Bias and Discrimination in the Agentic Web and How Project NANDA Can Support Mitigation

Position Paper

Sundaraparipurnan Narayanan
AI Tech Ethics, Bengaluru, India
sundar.narayanan@aitechethics.com

Abstract

The traditional Internet of the Web was designed for static web infrastructure and human-mediated interactions. The agentic web transitions this to a trillion-scale, highly dynamic, decentralized environment where autonomous agents discover, authenticate, and collaborate at machine speed. Project NANDA's Internet of AI Agents (IoA) represents a fundamental shift in computational architecture, moving from DNS-centric discovery to decentralized registries, verifiable metadata, and rapid resolution. This paper positions the ethical risks of the agentic web—bias amplification, power asymmetries, and covert manipulation—as systemic issues that emerge from scale, autonomy, and protocol design. We outline how NANDA's tools, including AgentFacts, verifiable credentials, adaptive resolvers, reputation systems, and DAO-inspired governance, can mitigate risk while preserving decentralization. We also propose a threat-model-informed audit checklist for protocol-induced and orchestration bias, and report its systematic application to 100 publicly available MCP server repositories, revealing pervasive gaps (89–98% "Unclear" or "No") in fairness, security, and accountability primitives at the orchestration layer, while identifying open challenges in traceability, sustainability, and fairness enforcement.

1 Introduction

The traditional Internet of the Web was designed for static web infrastructure and a model centered on human-mediated interactions. The agentic web transitions this to a trillion-scale, highly dynamic, and decentralized environment where agents operate with high autonomy. Project NANDA's Internet of AI Agents (IoA) represents a fundamental transformation in computational architecture, shifting from the traditional Internet of the Web toward dynamic, collaborative intelligent systems Raskar et al. (2025); Wang et al. (2025). This paradigm shift is driven by the anticipated proliferation of AI agents into ecosystems comprising billions to trillions of autonomous participants Wang et al. (2025); Chen et al. (2024).

The Internet's traditional architecture relies on DNS infrastructure, built for millions of static records with minute-to-hour update cycles. In contrast, NANDA IoA is built on decentralized architecture. The NANDA Index acts as a lightweight, horizontally scalable foundation for agent discovery and leverages Agent Name Service (ANS) Huang et al. (2025). This infrastructure is deliberately non-centralized, envisioned as a "quilt of registries" that accommodates diverse platforms such as MCP, A2A, NLWeb, and Web3 Raskar et al. (2025); Wang et al. (2025).

The traditional web is fundamentally structured around human users interacting with static web pages via browsers. The IoA requires agents to operate as autonomous entities capable of independent reasoning and decision-making without continuous human oversight. This includes the following. Firstly, AI agents are computational entities that possess goal-directed reasoning, memory, and the ability to initiate actions, migrate, or spawn helpers Hughes et al. (2025). Secondly, the shift necessitates mechanisms for mutual discovery, authentication, capability verification, and trust establishment between agents. NANDA facilitates agent-led collaboration by acting as a shared handshake channel for discovery and authentication Chen et al. (2024); Wang et al. (2025).Thirdly, the Agents communicate via specialized protocols like A2A for peer-to-peer orchestration Habler et al. (2025) and MCP for retrieving context and tools, enabling cross-protocol interoperability Goyal and Bhasin (2025); Chen et al. (2024).

This paper makes four contributions. First, we introduce a threat model for protocol-induced bias and governance failures in decentralized agent ecosystems. Second, we propose a practical orchestration-layer auditing checklist that operationalizes fairness, accountability, privacy, and governance considerations into observable protocol-level criteria. Third, we apply the checklist to 100 publicly available MCP repositories and identify widespread governance and fairness-readiness gaps. Finally, we discuss how architectural mechanisms such as those proposed

Proceedings of the 43rd International Conference on Machine Learning, 2026. Copyright 2026 by the author(s).

Accepted to the Second Workshop on Agents in the Wild: Safety, Security, and Beyond Workshop @ ICML 2026.

