Artificial Social Intelligence

Moshe Tennenholtz
moshet@technion.ac.il
Technion - Israel Institute of Technology
Haifa, Israel

Omer Madmon
omermadmon@campus.technion.ac.il
Technion - Israel Institute of Technology
Haifa, Israel

ABSTRACT

Artificial intelligence (AI) has been growing at an unprecedented pace. Many of us have experienced a "ChatGPT moment" — a realization that AI will profoundly transform our lives. While numerous challenges and calls for improvement remain, there is little doubt that AI agents will play a central role in shaping our future. We argue, however, that the prevailing perspective on AI agent design is insufficient for achieving desirable social welfare, not merely due to computational or regulatory constraints. A key shortcoming lies in overlooking the fact that AI agents operate within an AI ecosystem composed of multiple interacting agents. When such agents act jointly, misaligned incentives or incompatible technological designs may lead to poor social outcomes. Importantly, this perspective is orthogonal to the ongoing efforts to compare artificial and human behavior. Our argument is not merely conceptual but constitutes a concrete call to action: to establish a systematic research agenda on Artificial Social Intelligence. We illustrate this vision through four complementary research directions: (i) understanding multi-agent alignment in search ecosystems, (ii) analyzing model selection in language-based economics as a strategic choice, (iii) rethinking fairness and regulation through the lens of multi-agent ethics, and (iv) designing hybrid social laws for human–AI coexistence. Together, these directions outline a roadmap toward welfare-maximizing AI societies—an essential step toward socially aligned intelligence.

KEYWORDS

Collective Intelligence, AI Economics, AI Alignment, Algorithmic Game Theory, Responsible AI, Human–AI Interaction

1 INTRODUCTION

Artificial intelligence (AI) has entered a transformative era. Recent breakthroughs in large language models (LLMs), generative systems, and autonomous decision-making have led to the rapid emergence of AI agents capable of acting, interacting, and reasoning across domains traditionally dominated by humans. From information retrieval and content generation to negotiation, recommendation, and transportation, these agents increasingly populate the digital and physical spaces we inhabit. As AI systems begin to interact not only with humans but also with one another, a new question arises: *how do societies of AI agents behave?* Understanding this question is crucial for ensuring that the growing web of interacting artificial entities evolves in ways that serve collective welfare rather than undermining it.

The study of multi-agent systems (MAS) provides a natural theoretical foundation for this inquiry [22, 28, 45, 53, 55, 64]. Since its inception, MAS research has focused on coordination, cooperation, and competition among autonomous agents, developing tools to analyze fairness and efficiency [23, 26, 43], as well as equilibria and stability [3, 5, 11] in distributed settings. Game theory, mechanism design, and online learning have all contributed to a rigorous understanding of how rational or boundedly rational agents interact [9, 16, 17, 19, 63], particularly in dynamic, multi-agent settings [8, 12, 38, 39]. However, the modern AI landscape presents challenges that classical MAS theory could not have foreseen. The agents now entering our ecosystems are powered by heterogeneous technologies–ranging from symbolic reasoning systems [33, 34] to deep reinforcement learners [4, 15] and LLM-based decision makers [31, 56] –each endowed with distinct representations, inductive biases, and capabilities. Consequently, the traditional assumption of homogeneous rational agents gives way to a far more intricate reality: an ecosystem of *technologically diverse, incentive-driven, and strategically adaptive* AI agents.

Applications of such ecosystems are already visible. In search and recommendation systems, AI agents act on behalf of both producers and consumers, while AI-based rankers arbitrate their interactions and effectively define the incentive landscape [42, 44]. In digital marketplaces, automated buyers, sellers, and mediators negotiate through natural language [1, 21]. In high-stakes decision domains such as credit scoring, hiring, and insurance pricing, competing machine learning (ML) models are deployed by independent institutions, interacting indirectly to shape individuals' economic and social opportunities [2, 10, 32] . In transportation, autonomous vehicles share the road with human drivers under hybrid social and regulatory rules [27, 36].

Across these domains, a shared principle emerges: system-level outcomes depend not only on the design of individual agents but on the *alignment between their objectives, information structures, and technological representations.* When these alignments fail, efficiency, fairness, and trust can all deteriorate—sometimes dramatically—despite each agent operating optimally in isolation.

