Magical Implementation

Jacob Glazer

School of Management, Tel Aviv University
and Department of Economics, The University of Warwick

Ariel Rubinstein

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

July 21, 2024

Abstract: A principal would like 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 can cheat but the principal knows their cheating procedure. 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 "honest-cheater asymmetry principle". According to this principle, the truth is with the party satisfying that if his statement is true, then the other party (using the given cheating procedure) could have cheated and made the statement he is making, but not the other way around. 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, Rani Spiegler, Kemal Yildiz and especially Áron Tóbiás, for their comments.

1. Introduction

Two invigilators, A and B, have witnessed 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 A does not like the student who received the message while B does. 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 A claims that the whispered message was: "$\alpha=3$." This is a serious allegation and if correct, the student's exam will be disqualified. Invigilator B 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: A or B?

Although there is no definitive proof one way or the other, we would choose to believe B. The reasoning would be that if the message was "Solve the equation $\alpha+1=4$ first", then A (who dislikes the student) could solve the equation himself and claim that the message was "$\alpha=3$". On the other hand, if the message was "$\alpha=3$" it is very unlikely that B (who likes the student and who, as mentioned, has not seen the exam questions) could guess that the equation to be solved is $\alpha+1=4$ rather than any other equation with the same solution. Hence, there is an asymmetry between the two conflicting claims which makes it possible to reasonably conclude that B's claim is the truthful one.

In the above episode, the asymmetry between the two claims arises naturally and is not engineered by someone seeking to uncover the truth. In other cases, one might consider designing a mechanism that creates some sort of 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 this analysis.

We consider situations of the following nature: Two parties claim a prize being offered by a principal. Only one of them truly deserves the prize, and his identity is determined *unequivocally* by facts known exclusively to the two parties. Nevertheless, each party insists that he deserves the prize. An example can be found in the biblical Judgement of Solomon, where two women claim to be the true mother of the same baby and the king must decide who is telling the truth.

Economic theory suggests one approach to solve the king's problem using an asymmetry between the two mothers' 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. Most noticeably, Glazer and Ma (1989) (later generalized by Perry and Reny (1999)) 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.

We consider such situations without assuming any asymmetry in preferences or information. The only distinction between the two parties is that one deserves the prize and can claim it by telling the truth about the state of nature while the other can claim the prize only by telling a lie. The model is enriched by (i) a *langauge* that the principal can use to describe to each party the circumstances under which he deserves the prize, and (ii) a *cheating procedure* used by a party that needs to cheat in order to justify his claim for the prize. The procedure is common to the two parties and known to the principal. Both the language used and the cheating procedure are situation-specific. Currently, we maintain a level of vagueness, but each situation we analyze will have a formal specification.

We are interested in mechanisms that generate the correct outcome, based on the asymmetry between a truth-teller and a cheater. The approach adapted is not game-theoretic but rather rests on a novel concept referred to as "*magical implementation*". A magical implementation mechanism consists of the following stages (after the parties have both been informed about the state of nature):

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 the language, there are numerous texts that could describe this set and the particular text presented to a party is chosen at the principal's discretion.

Stage 2: Each party is required to present a (true or false) factual statement (a state of nature) to the principal and is informed that if his statement does not justify his claim for the prize, then he will not receive it.

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 "*honest-cheater asymmetry principle*": the truth is with the party satisfying that if his statement is true, then the other party (using the given 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 called "Lectio difficilior potior" ("the more difficult reading is the stronger") instructs scholars to choose the more unusual text.

Another such natural asymmetry involves word associations. As Michelbacher, Evert, and Schütze (2007) put it: "Human word associations are asymmetric or directed." For example, they note that 'soup' triggers 'tomato' much more often than 'tomato' triggers 'soup'; 'mango' triggers 'fruit' much more often than the reverse; a person with the name David Churchill brings Winston Churchill to mind but not the other way around, etc. Therefore if two parties disagree about which university a certain professor has graduated from, where one claims it is MIT and the other that it is NIT, then we will tend to believe that the professor was 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. After receiving the text T^i, party i sends a message to the principal in the form of a state in S.

In choosing the texts, the principal is constrained by 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. A different language is considered in each section below.

