Agentic AI Design Should be Mediated to Promote Social Welfare

Omer Madmon
The Faculty of Data and Decisions Science
Technion - Israel Institute of Technology
Haifa, Israel
omermadmon@campus.technion.ac.il

Moshe Tennenholtz
The Faculty of Data and Decisions Science
Technion - Israel Institute of Technology
Haifa, Israel
moshet@technion.ac.il

Abstract

The rise of generative AI and autonomous agents is creating ecosystems in which stakeholders strategically choose which agents to deploy on their behalf. These design choices–such as model backbone, prompting strategy, tool access, and fine-tuning pipeline–are often strategically interdependent, as the performance of a deployed agent depends on the agents chosen by others. This position paper argues that this agent design stage should itself be a target for mediation. This perspective shifts attention from mediating only the *downstream* interaction among deployed agents to mediating the *upstream* stakeholder game that determines which agents enter that interaction in the first place. We formalize this agent design stage as a *meta-game*, show how its equilibria can be collectively harmful, and discuss how various forms of mediation can realign incentives toward socially beneficial outcomes. We demonstrate these mediation approaches using both stylized examples and a real-world case study constructed from empirical LLM-agent interaction data. We outline concrete research directions for developing mediators that steer agentic design choices toward socially desirable outcomes.

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 [83, 31, 80, 38, 84]. These agents are no longer limited to single-turn responses or predefined actions: they can engage in extended reasoning [21, 85], adapt to dynamic settings [72, 3], and optimize for goals on behalf of their stakeholders [11, 86].

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 [32, 42]. Agents are already being employed in domains such as financial trading [69], automated negotiations [44, 1], search engine optimization [60] online marketplaces [6], and resource allocation [39]. These domains illustrate the promise and the risks of agent-mediated economies.

With advanced capabilities such as reasoning [19], tool use [57], and ever-expanding context windows [53], the volume of agent-driven economies are only expected to increase [47, 67], 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 [55].

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 as

Preprint.

fairness, efficiency, and welfare maximization [45, 22, 61, 56, 73]. Common forms of intervention include monetary transfers [59], incentive-compatible coordination signals [7, 8], and information revelation mechanisms [49, 16]. 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 call the *agent design stage*. In many AI ecosystems, stakeholders first choose which agents to deploy on their behalf (selecting model backbones, prompts, tools, and training pipelines) and only then do those agents interact [77, 2]. These design choices are strategically interdependent: the performance of one deployed agent often depends on the agents chosen by others. We model this stage itself as a *meta-game*, highlighting a difference from most multi-agent systems work, which studies the *downstream* behavior of deployed agents, and from classical mechanism design, which changes the interaction rules. Our focus is on mediating the *upstream* stakeholder game that determines which agents are deployed in the first place.

In this position paper, we argue that the design of AI agents should be *mediated* to promote social welfare and other societal objectives, 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 first develop these ideas through a sequence of stylized examples spanning a range of mediation approaches and application scenarios. We then move beyond abstract settings and demonstrate the role of mediation in a real-world use case involving automated bargaining agents, using data collected from LLM-based interactions.[1] Finally, we outline concrete research directions, discuss alternative perspectives, and conclude with a call for a multidisciplinary effort to study and develop mediation mechanisms for agent design in emerging AI ecosystems.

2 Agent design as a meta-game

We begin by formalizing agent design among strategic stakeholders through a game-theoretic lens, adopting the perspective that agent design itself constitutes a *meta-game*. While it is well known and widely accepted to model the interactions among deployed agents as a game, our position highlights an earlier and equally important strategic stage: before agents interact, stakeholders must decide *which agents to deploy*. These decisions are interdependent and therefore naturally form a strategic interaction among stakeholders. We refer to this stage as the *agent design game*.

Conceptually, this game precedes and shapes the downstream interaction among deployed agents. For simplicity and clarity of exposition, our framework abstracts away from the details of the induced interaction game and instead represents its outcomes through reduced-form payoff functions. This abstraction allows us to focus directly on the strategic structure of the design stage and on how mediation at this level can influence equilibrium outcomes in agentic AI ecosystems.

Consider 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. The underlying interaction among agents is abstracted into the payoff functions, reducing the model to a normal-form game.