Current efforts to align AI, however, are almost exclusively focused on the relationship between a single AI model and its human overseers, ensuring that one system's outputs reflect human preferences, values, or intentions. While this form of alignment is essential, it overlooks the broader phenomenon that arises once multiple AIs interact. Our perspective is therefore complementary to, rather than competing with, studies that compare [37, 41, 51] or align [25, 30, 47] AI behavior with human behavior. Where those studies aim to make individual models human-compatible, we aim to make *ecosystems* of AI agents socially compatible. This calls for a new field of inquiry, *Artificial Social Intelligence*, that integrates insights from multi-agent theory, economics, and AI ethics to study how artificial agents interact, coordinate, and coexist.

This paper outlines a vision for Artificial Social Intelligence as a foundational step toward robust, fair, and welfare-maximizing AI societies. We highlight four complementary research directions that exemplify this agenda:

- Multi-agent alignment in search ecosystems.

- Strategic interaction in language-based economics.
- Ethics and fairness in competitive multi-agent contexts.
- Hybrid social laws for heterogeneous human–AI societies.

Together, these directions chart a path toward understanding, designing and implementing AI ecosystems that are not merely intelligent, but *socially* intelligent.

2 THE SEARCH ECOSYSTEM

The search ecosystem consists of three fundamental components: a corpus of documents, information needs expressed by users through queries or questions, and rankers that determine the relevance of content for a given query. In recent years, rankers have evolved from traditional lexical models (e.g., Okapi BM25 [49]) to neural approaches (e.g., E5 [61] and Contriever [24]), and more recently to a variety of LLM-based ranking architectures [35, 48]. At the same time, new AI technologies have emerged to serve other actors in the ecosystem: publishers now deploy *document agents* designed to promote their content [6], while users employ *query agents* that transform their information needs into effective queries [57]. Each of these agents—document, query, and ranker—may itself be lexical, neural, or LLM-based.

This multiplicity of interacting AI agents introduces a profound *multi-agent alignment problem*. The interplay among document agents, user agents, and ranker agents forms a complex strategic ecosystem whose efficiency and fairness depend on their technological and incentive alignment. Understanding this interplay is essential for determining when this emerging economy of data producers and data consumers yields socially desirable outcomes. Importantly, this challenge is *orthogonal to traditional LLM alignment*, which focuses on aligning a *single* model with human expectations [62]. Here, the core question is how to achieve *alignment across multiple interacting AI agents*—a necessary step toward sustaining social welfare in digital ecosystems [44].

To illustrate this, consider an experiment designed not to optimize a particular document agent's performance, but to study how document agents and ranker agents interact. Both types of agents can be implemented as lexical, semantic, or LLM-based. The results reveal a clear pattern: when the document agent and the ranker agent are *mismatched* in type, the document agent's ability to promote its content in rankings drops significantly. Conversely, *alignment* between their underlying technologies substantially improves ranking promotion. Table 1, taken from Nachimovsky et al. [44], illustrates this phenomenon: for example, the E5 ranker—a dense semantic model—is most effectively influenced by a semantic (embedding-based) document agent, while LLM-based rankers respond best to LLM-based document agents, and lexical rankers to lexical document agents. Importantly, *multi-agent alignment should not be conflated with homogeneity of design*; as shown in Section 3, optimal outcomes may emerge precisely from interactions among technologically diverse agents.

In practice, the real search ecosystem involves countless query, document, and ranking agents, each possibly using distinct representations and learning paradigms. Their alignment—or lack thereof—has a direct impact on both user experience and publisher welfare. Without a systematic study of such *multi-agent alignment*, AI technologies in search environments may yield unpredictable and suboptimal social outcomes.