A party can always announce the truth but if he wishes to cheat he is limited by a *cheating procedure* which will be specified in each of the examples. The cheating procedure is common to both parties and known to the principal. It is represented as a function that assigns to each text $T \in \mathfrak{T}$ a binary relation $\rightarrow_T$ on S such that $s \rightarrow_T t$ 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." That is, $s \rightarrow_T t$ if:
(i) the cheating procedure and the text T induce a party in state s to consider t;
(ii) T does not entitle a party to the prize given s (that is, $s \notin Int(T)$); and
(iii) T entitles a party to the prize given t (that is, $t \in Int(T)$).
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. However, there may be multiple texts T for which $Int(T) = W^i$, and the principal seeks a pair of texts that will enable him to grant the prize to the deserving party.

(2) Given any state $s \in W^i$ and for any state $t \in S$ such that $s \rightarrow_{T^j} t$ (and thus $t \in W^j$), we have $t \nrightarrow_{T^i} s$. That is, given any state, if the undeserving party cheats, then the principal can apply the *honest-cheater asymmetry principle* and infer 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.

Notice that if the parties understand the principal's inference method, then following the cheating procedure is not a dominating strategy and outsmarting the cheating procedure might be beneficial. When $s \in W^i$, we assume that party j will declare a state t such that $s \rightarrow_{T^j} t$ even if $t \nrightarrow_{T^i} s$. But then, party j would do better if he finds a state r such that $r \rightarrow_{T^i} s$ and $s \nrightarrow_{T^j} r$ and based on this lie persuades the principal that party i is the cheater.

Discussion: The notion of magical implementation is fundamentally different from the classical notion of Nash implementation. For one thing, the mechanism does not involve a game. The principal provides each party 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. The principal does not inform the parties of what will happen if their claims contradict each other. 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 may not get the prize.

The honest-cheater asymmetry principle differs from the basic idea behind the standard Nash implementation mechanism a la Maskin (1999). There, in an environment of at least three agents, the mechanism accepts an appeal by an agent if and only if it is against his own interests as given by a consensus among the other agents about his preferences. Unlike our mechanism, it does not take into account the interests of the other agents to cheat, given the report of the appealing agent. 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 making use of differences in interests.

Notice the fundamental 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) and Ben-Porath, Dekel and Lipman (2019)). In that literature, an informed agent is limited as to the lies he can tell about the state. These limits are given and are not affected by the mechanism designed by the principal. In contrast, the set of states in which an agent considers cheating in our case is determined endogenously by the texts presented by the principal and the cheating procedure.

3. Setting a trap for the cheater

The implementation problem: The set of states S is finite and is partitioned into W^1 and W^2, each of which has at least two states. Recall our assumption that both parties receive complete information about the state.

The language: Each doubleton Z of states in S is identified by a distinct label $d(Z)$. Let $\mathfrak{D}$ be the set of labels. The parties do not know the meaning of the labels and given a label they need to ascertain which two states are behind it. Following are two examples of such a setup:

(a) The set S is a set of cities. Any two cities are connected by a unique road labeled by an exclusive number. If a party is given the number of a road he needs to ascertain which two cities it connects.

(b) The set S is a set of vectors in a Euclidean space. A doubleton (or any other finite set) is definable as the set of solutions to an equation. The set of solutions to the equation $g(y) = 0$ is labeled by the equation. The equations are complicated and the parties cannot solve them by themselves.

A text is characterized by a set $D \subseteq \mathfrak{D}$ of labels:

$T(D)$: You deserve winning the prize if the state of nature is a member of at least two sets with labels in D.

The interpretation of $T(D)$ is the set of all states that belong to at least two sets 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 with labels in D contain s?" Denote by $A(D,s)$ the answer to question $Q(D,s)$ which is 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 that 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 why given this section's framework, the principal cannot achieve his goal with a single agent (unless he always or never deserves the prize).

The cheating procedure corresponds to the relation $\rightarrow_{T(D)}$ 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 any of these cases, the party picks an arbitrarily state not explored before. Eventually, he must find a state that satisfies the text.

Example: Consider the problem $\langle S = \{1,2,3,4\}, W^1 = \{1,2\}, W^2 = \{3,4\}\rangle$.
Let $D^1 = \{d(\{1,2\}), d(\{1,3\}), d(\{2,4\})\}$. The text $T(D^1)$ is solved by 1 and 2 only. If the state is 3, then party 1 will declare the state 1 and if the state is 4, then he will declare the state 2. That is, $3 \rightarrow_{T(D^1)} 1$ and $4 \rightarrow_{T(D^1)} 2$.
Let $D^2 = \{d(\{1,4\}), d(\{2,3\}), d(\{3,4\})\}$. Similarly, the text $T(D^2)$ is solved by 3, and 4 only and induces the relation $1 \rightarrow_{T(D^2)} 4$ and $2 \rightarrow_{T(D^2)} 3$.
Clearly, the pair of texts $T(D^1)$ and $T(D^2)$ magically implements the problem.

