Mechanism Design and the Revelation Principle

CSCI 1440/2440

2025-02-05

First, we introduce the mechanism design formalism, assuming a Bayesian game setting. Then, we introduce the notion of an indirect mechanism, presenting English, Japanese, and Dutch auctions as examples. Finally, we cover the Revelation Principle,[1] which can transform any mechanism into a direct incentive compatible one, thereby converting a game-theoretic problem into a decision-theoretic one.

1 *Mechanism Design Framework*

Mechanism design has been referred to as the engineering branch of game theory. It is concerned with designing mechanisms (i.e., games) such that the outcomes that arise when the games are played by rational agents (i.e., the equilibria) achieve some desiderata.

The mechanism design framework thus consists of three parts: the mechanism formalism, which builds on Bayesian games; solution concepts, or equilibria, which serve to predict the outcome of the mechanism/game; and desiderata, or objectives.

Mechanisms The mechanism design paradigm transpires as follows: A designer selects a mechanism, meaning the rules of the game. After observing the mechanism/game, the participants make their decisions. The rules of the game then determine the outcome that is realized, as a function of the participants' choices. Furthermore, ensuing utilities depend on this outcome—in general, for both the mechanism designer and the participants.

This interaction between a mechanism designer and participants can be modeled as a multi-stage game. We restrict our present attention to two stages: the mechanism announces the game rules in the first stage, and then the participants play a simultaneous-move (i.e., one-shot) Bayesian game in the second stage.[2]

Recall that a Bayesian (i.e., incomplete information) game is given by $\mathcal{B} \doteq \langle [n], \{T_i\}_{i \in n}, \{A_i\}_{i \in n}, \{u_i\}_{i \in n}, F \rangle$. As usual, $[n] = \{1, \ldots, n\}$ is the set of players, and we write $T = \prod_{i=1}^{n} T_i$ and $A = \prod_{i=1}^{n} A_i$ to denote the denote the joint type and action spaces, respectively. Before play commences, each player is informed of her type (private information), sometimes called a signal, drawn from her set of available types T_i. Conditioned on this private information, she invokes a strategy, $s_i : T_i \to A_i$. Player i's **utility** $u_i : A \times T \to \mathbb{R}$ depends on both the players' collective actions and (in general) all players' types. The joint distribution F over all players' types is assumed to be common

[1] Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981; and Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979

[2] Such games are an instance of single-leader multiple-follower Stackelberg games.

Heinrich von Stackelberg. *Marktform und gleichgewicht*. Julius Springer, 1934 knowledge, known to both the players and, in a mechanism design setting, the mechanism designer as well.

A **mechanism** $\mathcal{M} \doteq \langle \mathcal{B}, \Omega, g \rangle$ builds on this definition with an **outcome space**, denoted Ω, and an **outcome function** $g : A \to \Omega$, which maps the players' collective actions to an outcome. That is, $g(\mathbf{s}(\mathbf{t}))$ is the **outcome** when player i of type t_i plays strategy s_i and the remaining players of type $\mathbf{t}_{-i}$ play strategy $\mathbf{s}_{-i}$.[3] Player i's **utility** $u_i : \Omega \times T \to \mathbb{R}$ in a mechanism thus depends on the outcome and, as usual, (in general) all players' types. Finally, the actions in a mechanism are called **reports**, or **messages**, as they are indeed messages, sent from the players to the mechanism designer, who is often called the **center**.

Equilibria As mechanisms encode an incomplete-information games, solutions typically take the form of joint strategy profiles $\mathbf{s}^*$ that the players are predicted to play. Dominant-strategy or ex-post Nash equilibria when they exist, and otherwise Bayes-Nash equilibria, are applied to make these predictions.

Desiderata There are two prevalent approaches to problems in mechanism design, one set-valued, and one numeric-valued.

First, it may be the designer's goal to **implement** a social choice correspondence $f : T \rightrightarrows \Omega$, so that the equilibria of the mechanism coincide with the social choices. Implementation can be strong or weak, depending on whether all or some equilibria coincide with the social choices. At one end of the spectrum, a design may deemed successful if $g(\mathbf{s}^*(\mathbf{t})) = f(\mathbf{t})$, for all type profiles $\mathbf{t}$ and *all* equilibria $\mathbf{s}^*(\mathbf{t})$.[4] Alternatively, a design may deemed successful if for all type profiles $\mathbf{t}$, there exists an equilibrium $\mathbf{s}^*(\mathbf{t})$ s.t. $g(\mathbf{s}^*(\mathbf{t})) \in f(\mathbf{t})$.