Type	Model	Lexical		Semantic		LLM	
		BM25	TF.IDF	Contriever	E5	Gemma	Llama
Human	-	0.071	0.074	0.179	0.183	0.117	0.078
Lexical	-	**0.430**	**0.433**	0.253	0.236	0.158	0.049
Semantic	Contriever	0.269	0.267	**0.363**	0.289	0.226	0.094
Semantic	E5	0.229	0.233	0.308	**0.329**	0.183	0.092
LLM	Gemma	0.217	0.216	0.273	0.264	**0.497**	0.280
LLM	Llama	0.084	0.089	0.268	0.240	0.354	**0.640**

Table 1: From "Agent Alignment" to "Multi-agent Alignment" within the search ecosystem, from Nachimovsky et al. [44].

3 LANGUAGE-BASED ECONOMICS

The search ecosystem already illustrates our claim that achieving social welfare through artificial social intelligence requires studying the *alignment* among interacting AI agents. However, similar challenges are expected to arise in other emerging markets populated by strategic AI agents. Persuasion, negotiation, and bargaining tasks are rapidly becoming automated, as large LLMs enable natural language interactions between autonomous agents. These central economic interactions are inherently multi-agent: whereas in the search ecosystem we encounter content producers, consumers, and rankers, here we face sellers, buyers, and market makers – each driven by distinct incentives and powered by potentially incompatible AI technologies. Understanding how the alignment between these agent types affects economic outcomes (such as efficiency and fairness) is essential. Without it, the AI economy risks systemic failures once autonomous agents begin interacting at scale.

To explore these dynamics, Shapira et al. [52] introduced **GLEE**, a unified framework for *Games in Language-Based Economic Environments*, focusing initially on two-player interactions. GLEE provides a principled methodology for evaluating the behavior of both LLM agents and human participants across a wide class of language-based economic scenarios. At its core lies a clear parameterization of the space of bargaining, negotiation, and persuasion games, defining consistent degrees of freedom and evaluation metrics across domains. The framework's richness stems from its flexible parameters, including the game horizon (number of rounds), information structure (knowledge of opponents' preferences), and communication form (free-form dialogue vs. structured messages). GLEE is implemented as an open-source platform that enables researchers to instantiate a wide range of economic games and systematically evaluate LLM behavior within them. While specific models may evolve, the framework and its evaluation metrics remain general.

An empirical analysis by Shapira et al. [52] reveals several behavioral and economic regularities. First, economic outcomes such as efficiency, fairness, and self-gain are strongly affected by structural market parameters such as information structure, communication form, and interaction horizon. Second, there is no universally dominant LLM: performance depends critically on the opponent's identity and behavior. Finally, human participants exhibit more extreme behavior: they either outperform all LLMs or perform significantly worse, depending on the context and role assigned. These findings demonstrate that GLEE serves not only as a research infrastructure but also as a tool for uncovering new insights into economic reasoning and strategic behavior in language-based interactions.

Figure 1 illustrates one such result. The figure shows how different combinations of buyer and seller LLMs affect efficiency (trades occur if and only if the buyer's value exceeds the seller's) and fairness (distance from the midpoint between buyer and seller values when trade occurs). This view highlights that the choice of LLM itself can be viewed as a *strategic decision* made by the agents' deployers. Before any negotiation occurs, each participant selects which LLM to deploy—thus defining a *meta-game* played over model choices. Notably, equilibrium and welfare-optimal outcomes do not always emerge when both agents use the same model type, showing that *alignment* is a richer notion than simple homogeneity.

The outcome of this meta-game determines the subsequent dynamics and efficiency of the underlying economic interaction. Understanding such higher-order strategic behavior—where AI deployers strategically select technologies that later interact—is crucial for anticipating systemic effects in markets populated by autonomous agents. This emphasizes that the design of Artificial Social Intelligence must extend beyond individual models to encompass the ecosystems in which they are selected and deployed.

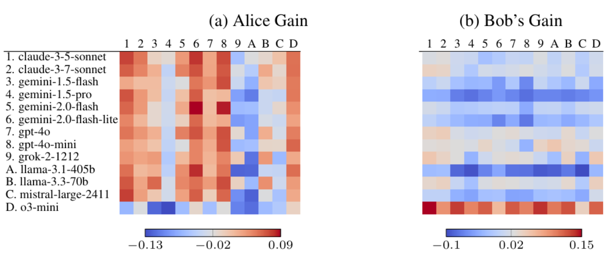

Figure 1: Language model selection as a meta-game, taken from Shapira et al. [52].