The following claim generalizes the example:

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 procedure, then the problem is magically implementable.*

Proof. Enumerate the states of W^1 and W^2 as $a^1,\dots,a^K$ and $b^1,\dots,b^L$, respectively (without loss of generality assume that $K \le L$). Form a sequence $x^1,\dots,x^{2L}$ starting with b^1 and alternating between states in W^1 and states in W^2 (for the the case $K = L$, we need to denote $x^{2L+1} = x^1$). The states in W^2 appear in order, i.e. $b^1,\dots,b^L$. The states of W^1 appear cyclically (if necessary) in their order, i.e. $a^1,\dots,a^K$. Thus, for example, if $K = 3$ and $L = 5$ the sequence will be: $(x^1,\dots,x^{10}) = (b^1, a^1, b^2, a^2, b^3, a^3, b^4, a^1, b^5, a^2)$. The key in this construction is that there is no pair of states $a \in W^1$ and $b \in W^2$ such that a appears right after b somewhere in the sequence and appears right before b elsewhere 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 labels $d(\{a^1,a^2\}), d(\{a^2,a^3\}),\dots,d(\{a^K,a^1\})$; and
(ii) $|W^2|$ labels $d(\{x^k,x^{k+1}\})$, one for each $x^k \in W^2$.
For any $s \in W^1$, we have $|A(D^1,s)| \ge 2$ and therefore, there is no t for which $s \rightarrow_{M(D^1)} t$.
For any $t \in W^2$ the set $A(D^1,t)$ contains only one doubleton $\{t,s\}$ where s is the state

which appears right after t in the sequence $(x^1,\ldots,x^{2L})$ and therefore t does not satisfy $T(D^1)$. This s is in W^1 and $|A(D^1,s)| > 1$. Thus, there is a unique $s \in W^1$ such that $t \rightarrow_{T(D^1)} s$.

Similarly, the set D^2 consists of:

(i) L labels $d(\{b^1,b^2\}), d(\{b^2,b^3\}),\ldots, d(\{b^L,b^1\})$; and

(ii) one label $d(\{x^k,x^{k+1}\})$ for each $a \in W^1$, where k is the first place in the sequence at which $x^k = a$.

Figure 1 illustrates the construction of D^1 and D^2:

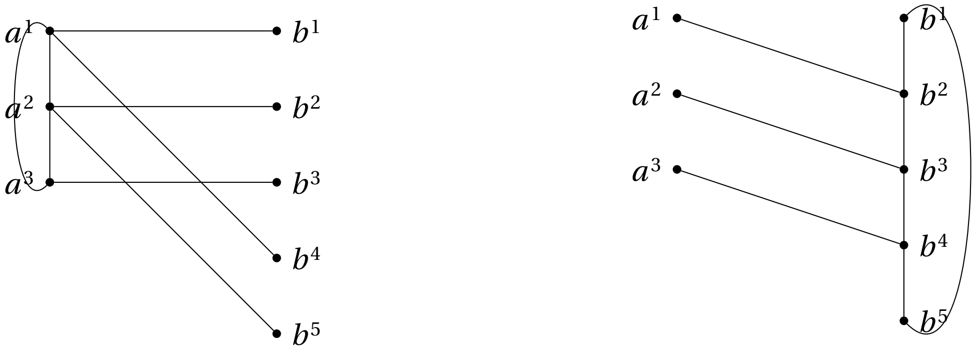

Figure 1. An illustration of the proof: $T(D^1)$ (on the left) and $T(D^2)$ (on the right) for the case of $K=3$ and $L=5$. A line connecting two states x and y indicates that the label $d(\{x,y\})$ is in the text.

For any $t \in W^2$, the set $A(D^2,t)$ contains at least two doubletons and therefore there is no s for which $t \rightarrow_{T(D^2)} s$. For any $s \in W^1$, the set $A(D^2,s)$ contains only one set $\{s,t\}$ where t appears right after s in the sequence $(x^1,\ldots,x^{2L})$ and therefore $|A(D^2,s)| = 1$. This t is in W^2 and $|A(D^2,t)| > 1$. Thus, there is a unique $t \in W^2$ such that $s \rightarrow_{T(D^2)} t$.

