Position: Agentic AI Design Should be Mediated to Promote Social Welfare

Omer Madmon and Moshe Tennenholtz

Technion - Israel Institute of Technology

December 9, 2025

Abstract

The rise of generative AI, large language models (LLMs), and autonomous agents has created ecosystems where strategically designed agents interact on behalf of stakeholders. While each agent aims to maximize its stakeholders' utilities, these interactions can lead to market failures and socially undesirable outcomes. This position paper argues that such systems should be mediated to promote social welfare. We model agent design as a strategic game, demonstrate how the game equilibria can be collectively harmful, and discuss how various forms of mediation can realign incentives toward socially beneficial outcomes. We further outline concrete research directions that pave the way for developing mediators capable of steering agentic design choices toward socially desirable equilibria. We conclude that integrating mediation principles into AI ecosystem design is essential to ensure that autonomous agents advance, rather than undermine, societal welfare.

1 Introduction

With the rapid rise of generative AI and autonomous agents, many real-world applications now involve interactions between agents endowed with increasingly sophisticated capabilities, including decision-making and strategic behavior (Xi et al., 2023; Fu et al., 2023; Wang et al., 2024; Guo et al., 2024; Xie et al., 2025). These agents are no longer limited to single-turn responses or predefined actions: they can engage in extended reasoning (Chen et al., 2025; Xu et al., 2025), adapt to dynamic settings (Sheilsspeigh et al., 2024; Akata et al., 2025), and optimize for goals on behalf of their stakeholders (Bandi et al., 2025; Yang et al., 2025).

Stakeholders deploy such agents to act on their behalf across a wide variety of economic environments. Decision-making processes that were once centralized in human hands are now being gradually delegated to autonomous systems, with direct implications for economic outcomes (Fügener et al., 2022; Hemmer et al., 2023). Agents are already being employed in domains such as financial trading (Sarin et al., 2024), automated negotiations (Hua et al., 2024; Abdelnabi et al., 2024), search engine optimization (Mordo et al., 2025a) online marketplaces (Aslesha et al., 2025), and resource allocation (Hady et al., 2025). These domains illustrate the promise and the risks of agent-mediated economies.

With advanced capabilities such as reasoning (Bilal et al., 2025), tool use (Masterman et al., 2024), and ever-expanding context windows (Liu et al., 2025), the volume and scope of agent-driven economies are only expected to increase (Joshi, 2025; Rothschild et al., 2025), making it urgent to study the extent to which agents can reach socially beneficial outcomes. While each agent is typically designed to maximize the utility of its stakeholder, it is well known that such self-interested interactions may lead to market failures and suboptimal outcomes (Mas-Colell et al., 1995).

Fortunately, these negative effects are not inevitable, as the literature on game theory and mechanism design offers a rich toolkit for addressing market failures by introducing mediators—entities that shape the rules of the game to guide self-interested behavior toward socially desirable outcomes such

omermadmon@campus.technion.ac.il
moshet@technion.ac.il

as fairness, efficiency, and welfare maximization (Hurwicz, 1960; Clarke; Myerson, 2008; Maskin, 2008; Sinha and Anastasopoulos, 2015). Common forms of intervention include monetary transfers (Monderer and Tennenholtz, 2003), incentive-compatible coordination signals (Aumann, 1974, 1987), and information revelation mechanisms (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019). Depending on the assumptions, these interventions can guide agent behavior toward improved collective outcomes.

This paper focuses on an earlier stage of the pipeline, which we refer to as the *agent design stage*. In real-world scenarios such as digital platforms, stakeholders usually design and deploy agents that will subsequently interact on their behalf (Tallam, 2025; Acharya et al., 2025). Agent design is a complex and inherently strategic task, with many degrees of freedom: prompt design, tool integration, model backbone, fine-tuning methods, and more. Crucially, the effectiveness of a given agent often depends on the design choices of competing agents. This strategic interdependence has long been recognized in pre-AI contexts: for example, in repeated Prisoner's Dilemma competitions, where an agent performs well only if its opponent is not explicitly designed to exploit its cooperation (Axelrod, 1980, 1981, 1987).

In this position paper, we argue that **the design of AI agents should be *mediated* to promote social welfare** and other socially beneficial outcomes, such as fairness, efficiency, and stability. Foundational ideas from economic theory can, and should, be adopted by ML researchers, practitioners, and regulators to design better collaborative AI environments. To this end, we adopt a game-theoretic framework to capture the strategic nature of agent design and to demonstrate how different forms of mediation can improve outcomes. We further discuss how these principles can extend beyond abstraction, offering concrete mechanisms for shaping real-world agent ecosystems toward societal good. We conclude by charting several research directions that are central to the development of socially beneficial mechanisms and mediators, surveying current progress and emphasizing gaps that invite further exploration.

2 Agent design as a game

We formalize the *agent design game* as a standard normal-form game. There are n players (stakeholders), each of whom must decide which agent to deploy. For every player $i \in \{1, \dots, n\}$, let A_i denote the set of agents available to them. This set may include, for example, choices over LLM backbones, prompting strategies, context window sizes, training hyperparameters, or tool integrations. A *strategy profile* is a tuple $a = (a_1, \dots, a_n) \in A \coloneqq \times_i A_i$.

Once a profile a is chosen, the deployed agents interact in the underlying environment. The outcome of this interaction for player i is captured by a payoff function $u_i : A \to \mathbb{R}$. Intuitively, $u_i(a)$ represents the expected utility that player i derives when the selected agents are deployed, where the expectation accounts for both the stochasticity of the agents' behavior and the uncertainty in the environment (e.g., market conditions, information structures). This utility can also incorporate costs incurred by the chosen agents, such as training expenses or inference costs, in addition to the utilities obtained from their interactions. Players are assumed to be rational and risk-neutral, meaning that their objective is to maximize expected utility given their competitors' choice of agents. In this way, the underlying interaction among agents is abstracted into the payoff functions $u_1, \dots, u_n$, and the agent design problem reduces to a normal-form game. Viewed through this lens, the agent design game serves as a *meta-game*: players strategically choose agents at the design stage, and the realized utilities reflect the downstream outcomes of their agents' interactions.

Under the standard assumption of rational self-interested behavior, the relevant solution concept for the agent design game is the *Nash equilibrium* (Nash Jr, 1950): a strategy profile $a^* \in A$ such that no player can gain by unilaterally deviating, i.e., $u_i(a^*) \geq u_i(a'_i, a^*_{-i})$ for all i and all $a'_i \in A_i$. In other words, each agent's design choice is a best response to the others, and the resulting outcome is stable against unilateral deviations. We adopt this as the benchmark prediction for how rational stakeholders are expected to design their agents. While there are various ways to measure the extent to which society benefits from the outcome of the game, we focus on *utilitarian social welfare*, defined simply as the sum of players' utilities, $\sum_i u_i(a)$. However, as we demonstrate in the sequel, Nash equilibria in agent design games may be misaligned with societal objectives, potentially yielding outcomes that are Pareto-dominated or otherwise detrimental to social welfare. This motivates the introduction of mediators—mechanisms that can alter information, incentives, or coordination—to steer equilibrium behavior toward more desirable outcomes.

Example 1 demonstrates how the framework can be applied to a concrete, simple setting that captures real considerations in agent design, and illustrates how, without interventions, strategic behavior may lead to suboptimal societal outcomes.