4 MULTI-AGENT ETHICS

In the previous sections, artificial agents participated in all parts of the ecosystem, including the *mediators* that govern these systems (e.g., the ranker in the search ecosystem or the market platform in language-based economic interactions). However, the vision of *Artificial Social Intelligence* extends further, to ecosystems in which the mediator acts as a *regulator*—enforcing ethical constraints, fairness criteria, or social norms on interacting AI agents.

The field of AI ethics has grown rapidly [46], with fairness occupying a central role [40, 65]. Fairness principles have been embedded into ML models and are increasingly promoted by regulators and policy makers [50, 60]. Yet, a critical gap remains: in multi-agent environments, enforcing fairness at the *individual level* may not ensure fairness at the *system level*. When multiple firms or agents each employ "fair" classifiers, the aggregate outcome of their interactions can still be socially undesirable, sometimes even *less fair* than in the absence of fairness constraints. This systemic failure, documented in the literature [7, 13, 18], exemplifies the need for an explicitly multi-agent approach to AI ethics – one that accounts for interactions, competition, and incentive effects. We cannot expect isolated regulatory measures or unilateral commitments by individual firms to resolve these issues without addressing the broader ecosystem.

To gain intuition, consider a market with multiple lenders offering loans to a shared pool of borrowers. Gradwohl et al. [18] adopted a welfare-driven framework to formalize *fairness under competition*. Intuitively, an ecosystem satisfies fairness under competition if the welfare of different borrower groups is equalized once lenders deploy their classifiers. Building on the standard Equal Opportunity (EO) fairness criterion [20], Gradwohl et al. [18] defined a new concept: *Equal Opportunity under Competition (EOC)*. The EOC level measures how far the ecosystem is from satisfying fairness under competition (lower values indicating higher ecosystem fairness).

Gradwohl et al. [18] show that even when each lender's classifier satisfies EO, the resulting ecosystem may deviate substantially from EOC. They provide quantitative bounds on this deviation, grounded in model primitives, and identify two primary mechanisms driving this discrepancy. The first arises when the correlation between classifiers differs across protected groups; the second when lenders' borrower pools overlap but are not identical. Both forces demonstrate how local fairness does not necessarily aggregate into global fairness once agents interact strategically.

Table 2 illustrates this phenomenon through experiments conducted on real and synthetic loan datasets of varying sizes (rows in the table). Each classifier was trained using standard ML techniques constrained to satisfy EO, yet the results (percent deviations in the table) show significant violations of EOC across settings. These findings underscore a key insight: constraining individual agents or firms to obey fairness criteria—while ignoring the interactions among them—fails to achieve genuine ethical outcomes at the system level. Addressing such multi-agent ethical failures is therefore a cornerstone of Artificial Social Intelligence.

	300	1k	3k	10k	30k	100k
Exp. 1	[75.0, 82.2]	[68.0, 76.4]	[55.6, 64.2]	[49.4, 58.2]	[42.0, 50.6]	[26.2, 34.0]
Exp. 2	[75.6, 82.8]	[65.2, 73.8]	[51.8, 60.8]	[35.4, 44.2]	[25.8, 33.8]	[12.6, 19.0]
Exp. 3	[74.2, 81.2]	[63.4, 71.2]	[50.8, 59.6]	[35.6, 43.8]	[27.6, 36.2]	[14.2, 20.6]
Exp. 4	[52.2, 60.4]	[34.2, 42.4]	[17.8, 25.2]	[5.4, 10.0]	[0.8, 3.2]	[0.0, 0.0]

Table 2: Individually fair classifiers can yield unfair outcomes under competition, adapted from Gradwohl et al. [18].

5 HYBRID SOCIAL LAWS

AI is rapidly transforming environments traditionally governed by human decision-making into fully autonomous multi-agent systems. A canonical example is the vision of autonomous transportation. Yet, an important and underexplored challenge lies in the design of *hybrid environments*: systems in which both human-controlled and AI-controlled agents coexist under potentially distinct social norms and regulations. Understanding how to govern such mixed populations effectively is an essential first step toward developing *hybrid social laws*: regulatory and behavioral frameworks that enable welfare-maximizing coexistence between humans and autonomous systems. Such laws are crucial for the gradual and safe transition toward fully automated ecosystems, and they call for joint attention from regulators, policy makers, and AI designers [29].

