

The Mechanism Design Toolkit

Architecting Human + AI Markets

Bellagio Human-AI Economy Convening · May 2026 · v1

Mechanism design is about designing the rules of a market or institution so that self-interested participants, each holding private information, are led to report truthfully and behave in ways that produce a desired outcome (Hurwicz 1960, 1972; Myerson 1979, 1981). The designer chooses the rules and participants their strategies. The goal is to make the rules such that the participants’ best strategy is to tell the truth.

Mechanism design addresses several recurring market problems:

- *Private information and adverse selection.* One side knows something the other cannot observe: a model provider knows its true capabilities, a content creator knows relevance to a query, a skill creator knows whether their instructions actually work. Low-quality participants crowd out high-quality ones when buyers cannot tell the difference.
- *Strategic manipulation and incentive compatibility.* Agents will misreport if it benefits them: shading bids, gaming benchmarks, flooding a marketplace with low-quality content. The mechanism must make truthful behaviour each agent’s best strategy.
- *Externalities and missing markets.* AI training data has public-good characteristics: it is hard to exclude and non-rival in use. No established markets exist for training data compensation, skill monetization, or grounding layer inclusion. When a market is missing entirely, mechanism design alone is not enough; institutions, rights frameworks, or regulation may be needed to create the conditions for a market to exist.
- *Market thickness and the chicken-and-egg problem.* A market needs enough participants on both sides to function. Mechanism design can create thickness where none exists: kidney exchanges pool incompatible donor-patient pairs into a thick matching market; Uber aggregates scattered riders and drivers so that batching and matching algorithms become possible. Most AI markets today are thin or non-existent. The skills marketplace needs both creators and users; the grounding layer needs both content providers and AI companies willing to pay. Building thickness is itself a design problem.
- *Principal-agent problems.* When one party (the principal) delegates decisions to another (the agent), their incentives may diverge. This is the defining feature of agentic AI markets: a user delegates shopping to an AI agent that optimizes a proxy for their true preference; a content creator delegates distribution to a platform that may not prioritize their interests. Moral hazard is one dimension of this: after delegation, the agent’s actions become hard to observe.

Core Concepts

Mechanism. A *mechanism* is a specification of how economic decisions are determined as a function of the information reported by individuals (Myerson, Palgrave). More concretely: participants send messages (reports, bids, claims), and a pre-specified rule assigns an *outcome* (an allocation of goods or services, and payments) for every collection of messages received (Hurwicz 1960). An auction, a matching platform, and a pricing algorithm are all mechanisms.

Private Information (Types). Each agent holds *private information* (their “type”) that others cannot directly observe. This may include their true valuation of a good, their cost of providing a service, or the quality of content they can supply. The mechanism designer must work *with* this asymmetry, not around it.

Incentive Compatibility (IC). A mechanism is *incentive-compatible* if it is a rational equilibrium for everyone to be both honest and obedient (Myerson 2008). Honest, because when the plan depends on private information that is hard to observe, people must have an incentive to share it truthfully (the *adverse selection* problem). Obedient, because when the plan requires actions that are hard to monitor, people must have an incentive to act as the plan prescribes (the *moral hazard* problem). In the strongest form (*dominant-strategy IC*), truthful reporting is optimal regardless of what others do; this is what the Vickrey auction achieves.

Individual Rationality (IR) / Participation Constraint. No agent should be made worse off by participating in the mechanism than by their *outside option* (Hurwicz 1972). If participation is voluntary, the mechanism must leave every agent at least as well off as non-participation. The stronger the outside option, the harder the design problem.

Revelation Principle. Any equilibrium outcome of *any* mechanism can be replicated by an incentive-compatible direct mechanism in which agents simply report their types truthfully (Gibbard 1973; Dasgupta, Hammond & Maskin 1979; Myerson 1979). The designer can restrict attention to the well-structured class of direct mechanisms when searching for the optimal design, and then “translate back” to a practical implementation.

The Toolkit

1. Auctions elicit agents’ private valuations through payment rules that make truth-telling incentive-compatible. In a *second-price (Vickrey) auction*, the highest bidder wins but pays the *second*-highest bid. Because your bid never determines your own price, truthful bidding is a dominant strategy. The Vickrey-Clarke-Groves (VCG) mechanism generalizes this: each agent “internalizes” the total social surplus, making truth-telling dominant (Vickrey 1961; Clarke 1971; Groves 1973). *Use when:* allocating scarce positions (ad slots, grounding layer inclusion, API access) where agents know their own value but the platform does not.