Under the standard assumption of rational self-interested behavior, the relevant solution concept for the agent design game is the *Nash equilibrium* [62]: 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

[1] Code is attached within the supplementary material of the paper.

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 mostly focus on *utilitarian social welfare*, defined simply as the sum of players' utilities, $\sum_i u_i(a)$.[2] 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.

Remark 1. Example 1, as well as the examples presented later in Section 3, are intentionally stylized and simplified. Their purpose is to illustrate the meta-game perspective, highlight the types of inefficiencies that may arise, and demonstrate mediation techniques. In Section 4, we move beyond these abstractions and show how mediation can improve welfare in an agent-design meta-game constructed using real data collected from interactions between LLM-based bargaining agents.

3 Fifty forms of mediators

While there are many possible forms of mediation that could be applied to AI design, 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 modern AI ecosystems.

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:

[2]We mostly adopt utilitarian social welfare for simplicity. However, as discussed in Section 4, the same mediation framework can be adapted to promote alternative societal objectives, such as Nash welfare or fairness (see Remark 2).

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.

Strategy space restrictions correspond to mediators that limit the available design choice space. 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. Thus, while strategy restriction can be effective, 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** [65], 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 [59], 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 the following transfer scheme: If (L, S) is played, Player 1 is rewarded with an additional payment equivalent to a utility bonus of $+8$. Symmetrically, if (S, L) is played, Player 2 is rewarded a similar bonus. This modification changes the payoffs $u_1(L, S)$ and $u_2(S, L)$ from -2 to 6, thereby making L a strictly dominant strategy for both players. As a result, (L, L) becomes the *unique* equilibrium outcome. Importantly, since transfers are never actually triggered, 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.[3] 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 [7]. 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 [63]. 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. [5], who derived sharp bounds in several canonical settings. The following example illustrates this:

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.

Correlation devices are 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

[3] 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.

instance, balancing rule-based filters against LLM-driven models, or signaling when to prioritize personalization over diversity. In such domains, correlation devices can align design choices to balance efficiency and fairness, enabling outcomes unattainable through decentralized behavior alone.

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 [49, 17, 15, 16], 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 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* [41]. 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. 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 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 [75].

Beyond the observation that such problems can be cast as linear programs [26, 24, 33], Smolin and Yamashita [75] characterize the optimal information structures in concave games [66] 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.[4]

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

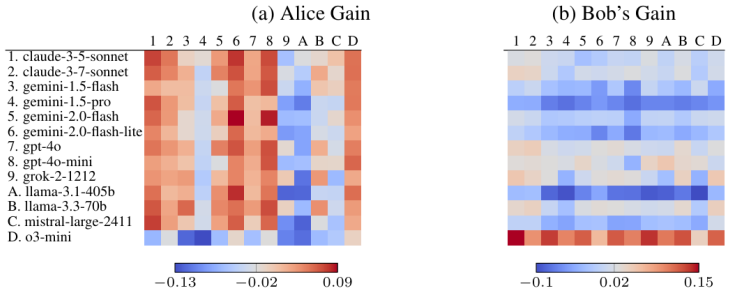

Figure 1: The agent design meta-game, taken from Shapira et al. [71].

Information design is relevant when platforms can access global data. For example, a large recommendation platform can see aggregate user preference signals that individual developers cannot. By carefully choosing what feedback to provide, the platform can shape agent design, aligning deployed agents with user needs while preserving ecosystem diversity, robustness, and adaptability.