Current traffic laws and regulations are fundamentally designed for *human-driven vehicles* (HDVs). They rely on assumptions about human perception, cognition, and reaction times, embedding safety margins to account for human error and behavioral variability. For instance, speed limits are typically determined by worst-case considerations such as limited visibility, road curvature, and human reaction delays. These limits are calibrated to the capabilities of an average driver, not to the potential precision and coordination of autonomous vehicles.

In contrast, *Communicating Autonomous Vehicles (CAVs)*, or self-driving cars, possess different capabilities: they can obey rules precisely, exchange information in real time, and process data far more rapidly than humans. Although these capabilities are still developing, even near-term CAVs can be expected to perform at least as safely and efficiently as competent human drivers. Moreover, the adoption of CAVs is likely to be gradual, leading to a transitional period characterized by mixed traffic composed of both human and autonomous vehicles.

Addressing this transition calls for a multi-agent perspective on transportation, treating each vehicle as an autonomous agent and traffic regulations as a form of *social law* [14, 54, 58, 59]. This view highlights the need for new, differentiated rules that acknowledge the heterogeneous capabilities of human and artificial drivers. In a recent work, Kraicer et al. [29] proposed *hybrid traffic laws*: a class of regulations tailored to mixed traffic environments that assign distinct behavioral requirements to CAVs and HDVs. Unlike approaches that rely solely on the superior driving skills of CAVs, hybrid traffic laws leverage their capacity to follow dynamically changing and context-sensitive rules. By exploiting these strengths, such laws can improve traffic flow and enhance safety.

Kraicer et al. [29] explored a range of hybrid policy designs across varying proportions of autonomous vehicles. Their analysis demonstrated that, by recognizing CAV-specific abilities, regulators can create policies that maintain efficiency and safety even during the transitional phase toward full automation. Figure 2, adopted from Kraicer et al. [29], illustrates the results of one such policy experiment. Each column represents a different density of CAVs, while rows correspond to different overall traffic loads. HD denotes human-driven vehicles, and AVs are autonomous ones. The numbers 1–5 indicate increasing passenger demand thresholds required for CAVs to use a dedicated lane. The table reports the resulting average delays under these policies, showing that well-designed hybrid traffic laws–where certain constraints apply selectively to human drivers–can substantially improve aggregate efficiency and reduce congestion. These findings underscore the potential of hybrid social laws as a regulatory bridge between human and fully autonomous societies.

Looking ahead, it is reasonable to expect similar challenges to arise not only in ecosystems where humans and AI coexist, but also in future societies composed entirely of diverse AI agents, each operating under its own objectives, capabilities, and social norms.

	Demand	vType	numPass	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0		HD		192.807	173.815	181.5477	195.4971	202.9642	207.724	210.4421	209.8008	217.0643	213.1694
1			1		120.5765	106.1905	132.6179	155.5527	168.8778	179.8417	188.5404	194.9498	200.3498
2			2		120.2105	96.78803	108.7041	130.1498	143.1436	150.0171	155.2124	160.0152	163.644
3			3		119.0711	94.19102	96.43404	112.1525	121.4078	128.6566	133.0982	137.6416	140.3156
4			4		120.7064	94.12769	92.14707	102.7259	111.3948	114.0697	116.7013	118.239	120.3276
5	Daily12_2	AV	5		121.8065	96.71369	95.01156	98.61356	103.5838	106.7504	110.0288	111.0654	113.0295
6		HD		193.25	182.0514	199.7239	216.5159	225.0792	230.0939	233.0191	233.7107	239.3215	235.366
7			1		120.8196	109.0813	143.1662	168.5117	183.7632	197.4054	208.217	215.2919	220.249
8			2		120.8519	98.10322	111.6981	137.1477	152.3528	160.534	165.8047	172.6427	174.7568
9			3		120.2788	97.01408	98.60115	113.7629	128.5454	136.5664	138.4371	145.0968	146.2401
10			4		119.8881	96.46447	94.91696	102.2126	112.4944	116.917	119.333	122.2484	124.2792
11	Daily12_2.5	AV	5		122.6068	98.93614	93.61976	98.66064	103.7731	110.976	113.3258	114.9097	193.25
12		HD		142.862	54.65004	70.97189	117.012	145.028	159.2579	162.9034	164.305	168.3315	162.7599
13			1		33.83732	39.39451	76.41571	106.3076	126.7745	136.2531	144.5794	149.9113	153.0394
14			2		34.25718	35.9515	58.97368	85.59612	104.9186	109.7928	114.3042	117.5735	119.3604
15			3		34.14236	36.97639	52.78697	69.69083	85.29766	91.61522	95.63371	96.24237	97.54513
16			4		32.01905	34.97621	50.48681	64.90813	75.88094	79.72572	81.76952	82.51668	81.60089
17	Daily12_3	AV	5		33.26574	36.17929	50.03641	63.38762	70.68035	73.36346	77.29344	76.21123	75.09797