Example 1 *Two stakeholders engage in automated contract bargaining. Player 1 chooses between a high-quality costly model E_1 and a cheap model C_1; Player 2 chooses a model $\{E_2, C_2\}$ and whether to enable a* market-price tool *T, resulting in four strategies $(E_2, \varnothing), (C_2, \varnothing), (E_2, T), (C_2, T)$.*

Agents jointly produce a surplus $S(a_1, a_2)$, divided according to bargaining shares $\alpha_i(a_1, a_2)$, while incurring design-specific costs $c_i(a_i)$, inducing a utility of:

$$u_i(a_1, a_2) = \alpha_i(a_1, a_2)S(a_1, a_2) - c_i(a_i),$$

Appendix A specifies a particular form of the surplus, bargaining share and cost functions, encoding that stronger models raise surplus, the tool both improves efficiency and shifts bargaining power toward Player 2, and sophisticated designs are costlier. The resulting payoffs are:

	$(E_2, \varnothing)$	$(C_2, \varnothing)$	(E_2, T)	(C_2, T)
E_1	(3.91, 4.23)	(3.41, 4.81)	(2.82, 4.64)	(2.51, 5.04)
C_1	(4.84, 2.24)	(4.62, 3.85)	(4.05, 2.36)	(3.90, 3.90)

The agent design game introduced in Example 1 has a unique Nash equilibrium at $(C_1, (C_2, T))$, yielding payoffs of (3.90, 3.90). This outcome arises because C_1 strictly dominates E_1 for Player 1, and (C_2, T) strictly dominates the alternatives for Player 2. Yet from a societal perspective, the equilibrium is inefficient: its total welfare of 7.80 falls short of what is attainable at other profiles. In particular, the profile $(E_1, (E_2, \varnothing))$ achieves a higher welfare of 8.14 and also *Pareto-dominates* the equilibrium, since both players earn strictly higher payoffs, (3.91, 4.23). Although both parties would be better off deploying stronger models and forgoing the tool, individual incentives drive them toward a collectively worse outcome, illustrating a market failure.

3 Fifty forms of mediators

While there are many possible forms of mediation that could be applied to AI interactions, in this section, we focus on several fundamental approaches that capture the essence of how outcomes in the agent design framework can be improved. These mediators differ in their capabilities and the extent to which they can intervene in the game, ranging from adjusting incentives through payments to shaping information flows and enforcing institutional rules that constrain the space of admissible agent designs. Our aim is twofold: first, to connect each mediator to its theoretical foundation in game theory and mechanism design, highlighting how it has been shown to resolve inefficiencies in strategic settings; and second, to illustrate how analogous principles could be instantiated in AI ecosystems, where agents are designed and deployed by stakeholders with conflicting interests. In doing so, we pave the path from abstract theory to concrete, implementable interventions that can guide agent interactions toward more desirable outcomes.

3.1 Strategy space restriction

We begin by exploring a natural and simple approach to mediation: restricting the strategy space available to agent designers. In this approach, the mediator limits the set of models, tools, or design features that stakeholders may deploy, thereby removing harmful options from the game altogether. This form of intervention directly shapes the strategic environment and can eliminate equilibria that are individually rational but socially undesirable. The idea is illustrated in the following example:

Example 2 *Returning to Example 1, suppose the platform forbids the use of the tool. The reduced game then includes only the strategies without the tool:*

	$(E_2, \varnothing)$	$(C_2, \varnothing)$
E_1	(3.91, 4.23)	(3.41, 4.81)
C_1	(4.84, 2.24)	(4.62, 3.85)

In this restricted setting, it is straightforward to see that the unique equilibrium is the profile in which both players choose the cheaper model (e.g., by dominant strategies elimination), yielding payoffs $(4.62,\ 3.85)$ *and total welfare* 8.47*. This represents an improvement over the equilibrium of the original game, where welfare was only* 7.80*.*

In practice, strategy space restrictions correspond to mediators that limit the design choices available to stakeholders. Such interventions are feasible in environments where the platform has direct control over the APIs, tools, or resources accessible to deployed agents. For instance, a trading platform may forbid certain order types that enable manipulative strategies, or a social media platform may restrict the use of engagement-boosting tools that generate harmful dynamics. In other contexts, however, such control is neither practical nor desirable: forbidding legitimate tools may stifle innovation, reduce efficiency, or incentivize circumvention by stakeholders. Thus, while strategy restriction can be an effective mediator in principle, its applicability depends critically on the institutional setting and the trade-off between control and flexibility.

3.2 Monetary payments

An alternative form of mediation is to offer monetary payments (or to impose additional costs) for playing specific strategies in the game. In agent design games, this corresponds to a mediator influencing the agents' incentives by adjusting their payoff structure without explicitly restricting their strategic options. Rather than forbidding undesirable actions, the mediator makes them less attractive through cost imposition or more appealing through subsidies.

This approach provides higher flexibility and finer control for the mediator compared to strategy space restriction. Indeed, from a mathematical perspective, the latter can be seen as an extreme case of monetary payments, where certain strategies are penalized with an additional cost of arbitrarily large magnitude, effectively removing them from consideration. Thus, monetary payments generalize the idea of restricting strategies by enabling smooth and modifications to incentives rather than binary ones. To illustrate this mediation approach, consider the following example.

Example 3 *Two stakeholders must choose a communication format for their negotiation agents: free-form Language (L) or Structured (S). If they adopt different formats, the agents are unable to interact effectively, leading to costly miscommunication for both parties. The stakeholders, however, differ in their underlying technological and budgetary constraints. Player 1, who has access to a powerful proprietary LLM, benefits more from flexible natural language communication and therefore favors L. Player 2, by contrast, relies on a more limited open-source model and faces higher expenses for processing language tokens, making a more rigid structured protocol preferable. This asymmetry in preferences, combined with the high cost of misalignment, motivates the following payoff structure:*

	L	S
L	$(5,\ 3)$	$(-2,\ -2)$
S	$(-2,\ -2)$	$(3,\ 5)$

Note that the resulting interaction corresponds to the well-known ***Battle of the Sexes*** *(Rapoport, 1966), which admits three Nash equilibria: two pure equilibria* (L, L) *and* (S, S)*, each yielding an optimal social welfare of* 8*, and one mixed equilibrium. In the mixed equilibrium, both players randomize over their preferred actions—Player 1 choosing language-based communication with probability* $7/12$ *and Player 2 with probability* $5/12$*—so that misalignment occurs with positive probability. This stochastic coordination failure reduces the expected social welfare to* $\frac{11}{6}$*, which is strictly below the welfare of the pure equilibria. Thus, only the pure equilibria are socially efficient.*

As the game admits three Nash equilibria, the eventual outcome is indeterminate. A welfare-maximizing mediator thus aims to steer the agents toward one of the efficient equilibria. This can be achieved by leveraging the result of Monderer and Tennenholtz (2003), which shows that any equilibrium can be implemented in dominant strategies by appropriately designing monetary transfers. Crucially, these transfers need not actually be executed in equilibrium: their mere presence in the strategic environment is sufficient to steer behavior towards the desired outcome.

In Example 3, the mediator could commit to a transfer scheme that rewards Player 1 with an additional payment equivalent to a utility bonus of +6 whenever Player 1 plays L while Player 2 plays S. This modification changes the payoff $u_1(L, S)$ from -2 to 4, thereby making L a strictly dominant strategy for Player 1. Anticipating this, Player 2's best response is to also play L, so that the unique equilibrium outcome becomes (L, L). Importantly, since the threatened transfer is never actually triggered in equilibrium, the mediator achieves coordination on the efficient outcome without incurring any real monetary cost. This demonstrates the power of monetary payments to guide design choices towards socially desirable equilibria without restricting strategies outright or bearing any real cost.

The applicability of monetary-payment mediation in real-world agent design depends strongly on the system context. In commercial platforms where interactions already involve priced resources—such as AI services that charge for API calls or tokens, or marketplace platforms where bandwidth and compute budgets are explicitly metered—the mediator can plausibly implement transfers by adjusting usage costs or providing targeted subsidies. For instance, an online travel platform or financial trading venue could adjust token pricing or API fees to encourage coordination on communication protocols that improve efficiency. By contrast, in open-source agent frameworks, decentralized multi-agent environments, or collaborative research settings, there is often no central authority capable of imposing or enforcing payments. Thus, while monetary payments offer strong guarantees, their practical deployment is limited to domains where pricing mechanisms and enforceable resource accounting are already embedded in the system.