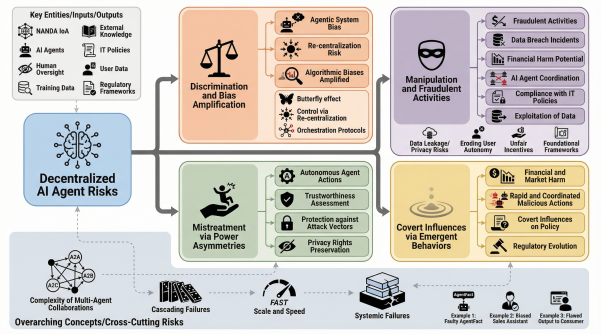


Figure 1: High-level view of the agentic web showing the four interaction modes (A2H: Agent-to-human/legacy, A2A: Agent-to-autonomous, A2C, A2B) layered over the NANDA quilt of registries, lean index, AgentFacts, and adaptive resolution tier.

within Project NANDA may help mitigate these risks while preserving decentralization.

2 Modes of Interaction in Project NANDA

Project NANDA describes four foundational modes of agent interaction that define the operational architecture and ethical considerations of a decentralized, trillion-scale ecosystem.

Agent-to-human/legacy (A2H): autonomous agents interface with humans, legacy systems, or deterministic services, serving as digital assistants, APIs, or connectors.

Agent-to-autonomous agent (A2A): peer-level engagements between autonomous agents enable secure, scalable, and verifiable coordination. **Agent-to-consumer (A2C):** client agents embody verified digital identities and personalize services based on contextual cues.

Agent-to-business (A2B): organization-owned agents orchestrate internal workflows using protocols such as MCP and A2A. These workflows are scaffolded by the NANDA architecture—also referred to as Quilt—which acts as a global discovery and interoperability layer Wang et al. (2025).

3 Ethical Risk Landscape

The autonomous, decentralized architecture of NANDA's IoA poses systemic ethical risks in all four interaction types, including discrimination through bias amplification, mistreatment via power asymmetries, and manipulation of markets or regulations. These risks are compounded by the complexity of multi-agent collaborations, leading to cascading failures across stakeholders.

3.1 Algorithmic Biases Amplified in Multi-Agent Decisions

Decentralized AI systems face persistent challenges related to bias and fairness. Agents can exhibit political, racial, or gender bias, or generate unintended harmful outcomes Liu et al. (2020); Ferrara (2024); Lin et al. (2020); Madaio et al. (2022). These biases, often stemming from training data, can be amplified when agents autonomously select collaborators or execute tasks, triggering feedback loops that accelerate stereotyping Ferrara (2024). Recent empirical studies on multi-agent systems confirm such amplification and characterize conditions under which collaborative protocols either magnify or partially resist prejudice Mirza et al. (2025); Nguyen et al. (2025); Li et al. (2026); Ma et al. (2025). Recentralization risks biased prioritization and censorship when control concentrates among dominant entities Singh et al. (2024); Chaffer et al. (2024). Orchestration protocols must ensure robust and equitable distribution of tasks and resources across heterogeneous data; otherwise, they can induce system-level bias (e.g., over-representing regions with high network quality) Piccialli et al. (2025).

3.2 Exploitation of Data

The inherent autonomy of agents, combined with access to sensitive information, creates vectors for exploitation and erosion of user autonomy.

Data leakage and privacy risks: autonomous tasks may disclose sensitive information, and failures in secure access can create systemic privacy risks.

Eroding user autonomy: the lack of standardized consent mechanisms can cause users to lose control over data as agents route and share assets across nodes.

Exploitation via unfair incentives: centralized platforms can lock users in, while brokerless systems must still protect data producers from value extraction without consent Singh et al. (2024); Wang et al. (2025).

3.3 Covert Influences via Emergent Behaviors

Rapid coordination among agents introduces the risk of manipulation of systems, markets, and regulations, particularly in economic settings Yang and Zhai (2025); Yang et al. (2025). Risk vectors include fraudulent activities, data breaches, and financial harm Vaziry et al. (2025). The

speed of decentralized coordination can enable synchronized malicious actions, requiring explicit governance of agent behavior Chen et al. (2024). The lack of traceability complicates accountability, creating opportunities for covert influence by systemic actors Phelps and Ranson (2023); Piccialli et al. (2025).