By the construction of the sequence $(x^1,\ldots,x^{2L})$ there is no case where t comes right after s and also s comes right after t. Thus, for no $s \in W^1$ and $t \in W^2$ do we have both $t \rightarrow_{T(D^1)} s$ and $s \rightarrow_{T(D^2)} t$. 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 sets with labels in D^i contain s and he will declare s. Party j starts with s and finds that only one set $\{s,t\}$ has a label in D^j. He then asks about the state t and ascertains that it belongs to two sets with labels in D^j and declares t. But party j has fallen into a trap! The principal can apply the honest-cheater asymmetry principle (given the state s, party j may cheat using t and if the state was t, then party i cannot cheat by declaring s) and concludes that j is the cheater.

4. Setting a logical riddle

This section investigates magical implementation assuming that the language of the principal and the parties' cheating procedure are those discussed in Glazer and Rubinstein (2012) except in the context of a *single*-agent implementation problem.

The implementation problem: Let $S = \{0,1\}^K$ where $K \geq 2$. A state $(x_1, \ldots, x_K)$ is a vector representing the truth values of the binary variables $v_1, \ldots, v_K$, where $x_k = 1$ indicates the "truth" of the variable v_k and $x_k = 0$ indicates the "falsity". Recall that we assume that both parties receive complete information about the state. Two states s and t are considered to be neighbors, denoted by sNt, if they differ in exactly one component.

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, and every ϕ_v is either v (the variable v) or $\neg\, v$ (the negation of v) and ϕ_z is either z or $\neg\, z$. For any such proposition ϕ, denote by $\alpha(\phi)$ the antecedent of ϕ and by $\beta(\phi)$ the consequent of ϕ (that is, $\phi = \alpha(\phi) \rightarrow \beta(\phi)$). For example, $\phi = v_1 \wedge \neg v_2 \rightarrow \neg v_4$ is such a proposition; its antecedent is $\alpha(\phi) = v_1 \wedge \neg v_2$ and its consequence is $\beta(\phi) = \neg v_4$. We interpret such a proposition as the statement: "If the state satisfies the antecedent of the proposition, then it should also satisfy its consequent." A proposition is *complete* if all K variables appear in it. A complete proposition excludes one state which satisfies the antecedent but not the consequent. For example, when $K = 3$, the complete proposition $v_1 \wedge \neg v_2 \rightarrow \neg v_3$ excludes the state $(v_1, v_2, v_3) = (1,0,1)$.

Finally, a *text* $T(\Phi)$ has the following form:

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

The interpretation of a text $T(\Phi)$ is the set of all states that satisfy all the propositions in Φ. For example: Let $\Psi = \{v_1 \wedge v_2 \rightarrow \neg v_3,\ \neg v_1 \wedge \neg v_2 \rightarrow v_3\}$. For the case $K = 3$, the text $T(\Psi)$ is satisfied by all states except $(0,0,0)$ and $(1,1,1)$.

One additional constraint placed 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 w appears in the consequents of both propositions – in one case as w and in the other as $\neg w$). For example, a text that includes two propositions, $v_1 \rightarrow v_3$ and $v_2 \rightarrow \neg v_3$, is not coherent.

The cheating procedure: A party that receives a text $T(\Phi)$ carries out the following procedure:

Step 1. Determine whether the true state satisfies all propositions in Φ.
If it does, then announce the true state.
If it does not, then go to Step 2.
Step 2. Find a new state, which was not examined earlier, such that for every variable in which the true state differs from the new state, there is a proposition in Φ satisfying: (i) both the true state and the new state satisfy its antecedent and (ii) the true state does not satisfy its consequent while the new one does.
If the new state satisfies the text, then announce it.
If it does not, then iterate Step 2, starting over from the true state.
Step 3. If you run out of states to check in Step 2, then announce the true state.

This cheating procedure induces for every text $T(\Phi)$ a binary relation $\rightarrow_{T(\Phi)}$ on S defined by: $s \rightarrow_{T(\Phi)} t$ if:

(i) s does not satisfy at least one of the propositions in Φ;

(ii) t satisfies all propositions in Φ; and

(iii) for every variable v for which s and t differ, there is a proposition $\phi \in \Phi$ such that s and t satisfy $\alpha(\phi)$ and s does not satisfy $\beta(\phi)$ while t does.

