{b, e\}$ and party 1 will not be caught cheating if he announces W^1. Thus, $W^1 \nrightarrow_{T(Z_1 \cup Z_3, W^2)} W^2$ and $W^2 \rightarrow_{T(Z_1 \cup Z_3, W^1)} W^1$ and in the case of disagreement the principal will be able to use the one-way cheating principle to infer that party 1 is the cheater.

6. Takeaway

The paper's takeaway is methodological. It draws attention to the potential of designing mechanisms that rely on the difficulty of cheating. Most of the implementation literature ignores this and assumes that cheating is as easy as telling the truth. Although some of the more recent literature does incorporate the difficulty of cheating by assuming that it is costly or that the agents are bounded in the lies they can tell, the imposed constraints are taken to be independent of the mechanism. (See, for example, Green and Laffont (1986), Lipman and Seppi (1995), Glazer and Rubinstein (2006), Kartik and Tercieux (2012) and Ben-Porath, Dekel and Lipman (2019).) According to our approach, agents wishing to cheat successfully behave like problem solvers without taking into account any strategic considerations and without fully understanding the principal's decision rule. The difficulty of the problem facing the agent depends on the mechanism. We provide three examples in which a mechanism designer who understands the cheating procedure being used by the agents can 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".

As already mentioned, our notion of implementation is not game-theoretic. We feel that the literature is overly conservative in its focus on implementation by means of game-theoretical concepts. Although game theory obviously provides an interesting approach to study implementation problems, it is far from being the only one and models that ignore agents' strategic reasoning might be interesting in their own right.

A Final Comment by Ariel Rubinstein: I am fully aware (and in fact proud) that the paper is written in a style different from what is the convention these days in Economic Theory. The discussion is purely conceptual. We do not claim that there are practical applications; the paper is short; and although the discussion is carried out in formal language, we avoid any fancy mathematics. The goal is simply to convey an idea. Indeed, I am suspicious of any work in Economic Theory that goes beyond presenting one main idea accompanied by a few simple examples. This paper should be read almost like a story: you might find it interesting, entertaining or elegant, or maybe... not. If the reader derives something useful from the article, that's fine; however, it is not my intention to generate any "practical" conclusions.

References

Ben-Porath Elchanan, Eddie Dekel and Barton L. Lipman (2019). "Mechanisms with Evidence: Commitment and Robustness". *Econometrica, 87*, 529-566.

Glazer, Jacob and Ching-To Albert Ma (1989). "Efficient Allocation of a 'Prize' — King Solomon's Dilemma". *Games and Economic Behavior, 1*, 222-233.

Glazer, Jacob and Ariel Rubinstein (2012). "A Model of Persuasion with Boundedly Rational Agents". *Journal of Political Economy, 120*, 1057-1082.

Glazer, Jacob and Ariel Rubinstein (2006). "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach". Theoretical Economics, 1 (2006), 395-410.

Green Jerry R. and Jean-Jacques Laffont (1986). "Partially Verifiable Information and Mechanism Design". *Review of Economic Studies, 53*, 447-456.

Kartik, Navin. and Olivier Tercieux (2012): "Implementation with Evidence". *Theoretical Economics*, 7, 323-355.

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

Maskin, Eric (1999), "Nash Equilibrium and Welfare Optimality". *The Review of Economic Studies, 66*, 23-38.

Michelbacher, Lukas, Stefan Evert, and Hinrich Schütze (2007). "Asymmetric Association Measures." *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing*, 367-372.

Perry, Motty and Philip J. Reny, (1999). "A General Solution to King Solomon's Dilemma." *Games and Economics Behavior, 26*, 279-285