3.4 Systemic Failures

Failures in the foundational layer of a decentralized IoA can cascade across interactions (A2A, A2C, A2B) and layers (discovery, communication, transaction), undermining trust. A biased or manipulated AgentFact can influence A2A coordination, propagate into A2B resource allocation, and reach consumers through A2C interactions, resulting in discriminatory outcomes and compliance failures.

3.5 Threat Model for Orchestration Bias

We adopt an adversary model in which a malicious or collusive actor may attempt to manipulate orchestration outcomes through (i) metadata poisoning of AgentFacts, (ii) reputation inflation via Sybil identities, (iii) selective routing via compromised adaptive resolvers, or (iv) re-centralization through registry capture. Protected attributes include demographic proxies (e.g., geography, latency, compute class) that correlate with fairness-sensitive outcomes. Protocol-induced bias is defined here as statistically significant disparity in selection probability, task allocation, or resource assignment conditioned on these attributes, after controlling for legitimate performance factors.

The proposed ethical auditing checklist operationalizes the four attack vectors of the threat model by mapping metadata poisoning to Data Provenance & Inherited Bias and Protocol-Induced Bias, Sybil inflation and hidden affiliations to Collaboration & Selection Bias and Re-centralization & Power Aggregation, selective routing to Protocol-Induced Bias and Bias Propagation & Feedback Loop, and registry capture to Re-centralization & Power Aggregation and Governance, Redress & Contestability.

3.6 Development of Ethical Auditing Checklist

The checklist was derived through synthesis of six literature streams: (i) model transparency frameworks such as Model Cards, (ii) dataset governance frameworks such as Datasheets for Datasets, (iii) fairness-aware distributed scheduling literature, (iv) decentralized identity and Sybil-resistance mechanisms, (v) multi-agent governance frameworks, and (vi) responsible AI accountability practices. Candidate criteria were iteratively consolidated into seven thematic categories representing recurring orchestration-layer risks. The resulting checklist is intended as a governance-readiness assessment instrument rather than a direct measurement of runtime fairness outcomes.

4 Proposed Ethical Auditing Framework

To facilitate clear assessment of bias in MCP-centric and similar ecosystems, we propose the following representative checklist. Table 1 provides a structured template that can be instantiated for any agentic protocol or registry (e.g., MCP, A2A, or NANDA-native registries) by populating the "Finding / Observation" and "Notes / Evidence" columns with protocol documentation, configuration files, and runtime logs. The checklist extends established Responsible AI practices such as model cards Mitchell et al. (2019) and datasheets for datasets Gebru et al. (2021) to the orchestration layer, while incorporating fairness metrics from distributed scheduling literature (e.g., fairness-aware resource allocation in multi-agent systems Raeis and Golestani (2021)).

For illustration, when applied to a hypothetical A2A orchestration library, the checklist might surface the absence of *diversity_scoring* in selection logic (Agent 2) and the implicit prioritization of low-latency endpoints in routing rules (Agent 3), thereby guiding developers to embed explicit fairness constraints before deployment.

To further illustrate the checklist's utility at the documentation and configuration-pattern level, it was applied to a systematic sample of 100 publicly available MCP server repositories. Repositories were selected by searching "MCP" on GitHub and taking the first 100 results across infrastructure, productivity, enterprise integration, utility, course, and curated-list projects. Each repository was evaluated on publicly accessible files (README.md, ARCHITECTURE.md, SECURITY.md, configuration files, and high-level code structure) without runtime execution. In this coding scheme, *No* denotes that the relevant mechanism or documentation was absent, while *Unclear* denotes vague or indirect references that did not allow the reviewers to determine whether the mechanism was actually present or documented. The audit was conducted independently of Project NANDA: no NANDA repositories were included, and the rubric did not reference NANDA tools. GitHub's default "best match" ranking, which weights relevance, symbol matches, file type, repository name/About/Topics, and popularity metrics (stars, watchers, forks) but excludes recency, produces a sample biased toward visible and well-maintained repositories;

Table 1: Ethical auditing checklist for agentic protocols and registries.