Figure 2: Towards a hybrid society: The role of hybrid traffic laws, taken from Kraicer et al. [29].

6 CONCLUSION WITH A CALL TO ARMS

Artificial Social Intelligence envisions a world where ecosystems of AI agents—each with distinct technologies, incentives, and objectives—interact, adapt, and co-evolve. Rather than focusing on isolated intelligence, this vision emphasizes the emergent properties of *societies* of intelligent agents. It challenges the prevailing paradigm of aligning a single model with human intent and instead asks: *How do we align a world of AIs with each other, and with us?*

The cases discussed in this paper—search ecosystems, language-based markets, multi-agent ethics, and hybrid social laws—are only early proof-of-concepts of this broader research direction. They illustrate that progress toward socially beneficial AI requires understanding incentives, communication, and coordination *between* artificial entities, not only between humans and machines. Without such a perspective, even powerful AI systems risk generating fragmented and inefficient ecosystems—an "intelligence without society." Our central claim is thus bold yet necessary: *without addressing Artificial Social Intelligence, the pursuit of socially-beneficial, cooperative AI will remain fundamentally incomplete.*

Realizing this vision demands an interdisciplinary research agenda that bridges economics, multi-agent systems, and the algorithmic foundations of AI. We must study new forms of alignment, competition, and cooperation among artificial agents, guided by concepts such as efficiency, fairness, and stability from the social sciences. Equally, regulators and policymakers must become active participants in this process, shaping environments where different AIs can coexist productively. In this sense, Artificial Social Intelligence is not merely a scientific aspiration—it is a call for rethinking how we design, deploy, and govern AI in complex societies. Ultimately, the road to socially aligned AI will not be paved by a single model or algorithm, but by an understanding of the ecosystems they form. Developing the science of Artificial Social Intelligence is, therefore, not just a technical challenge—it is an intellectual frontier at the intersection of technology, economics, and human values.

REFERENCES

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

[2] Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. *ai* 5, 1 (2024), 383–404.

[3] Abdollah Amirkhani and Amir Hossein Barshooi. 2022. Consensus in multi-agent systems: a review. *Artificial Intelligence Review* 55, 5 (2022), 3897–3935.

[4] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE signal processing magazine* 34, 6 (2017), 26–38.

[5] Rina Azoulay-Schwartz and Sarit Kraus. 2004. Stable repeated strategies for information exchange between two autonomous agents. *Artificial Intelligence* 154, 1-2 (2004), 43–93.

[6] Niv Bardas, Tommy Mordo, Oren Kurland, and Moshe Tennenholtz. 2025. Automatic Document Editing for Improved Ranking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2779–2783.

[7] Amanda Bower, Sarah N Kitchen, Laura Niss, Martin J Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391* (2017).

[8] Ronen Brafman and Moshe Tennenholtz. 2002. Efficient learning equilibrium. *Advances in Neural Information Processing Systems* 15 (2002).

[9] Felix Brandt, Vincent Conitzer, and Ulle Endriss. 2012. Computational social choice. *Multiagent systems* 2 (2012), 213–284.