2. Information Cascades & Herding. When agents make decisions *sequentially* and observe prior choices but not prior private signals, rational Bayesian updating can lead agents to ignore their own private information and follow the crowd. A few early signals, even incorrect ones, can lock an entire market onto the wrong product or technology. This is rational at the individual level but collectively pathological. *Use when:* AI agents purchasing on behalf of users, product recommendations, technology adoption. *Risk:* homogeneous AI models cascade faster and more uniformly than humans.

3. Search & Discovery (Explore vs. Exploit). Agents face a fundamental tradeoff: *exploit* the best-known option or *explore* unknown alternatives that may yield higher long-run payoff. If all agents exploit, promising new options are never discovered. **Congestion** compounds the problem: when everyone flocks to the “known good,” its quality degrades. *Use when:* there are too many options to sift through and agents need help finding the right one. Applies to skill marketplaces (which skill fits this task?), content recommendation (which source is most relevant?), and any market where new entrants struggle to get noticed. Mechanisms can incentivize exploration through subsidized trials, diversified recommendations, or batching and matching (as Uber/Lyft do with riders and drivers).

4. Shapley Value. The Shapley value assigns each contributor a payoff equal to their average marginal contribution across all possible orderings in which the coalition could have formed. It is the *unique* division satisfying efficiency, symmetry, and additivity axioms. In data markets, any single data point has a tiny marginal value, but data collectively is enormously valuable; Shapley provides a principled resolution. Computing exact Shapley values is intractable at scale, but practical implementations use Truncated Monte Carlo or KNN-Shapley approximations (Jia et al. 2019). *Use when:* training data attribution (cf. Data Shapley, Ghorbani & Zou 2019), content creator compensation (cf. user-centric streaming models), fair division of jointly created value.

AI Market Design Challenges

Apply the mechanism design toolkit to four emerging AI markets

Analysis Framework

For your assigned market, work through these questions. The goal is to design a mechanism, a set of rules for allocation and payments, that aligns individual incentives with good collective outcomes.

- 1. Agents.** Who are the participants? What does each want to maximize? Who is the mechanism designer?
 - 2. Private information.** What does each agent know that others cannot observe? (Values, costs, quality, preferences, capabilities.)
 - 3. Outside option (IR).** What is each agent's alternative to participating? How strong is it? When does the participation constraint bind?
 - 4. Gaming (IC).** How could agents manipulate the mechanism? What would they misreport, and why?
 - 5. Your mechanism.** What allocation rule and payment rule would make truthful participation each agent's best strategy? Which tool from the toolkit is most relevant?
 - 6. Failure modes.** What could still go wrong? Consider: information cascades, congestion, winner-take-all dynamics, lock-in, manipulation, and the fundamental efficiency limit (private information precludes full efficiency).
-

1. Information Disclosure in LLM Markets

A small number of foundation model providers (OpenAI, Anthropic, Google, Meta) make capability, safety, and cost claims to buyers who cannot verify them before committing. How do we elicit truthful disclosure?

Private info: Providers know their true capabilities, training data composition, safety properties, and costs. Buyers know their use-case requirements but cannot independently verify provider claims.

Design challenges:

- *Gameable benchmarks.* Benchmarks are the primary disclosure mechanism, but models can be optimized to score well on known benchmarks without genuine capability improvement (Goodhart's Law). What makes a benchmark credible?
- *Proper scoring rules.* A scoring rule pays agents based on the accuracy of their forecasts against realized outcomes. A *proper* scoring rule (Brier, Savage) makes truthful reporting the payoff-maximizing strategy. Could a benchmark consortium require labs to commit predictions on held-out tasks, with payoffs scored against results?
- *Structural concentration.* High fixed costs (training) and high marginal costs (inference) limit the field to a few providers. This makes the disclosure problem harder: with few players, each has more incentive to strategically misrepresent.

Primary tools: Proper scoring rules, signaling.

2. The Grounding Layer

AI companies (OpenAI, Google, Anthropic) need high-quality content for retrieval and context. **Content creators and publishers** (newspapers, academics, artists, databases) can license that content, or withhold it and keep their direct audience.

Private info: Each side holds information the other lacks. *Creators* know the quality, provenance, and production cost of their content (original reporting vs. aggregation, expert vs. amateur, fact-checked or not). *AI companies* know query patterns, volume, and how much value each content source actually adds to their system.