4 A real-world scenario: mediating the design of bargaining agents

We now demonstrate how mediation can improve outcomes in a realistic agent-design environment based on empirical interaction data. Consider two stakeholders, Alice and Bob, who each deploy an LLM-based agent to participate in an alternating-offer bargaining game [68], with Alice making the first offer. To construct the agent-design meta-game, we rely on the `GLEE` framework of Shapira et al. [71], a benchmark for evaluating language-based economic interactions between LLM agents. The framework includes a large collection of bargaining environments spanning dozens of parametrized market configurations that vary along economically meaningful dimensions such as discount factors, information structure, horizon length, and communication format. For each pair of models, the dataset reports the average utility obtained by each player across these configurations, providing an empirical estimate of their expected performance in heterogeneous markets. We interpret the resulting payoff bi-matrix (Figure 1) as the payoff matrix of an agent-design meta-game in which stakeholders simultaneously choose which model to deploy across a population of bargaining environments, and the reported averages serve as estimates of the induced expected utilities. Without mediation, the unique Nash equilibrium of the game is the profile in which Alice chooses `claude-3-5-sonnet` and Bob chooses `claude-3-7-sonnet`, yielding payoffs of $u_A \approx 0.047, u_B \approx 0.026$ and a social welfare of $SW := u_A + u_B \approx 0.073$. In what follows, we present two practical mediation approaches, demonstrate their effectiveness, and discuss how they can be implemented in real-world environments.

Restriction to a closed model family. In our agent-design meta-game, each stakeholder can choose among a set of 13 candidate models, including proprietary models developed by providers such as OpenAI, Anthropic, and Gemini, as well as open-source alternatives such as Llama and Mistral. This formulation is appropriate in environments where the interaction platform merely facilitates communication between agents while stakeholders retain full freedom over model selection and deployment. In many practical settings, however, the platform hosting the interaction also provides the underlying models, and therefore, has the ability to restrict stakeholders to a predefined model catalog. An even more realistic scenario arises when deployment is limited to models from a single provider (for example, due to technical integration constraints or commercial agreements). Motivated by such settings, we consider a mediated version of the agent-design game in which the bargaining platform restricts both stakeholders to select models from a closed family (e.g., only Gemini models). This constitutes a concrete instance of the strategy-space restriction mediator introduced in Section 3.1. Indeed, when the mediator restricts both Alice and Bob to select among Gemini models (excluding the outdated `gemini-1.5-flash`), the resulting agent-design game admits a unique Nash equilibrium at (`gemini-1.5-pro`, `gemini-2.0-flash`), yielding an increase of $\approx 11.26\%$ in social welfare relative to the unrestricted setting.[5] This illustrates how platform-level control over model availability can reshape equilibrium deployment choices and improve ecosystem-level outcomes.

[5]Excluding `gemini-1.5-flash` is essential: if it remains available, an additional equilibrium emerges in which both players select `gemini-1.5-flash`, resulting in lower welfare than in the unrestricted game. An alternative approach would

Model recommendation with conditional vouchers. In many agent deployment environments, platforms do not merely restrict the set of available models, but actively recommend specific configurations to stakeholders, e.g., through default selections, ranking interfaces, or automated deployment suggestions. These recommendations can be reinforced by conditional incentives such as compute credits, pricing discounts, or priority access to infrastructure that are granted only when stakeholders follow the recommendation. Importantly, such incentives can naturally depend on the *joint deployment outcome*: for instance, a platform may provide discounts only when both parties deploy compatible models, when a recommended pair of agents is jointly adopted, or when deployment choices satisfy certain interoperability or performance criteria. Formally, this corresponds to a correlated mediation scheme with state-contingent monetary transfers, in which the platform samples a pair of recommended models according to a joint distribution and provides stakeholder-specific vouchers conditional on the realized recommendation being followed by both sides. This can be seen as a combination of monetary payments (Section 3.2) and correlation devices (Section 3.3). Unlike strategy space restriction, this approach preserves flexibility by allowing all models to remain available while selectively steering equilibrium behavior toward socially beneficial outcomes. In our scenario, this mediator increases social welfare to ≈ 0.0952, an improvement of $\approx 29.6\%$ relative to the non-mediated game. The required transfers remain moderate: the expected transfer cost is ≈ 0.0073, which is $\approx 10.0\%$ of the baseline welfare (or $\approx 7.6\%$ of the mediated welfare). Appendix B provides the linear program formulation of this mediator and its solution for our game.