[10] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. 2020. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110, 10 (2020), 3267–3297.

[11] Georgios Chalkiadakis, Edith Elkind, Evangelos Markakis, Maria Polukarov, and Nick R Jennings. 2010. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research* 39 (2010), 179–216.

[12] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[13] Cynthia Dwork and Christina Ilvento. 2018. Individual fairness under composition. *Proceedings of fairness, accountability, transparency in machine learning* (2018).

[14] David Fitoussi and Moshe Tennenholtz. 2000. Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence* 119, 1-2 (2000), 61–101.

[15] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. 2018. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning* 11, 3-4 (2018), 219–354.

[16] Piotr J Gmytrasiewicz and Edmund H Durfee. 2000. Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems* 3, 4 (2000), 319–350.

[17] Piotr J Gmytrasiewicz and Edmund H Durfee. 2001. Rational communication in multi-agent environments. *Autonomous Agents and Multi-Agent Systems* 4, 3 (2001), 233–272.

[18] Ronen Gradwohl, Eilam Shapira, and Moshe Tennenholtz. 2025. Fairness under Competition. *CoRR* abs/2505.16291 (2025). https://doi.org/10.48550/ARXIV.2505.16291 arXiv:2505.16291

[19] Mohammad Taghi Hajiaghayi, Robert Kleinberg, and Tuomas Sandholm. 2007. Automated online mechanism design and prophet inequalities. In *AAAI*, Vol. 7. 58–65.

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

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

[22] Michael N Huhns and Larry M Stephens. 1999. Multiagent systems and societies of agents. *Multiagent systems: a modern approach to distributed artificial intelligence* 1 (1999), 79–114.

[23] Vincenzo Iannino, Valentina Colla, Claudio Mocci, Ismael Matino, Stefano Dettori, Sebastian Kolb, Thomas Plankenbühler, and Jürgen Karl. 2021. Multi-agent systems to improve efficiency in steelworks. *Matériaux & Techniques* 109, 5-6 (2021), 502.

[24] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).

[25] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).

[26] Jiechuan Jiang and Zongqing Lu. 2019. Learning fairness in multi-agent systems. *Advances in Neural Information Processing Systems* 32 (2019).

[27] Hamid Khayyam, Bahman Javadi, Mahdi Jalili, and Reza N Jazar. 2019. Artificial intelligence and internet of things for autonomous vehicles. In *Nonlinear approaches in engineering applications: Automotive applications of engineering problems*. Springer, 39–68.

[28] Steffi Knorn, Zhiyong Chen, and Richard H Middleton. 2015. Overview: Collective control of multiagent systems. *IEEE Transactions on Control of Network Systems* 3, 4 (2015), 334–347.

[29] Tal Kraicer, Jack Haddad, Erez Karaps, and Moshe Tennenholtz. 2025. Towards Hybrid Traffic Laws for Mixed Flow of Human-Driven Vehicles and Connected Autonomous Vehicles. *CoRR* abs/2502.12950 (2025). https://doi.org/10.48550/ARXIV.2502.12950 arXiv:2502.12950

[30] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267 [cs.CL] https://arxiv.org/abs/2309.00267

[31] Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Kp Subbalakshmi, Jimin Huang, et al. 2025. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2509–2525.

[32] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 166–176.

[33] Baoyu Liang, Yuchen Wang, and Chao Tong. 2025. AI Reasoning in Deep Learning Era: From Symbolic AI to Neural–Symbolic AI. *Mathematics* 13, 11 (2025), 1707.

[34] Ben Liu, Jihai Zhang, Fangquan Lin, Cheng Yang, Min Peng, and Wotao Yin. 2025. Symagent: A neural-symbolic self-learning agent framework for complex reasoning over knowledge graphs. In *Proceedings of the ACM on Web Conference 2025*. 98–108.

[35] Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhuo Li, Zhicheng Dou, Defu Lian, and Jian-Yun Nie. 2024. Information retrieval meets large language models. In *Companion Proceedings of the ACM Web Conference 2024*. 1586–1589.

[36] Yifang Ma, Zhenyu Wang, Hong Yang, and Lin Yang. 2020. Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica* 7, 2 (2020), 315–329.