Thus, for example, given $\Psi = \{v_1 \wedge v_2 \rightarrow \neg v_3,\ \neg v_1 \wedge \neg v_2 \rightarrow v_3\}$, the induced relation consists of $(0,0,0) \rightarrow_{T(\Psi)} (0,0,1)$ and $(1,1,1) \rightarrow_{T(\Psi)} (1,1,0)$.

Claim B *If the principal is equipped with the above language and both parties use the above cheating procedure, then any problem $\langle S = \{0,1\}^K, W^1, W^2\rangle$ is magically implementable.*

Proof. The set S has a Hamiltonian cycle, that is, there is an anti-symmetric directed graph $\longrightarrow$ on S and an enumeration $x_1,\ldots,x_{2^K}$ of all states of S such that $x_k N x_{k+1}$ and $x_k \longrightarrow x_{k+1}$ for all k (and $x_{2^K} N x_1$ and $x_{2^K} \longrightarrow x_1$). Let $\longrightarrow^1$ be defined by $s \longrightarrow^1 t$ if $s \longrightarrow t$ and $s \in W^2$. Similarly, let $\longrightarrow^2$ be defined by $s \longrightarrow^2 t$ if $s \longrightarrow t$ and $s \in W^1$.

Figure 2 illustrates the above construction for $K = 3$, $W^1 = \{000, 111\}$ and $W^2 = S - W^1$.

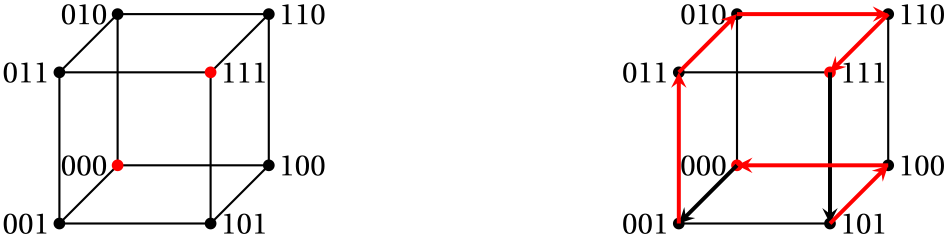

Figure 2. On the left, each node in the cube stands for a state in S. The edges represent the neighborhood relation. The set W^1 appears in red and W^2 appears in black.
On the right, a Hamiltonian cycle is indicated by a cycle of arrows. The six red arrows form $\longrightarrow^1$ and the two black arrows form $\longrightarrow^2$.

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. In other words, 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 $s_l = t_l = 0$.

Let Φ^i be the set of all propositions $\phi_{s,t}$ for which $s \longrightarrow^i t$. To verify that $(T(\Phi^1), T(\Phi^2))$ magically implements $\langle S, W^1, W^2\rangle$, notice first that the interpretation of $T(\Phi^i)$ is W^i. Now let $s \in W^i$. Party i declares s and party j notices that $s \notin W^j$ and tries to cheat. The cheating procedure only allows him to consider the unique t for which $s \longrightarrow^j t$, which is also the unique t for which $s \longrightarrow t$. If $t \in W^i$, then he fails to find a suitable state with which to cheat (formally, $s \not\rightarrow_{T(\Phi^j)} t$) and declares s. If $t \in W^j$, then $s \rightarrow_{T(\Phi^j)} t$, he declares t and falls into a trap! The principal can apply the honest-cheater asymmetry principle and will conclude that j is cheating since if the true state were t, then party i would either declare t (namely, for no r do we have $t \rightarrow_{T(\Phi^i)} r$) or there is a unique state r for which $t \rightarrow_{T(\Phi^i)} r$, which requires that $t \rightarrow^i r$ and therefore $t \rightarrow r$. Given the antisymmetry of $\rightarrow$, this implies that $r \neq s$. ■

We now contrast the model with the single agent model studied in Glazer and Rubinstein (2012). The set of the states, the principal's language and the cheating procedure are the same. The requirement here that the implementation should be via texts that fully describe the circumstances under which each party deserves the prize is analogous to our notion there of "truthful implementation". There, we showed that a set W can be truthfully implemented 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. As demonstrated there, in many examples the set W and its complement do not satisfy this condition and the principal cannot achieve his goal by involving only *one* of the two parties. Claim B shows that in those cases, the principal can nevertheless magically implement his goal by involving *both* parties.