Alternatively, the designer may formulate her goals in terms of a numeric function of a solution to the induced game that it seeks to maximize,[5] such as **expected welfare** $\mathbb{E}_{\mathbf{t} \sim F}\left[W(g(\mathbf{s}^*(\mathbf{t}))\right]$, or **expected revenue** $\mathbb{E}_{\mathbf{t} \sim F}\left[R(g(\mathbf{s}^*(\mathbf{t}))\right]$.[6]

In much of the mechanism design literature, the problem is greatly simplified by reliance on the revelation principle, which argues that the strategic outcome of any mechanism can be replicated by a direct mechanism (i.e., a mechanism in which agents simply report types).

2 *Direct vs. Indirect Mechanisms*

Direct mechanisms are potentially much simpler for participants than indirect mechanisms, as they simply seek to elicit their private information, rather than some function of that information.

[3] For example, the outcome function of an auction maps a profile of bids to an outcome, which is described by an allocation and a payment rule.

[4] Abusing notation, here $g(\mathbf{s}^*(\mathbf{t}))$ denotes the set of all outcomes corresponding to any equilibrium $\mathbf{s}^*(\mathbf{t})$.

[5] Implicit in this goal is an equilibrium selection problem that cannot be overlooked; in case the predicted solution is not unique, welfare/revenue could, for example, be computed in either the worst or the average case.

[6] Note that within the MD framework these goals could easily be relaxed so that, for example, welfare/revenue exceeds some threshold value (or is maximized) with high probability.

Definition 2.1. A **direct mechanism** is one in which the space of possible reports is equal to the space of possible types. All other mechanisms are called **indirect**.

First-, second-, third, etc.-price and all-pay auctions are all examples of direct mechanisms, when the space of possible bids is restricted to the space of possible types, a natural assumption. Examples of indirect mechanisms include the **English** and **Japanese**, both **ascending**, auctions; and the **Dutch**, a **descending**, auction.

Example 2.2. The Japanese auction consists of a number of rounds. On round $k = 1, 2, \ldots$, the auctioneer offers the good at price $p = k\epsilon$, for some small $\epsilon > 0$, asking all bidders if they are interested in the good at that price. The auction continues so long as more than one bidder is interested. The auction terminates, say at round t, when one or fewer bidders remain interested. If there is one interested bidder at round t, then she wins, paying $t\epsilon$; if there are no interested bidders then a winner is selected at random from the set of interested bidders during round $t - 1$. This winner pays $(t - 1)\epsilon$.

In this auction, actions consist of t binary answers to queries "Would you like the good at price p?". In practice, it may be easier for bidders to answer queries like this one, rather than articulate an exact value for a good, as is required in a sealed-bid auction. English auctions, perhaps the most widely used ascending auctions,[7] offer a compromise between Japanese ascending auctions and sealed-bid second-price. In English auctions, bidders respond to the query "Name a price higher than p at which you would like the good."

Example 2.3. The Dutch auction also consists of a number of rounds, but in this case, the price p is initialized high enough so that no bidders are interested. The price is then decremented successively by ϵ—at a known clock speed—until a bidder (or a set of bidders) declares interest in the good. That bidder is then declared the winner;[8] the winner receives the good and pays the final price.

Not surprisingly, Dutch auctions are popular in the Netherlands, where they are used to sell flowers—perishable goods—where the clock speed dictates a worst-case end time for an auction.

3 *Incentive Compatibility*

It would be difficult for a mechanism that operates under misleading or incorrect information to achieve its desiderata. So for a direct mechanism to be successful, it should incentivize its participants to report their private information *truthfully*. In other words, we are

[7] This is not to say that second-price (i.e., Vickrey) auctions are not used in practice. On the contrary, stamp auctioneers used this mechanism to sell stamps by mail as early as the late 1800s, before Vickrey was born!

[8] (or a tie is broken randomly)

specifically interested in designing direct mechanisms for which truthtelling behavior is an equilibrium. When this condition holds of a direct mechanism, it is called incentive compatible.