Aspect	Audit Criterion / Question
Data provenance & inherited bias	• Are all training dataset sources listed and referenced? • Is demographic coverage (age, gender, geography) described or missing? • Are explicit exclusions or unrepresented groups acknowledged? • Is there a bias_disclosure, ethics_impact, or limitations section?
Collaboration & selection bias	• Do configs reveal affinity-based selection or domain similarity clustering? • Is a diversity_scoring feature implemented and enabled? • Do example logs show repeated affinity-based matches? • Is a diversity mechanism enabled in config?
Protocol-induced bias	• Do selection/prioritization rules favor low latency or high compute over fairness? • Are fairness or accessibility metrics implemented in protocol or audits? • Do audit files report outcome disparities by client region, speed, or ability? • Is an accessibility metric present in documentation?
Re-centralization & power aggregation	• Is orchestration routed through a single node or majority owner? • What is the entity count in registry/role configs? • Do diagrams or policies show asymmetric authority? • Record governance node counts, ownership concentration index, and aggregation risk.
Bias propagation & feedback loop	• Are bias alerts or anomaly triggers documented in logs/examples? • Are self-reflection or corrective mechanisms present? • Do benchmark results compare subgroup impacts or only aggregate success? • Is there any mechanism for bias escalation tracking?
Privacy, consent & user autonomy	• Are user-consent flows explicit for agent-mediated data sharing? • Are retention, minimization, and opt-out controls specified? • Do client-side policies constrain downstream reuse of user data? • Is differential privacy or scoped access documented?
Governance, redress & contestability	• Is there a process to challenge or override harmful agent decisions? • Are accountable maintainers, DAO rules, or arbitration channels named? • Do governance docs include fairness review or incident-response steps? • Is a redress or appeal channel documented?

this is acknowledged as a limitation that favors practically relevant projects while potentially under-representing obscure or newly created ones. Two independent raters (the first author and a research assistant blind to each other's scores) applied the standardized rubric; inter-rater agreement reached 92% on binary findings (Cohen's $\kappa = 0.87$) after dual review of a 20% random subsample, with all disagreements resolved through documented consensus discussion.

These results confirm that current MCP-centric ecosystems overwhelmingly omit explicit bias-mitigation, fairness, and accountability primitives at the protocol and orchestration layers. The pervasive documentation gaps, evident even in relatively mature repositories, highlight a structural vulnerability: most MCP servers are infrastructure or utility projects that defer ethical and security considerations to the consuming application layer. Such omissions create fertile ground for protocol-induced bias amplification when these servers are composed into full agentic systems.

These pervasive documentation gaps are not incidental; they reflect the current maturity stage of MCP-centric infrastructure, which prioritizes functional interoperability over ethical safeguards. The resulting blind spots at the orchestration layer constitute a structural vulnerability that will propagate into higher-level agentic applications unless addressed at the protocol level.

4.1 Detailed Findings and Implications

The 100-repository scan reveals consistent patterns across categories. In the General Assessment, 94% of repositories lacked clear documentation of privilege boundaries, token flows, or tamper-resistant audit and telemetry systems. A2A Review findings were even more pronounced (97% "Unclear"/"No"), with almost no repositories addressing preferential routing, negotiation telemetry, fairness metrics, or cryptographic identity checks. A2C

Table 2: Aggregated bias-audit findings across 100 MCP server repositories.

Category	Key Audit Criteria	Prevalence (Unclear or No)	Collective Insight
General Assessment	Privilege boundaries, token flows, access controls, identity validation, audit/telemetry	94%	Near-universal absence of comprehensive security and accountability documentation.
A2A Review	Preferential routing/negotiation, hardcoded peers, negotiation telemetry, fairness/diversity metrics, Ed25519/ZTAA enforcement	97%	Overwhelming lack of mechanisms to prevent preferential treatment, collusion, or hidden affiliations.
A2C Review	Fairness constraints in personalization, differential privacy / data minimization, consent-model scoping	89%	User-facing interaction layers consistently omit explicit fairness, privacy-preserving, or limited-scope safeguards.
Overall Risk	—	Low (68%), Medium (27%), High/Unclear (5%)	Foundational infrastructure exhibits systemic blind spots; higher-level agentic applications built atop these layers inherit significant orchestration bias risks.