3.3 Correlation devices

Another form of mediation arises through the use of *correlation devices*, which lead naturally to the solution concept of *correlated equilibrium*, proposed by Aumann (1974). In a correlated equilibrium, a mediator draws a joint signal from a publicly known distribution and privately recommends an action to each player. Given the recommendation, no player has an incentive to deviate unilaterally, provided that others follow their own recommendations. This concept extends Nash equilibrium by allowing coordination on correlated strategies, thereby enabling outcomes that are otherwise unreachable. Importantly, computing a correlated equilibrium can be formulated as a linear programming problem and thus solved in polynomial time (Papadimitriou and Roughgarden, 2008). Moreover, in certain classes of games it is known how much correlation can improve social welfare: the *value of correlation* (defined as the ratio between the optimal correlated equilibrium welfare and the best Nash welfare) has been studied by Ashlagi et al. (2008), who derived sharp bounds in several canonical settings. The following example illustrates this mediation approach:

Example 4 *Consider the agent design game defined in Example 3. Recall that the game admits three Nash equilibria: two pure equilibria, (L, L) and (S, S), which are efficient but asymmetric and thus unfair, and one symmetric mixed equilibrium, which is fair but inefficient. Suppose that a mediator seeks to implement an outcome that achieves both fairness and efficiency. By employing a correlation device, the mediator can recommend (L, L) with probability $1/2$ and (S, S) with probability $1/2$, thereby guaranteeing each player the same expected payoff while preserving efficiency.*

This distribution is a correlated equilibrium because no player can benefit from deviating from the mediator's recommendation. For instance, if Player 1 is recommended to play L, she knows that Player 2 is also recommended L, yielding her a payoff of 5. Deviating to S instead would lead to the outcome (S, L), which gives her only -2. A symmetric argument holds when the recommendation is S: by following it, Player 1 secures a payoff of 3, whereas deviating to L would result in (L, S) and a payoff of -2. The same reasoning applies to Player 2.

Thus, the proposed correlated strategy profile is self-enforcing. It resolves the tension between efficiency and fairness: each player receives the same expected payoff while the overall welfare remains maximal. This example highlights how mediators can exploit correlation to achieve desirable equilibrium outcomes that are unattainable under rational and independent behavior alone.

Beyond stylized examples, correlated equilibria are directly applicable to real-world AI agent design. In many multi-agent systems, a mediator can coordinate the design choices of agents, made by the stakeholders. For example, in multi-agent reinforcement learning, a central training platform may recommend exploration schedules or hyperparameter settings that diversify agent behavior while avoiding inefficient uniformity. In online platforms, mediators can guide the design of moderation or recommendation agents—for instance, balancing rule-based filters against LLM-driven models, or signaling when to prioritize personalization over diversity. In such domains, correlation devices provide a mechanism for aligning design choices in a way that reconciles efficiency with fairness, yielding outcomes that decentralized behavior alone cannot achieve.

3.4 Information design

A further form of mediation arises through the strategic control of information flows, known in the economic literature as *information design*. Instead of restricting actions or modifying payoffs, a mediator can influence the behavior of stakeholders by determining what information about the environment is revealed to them. This perspective—formalized in the literature on Bayesian persuasion and information design (Kamenica and Gentzkow, 2011; Bergemann et al., 2015; Bergemann and Morris, 2016, 2019)—treats the mediator as an information designer, who observes the state of the world and commits to an information structure that guides the stakeholders' beliefs and therefore their equilibrium actions. Such interventions are especially relevant in AI ecosystems, where platforms often possess richer knowledge about user preferences, system dynamics, or global conditions than individual stakeholders, and can therefore shape outcomes by selectively revealing this knowledge.

Formally, an information-design problem is naturally modeled as a *Bayesian game* (Harsanyi, 1967). There are N players, each choosing an action $a_i \in A_i$. A state of the world $\theta \in \Omega$ is drawn from a common prior distribution μ_0. Payoffs depend on both actions and the state, $u_i(a, \theta)$ for each player and $v(a, \theta)$ for the mediator (e.g., social welfare as before, or another objective aligned with the mediator's incentives). Players do not observe θ directly. Instead, the mediator commits to an *information structure*, consisting of a signal distribution $\pi \in \Delta(\Omega \times S)$ with marginal μ_0. Each player receives a private signal, updates her belief about θ, and chooses a strategy $\sigma_i : S_i \to \Delta(A_i)$. A *Bayes–Nash equilibrium* is a strategy profile such that, given the induced beliefs, no player can profitably deviate from her signal-contingent action. The mediator's problem is to choose an information structure that maximizes expected payoff in equilibrium. To illustrate, let us consider the following example, in which stakeholders deploy agents with hyperparameters that must be tuned to uncertain user preferences.

Example 5 *Consider N stakeholders, each designing an AI agent to be deployed on a platform. Each stakeholder must set a hyperparameter $a_i \in \mathbb{R}$, such as the LLM temperature or an exploration–exploitation parameter. There exists an unknown* state of the world *$\theta \in \mathbb{R}$, representing the user's latent preference (e.g., how exploratory or creative outputs should be). The state θ is drawn from a prior distribution $\mathcal{N}(0, 1)$, observed by the platform but not by the stakeholders. Each stakeholder aims to align their choice with θ, with payoff $u_i(a, \theta) = -\frac{1}{2}(a_i - \theta)^2$. Thus, in the absence of mediated information, stakeholders optimally respond to their prior, leading to inefficient misalignment.*

The platform (mediator) seeks both (i) to ensure that the average *hyperparameter choice reflects the user's true preference, and (ii) to preserve* diversity *across agents, which may increase robustness, allow adaptation to future users, or generate heterogeneous training data. Its objective is:*

$$v(a, \theta) \;=\; \frac{1}{N} \sum_{i=1}^{N} a_i \cdot \theta \;-\; \frac{\rho}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} a_i a_j,$$

where $\rho > 0$ governs the relative weight on diversity. In this setting, the mediator's task is to design an information structure (signals about θ) that induces equilibrium play consistent with its objective. The resulting game is the prediction game analyzed by Smolin and Yamashita (2022).

Beyond the observation that such problems can be cast as linear programs (Dughmi and Xu, 2016; Cummings et al., 2020; Galperti et al., 2024), Smolin and Yamashita (2022) characterize the optimal information structures in concave games (i.e., games in which players' utilities are concave in their own actions) using duality arguments. In Example 5, the optimal information structure recommends:

$$a_i(\theta) \;=\; \Big(\tfrac{1}{2\rho} + \tfrac{1}{2N}\Big)\theta \;+\; \varepsilon_i \;-\; \frac{1}{N-1} \sum_{j \neq i} \varepsilon_j,$$

where the $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are Gaussian noise terms that are independent of θ but negatively correlated across players, for some carefully chosen σ_ε^2. Intuitively, this policy achieves two goals simultaneously: the *average* of the recommended hyperparameters, $\frac{1}{N} \sum_i a_i(\theta)$, is informative about the state θ, ensuring aggregate alignment with user preferences. At the same time, the negatively correlated noise ensures that individual recommendations are dispersed, preserving diversity. The variance σ_ε^2 is chosen so as to optimally trade off these two objectives, with stronger anticoordination motives (larger ρ) corresponding to larger dispersion.[1]

Information design is relevant in ecosystems where platforms can access global data. For instance, a large-scale recommendation platform may observe rich aggregate signals about user preferences, which individual developers cannot access. By carefully designing what feedback or guidance is provided, the platform can shape agent design choices, and align deployed agents with user needs while preserving ecosystem diversity, robustness, and adaptability.

4 Research directions

In this section, we outline research directions that emerge from our perspective, spanning both theoretical and applied domains. We discuss relevant existing work and identify the gaps and opportunities that motivate future progress.

4.1 Models of agent design interactions

While this paper provides several stylized examples of how agent-design choices among competing stakeholders can be modeled within a game-theoretic framework, these examples are intentionally simplified, serving mainly to illustrate the various concepts of mediation rather than to capture the full complexity of real-world settings Dean et al. (2025). Game theory has recently been used to model interactions among autonomous agents in diverse settings, including competition among content creators in recommendation systems (Ben-Porat and Tennenholtz, 2018; Hron et al., 2023; Jagadeesan et al., 2023a; Yao et al., 2023a,b, 2024a,b,c,d,e) and search engines (Raifer et al., 2017; Kurland and Tennenholtz, 2022; Nachimovsky et al., 2024; Nachimovsky and Tennenholtz, 2025; Madmon et al., 2025a,b), strategic data-sharing (Gradwohl and Tennenholtz, 2022, 2023a,b; Tsoy and Konstantinov, 2023; Hossain and Chen, 2023; Falconer et al., 2025; Taitler et al., 2025), interactions between strategic users and learning algorithms (Hardt et al., 2016; Dong et al., 2018; Eilat et al., 2022; Rosenfeld and Rosenfeld, 2023; Horowitz et al., 2024; Trachtenberg and Rosenfeld, 2025; Saig and Rosenfeld, 2025), and the societal impact of generative AI (Esmaeili et al., 2024; Taitler and Ben-Porat, 2025a,b,c).