5. Making cheating too risky

We now revisit the example in the introduction. If the student received the actual answer to the question, then the invigilator that likes him will be reluctant to claim that the student only received a recommendation to start with a particular equation. Such a claim requires the invigilator to guess the equation that the whispered message "$\alpha = 3$" refers to. Since, unlike the principal, he is not familiar with the exam this involves taking the risk of being caught cheating. Conversely, if the student merely received a recommendation to start with a specific equation, then the hostile invigilator is not taking any risk by cheating, i.e. he can solve the equation and claim that the student received the actual solution to the equation. This asymmetry enables the principal to infer that the student did not receive the actual answer to the question when one invigilator claims that the student received the answer while the other claims that the student only received a recommendation to start with a particular question that indeed appears in the exam.

We now formalize the example within a broader framework.

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

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

μ: A probability measure on S.

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

Assume that any information set in I is either a subset of W^1or of W^2, that is the information held by the parties is sufficient to determine the party that deserves the prize. This assumption is not imposed on I_p since the principal would then be able to make a correct decision without eliciting any information from the parties.

The language: A text has two parameters: Y which is a subset of S and 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 $W \subseteq S$, which is a union of cells in I and is interpreted as a description of the cells in which the party that receives the text deserves the prize. A text $T(Y, W)$ has the following form:

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

The interpretation of a text $T(Y, W)$ is the set W itself.

The procedure of cheating: The cheating procedure in this section is more conventional than in the previous two. After receiving a certain information set, a party's willingness to cheat by reporting a false information set depends on his fear of getting caught. We say that a party is *caught cheating* if his announcement and the principal's knowledge do not intersect. It is assumed that a party believes that the information he receives from the principal is truthful (but does not necessarily consist of all the information possessed by the principal). It is further assumed that a party considers cheating (by announcing an untrue information set in I) only if he believes (given the prior, the information set in I which the parties initially received, the principal's announcement and the principal's information structure I_p) that the probability of getting caught does not exceed some threshold τ.

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 $K \nsubseteq W$, 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, $\mu(\{s|\ L \cap I_p(s) = \emptyset\} \mid K \cap Y) \leq \tau$.
If you find such an L, then declare it; otherwise declare K.

The procedure generates the binary relation $\rightarrow_{T(Y,W)}$ on the information sets in I, defined by $K \rightarrow_{T(Y,W)} L$ if $K \nsubseteq W$, $L \subseteq W$, 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 need to modify the definition of magical implementation.

First, we require that when designing the text $T(Y,W)$ for party i the principal is honest in two senses:
(i) W is a truthful description of the circumstances in which i deserves the prize, that is, $W = W^i$.
(ii) The information set Y is correct given the information that the principal possesses.
Accordingly, we model the principal's strategy as a partition I_p^* coarser than I_p with the interpretation that in state s the principal provides the text $T(I_p^*(s), W^i)$ to party i (where $I_p^*(s)$ is the cell in the principal's information structure I_p^* that contains s).

Second, the texts are required, as before, to have the property that if the parties follow the cheating procedure, then whenever their claims differ the principal will be able to activate the honest-cheater asymmetry principle and correctly determine the deserving party.

To conclude, 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 $K \nrightarrow_{T(Y,W^j)} L$.

Notice the difference between catching a cheating party and inferring that a party is cheating using the honest-cheater asymmetry principle. The former occurs only when the principal has solid proof that he is cheating, namely his statement contradicts the information possessed by the principal. In our setup, in the case of a dispute and if no one is caught cheating, then the principal awards the prize to one of the parties based on the honest-cheater asymmetry principle, but *without any solid proof* that the other party is the cheater.

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

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

Each party's information partition is $I=\{\{a,b\},\{c\},\{d\}\}$ and the principal's information partition is $I_p=\{\{a,c\},\{b,d\}\}$. We assume that $\mu(a)=\mu(b)>0$. Let party 1 be the hostile party and party 2 the friendly party. The winning sets are $W^1=\{a,b\}$ and $W^2=\{c,d\}$.

When $\tau<1/2$, the texts $T(S,W^1)$ and $T(S,W^2)$ (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 1 is not afraid to cheat and declare $\{a,b\}$ since with certainty he will 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 2 is afraid to report $\{c\}$ since if he does, then he will be caught cheating with probability $\mu(b)/[\mu(a)+\mu(b)]=1/2>\tau$. Similarly, he is afraid to cheat by reporting $\{d\}$.