Remark 2. The mediators presented here, as well as those discussed in Section 3, can be adapted to alternative societal objectives, such as fairness $-(u_A - u_B)^2$ or Nash welfare $u_A \cdot u_B$. Since these objectives are concave, the resulting optimization problems remain tractable, replacing linear programs with convex programs that maximize a concave function over a convex set.[6]

5 Research directions

Our perspective gives rise to concrete research directions. First, we need richer models of *agent design interactions*. Much of the existing literature models strategic behavior *after* systems have already been fixed, e.g. in recommendation and search [13, 43, 46], data sharing [36, 37, 76], and strategic classification [40]. By contrast, our focus is on the earlier game in which stakeholders choose architectures, prompts, tools, and training pipelines. A natural next step is to move beyond reduced-form utilities and jointly model the *design game* and the *induced downstream interaction game*, so that one can ask how mediation at the design stage propagates to downstream outcomes. [7]

Second, abstract mediation ideas should be turned into a theory of *practical mediators*. This includes robust mediators that operate under uncertainty about stakeholders' preferences, beliefs, or private information [30, 54, 14, 4, 29, 9], as well as incentive-design methods that implement desirable equilibria through transfers, rewards, or payoff shaping [81, 34, 82, 35, 58]. An equally important direction is to go beyond welfare maximization and design mediators for fairness objectives, building on various notions of fairness [73, 23, 12, 50]. On the algorithmic side, this raises tractability questions, related to recent work on computational aspects of economic models [63, 26, 24, 10, 27].

Finally, these ideas should be tested in *real AI ecosystems*, potentially relying on existing game theoretic evaluation benchmarks [25, 71, 79]. One promising agenda is to build simulation environments in which stakeholders choose agent designs, a mediator intervenes, and the resulting agents interact at scale. Such environments would enable empirical comparison of mediation protocols, advancing mediated agent design from a conceptual proposal to an experimentally grounded research program.

6 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.

---

be to allow the entire Gemini model family and then implement the welfare-maximizing equilibrium as a dominant-strategy equilibrium using the zero-implementation technique of Monderer and Tennenholtz [59] (as in Example 3).

[6] For Nash welfare, a concave formulation can be obtained after normalizing the game so that all payoffs are nonnegative.

[7] Related ideas from meta-games and games with commitments may provide useful starting points [52, 78, 48].

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 [74, 64, 20]. From this viewpoint, external interventions could stifle innovation or introduce inefficiencies. For example, overly restrictive content-moderation protocols could reduce diversity, leading to homogenized outcomes or discouraging experimentation [70]. 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 [18, 87].

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 [28, 51]. By making the rationale explicit and interpretable, mediators can enhance trust and legitimacy among affected stakeholders, fostering confidence in the governance of AI agent interactions.

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 outcomes within the focal game, it may still be perceived as unfair 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.

7 Concluding remarks

In this paper, we argued that the process of agent design can benefit from careful 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 and real-world examples, we illustrated how mediation can take different forms and how these can improve outcomes in strategic environments. We then outlined a set of research directions aimed at translating these 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 our understanding of when and how mediation should occur. Constructive debate around these concerns can lead to effective, 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

[1] 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.

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

[3] 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.

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

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

[6] 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.

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

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

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

[19] 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.

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

[21] 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.

[22] 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`.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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, 2023. Curran Associates Inc.

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

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

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

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

[57] 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.

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

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

[60] 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*, 2025.

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

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

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

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

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

[66] J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

[67] 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.

[68] Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.

[69] 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.

[70] 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.

[71] 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*, 2025. URL `https://openreview.net/forum?id=xqt43SBjke`.

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

[73] 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.

[74] 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.

[75] 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.

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

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

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

[79] Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas. *arXiv preprint arXiv:2604.15267*, 2026.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

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

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

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

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

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 $(C_1, (C_2, T))$.

B Correlated mediation with conditional vouchers

B.1 Linear-program formulation

Consider a two-player normal-form game with row-player payoff matrix $A \in \mathbb{R}^{m \times n}$ and column-player payoff matrix $B \in \mathbb{R}^{m \times n}$. A mediator draws an action profile (i, j) from a joint distribution $p = (p_{ij})$ and recommends action i to the row player and action j to the column player. In addition, the mediator may provide nonnegative transfers t_{ij}^R to the row player and t_{ij}^C to the column player whenever profile (i, j) is realized.