However, the space of modeling the *design* of AI agents as a strategic game, rather than their *behavior* within a fixed interaction, remains relatively underexplored. Most existing models treat design choices (architectures, training data, prompting strategies) as exogenous, analyzing equilibrium behavior under fixed designs. Notable steps in this direction include studies of competition among learning algorithms (Ben-Porat and Tennenholtz, 2017, 2019; Feng et al., 2022; Gafni et al., 2024; Dvorkin, 2025), and particularly, analyses of how such competition affects societal outcomes such as welfare and fairness (Jagadeesan et al., 2023b; Gradwohl et al., 2025; Einav and Rosenfeld, 2025). Yet these works typically focus on relatively simple classification tasks, whereas modern agentic AI systems exhibit a far richer set of design degrees of freedom (objectives, architectures, and communication protocols), inducing a broader family of games whose strategic and welfare properties warrant deeper analysis.

One promising direction is to *jointly* model the meta-game of agent design and the induced game played by the resulting agents, thereby capturing the feedback between design decisions and emergent behaviors. Related ideas have been explored in the use of meta-games to evaluate reinforcement learning dynamics (Li and Wellman, 2024), suggesting that such hierarchical formulations can provide valuable insights into strategic adaptation across levels of abstraction. Unlike our illustrative examples, in which the behavior of the designed agents was abstracted into the meta-game utility, integrated frameworks could enable the study of how design incentives propagate through the induced game, shaping equilibrium outcomes and the role of mediation.[2]

[1] It can also be shown that this recommendation is incentive-compatible, namely, no player can benefit from unilateral deviation from playing the recommended action.

[2] Such models could be analyzed, e.g., through the lens of the existing literature on games with commitments (Tennenholtz, 2004; Kalai et al., 2010), which formalizes how agents can strategically commit to policies or programs that influence others' responses.

4.2 Novel mediation mechanisms and algorithms

While theoretical models clarify how mediation can improve equilibrium outcomes, realizing these ideas in practice requires the design of concrete algorithms and mechanisms that can operate under uncertainty, align heterogeneous incentives, and remain computationally tractable. Examples of such mediation arise across diverse real-world contexts, including ad auctions (Borgs et al., 2007; Aggarwal et al., 2022; Marotta et al., 2022), reputation systems (Che and Hörner, 2018; Romanyuk and Smolin, 2019; Acemoglu et al., 2022; Lorecchio and Monte, 2023; Arieli et al., 2024), allocation mechanisms (Budish et al., 2013; Blumrosen and Dobzinski, 2021; Danino et al., 2025), and strategic communication (Abraham et al., 2019; Arieli et al., 2023). These examples highlight the range of ways mediators align the incentives of individual decision-makers with societal outcomes. In the context of agentic-AI design, mediators must similarly capture the characteristics of the design environment to guide the evolution of deployed agents toward desirable outcomes.

AI design mediators must operate effectively under uncertainty about key components of the ecosystem—such as agents' preferences, beliefs, and private information. Insights from the literature on robust mechanism design (Feige and Tennenholtz, 2011; Lopomo et al., 2021; Bergemann and Morris, 2005), which studies how to achieve stable outcomes under incomplete information, can guide the design of such mediators. Related work has explored robustness to uncertainty in diverse economic settings, including pricing (Bergemann and Schlag, 2008, 2011), trading (Cesa-Bianchi et al., 2021, 2024; Bernasconi et al., 2024), auctions (Noussair and Silver, 2006; Golrezaei et al., 2019; Ausubel and Baranov, 2020), and persuasion (Arieli et al., 2022; Dworczak and Pavan, 2022; Babichenko et al., 2022; Bacchiocchi et al., 2024; Arieli et al., 2025). Extending these approaches to agent design mediators could enable reliable system performance even when information is incomplete or misspecified.

Another line of research explores how different notions of equilibrium could be implemented using incentive design techniques such as side payments (Wu et al., 2023; Geffner and Tennenholtz, 2024; Wu et al., 2024; Geffner et al., 2025; McMahan et al., 2025), demonstrating how reward or payoff shaping can improve welfare and stability. Adopting these ideas into the agentic-AI design framework holds promise in developing mediators that steer the ecosystem towards socially-desirable design choices.

Beyond welfare maximization, a growing body of work examines mediators that promote other social objectives such as fairness. These include mechanism-design approaches for pricing, trading, and allocation (Sinha and Anastasopoulos, 2015; Cohen et al., 2022; Banerjee et al., 2024, 2025) as well as machine-learning formulations for fair classification and clustering (Chierichetti et al., 2017; Kleinberg et al., 2018; Bera et al., 2019; Zafar et al., 2019; Huang and Vishnoi, 2019; Esmaeili et al., 2020). Such perspectives expand the role of mediation from optimizing welfare to achieving socially balanced outcomes, and could also be adopted to mediators in the context of agentic-AI design. From an algorithmic perspective, recent advancements at the intersection of economic theory and computational complexity (Papadimitriou and Roughgarden, 2008; Dughmi and Xu, 2016; Cummings et al., 2020; Babichenko et al., 2023; Dütting et al., 2024; Guo et al., 2025) provide a foundation for studying the computational tractability of mediation—an essential step towards developing practical mediators.

A promising direction for future work lies in developing mediators tailored to agent design interactions, where agents differ in their computational, informational, or communicative capabilities. Building on the concept of social laws (Shoham and Tennenholtz, 1992, 1995, 1997), such mediators could establish adaptive coordination rules that account for these heterogeneities, ensuring efficient and stable interaction across diverse agent types. For example, in autonomous driving, hybrid traffic laws that differentiate between human and AI drivers (Kraicer et al., 2025) illustrate how mediators can shape system-wide behavior to improve safety and reduce congestion. Extending this perspective, mediators in digital ecosystems could dynamically adjust rules and incentives to the capacities of participating agents, paving the way toward fairer and more efficient AI-driven environments.

4.3 Mediators in real-world environments

While theoretical models provide conceptual foundations, the next frontier lies in implementing and evaluating mediators in real-world AI ecosystems. Doing so requires bridging the gap between abstract formulations and the complexity of socio-technical environments in which AI agents operate. Recent progress in ML-based mechanism design offers a starting point for such implementations. Deep and differentiable approaches to mechanism and policy optimization demonstrate how learning algorithms can approximate theoretically optimal mechanisms in data-driven settings (Dütting et al., 2019; Nahum et al., 2024; Ravindranath et al., 2024).

Another direction is to leverage game-theoretic evaluation frameworks to empirically assess AI systems and their interactions (Duan et al., 2024; Fan et al., 2024; Park et al., 2024; Shapira et al., 2025c). Such frameworks provide systematic ways to measure strategic behaviors and welfare outcomes among learning agents. These research directions could be extended to design and evaluate mediators in the context of agentic-AI design. Examples of such attempts are found in information design, where mediators have been used for algorithmic recourse (Harris et al., 2022), dynamic pricing (Agrawal et al., 2023), and AI alignment (Bai et al., 2024).

In parallel, recent studies explore *AI-driven mediators* that directly interact with decision makers (Tan et al., 2024; Koçak et al., 2024; Duetting et al., 2025). While these works focus on mediation at the interaction level, similar approaches could be applied at the *design level*, where mediators guide the pre-deployment choice of model architectures, training data, or alignment objectives. This vision connects with emerging applications such as auction design for AI-generated content (Duetting et al., 2024), digital advertising with LLMs (Bergemann et al., 2025), and sponsored question answering (Mordo et al., 2024), as well as orchestrating AI agents with contracts (Ivanov et al., 2024).