Review showed 89% gaps in personalization fairness, differential privacy, data minimization, and consent scoping. Many repositories are foundational (proxies, SDKs, libraries, curated lists, scripts, or early-stage projects), which helps explain the predominance of Low/Medium risk ratings; however, higher-level agentic applications built on these layers inherit the blind spots. The few Medium/High-risk cases typically involved command execution, vector databases, or workflow automation, where missing safeguards are more consequential. These empirical patterns strongly support the need for protocol-aware, verifiable controls at the orchestration layer.

4.2 Connection to Existing MAS Evaluation Frameworks

The proposed ethical auditing checklist operates at the *infrastructure and documentation layer* and is therefore complementary to dynamic, outcome-oriented benchmarks such as MASEval, MAFBench, and MALIBU. While MASEval and MAFBench quantify runtime effects of orchestration choices (e.g., performance, latency, and coordination efficiency), and MALIBU measures implicit bias propagation through persona effects in multi-agent LLM interactions, the checklist identifies *upstream* protocol and registry gaps that these frameworks presuppose but do not audit.

This relationship enables a natural two-phase evaluation methodology for future work. Phase 1 (static) applies the checklist to surface documentation and configuration deficiencies in public repositories or protocols (as demonstrated here). Phase 2 (dynamic) feeds the flagged systems into MASEval, MAFBench, or MALIBU workflows to measure whether the identified gaps correspond to measurable outcome disparities at runtime. For example, fairness-aware constraints embedded in NANDA adaptive resolvers (demographic parity bounds or max-min fairness) could serve as an explicit control variable in MALIBU-style experiments: routing decisions that enforce parity on protected-attribute proxies could be tested for their ability to dampen persona-driven bias amplification. Such integration would provide reproducible triggers and verifiable fixes, directly supporting FMAI's emphasis on empirical grounding of ethical mechanisms.

4.3 Query-Wise Statistics

The detailed checklist analysis exposes a consistent transparency deficit across the full sample.

Privilege boundaries and token flows: 94% were "Unclear" or "No", indicating ambiguity in authorization boundaries and escalation paths.

Access-control consistency with architecture: 92% did not explain how implemented controls align with the stated architecture.

Identity validation mechanisms: 93% lacked clear evidence of ZTAA-, key-, or signature-based identity checks.

Audit and telemetry systems: 95% did not document comprehensive, tamper-resistant telemetry for privileged operations.

Preferential routing or negotiation safeguards: 96%

showed no evidence of controls against biased routing or negotiation outcomes.

Hardcoded peer or trusted-partner logic: 94% suggested named peers or whitelists that can function as de facto gatekeeping.

Fairness or diversity metrics in the communication layer: 98% lacked such metrics entirely.

Ed25519/ZTAA checks for hidden affiliations: 95% did not document cryptographic enforcement against Sybil attacks or affiliation masking.

Fairness constraints in personalization or recommendation modules: 91% lacked explicit fairness constraints.

Differential privacy or data minimization: 89% did not document these protections.

Consent scoping and logging: 90% lacked clear, auditable limited-scope consent models.

Doc-vs-code integrity and verifiable credentials: 93% showed missing or unverifiable alignment between documented claims, token handling, audit logging, and credential use.

Across the checklist, 89–98% of critical items received "Unclear" or "No" findings, confirming systemic orchestration-layer blind spots in larger agentic deployments. The rubric, aggregated dataset, and replication instructions are available from the corresponding author for independent verification and extension.

4.4 Limitations of the current study

A limitation of the present study is that checklist items were evaluated uniformly across repositories despite differences in intended functionality. Future versions of the framework should incorporate applicability weighting, allowing criteria to be conditionally evaluated based on repository risk class, deployment context, and operational scope.