Definition 3.1. In a direct mechanism, reporting (i.e., "playing") one's true type is called **truthtelling**: i.e., $s_i(\mathbf{t}_i) = \mathbf{t}_i$, for all $i \in [n]$.

Just as there are various notions of equilibrium in Bayesian games (BNE, EPNE, and DSE), there are corresponding notions of incentive compatibility (BIC, EPIC, and DSIC).

In the following definitions, we denote truthtelling by $\mathbf{s}^*$, and true types by $\mathbf{t}^*$: i.e., $\mathbf{s}^*(\mathbf{t}^*) = \mathbf{t}^*$. Moreover, because strategies map from types to types in direct mechanisms, quantifying over all other-agent strategies $\mathbf{s}_{-i}(\mathbf{t}_{-i})$ is equivalent to quantifying over all types $\mathbf{t}_{-i}$.

Definition 3.2. A (direct) mechanism is said to be **Bayesian incentive compatible (BIC)** iff truthtelling is a BNE: i.e., $\mathbf{s}^*$ is s.t.

$$\mathop{\mathbb{E}}_{\mathbf{t}_{-i} \sim F_{\mathbf{t}_{-i} \mid t_i}} \left[u_i(g(s_i^*(t_i), \mathbf{s}_{-i}^*(\mathbf{t}_{-i})); \mathbf{t}) \right] \geq \mathop{\mathbb{E}}_{\mathbf{t}_{-i} \sim F_{\mathbf{t}_{-i} \mid t_i}} \left[u_i(g(t_i', \mathbf{s}_{-i}^*(\mathbf{t}_{-i})); \mathbf{t}) \right], \quad \forall i \in [n], \forall t_i, t_i' \in T_i.$$

Equivalently, $\mathbf{t}^*$ is s.t.

$$\mathop{\mathbb{E}}_{\mathbf{t}_{-i} \sim F_{\mathbf{t}_{-i} \mid t_i^*}} \left[u_i(g(t_i^*, \mathbf{t}_{-i}^*); \mathbf{t}) \right] \geq \mathop{\mathbb{E}}_{\mathbf{t}_{-i} \sim F_{\mathbf{t}_{-i} \mid t_i^*}} \left[u_i(g(t_i', \mathbf{t}_{-i}^*); \mathbf{t}) \right], \quad \forall i \in [n], \forall t_i' \in T_i.$$

Definition 3.3. A (direct) mechanism is **ex-post Nash incentive compatible (EPIC)** iff truthtelling is an EPNE: i.e., $\mathbf{s}^* = (s_i^*, \mathbf{s}_{-i}^*)$ is s.t.

$$u_i(g(s_i^*(t_i), \mathbf{s}_{-i}^*(\mathbf{t}_{-i})); \mathbf{t}) \geq u_i(g(t_i', \mathbf{s}_{-i}^*(\mathbf{t}_{-i})); \mathbf{t}), \quad \forall i \in [n], \forall t_i, t_i' \in T_i, \forall \mathbf{t}_{-i} \in T_{-i}.$$

Equivalently, $\mathbf{t}^*$ is s.t.

$$u_i(g(t_i^*, \mathbf{t}_{-i}^*); \mathbf{t}) \geq u_i(g(t_i', \mathbf{t}_{-i}^*); \mathbf{t}), \quad \forall i \in [n], \forall t_i' \in T_i.$$

Definition 3.4. A (direct) mechanism is **dominant strategy incentive compatible (DSIC)** iff truthtelling is a DSE: i.e., $\mathbf{s}^*$ comprises n strategies s_i^*, one per player i, s.t.

$$u_i(g(s_i^*(t_i), \mathbf{s}_{-i}(\mathbf{t}_{-i})); \mathbf{t}) \geq u_i(g(t_i', \mathbf{s}_{-i}(\mathbf{t}_{-i})); \mathbf{t}), \quad \forall i \in [n], \forall t_i, t_i' \in T_i, \forall \mathbf{s}_{-i} \in S_{-i}, \forall \mathbf{t}_{-i} \in T_{-i}.$$

Equivalently, $\mathbf{t}^*$ is s.t.

$$u_i(g(t_i^*, \mathbf{t}_{-i}); \mathbf{t}) \geq u_i(g(t_i', \mathbf{t}_{-i}); \mathbf{t}), \quad \forall i \in [n], \forall t_i, t_i' \in T_i, \forall \mathbf{t}_{-i} \in T_{-i}.$$

As it turns out, the EPIC and DSIC equilibrium concepts coincide in direct mechanisms. A proof of the following proposition can be found in the appendix.