Evaluating mediation mechanisms in dynamic, large-scale environments also demands suitable simulation infrastructures. Recent work in evaluating search and recommendation systems (Mordo et al., 2025b; Ye et al., 2025) illustrates how controlled, reproducible simulation environments can capture evolving user, agent, and platform behaviors. Such environments could be repurposed for testing both *agent design strategies* and *mediation protocols*, especially in settings where recommendation or ranking functions already act as implicit mediators between competing agents.

Finally, while recent work has demonstrated notable success in predicting human choices in economic decision-making (Apel et al., 2022; Plonsky et al., 2025; Shapira et al., 2025a,b), similar approaches could be applied to predict the behavior of AI agents and stakeholders in design settings. By modeling how agents or their designers respond to incentives, feedback, or information, mediators could leverage such predictive tools to anticipate the downstream effects of different design choices and guide the system toward more efficient, fair, and socially aligned equilibria.

5 Alternative views

While our perspective emphasizes the importance of mediating AI interactions through various forms of intervention, there are natural and reasonable arguments against this approach. In what follows, we outline these alternative views and address their implications in relation to our framework.

5.1 Self-correcting markets and intervention risk

A central counterargument draws inspiration from classical economic reasoning: one could claim that, much like competitive markets, AI ecosystems tend to self-correct through feedback mechanisms, adaptation, and evolution of incentives (Smith, 1776; Persky, 1989; Bowles et al., 1993). From this longstanding viewpoint, external interventions or mediations could distort natural equilibria, stifle innovation, or introduce inefficiencies. For example, overly restrictive content-moderation protocols could reduce diversity, leading to homogenized outcomes or discouraging experimentation (Schwemer et al., 2023). Similarly, in algorithmic markets, ill-designed fairness or exposure adjustments may reduce aggregate welfare by disrupting natural competition dynamics or by inducing unintended strategic responses from agents (Bertsimas et al., 2011; Yang et al., 2024).

This perspective highlights a valid concern: not every interaction requires mediation, and in some environments, intervention may indeed cause more harm than benefit. Our position, therefore, is not that *mediation is always needed*, but rather that *we should develop principled methods to determine when and how mediation is needed.* Research in this direction should aim to identify the structural features of environments—such as asymmetries in information, power, or computational capabilities—that justify mediation, and to design diagnostic and empirical tools capable of detecting these conditions in practice. Such an agenda would parallel the role of welfare and efficiency analyses in economics: understanding when markets fail and when corrective mechanisms are socially desirable. This concern also calls for the development of *transparent and explainable mediation mechanisms*, aligning with the broader movement toward explainable and accountable AI (Dwivedi et al., 2023; Li et al., 2023). By making the rationale behind interventions explicit and interpretable, mediators can enhance trust and legitimacy among affected stakeholders, mitigating concerns about arbitrariness and fostering confidence in the governance of AI agent interactions.

5.2 Fairness and distributional concerns in mediation

Another concern arises from fairness considerations. Mediation, by design, alters the strategic landscape, potentially changing the distribution of utilities among stakeholders. In some cases, a stakeholder that benefits under unmediated conditions may lose relative power or utility once mediation is introduced—raising questions about whether societal welfare improvements justify individual sacrifices. For example, if a dominant agent designer or platform faces reduced advantage under a fairness-enforcing mediator, one might view this as unfair redistribution rather than progress. Moreover, mediation may have external effects on entities outside the modeled interaction. For instance, a mediation mechanism that penalizes certain design choices might indirectly disadvantage specific tool providers, data curators, or model developers whose technologies align with those disfavored strategies. In such cases, even if the mediation improves aggregate outcomes within the focal game, it may still be perceived as unfair or harmful at the ecosystem level.

We believe the resolution lies in *careful modeling*. Fair mediation frameworks should explicitly incorporate these concerns into their design objectives and constraints. Depending on the context, one may impose Pareto-improvement requirements, add regularization terms penalizing excessive losses to individual agents, or relax equilibrium concepts to accommodate bounded rationality and fairness trade-offs. Likewise, when tool providers or other external entities are significantly affected, they can be formally modeled as players or stakeholders within the same game-theoretic environment. Such extensions would ensure that mediation mechanisms remain context-aware and socially legitimate, aligning with both efficiency and equity principles.

6 Concluding remarks

In this paper, we argued that the process of agent design can substantially benefit from careful and responsible mediation. By steering outcomes toward socially desirable goals, mediation can help account for the broader economic and societal effects of deployed agents. Through a series of stylized examples, we illustrated key principles of how mediation can take different forms and how these can improve welfare and fairness in strategic environments. We then outlined a set of research directions aimed at translating these simplified principles into real-world systems capable of mediating agent design in practice. We discussed critical perspectives on our approach, emphasizing that opposing views play an essential role in shaping a more robust understanding of when and how mediation should occur.

Constructive debate around these concerns can lead to better, more transparent, and trustworthy mediators that balance intervention with autonomy in strategic interactions. We conclude with a *call to action* for both practitioners and researchers, within and beyond the ML community, to join forces in addressing the challenges of responsible, socially-beneficial AI mediation. By embracing mediation as a central design principle, we can ensure that the next generation of intelligent systems is aligned with the collective good.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

Ittai Abraham, Danny Dolev, Ivan Geffner, and Joseph Y Halpern. Implementing mediators with asynchronous cheap talk. In *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing*, pages 501–510, 2019.

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Learning from reviews: The selection effect and the speed of learning. *Econometrica*, 90(6):2857–2899, 2022.

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEe Access*, 2025.

Gagan Aggarwal, Kshipra Bhawalkar, Aranyak Mehta, Divyarthi Mohan, and Alexandros Psomas. Simple mechanisms for welfare maximization in rich advertising auctions. *Advances in Neural Information Processing Systems*, 35:28280–28292, 2022.

Shipra Agrawal, Yiding Feng, and Wei Tang. Dynamic pricing and learning with bayesian persuasion. *Advances in Neural Information Processing Systems*, 36:59273–59285, 2023.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, pages 1–11, 2025.

Reut Apel, Ido Erev, Roi Reichart, and Moshe Tennenholtz. Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, 73:1025–1091, 2022.

Itai Arieli, Yakov Babichenko, and Fedor Sandomirskiy. Bayesian persuasion with mediators. *arXiv preprint arXiv:2203.04285*, 2022.

Itai Arieli, Ivan Geffner, and Moshe Tennenholtz. Mediated cheap talk design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5456–5463, 2023.

Itai Arieli, Omer Madmon, and Moshe Tennenholtz. Reputation-based persuasion platforms. *Games and Economic Behavior*, 147:128–147, 2024.

Itai Arieli, Yakov Babichenko, Omer Madmon, and Moshe Tennenholtz. Robust price discrimination. *Games and Economic Behavior*, 154:377–395, 2025.

Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. On the value of correlation. *Journal of Artificial Intelligence Research*, 33:575–613, 2008.

Chilakamarri L Aslesha, D Kavyasree, G Sai Gayatri, I Ashajyothi, Buddha Poorna, and A Thanuja. Ai agent marketplace. *TechPioneer Journal of Engineering and Sciences*, 2(1):36–47, 2025.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

Lawrence M Ausubel and Oleg Baranov. Core-selecting auctions with incomplete information. *International Journal of Game Theory*, 49(1):251–273, 2020.

Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24(1):3–25, 1980.

Robert Axelrod. The emergence of cooperation among egoists. *American political science review*, 75 (2):306–318, 1981.

Robert Axelrod. The evolution of strategies in the iterated prisoner's dilemma. *The dynamics of norms*, 1(1), 1987.

Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Regret-minimizing bayesian persuasion. *Games and Economic Behavior*, 136:226–248, 2022.

Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Algorithmic cheap talk. *arXiv preprint arXiv:2311.09011*, 2023.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion without a clue. *Advances in Neural Information Processing Systems*, 37:76404–76449, 2024.