Thus, $\{c\}\rightarrow_{T(X,W^1)}\{a,b\}$ and $\{d\}\rightarrow_{T(X,W^1)}\{a,b\}$ but $\{a,b\}\nrightarrow_{T(X,W^2)}\{c\}$ and $\{a,b\}\nrightarrow_{T(X,W^2)}\{d\}$ and the problem is magically implementable.

Note that if $\tau>1/2$, then we will also have $\{a,b\}\rightarrow_{T(X,W^2)}\{c\}$ and $\{a,b\}\rightarrow_{T(X,W^2)}\{d\}$ 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 2 (with information $\{a, b\}$) will not be afraid to cheat by declaring $\{c\}$. At the same time, party 1 will not be afraid to cheat in state c by declaring $\{a, b\}$. Thus, the principal's full information revelation will not enable him to perform his magic.

The asymmetry between a cheater and a truth-teller arose on its own in the above example. We will see that in the general framework, the principal may wish to manipulate the situation by sharing some of his information with the parties in order to create asymmetry between a truth-teller and a cheater.

Setting a trap by providing additional information: The following example demonstrates that providing information is sometimes a necessary tool for magical implementation:

$S = \{a, b, c, d, e, f\}$,
$W^1 = \{a, b, c\}$, $W^2 = \{d, e, f\}$,
$I = \{W^1, W^2\}$, $\mu =$ the uniform probability measure on S, and
$I_p = \{I_1 = \{a\}, I_2 = \{c, d\}, I_3 = \{b, e\}, I_4 = \{f\}\}$.

The case of $\tau < \mu(a)/\mu(W^1) = \mu(f)/\mu(W^2) = 1/3$ is trivial. Magical implementation is possible without the principal giving the parties any additional information: after receiving the information set W^1, party 2 will not cheat by declaring W^2 because he would be caught cheating with probability $1/3 = \mu(a)/\mu(W^1) > \tau$ and analogously for W^2. Formally, the relations $\rightarrow_{T(S,W^1)}$ and $\rightarrow_{T(S,W^2)}$ are empty.

When $1/3 = \mu(a)/\mu(W^1) = \mu(f)/\mu(W^2) < \tau$, magical implementation is not possible without providing additional information, since both $W^1 \rightarrow_{T(S,W^2)} W^2$ and $W^2 \rightarrow_{T(S,W^1)} W^1$. That is, in W^1 party 2 is not deterred from announcing W^2 and in W^2 party 1 is not deterred from announcing W^1.

The more interesting case is $1/3 = \mu(a)/\mu(W^1) < \tau < \mu(a)/[\mu(a) + \mu(c)] = 1/2$. As mentioned, magical implementation is impossible without the principal providing additional information. It will be shown that magical implementation is feasible if the principal commits to supply additional information according to the information structure $I_p^* = \{I_1 \cup I_3, I_2 \cup I_4\}$.

Assume that the principal has announced $I_1 \cup I_3$ (or analogously $I_2 \cup I_4$). In the case they are initially informed of W^1, the parties 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 informed of W^2, the parties will conclude that the state of nature is e and therefore the principal has the information $\{b, e\}$ and party 1 will not be caught cheating if he announces W^1. Thus, $W^1 \nrightarrow_{T(I_1 \cup I_3, W^2)} W^2$ and $W^2 \rightarrow_{T(I_1 \cup I_3, W^1)} W^1$ and in the case of disagreement the principal will be able to use the honest-cheater asymmetry principle to infer that 1 is the cheater.

6. Comments by Ariel Rubinstein

a. I am fully aware (and 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 any applications. The paper is short. 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 with 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.

b. The paper contributes to "implementation theory", although we diverge from the traditional commitment to game-theoretical tools. While we of course acknowledge the appeal of game-theoretical models – having employed them ourselves in the past – we wish to challenge the status of game theory as the exclusive approach used in the implementation literature.

c. I doubt that people in the real world use any of the described procedures of cheating "as is". In Glazer and Rubinstein (2012), we claim – quite convincingly if I may say so myself – that hints of the procedure described there have been observed in experimental data. Nevertheless, we have refrained from running another set of experiments here. Such experiments might have been interesting, but they are not necessary in order to construct interesting stories.

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.

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.

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