Proposition 3.5. *Truthtelling in a direct mechanism is a dominant strategy equilibrium (DSE) iff it is an ex-post Nash equilibrium (EPNE).*

This equivalence does not carry over to indirect mechanisms, however, where the DSIC property will prove too strong to hope for, leaving us to settle for EPIC mechanisms.[9]

[9] Haha! Imagine having to settle for an EPIC mechanism!

4 The Revelation Principle

In our search for mechanisms that satisfy certain desiderata, the revelation principle allows us to restrict our attention to direct mechanisms for which truthtelling is an equilibrium. The principle follows via construction: we construct a direct mechanism in which "the agents don't have to lie, [because] the mechanism lies for them."

Theorem 4.1 (Revelation Principle)**.** *If a (possibly indirect) mechanism $\mathcal{M}$ implements a social choice function*[10] *f in dominant strategies (resp. via a Bayes-Nash equilibrium), then there exists a DSIC (resp. BIC) direct mechanism that likewise implements f.*

Proof. Given a (possibly indirect) mechanism $\mathcal{M}$ that implements the social choice function f at equilibrium $\mathbf{s}$, we construct a truthful (i.e., incentive compatible) direct mechanism $\mathcal{M}^*$ as follows:

- Elicit types $t_1, t_2, \ldots, t_n$ from all the agents.
- Simulate $\mathcal{M}$ by performing i's equilibrium action $s_i(t_i)$ on her behalf, given her reported type t_i.
- Return the outcome produced by $\mathcal{M}$, namely $g(\mathbf{s}(\mathbf{t}))$ (which, by assumption, equals $f(\mathbf{t})$).

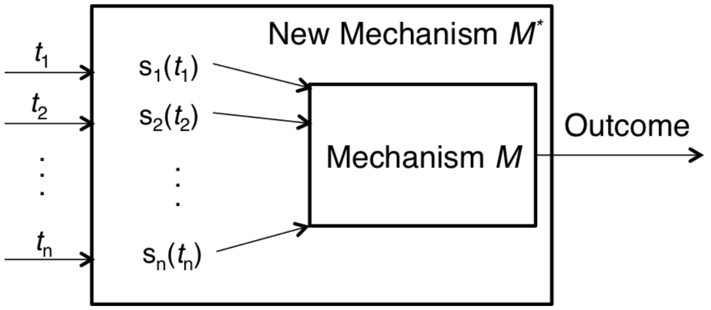

Figure 1: The Revelation Principle

We can think of the construction (see Figure 1) as a machine that first asks all agents for their types, and then runs the equilibrium strategy on their behalf. Each agent reports a (possibly false) type t_i to the direct mechanism $\mathcal{M}^*$, which simulates $s_i(t_i)$. If agent i lies to the machine (and if she is the only one lying), the machine will run everyone else's equilibrium strategy based on their true types, except for agent i's. The outcome will be exactly the same outcome as running $\mathcal{M}$, assuming i deviates from its equilibrium strategy. But this deviation was not in i's best interest in $\mathcal{M}$, so likewise, it is not in i's best interest in $\mathcal{M}^*$. Therefore, $\mathcal{M}^*$ is truthtelling (i.e.: the agents are incentivized to report their true types).

[10] A **social choice function** $f : T \to \Omega$ assigns unique outcomes to types.

By construction, the direct mechanism $\mathcal{M}^*$ implements the social choice function f. Further, it is DSIC, if $\mathbf{s}$ is a dominant-strategy equilibrium, and BIC, otherwise.

□

Example 4.2. Consider a modified second-price auction $\mathcal{M}$ in which the winner is the highest bidder, and she pays *twice* the second-highest bidder's bid. This auction has a DSE in which the bidders bid half their values. (Why?)

Given $\mathcal{M}$ and the aforementioned DSE, the mechanism $\mathcal{M}^*$ constructed according to the revelation principle, works as follows:

- Elicit all bidders' values $v_1, v_2, \ldots, v_n$.
- For each bidder i, submit the sealed bid $v_i/2$.
- Return the outcome produced by original auction $\mathcal{M}$, namely the highest bidder wins and pays *twice* the second-highest bid.