Fengshuo Bai, Mingzhi Wang, Zhaowei Zhang, Boyuan Chen, Yinda Xu, Ying Wen, and Yaodong Yang. Efficient model-agnostic alignment via bayesian persuasion. *arXiv preprint arXiv:2405.18718*, 2024.

Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor, and Sri Vidya Vilipala. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9):404, 2025.

Siddhartha Banerjee, Kamesh Munagala, Yiheng Shen, and Kangning Wang. Fair price discrimination. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2679–2703. SIAM, 2024.

Siddhartha Banerjee, Kamesh Munagala, Yiheng Shen, and Kangning Wang. Majorized bayesian persuasion and fair selection. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1837–1856. SIAM, 2025.

Omer Ben-Porat and Moshe Tennenholtz. Best response regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. *Advances in Neural Information Processing Systems*, 31, 2018.

Omer Ben-Porat and Moshe Tennenholtz. Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 173–191, 2019.

Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, pages 1771–1813, 2005.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Dirk Bergemann and Karl Schlag. Robust monopoly pricing. *Journal of Economic Theory*, 146(6): 2527–2543, 2011.

Dirk Bergemann and Karl H Schlag. Pricing without priors. *Journal of the European Economic Association*, 6(2-3):560–569, 2008.

Dirk Bergemann, Benjamin Brooks, and Stephen Morris. The limits of price discrimination. *American Economic Review*, 105(3):921–957, 2015.

Dirk Bergemann, Marek Bojko, Paul Duetting, Renato Paes Leme, Haifeng Xu, and Song Zuo. Data-driven mechanism design: Jointly eliciting preferences and information. In *Proceedings of the 26th ACM Conference on Economics and Computation*, EC '25, page 507, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400719431. doi: 10.1145/3736252.3742578. URL https://doi.org/10.1145/3736252.3742578.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. No-regret learning in bilateral trade via global budget balance. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 247–258, 2024.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations research*, 59(1):17–31, 2011.

Ahsan Bilal, Muhammad Ahmed Mohsin, Muhammad Umer, Muhammad Awais Khan Bangash, and Muhammad Ali Jamshed. Meta-thinking in llms via multi-agent reinforcement learning: A survey. *arXiv preprint arXiv:2504.14520*, 2025.

Liad Blumrosen and Shahar Dobzinski. (almost) efficient mechanisms for bilateral trading. *Games and Economic Behavior*, 130:369–383, 2021.

Christian Borgs, Jennifer Chayes, Nicole Immorlica, Kamal Jain, Omid Etesami, and Mohammad Mahdian. Dynamics of bid optimization in online advertisement auctions. In *Proceedings of the 16th international conference on World Wide Web*, pages 531–540, 2007.

Samuel Bowles, Richard Edwards, and Frank Roosevelt. *Understanding capitalism.* Harper Collins College, 1993.

Eric Budish, Yeon-Koo Che, Fuhito Kojima, and Paul Milgrom. Designing random allocation mechanisms: Theory and applications. *The American Economic Review*, 103(2):585–623, 2013. ISSN 00028282. URL http://www.jstor.org/stable/23469677.

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 289–309, 2021.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Bilateral trade: A regret minimization perspective. *Mathematics of Operations Research*, 49(1): 171–203, 2024.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2018.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

Edward H. Clarke. Incentives in public decision-making. 35(3):379–382. ISSN 1573-7101. doi: 10.1007/BF00124449. URL https://doi.org/10.1007/BF00124449.

Maxime C Cohen, Adam N Elmachtoub, and Xiao Lei. Price discrimination with fairness constraints. *Management Science*, 68(12):8536–8552, 2022.

Rachel Cummings, Nikhil R Devanur, Zhiyi Huang, and Xiangning Wang. Algorithmic price discrimination. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2432–2451. SIAM, 2020.

Gal Danino, Moran Koren, and Omer Madmon. The multi-bmby mechanism: Proportionality-preserving and strategyproof ownership restructuring in private companies. *Journal of Economics & Management Strategy*, 2025.

Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. Accounting for AI and users shaping one another: The role of mathematical models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=UkP4DhrJt1.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. Gtbench: Uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations. *Advances in Neural Information Processing Systems*, 37:28219–28253, 2024.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

Paul Duetting, Safwan Hossain, Tao Lin, Renato Paes Leme, Sai Srivatsa Ravindranath, Haifeng Xu, and Song Zuo. Information design with large language models. *arXiv preprint arXiv:2509.25565*, 2025.

Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, 2016.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pages 1706–1715. PMLR, 2019.

Paul Dütting, Michal Feldman, Inbal Talgam-Cohen, et al. Algorithmic contract theory: A survey. *Foundations and Trends® in Theoretical Computer Science*, 16(3-4):211–412, 2024.

Vladimir Dvorkin. Regression equilibrium in electricity markets. *IEEE Transactions on Energy Markets, Policy and Regulation*, 2025.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33, 2023.

Piotr Dworczak and Alessandro Pavan. Preparing for the worst but hoping for the best: Robust (bayesian) persuasion. *Econometrica*, 90(5):2017–2051, 2022.

Itay Eilat, Ben Finkelshtein, Chaim Baskin, and Nir Rosenfeld. Strategic classification with graph neural networks. *arXiv preprint arXiv:2205.15765*, 2022.

Ohad Einav and Nir Rosenfeld. A market for accuracy: Classification under competition. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RPPBhhRddB.

Seyed Esmaeili, Brian Brubach, Leonidas Tsepenekas, and John Dickerson. Probabilistic fair clustering. *Advances in Neural Information Processing Systems*, 33:12743–12755, 2020.

Seyed A Esmaeili, Kevin Lim, Kshipra Bhawalkar, Zhe Feng, Di Wang, and Haifeng Xu. How to strategize human content creation in the era of genai? *arXiv preprint arXiv:2406.05187*, 2024.

Thomas Falconer, Anubhav Ratha, Jalal Kazempour, Pierre Pinson, and Maryam Kamgarpour. Selling information in games with externalities. *arXiv preprint arXiv:2505.00405*, 2025.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17960–17967, 2024.

Uriel Feige and Moshe Tennenholtz. Mechanism design with uncertain inputs: (to err is human, to forgive divine). In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 549–558, 2011.

Yiding Feng, Ronen Gradwohl, Jason Hartline, Aleck Johnsen, and Denis Nekipelov. Bias-variance games. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 328–329, 2022.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

Yotam Gafni, Ronen Gradwohl, and Moshe Tennenholtz. Prediction-sharing during training and inference. In *International Symposium on Algorithmic Game Theory*, pages 425–442. Springer, 2024.

Simone Galperti, Aleksandr Levkun, and Jacopo Perego. The value of data records. *Review of Economic Studies*, 91(2):1007–1038, 2024.

Ivan Geffner and Moshe Tennenholtz. Making a nash equilibrium resilient to coalitions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 213–238, 2024.

Ivan Geffner, Caspar Oesterheld, and Vincent Conitzer. Maximizing social welfare with side payments. *arXiv preprint arXiv:2508.07147*, 2025.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Ronen Gradwohl and Moshe Tennenholtz. Pareto-improving data-sharing. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 197–198, 2022.

Ronen Gradwohl and Moshe Tennenholtz. Coopetition against an amazon. *Journal of Artificial Intelligence Research*, 76:1077–1116, 2023a.

Ronen Gradwohl and Moshe Tennenholtz. Selling data to a competitor. *arXiv preprint arXiv:2302.00285*, 2023b.

Ronen Gradwohl, Eilam Shapira, and Moshe Tennenholtz. Fairness under competition. *arXiv preprint arXiv:2505.16291*, 2025.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*, 2024.

Yongkang Guo, Jason D. Hartline, Zhihuan Huang, Yuqing Kong, Anant Shah, and Fang-Yi Yu. *Algorithmic Robust Forecast Aggregation*, page 1110–1129. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400719431. URL https://doi.org/10.1145/3736252.3742674.

Mohamad A Hady, Siyi Hu, Mahardhika Pratama, Zehong Cao, and Ryszard Kowalczyk. Multi-agent reinforcement learning for resources allocation optimization: a survey. *Artificial Intelligence Review*, 58(11):354, 2025.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Keegan Harris, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Z Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022.