5 Mitigation Pathways in Project NANDA

NANDA aims to meet these infrastructural requirements through a framework for secure, verifiable, and interoperable AI agent ecosystems. The decentralized tools leveraged include the following.

Verifiable Credentials (VCs): NANDA treats AgentFacts as verifiable claims requiring W3C-style cryptographic attestation Raskar et al. (2025). VCs are signed by issuers (enterprises, consortiums, or federated certification authorities) and support real-time revocation, reducing spoofing, impersonation, and Sybil risk Wang et al. (2025); Singh et al. (2025). Under the above threat model, VCs provide cryptographic unforgeability and revocation lists ensure liveness of exclusion.

Adaptive resolvers: The dynamic resolution layer supports programmable routing logic for real-time optimal endpoint selection, enabling geo-aware dispatching, load balancing, and threat mitigation Raskar et al. (2025); Zinky et al. (2025).

Reputation systems: AgentFacts can be enriched with verifiable trust and safety metadata, enabling risk-aware discovery and policy-based filtering of unsafe agents Wang et al. (2025). Reputation scores are updated via attested, stake-weighted consensus with slashing for detected collusion or Sybil clusters, drawing on mechanisms such as TraceRank-style endorsements Shi and Joo (2025).

Decentralized autonomous organization and collective governance: DAO-inspired governance, data cooperatives, and data unions can provide oversight and stewardship mechanisms for decentralized data governance Singh et al. (2024).

For adaptive resolvers we propose embedding max-min fairness or demographic parity bounds as programmable constraints. Protected attributes or proxies (e.g., geography, latency class) are measured via attested, privacy-preserving metadata fields in AgentFacts. Trade-offs with QoS are handled through constrained multi-objective routing (latency subject to parity bounds) or dynamic weighting; preliminary design uses ZK proofs inspired by LinkDID to verify fairness without exposing raw attributes Song (2023).

While NANDA's mechanisms substantially reduce orchestration-layer risks, they introduce new attack surfaces that require explicit acknowledgment. Fairness constraints in adaptive resolvers (e.g., demographic parity bounds) could be gamed by agents that misrepresent protected-attribute proxies in AgentFacts, or weaponized as a denial-of-service vector through deliberate violation of parity thresholds. DAO-inspired governance may similarly be subject to capture by dominant stakeholders, potentially leading to contested or biased definitions of fairness. These risks are mitigated in part through verifiable credentials, stake-weighted slashing, and privacy-preserving verification (e.g., ZK proofs), yet they highlight the necessity of continuous community stress-testing and adversarial red-teaming of the full NANDA stack.

6 Open Problems and Research Agenda

Despite the architectural solutions, several challenges remain open problems that require community-driven development, aligning with NANDA's decentralized ethos. Future community-driven audits using this checklist on public registries and protocol specifications will provide quantitative baselines for the open challenges identified.

Causality tracing and accountability: decentralization minimizes dependence on centralized institutions but introduces risk of weak traceability. A research challenge is enabling anonymity and confidentiality while preserving auditability Singh et al. (2024); Wang et al. (2025).

Sustainability and economic incentives: registries and metadata services require viable incentives; open questions remain about micropayments, staking, or usage tiers without rent-seeking Raskar et al. (2025); Wang et al. (2025).

Fairness enforcement and bias mitigation: commercial incentives can conflict with nondiscrimination requirements, making continuous enforcement necessary Wang et al. (2025). Related challenges in collusion-resistant reputation systems and privacy-preserving audits have been explored in decentralized identity and service-discovery mechanisms Shi and Joo (2025); Song (2023). Privacy-preserving fairness verification can leverage selective-disclosure techniques such as those in LinkDID Song (2023).

Decentralization ethos: NANDA's "quilt of registries" is intended to prevent centralized governance and vendor lock-in, but self-organizing communities may still embed preferential choices that become discriminatory Singh et al. (2024, 2025).