The mechanism $\mathcal{M}^*$ has the following three properties:

1. DSIC: Truthtelling is a dominant-strategy equilibrium.
2. The highest bidder in $\mathcal{M}^*$ (who by 1, has the highest value) wins.
3. This winner pays twice the second-highest bid in $\mathcal{M}$, which by 1, is *twice half* of the second-highest value, i.e., the second-highest value, while no other bidders make any payments.

Therefore, $\mathcal{M}^*$ is the second-price auction! Indeed, there are no DSIC auctions for this setting other than the second-price auction.[11]

By contraposition, the revelation principle states: if a social choice function cannot be implemented by a DSIC (resp. BIC) direct mechanism, then it cannot be implemented in dominant strategies (resp. via a Bayes-Nash equilibrium) by *any* (even indirect) mechanism. It is thus useful as a theoretical tool, because it allows us to explore the limits of possibility in our search over mechanisms, by ruling out as potential candidates *all* indirect mechanisms,[12] like the English, Japanese, and Dutch auctions, as well as direct mechanisms, like the first-price auction,[13] where truthtelling is not an equilibrium.

A DSIC and EPIC Coincide in Direct Mechanisms

Remark A.1. In a direct mechanism, EPIC is equivalent to DSIC.

Proof. DSIC implies EPIC, so it suffices to show that EPIC also implies DSIC. Let M be an EPIC, direct mechanism in which the space of possible actions equals the space of possible types.

[11] modulo possible additive offsets to the payment rule to satisfy individual rationality: i.e., to ensure $u_i(\omega, t_i) \geq 0$, for all players $i \in [n]$, outcomes $\omega \in \Omega$, and types $t_i \in T_i$

[12] We will nonetheless return to the study of indirect mechanisms later on in the course.

[13] Ad auctions, which were based on a second-price model for a decade or so, recently migrated to a first-price model. In other words, first-price auctions remain highly relevant in practice.

Since M is EPIC, for all bidders $i \in [n]$ and for all (true) type profiles $\mathbf{t} \in T$, truthful bidding satisfies

$$u_i(g(t_i, \mathbf{t}_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{t}_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i.$$

Our goal is to show that M is also DSIC: i.e., for all bidders $i \in [n]$ and for all (true) type profiles $\mathbf{t} \in T$,

$$u_i(g(t_i, \mathbf{b}_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{b}_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i, \forall \mathbf{b}_{-i} \in B_{-i}.$$

By assumption the mechanism is direct: i.e., $B_i = T_i$ for all bidders $i \in [n]$. It thus suffices to show: for all bidders $i \in [n]$ and for all (true) types $t_i \in T_i$,

$$u_i(g(t_i, \mathbf{t}_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{t}_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i, \forall \mathbf{t}_{-i} \in T_{-i}.$$

Fix a bidder i with true type t_i. EPIC implies that truthful bidding is a best response for bidder i, assuming the others are also bidding truthfully. In particular, when the other players' true type profile is $\mathbf{t}'_{-i} \in T_{-i}$, it holds that:

$$u_i(g(t_i, \mathbf{t}'_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{t}'_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i.$$

Similarly, when the other players' true type profile is $\mathbf{t}''_{-i} \in T_{-i}$, it holds that:

$$u_i(g(t_i, \mathbf{t}''_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{t}''_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i.$$

In other words, truthful bidding is a best response for bidder i *regardless of the other bidders' type profiles*. That is, truthful bidding is optimal for bidder i, for *all* other-agent type profiles: i.e.,

$$u_i(g(t_i, \mathbf{t}_{-i}); \mathbf{t}) \geq u_i(g(t'_i, \mathbf{t}_{-i}); \mathbf{t}), \quad \forall t'_i \in T_i, \forall \mathbf{t}_{-i} \in T_{-i}.$$

Since bidder i was arbitrary, truthful bidding is a dominant-strategy equilibrium (i.e., it is optimal for all bidders $i \in [n]$). □

References

[1] Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979.

[2] Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

[3] Heinrich von Stackelberg. *Marktform und gleichgewicht*. Julius Springer, 1934.