John C Harsanyi. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967.

Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Timm Vetter, Michael Vossing, and Gerhard Satzger. Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023. URL https://api.semanticscholar.org/CorpusID:257557813.

Guy Horowitz, Yonatan Sommer, Moran Koren, and Nir Rosenfeld. Classification under strategic self-selection. *arXiv preprint arXiv:2402.15274*, 2024.

Safwan Hossain and Yiling Chen. Equilibrium of data markets with externality. *arXiv preprint arXiv:2302.08012*, 2023.

Jiri Hron, Karl Krauth, Michael Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=l6CpxixmUg.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, et al. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024.

Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019.

Leonid Hurwicz. Optimality and informational efficiency in resource allocation processes. *Mathematical methods in the social sciences*, 1960.

Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C Parkes. Principal-agent reinforcement learning: Orchestrating ai agents with contracts. *arXiv preprint arXiv:2407.18074*, 2024.

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Meena Jagadeesan, Michael Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved bayes risk can yield reduced social welfare under competition. *Advances in Neural Information Processing Systems*, 36:66940–66952, 2023b.

Satyadhar Joshi. The transformative role of agentic genai in shaping workforce development and education in the us. *Available at SSRN 5133376*, 2025.

Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, and Dov Samet. A commitment folk theorem. *Games and Economic Behavior*, 69(1):127–137, 2010.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2018.

Özgecan Koçak, Phanish Puranam, and Afşar Yegin. Llms as mediators: Can they diagnose conflicts accurately? *arXiv preprint arXiv:2412.14675*, 2024.

Tal Kraicer, Jack Haddad, Erez Karaps, and Moshe Tennenholtz. Towards hybrid traffic laws for mixed flow of human-driven vehicles and connected autonomous vehicles. *arXiv preprint arXiv:2502.12950*, 2025.

Oren Kurland and Moshe Tennenholtz. Competitive search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2838–2849, 2022.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Zun Li and Michael P. Wellman. A meta-game evaluation framework for deep multiagent reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/17. URL https://doi.org/10.24963/ijcai.2024/17.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.

Giuseppe Lopomo, Luca Rigotti, and Chris Shannon. Uncertainty in mechanism design. *ERN: Other Microeconomics: Decision-Making under Risk & Uncertainty (Topic)*, 2021. URL https://api.semanticscholar.org/CorpusID:15972635.

Caio Lorecchio and Daniel Monte. Bad reputation with simple rating systems. *Games and Economic Behavior*, 142:150–178, 2023.

Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. On the convergence of no-regret dynamics in information retrieval games with proportional ranking functions. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=jJXZvPe5z0.

Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. The search for stability: Learning dynamics of strategic publishers with initial documents. *Journal of Artificial Intelligence Research*, 83, 2025b.

Veronica Marotta, Yue Wu, Kaifu Zhang, and Alessandro Acquisti. The welfare impact of targeted advertising technologies. *Information Systems Research*, 33(1):131–151, 2022.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Eric S Maskin. Mechanism design: How to implement social goals. *American Economic Review*, 98 (3):567–576, 2008.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*, 2024.

Jeremy McMahan, Young Wu, Yudong Chen, Xiaojin Zhu, and Qiaomin Xie. Optimally installing strict equilibria. *arXiv preprint arXiv:2503.03676*, 2025.

Dov Monderer and Moshe Tennenholtz. k-implementation. In *Proceedings of the 4th ACM conference on Electronic Commerce*, pages 19–28, 2003.

Tommy Mordo, Moshe Tennenholtz, and Oren Kurland. Sponsored question answering. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–173, 2024.

Tommy Mordo, Sagie Dekel, Omer Madmon, Moshe Tennenholtz, and Oren Kurland. Rlrf: Competitive search agent design via reinforcement learning from ranker feedback. *arXiv preprint arXiv:2510.04096*, 2025a.

Tommy Mordo, Tomer Kordonsky, Haya Nachimovsky, Moshe Tennenholtz, and Oren Kurland. Lemss: Llm-based platform for multi-agent competitive search simulation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3595–3605, 2025b.

Roger B Myerson. Perspectives on mechanism design in economic theory. *American Economic Review*, 98(3):586–603, 2008.

Haya Nachimovsky and Moshe Tennenholtz. On the power of strategic corpus enrichment in content creation games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14019–14026, 2025.

Haya Nachimovsky, Moshe Tennenholtz, Fiana Raiber, and Oren Kurland. Ranking-incentivized document manipulations for multiple queries. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 61–70, 2024.

Omer Nahum, Gali Noti, David C. Parkes, and Nir Rosenfeld. Decongestion by representation: Learning to improve economic welfare in marketplaces. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=coIaBY8EVF.

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

Charles Noussair and Jonathon Silver. Behavior in all-pay auctions with incomplete information. *Games and Economic Behavior*, 55(1):189–206, 2006.

Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. *arXiv preprint arXiv:2403.16843*, 2024.

Joseph Persky. Retrospectives: Adam smith's invisible hands. *Journal of Economic Perspectives*, 3 (4):195–201, 1989.

Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C Peterson, Daniel Reichman, Thomas L Griffiths, Stuart J Russell, Even C Carter, et al. Predicting human decisions with behavioural theories and machine learning. *Nature Human Behaviour*, pages 1–14, 2025.

Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. Information retrieval meets game theory: The ranking competition between documents' authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–474, 2017.

Anatol Rapoport. *Two-Person Game Theory: The Essential Ideas*. University of Michigan Press, Ann Arbor, MI, 1966.

Sai Srivatsa Ravindranath, Zhe Feng, Di Wang, Manzil Zaheer, Aranyak Mehta, and David C Parkes. Deep reinforcement learning for sequential combinatorial auctions. *arXiv preprint arXiv:2407.08022*, 2024.

Gleb Romanyuk and Alex Smolin. Cream skimming and information design in matching markets. *American Economic Journal: Microeconomics*, 11(2):250–276, 2019.

Elan Rosenfeld and Nir Rosenfeld. One-shot strategic classification under unknown costs. *arXiv preprint arXiv:2311.02761*, 2023.

David M Rothschild, Markus Mobius, Jake M Hofman, Eleanor W Dillon, Daniel G Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel. The agentic economy. *arXiv preprint arXiv:2505.15799*, 2025.

Eden Saig and Nir Rosenfeld. Evolutionary prediction games. *arXiv preprint arXiv:2503.03401*, 2025.

Saket Sarin, Sunil K Singh, Sudhakar Kumar, Shivam Goyal, Brij Bhooshan Gupta, Wadee Alhalabi, and Varsha Arya. Unleashing the power of multi-agent reinforcement learning for algorithmic trading in the digital financial frontier and enterprise information systems. *Computers, Materials & Continua*, 80(2), 2024.

Sebastian Felix Schwemer, Christian Katzenbach, Daria Dergacheva, Thomas Riis, and João Pedro Quintais. Impact of content moderation practices and technologies on access and diversity. *Available at SSRN 4380345*, 2023.

Eilam Shapira, Omer Madmon, Reut Apel, Moshe Tennenholtz, and Roi Reichart. Human choice prediction in language-based persuasion games: Simulation-based off-policy evaluation. *Transactions of the Association for Computational Linguistics*, 13:980–1006, 2025a.

Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. Can LLMs replace economic choice prediction labs? the case of language-based persuasion games. In *Will Synthetic Data Finally Solve the Data Access Problem?*, 2025b. URL https://openreview.net/forum?id=wfiqCxp7kd.

Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. GLEE: A unified framework and benchmark for language-based economic environments. In *Workshop on Scaling Environments for Agents*, 2025c. URL https://openreview.net/forum?id=xqt43SBjke.

Patrick Sheilsspeigh, Mattias Larkspur, Simeon Carver, and Silvester Longmore. Dynamic context shaping: A new approach to adaptive representation learning in large language models. 2024.

Yoav Shoham and Moshe Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, page 276–281. AAAI Press, 1992. ISBN 0262510634.

Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73(1):231–252, 1995. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(94)00007-N. URL https://www.sciencedirect.com/science/article/pii/000437029400007N. Computational Research on Interaction and Agency, Part 2.