Design challenges:

- *Fair compensation.* Millions of fragments retrieved; each creator's marginal contribution is small but content collectively is essential. A Shapley value problem.
- *Strong outside options.* Creators can withhold content entirely, maintain their direct audience, or license exclusively elsewhere. The IR constraint binds hard. Meanwhile AI companies want token-efficient, cost-minimizing retrieval. The mechanism must make participation worthwhile for creators despite this asymmetry in bargaining position.
- *Quality vs. volume.* How to prevent low-quality content flooding the grounding layer?

Primary tools: Shapley value (attribution), revenue sharing and minimum guarantees (participation), collective licensing.

3. The Market for Skills

People create skills (structured instructions, workflows, prompts) that tell AI agents how to perform tasks. This is an explore-exploit problem: users default to popular skills, so new and potentially better ones are never tried.

Private info: Skill creators know how well their instructions actually work. Users know their task requirements and willingness to pay, but cannot verify skill quality before committing.

Design challenges:

- *Explore vs. exploit and monetization.* Information congestion means users default to popular skills, concentrating revenue on a few well-known creators. The platform can subsidize exploration of new skills against the gravitational pull of popularity using incentive-compatible bandit mechanisms (cf. Babaioff, Sharma & Slivkins 2009; Mansour, Slivkins & Syrgkanis 2015). The discovery mechanism directly shapes who gets paid and how much.
- *Signaling with bite.* A skill's quality is hard to verify before use. Signals only sort types if the cost is differentially higher for low-quality producers (Spence 1973). For skills, this means: staked reputation (deposits forfeited on poor reviews), verified usage logs (the skill demonstrably ran successfully on tasks of declared type), or third-party certification with skin in the game.
- *Attribution boundary.* When AI itself writes and modifies skills, the line between human-authored and machine-generated blurs. Who owns a skill that was human-designed but AI-refined? This complicates both monetization and quality assurance.

Primary tools: Bandit mechanisms (explore/exploit), signaling with costly verification.

4. AI Agent Marketplaces

What happens when AI agents, not humans, are the primary buyers on platforms like Amazon, travel sites, or financial markets? Agents act autonomously on behalf of human principals, and their algorithms are opaque to the platform.

Private info: Agent algorithms and training data are unobservable. The platform does not know the type distribution of its participants (because AI agents are opaque), and principals' true preferences may differ from what they can articulate to their agent.

Design challenges:

- *Amplified herding and information design.* Similar models cascade faster and more uniformly than diverse humans. The platform can counter this by designing the information environment to inject diversity: controlling what signals agents observe, when, and in what order (cf. Kamenica & Gentzkow 2011; Bergemann & Morris 2019 on information design).
- *Robust mechanism design.* When the platform cannot observe agent type distributions (because agents are opaque algorithms), mechanisms must be designed to work across distributions, not just for a known one (Bergemann & Morris 2005). This is also the setting for automated/computational mechanism design (Conitzer & Sandholm; Procaccia), where the mechanism itself is learned.
- *Principal-agent misalignment.* The AI optimizes a proxy, not the human's true preference. Subtle misalignment compounds across millions of transactions.
- *Reputation.* Traditional reputation relies on social consequences. For algorithmic buyers, reputation may require verifiable action histories or staked deposits forfeited on detected misuse.

Primary tools: Information design, robust mechanism design, computational mechanism design.

Designing Your Mechanism

A mechanism has three moving parts and one property to check.

1. What do you ask each participant to report? Their value, their quality, a bid, a prediction, a preference ranking?

2. Based on those reports, who gets what? This is the *allocation rule*. Highest bidder wins? Best match gets selected? Everyone who qualifies gets included?

3. Based on those reports, who pays or gets paid, and how much? This is the *payment rule*. Pay your own bid? Pay the second-highest? Revenue share proportional to usage? Shapley attribution?

The check: Why would each participant report truthfully *and choose to participate*? This is the IC and IR check. Can anyone do better by lying, exaggerating, or holding back information? Is anyone better off walking away entirely?

Building blocks you can combine: *second-price auction* (allocate scarce slots; truthful bidding), *scoring rule* (elicit truthful capability claims against verifiable outcomes), *revenue sharing* (induce participation against strong outside options), *bandit mechanism* (subsidize exploration of new options), *Shapley attribution* (divide jointly created value fairly), *staked deposit* (signal quality; forfeit on poor performance).