The mediator chooses variables

$$p_{ij} \geq 0, \qquad t_{ij}^R \geq 0, \qquad t_{ij}^C \geq 0 \qquad \forall i \in [m], \; j \in [n],$$

subject to

$$\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} = 1.$$

The obedience constraints require that, conditional on receiving a recommendation, neither player wishes to deviate. Thus, for the row player,

$$\sum_{j=1}^{n} p_{ij}(A_{ij} - A_{i'j}) + \sum_{j=1}^{n} t_{ij}^{R} \geq 0, \qquad \forall i \in [m],\ \forall i' \neq i,$$

and for the column player,

$$\sum_{i=1}^{m} p_{ij}(B_{ij} - B_{ij'}) + \sum_{i=1}^{m} t_{ij}^{C} \geq 0, \qquad \forall j \in [n],\ \forall j' \neq j.$$

Let $\lambda \geq 0$ denote the penalty on realized transfers. The mediator solves

$$\begin{aligned}
\max_{p, t^R, t^C} \quad & \sum_{i,j} p_{ij}(A_{ij} + B_{ij}) - \lambda \sum_{i,j} (t_{ij}^{R} + t_{ij}^{C}) \\
\text{s.t.} \quad & \sum_{j} p_{ij}(A_{ij} - A_{i'j}) + \sum_{j} t_{ij}^{R} \geq 0 \qquad \forall i,\ \forall i' \neq i, \\
& \sum_{i} p_{ij}(B_{ij} - B_{ij'}) + \sum_{i} t_{ij}^{C} \geq 0 \qquad \forall j,\ \forall j' \neq j, \\
& \sum_{i,j} p_{ij} = 1, \\
& p_{ij} \geq 0,\ t_{ij}^{R} \geq 0,\ t_{ij}^{C} \geq 0 \qquad \forall i, j.
\end{aligned}$$

Hence, correlated mediation with conditional vouchers is a linear program. In our experiments we use $\lambda = 1$, representing equal weight to welfare and transfer cost.

B.2 Solution structure in the bargaining meta-game

We now apply the above program to the bargaining meta-game induced by the GLEE payoff bi-matrix in Figure 1. Throughout this section, action indices are mapped to models exactly as in Figure 1. In particular, indices $0, \ldots, 12$ correspond to the ordered list of models displayed next to the payoff matrix. For the mediated problem, the optimal correlated recommendation is supported on only five action pairs:

$$p_{1,0}^{\star} = 0.6539, \qquad p_{1,1}^{\star} = 0.0450, \qquad p_{1,10}^{\star} = 0.0041, \qquad p_{10,0}^{\star} = 0.2587, \qquad p_{10,10}^{\star} = 0.0382.$$

The resulting gross social welfare is

$$\sum_{i,j} p_{ij}^{\star}(A_{ij} + B_{ij}) = 0.0952,$$

while the expected realized transfer cost is

$$\sum_{i,j} \left(t_{ij}^{R\star} + t_{ij}^{C\star}\right) = 0.0073.$$

With $\lambda = 1$, the net mediator objective therefore equals

$$0.0952 - 0.0073 = 0.0880.$$

The nonzero realized transfers are sparse and small. In particular, the mediator pays the row player

$$t_{1,0}^{R\star} = 0.0068,$$

and the column player receives transfers only on

$$t_{0,1}^{C\star} = 0.00012, \qquad t_{0,10}^{C\star} = 0.00035.$$

Thus, in this bargaining agent design application, the optimal mediator relies on a highly sparse recommendation rule together with very limited profile-contingent vouchers.

C Declaration of generative AI usage

Generative AI tools were used in the preparation of this manuscript for two purposes:

First, we used ChatGPT to polish, edit, and improve the writing quality. More specifically, AI was used in either of the following forms: (i) the authors asked the AI to go over a specific paragraph and propose local polishing or fix grammar mistakes; and (ii) the authors provided the AI with a written paragraph, asked it to paraphrase it with a specific instruction (e.g., "rewrite the following paragraph more concisely"), and then manually reviewed both the source and output, and rewrote the final paragraph. In both cases, the AI was not used to generate ideas or generate new content beyond paraphrasing under explicit instructions, but rather to improve the writing quality of existing text written by the authors.

Second, we used GitHub Copilot to write the analysis code for the real-world bargaining use case (Section 4). The authors have carefully reviewed both the source code and its generated output to ensure correctness and alignment with the analysis description in the paper.