Yoav Shoham and Moshe Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1):139–166, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00028-3. URL https://www.sciencedirect.com/science/article/pii/S0004370297000283. Economic Principles of Multi-Agent Systems.

Abhinav Sinha and Achilleas Anastasopoulos. Mechanism design for fair allocation. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 467–473. IEEE, 2015.

Adam Smith. In *An inquiry into the nature and causes of the wealth of nations: Volume One*. London: printed for W. Strahan; and T. Cadell, 1776., 1776.

Alex Smolin and Takuro Yamashita. Information design in concave games. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 870, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538303. URL https://doi.org/10.1145/3490486.3538303.

Boaz Taitler and Omer Ben-Porat. Braess's paradox of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14139–14147, 2025a.

Boaz Taitler and Omer Ben-Porat. Collaborating with genai: Incentives and replacements. *arXiv preprint arXiv:2508.20213*, 2025b.

Boaz Taitler and Omer Ben-Porat. Selective response strategies for genai. *arXiv preprint arXiv:2502.00729*, 2025c.

Boaz Taitler, Omer Madmon, Moshe Tennenholtz, and Omer Ben-Porat. Data sharing with a generative ai competitor. *arXiv preprint arXiv:2505.12386*, 2025.

Krti Tallam. From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence. *arXiv preprint arXiv:2503.13754*, 2025.

Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromír Šavelka, Sébastien Meeùs, Mia Godet, and Karim Benyekhlef. Robots in the middle: Evaluating llms in dispute resolution. *arXiv preprint arXiv:2410.07053*, 2024.

Moshe Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, 2004.

Benyamin Trachtenberg and Nir Rosenfeld. Strategic classification with non-linear classifiers. *arXiv preprint arXiv:2505.23443*, 2025.

Nikita Tsoy and Nikola Konstantinov. Strategic data sharing between competitors. *Advances in Neural Information Processing Systems*, 36:16483–16514, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, March 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL https://doi.org/10.1007/s11704-024-40231-1.

Young Wu, Jeremy McMahan, Yiding Chen, Yudong Chen, Xiaojin Zhu, and Qiaomin Xie. Minimally modifying a markov game to achieve any nash equilibrium and value. *arXiv preprint arXiv:2311.00582*, 2023.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Data poisoning to fake a nash equilibria for markov games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15979–15987, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023.

Tian Xie, Pavan Rauch, and Xueru Zhang. How strategic agents respond: Comparing analytical models with llm-generated responses in strategic classification. *arXiv preprint arXiv:2501.16355*, 2025.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

Yingxuan Yang, Mulei Ma, Yuxuan Huang, Huacan Chai, Chenyu Gong, Haoran Geng, Yuanjian Zhou, Ying Wen, Meng Fang, Muhao Chen, et al. Agentic web: Weaving the next web with ai agents. *arXiv preprint arXiv:2507.21206*, 2025.

Zongsen Yang, Xingyu Fu, Pin Gao, and Ying-Ju Chen. Fairness regulation of prices in competitive markets. *Manufacturing & Service Operations Management*, 26(5):1897–1917, 2024.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? In *International Conference on Machine Learning*, pages 39674–39701. PMLR, 2023a.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39674–39701. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/yao23b.html.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative ai in content creation competition: Symbiosis or conflict? *arXiv preprint arXiv:2402.15467*, 2024a.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative AI in content creation competition: Symbiosis or conflict? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56885–56913. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/yao24b.html.

Fan Yao, Chuanhao Li, Karthik Abinav Sankararaman, Yiming Liao, Yan Zhu, Qifan Wang, Hongning Wang, and Haifeng Xu. Rethinking incentives in recommender systems: Are monotone rewards always beneficial? *Advances in Neural Information Processing Systems*, 36, 2024c.

Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Jingzhou Liu, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3874–3885, 2024d.

Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Jingzhou Liu, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3874–3885, New York, NY, USA, 2024e. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672021. URL https://doi.org/10.1145/3637528.3672021.

Xiaopeng Ye, Chen Xu, Zhongxiang Sun, Jun Xu, Gang Wang, Zhenhua Dong, and Ji-Rong Wen. Llm-empowered creator simulation for long-term evaluation of recommender systems under information asymmetry. In *Proceedings of the 48th International ACM SIGIR Conference on Research* *and Development in Information Retrieval*, SIGIR '25, page 201–211, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730026. URL https://doi.org/10.1145/3726302.3730026.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75):1–42, 2019.

A Example 1: Detailed utility specification

This appendix details the full construction of the utilities used in Example 1. The aim is to model, in a minimal yet interpretable way, a setting where two stakeholders design negotiation agents that differ in sophistication and cost, and where one side can optionally use a pricing tool that improves efficiency but shifts bargaining power.

Total surplus The joint surplus from deploying the two agents is modeled as

$$\begin{aligned} S(a_1, a_2) = s_0 + s_1\,\mathbf{1}[a_1 = E_1] + s_2\,\mathbf{1}[\text{model}(a_2) = E_2] \\ + s_T\,\mathbf{1}[\text{tool}(a_2) = T] + s_{12}\,\mathbf{1}[a_1 = E_1, \text{model}(a_2) = E_2]. \end{aligned}$$

Here, s_0 is a baseline level of expected value produced even by low-end models. Coefficients $s_1 > 0$ and $s_2 > 0$ capture the marginal benefit from each player upgrading to a high-quality model. The term $s_T > 0$ measures the efficiency gain from activating the pricing tool T, while $s_{12} > 0$ reflects complementarities when both agents are sophisticated. This specification mirrors real bilateral negotiations, where improved algorithms or tools raise joint value.

Bargaining shares To capture the distribution of the generated surplus, we assign Player 1 a bargaining weight

$$\begin{aligned} \alpha_1(a_1, a_2) = \alpha_0 + b_1\,\mathbf{1}[a_1 = E_1] - b_2\,\mathbf{1}[\text{model}(a_2) = E_2] \\ - b_T\,\mathbf{1}[\text{tool}(a_2) = T] + b_{12}\,\mathbf{1}[a_1 = E_1, \text{model}(a_2) = E_2]. \end{aligned}$$

with the remaining $1 - \alpha(a_1, a_2)$ as the bargaining weight of Player 2. The baseline α_0 represents roughly symmetric bargaining power. When Player 1 deploys the stronger model ($b_1 > 0$), their share rises slightly; conversely, when Player 2 uses a stronger model ($b_2 > 0$) or activates the pricing tool ($b_T > 0$), the advantage shifts toward Player 2. The interaction term b_{12} ensures that joint investment moderates these effects. This structure mirrors many real negotiation settings, where technological advantages or exclusive access to analytic tools can influence leverage during automated bargaining.

Deployment costs Each player pays design-specific costs,

$$\begin{aligned} c_1(E_1) &= c_{1E}, & c_1(C_1) &= c_{1C}, \\ c_2(E_2, \varnothing) &= c_{2E}, & c_2(C_2, \varnothing) &= c_{2C}, \\ c_2(E_2, T) &= c_{2E} + c_T, & c_2(C_2, T) &= c_{2C} + c_T. \end{aligned}$$

Costs increase with model sophistication and with enabling the tool, capturing compute and engineering expenses associated with advanced deployments.

Parameter values We used the following parameters for Example 1:

$$\begin{aligned} &s_0 = 10, \;\; s_1 = 2.1469, \;\; s_2 = 0.7760, \\ &s_T = 0.2433, \;\; s_{12} = 1.3017, \;\; \alpha_0 = 0.55, \;\; b_1 = 0.00015, \\ &b_2 = 0.01923, \;\; b_T = 0.08344, \;\; b_{12} = -0.02606, \;\; c_{1E} = 3.2748, \\ &c_{1C} = 0.8786, \;\; c_{2E} = 2.8131, \;\; c_{2C} = 0.6545, \;\; c_T = 0.9144. \end{aligned}$$

These values induce a game in which the efficient configuration $(E_1, (E_2, \varnothing))$ maximizes total surplus, but individual incentives favor the cheaper configuration $(C_1, (C_2, T))$.