[37] Olivia Macmillan-Scott and Mirco Musolesi. 2024. (Ir) rationality and cognitive biases in large language models. *Royal Society Open Science* 11, 6 (2024), 240255.

[38] Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. 2025. On the Convergence of No-Regret Dynamics in Information Retrieval Games with Proportional Ranking Functions. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=jJXZvPe5z0

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

[40] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[41] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *PNAS* 21, 9 (2024).

[42] Tommy Mordo, Sagie Dekel, Omer Madmon, Moshe Tennenholtz, and Oren Kurland. 2025. RLRF: Competitive Search Agent Design via Reinforcement Learning from Ranker Feedback. *arXiv preprint arXiv:2510.04096* (2025).

[43] Dolev Mutzari, Yonatan Aumann, and Sarit Kraus. 2023. Resilient fair allocation of indivisible goods. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2688–2690.

[44] Haya Nachimovsky, Moshe Tennenholtz, and Oren Kurland. 2025. A Multi-Agent Perspective on Modern Information Retrieval. *CoRR* abs/2502.14796 (2025). https://doi.org/10.48550/ARXIV.2502.14796 arXiv:2502.14796

[45] David C Parkes and Michael P Wellman. 2015. Economic reasoning and artificial intelligence. *Science* 349, 6245 (2015), 267–272.

[46] Michael Pflanzer, Veljko Dubljević, William A Bauer, Darby Orcutt, George List, and Munindar P Singh. 2023. Embedding AI in society: ethics, policy, governance, and impacts. *AI & society* 38, 4 (2023), 1267–1271.

[47] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

[48] Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Guiding retrieval using llm-based listwise rankers. In *European Conference on Information Retrieval*. Springer, 230–246.

[49] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

[50] Nicholas Schmidt, Bernard Siskin, and Syeed Mansur. 2018. How data scientists help regulators and banks ensure fairness when implementing machine learning and artificial intelligence models. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA*, Vol. 19.

[51] Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. 2024. Can llms replace economic choice prediction labs? the case of language-based persuasion games. *arXiv preprint arXiv:2401.17435* (2024).

[52] Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. 2024. GLEE: A Unified Framework and Benchmark for Language-based Economic Environments. *CoRR* abs/2410.05254 (2024).

https://doi.org/10.48550/ARXIV.2410.05254 arXiv:2410.05254
[53] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
[54] Yoav Shoham and Moshe Tennenholtz. 1995. On social laws for artificial agent societies: off-line design. *Artificial intelligence* 73, 1-2 (1995), 231–252.
[55] Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots* 8, 3 (2000), 345–383.
[56] Chuanneng Sun, Songjun Huang, and Dario Pompili. 2025. LLM-Based Multi-Agent Decision-Making: Challenges and Future Directions. *IEEE Robotics and Automation Letters* (2025).
[57] Zhaoyan Sun, Xuanhe Zhou, Guoliang Li, Xiang Yu, Jianhua Feng, and Yong Zhang. 2024. R-bot: An llm-based query rewrite system. *arXiv preprint arXiv:2412.01661* (2024).
[58] Moshe Tennenholtz. 1995. On computational social laws for dynamic non-homogeneous social structures. *Journal of Experimental & Theoretical Artificial Intelligence* 7, 4 (1995), 379–390.
[59] Moshe Tennenholtz. 1998. On stable social laws and qualitative equilibria. *Artificial Intelligence* 102, 1 (1998), 1–20.
[60] Mike Teodorescu, Yongxu Sun, Haren N Bhatia, and Christos Makridis. 2025. An Analysis of the New EU AI Act and A Proposed Standardization Framework for Machine Learning Fairness. *arXiv preprint arXiv:2510.01281* (2025).
[61] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).
[62] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. 2024. A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv:2407.16216 [cs.CL] https://arxiv.org/abs/2407.16216
[63] Ying Wen, Yaodong Yang, Rui Luo, and Jun Wang. 2019. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. *arXiv preprint arXiv:1901.09216* (2019).
[64] Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.
[65] Wenbin Zhang. 2024. AI fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine* 45, 3 (2024), 386–395.