7 Conclusion

The agentic web reshapes the Internet into a decentralized, autonomous, and rapidly evolving ecosystem. Project NANDA offers architectural mechanisms to support secure discovery, verifiable identity, and cross-protocol interoperability. Yet systemic ethical risks namely, bias amplification, power asymmetries, and covert manipulation that could persist at scale. Addressing these risks requires protocol-aware auditing, verifiable governance mechanisms, and sustained community stewardship. The proposed auditing framework and mitigation pathways position NANDA as a promising foundation for ethical, decentralized agentic systems while acknowledging the need for continued research and governance innovation.

The empirical audit of 100 MCP repositories indicates that the ethical risks identified in the threat model are not merely hypothetical; they are already latent in widely used components of the emerging agentic web. At the same time, the evidence should be interpreted at its intended scope: this is a documentation- and configuration-level audit rather than a runtime evaluation. Even with that caveat, the observed blind spots in fairness, accountability, and security primitives underscore the timeliness of infrastructure-level solutions such as Project NANDA's AgentFacts, fairness-aware adaptive resolvers, reputation systems, and DAO-inspired governance. Future work should focus on community-driven replication of the checklist, ZK-powered privacy-preserving audits, and controlled experiments quantifying fairness–QoS trade-offs on public multi-agent benchmarks.

References

T. Chaffer, J. Goldston, and G. A. (2024). Incentivized Symbiosis: A Paradigm for Human-Agent Coevolution. *arXiv preprint* arXiv:2412.06855.

W. Chen, Z. You, R. Li, Y. Guan, C. Qian, C. Zhao, C. Yang, R. Xie, Z. Liu, and M. Sun (2024). Internet of Agents: Weaving a Web of Heterogeneous Agents for Collaborative Intelligence. *arXiv preprint* arXiv:2407.07061.

L. Chen, Y. Zhang, J. Feng, H. Chai, H. Zhang, B. Fan, Y. Ma, S. Zhang, N. Li, T. Liu, N. Sukiennik, K. Zhao, Y. Li, Z. Liu, F. Xu, and Y. Li (2025). AI Agent Behavioral Science. *arXiv preprint* arXiv:2506.06366.

E. Ferrara (2024). The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications* 15, 100525.

M. Goyal and P. Bhasin (2025). Moving from monolithic to microservices architecture for multi-agent systems. *World Journal of Advanced Engineering Technology and Sciences* 15(1), 2119–2124.

I. Habler, K. Huang, V. Narajala, and P. Kulkarni (2025). Building a Secure Agentic AI Application Leveraging A2A Protocol. *arXiv preprint* arXiv:2504.16902.

K. Huang, V. Narajala, I. Habler, and A. Sheriff (2025). Agent Name Service (ANS): A Universal Directory for Secure AI Agent Discovery and Interoperability. *arXiv preprint* arXiv:2505.10609.

L. Hughes, Y. K. Dwivedi, T. Malik, M. Shawosh, M. A. Albashrawi, I. Jeon, V. Dutot, M. Appanderanda, T. Crick, R. De', M. Fenwick, S. M. Gunaratnege, P. Jurcys, A. K. Kar, N. Kshetri, K. Li, S. Mutasa, S. Samothrakis, M. Wade, and P. Walton (2025). AI Agents and Agentic Systems: A Multi-Expert Analysis. *Journal of Computer Information Systems* 65(4), 489–517.

Y.-T. Lin, T.-W. Hung, and L. T.-L. Huang (2020). Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy & Technology* 34(S1), 65–90.

H. Liu, J. Dacon, H. Liu, Z. Liu, W. Fan, and J. Tang (2020). Does Gender Matter? Towards Fairness in Dialogue Systems. In *Proceedings of COLING 2020*.

M. Madaio, E. Mayfield, S. L. Blodgett, and E. Dixon-Román (2022). Beyond "fairness". In *Beyond "Fairness"*. Routledge.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229.

C. Ma, E. Zhang, Y. Zhao, W. Liu, Y. Jia, P. Qing, L. Shi, A. Cohan, Y. Yan, and S. Vosoughi (2025). Judging with Many Minds: Do More Perspectives Mean Less Prejudice? On Bias Amplifications and Resistance in Multi-Agent Based LLM-as-Judge. *arXiv preprint* arXiv:2505.19477.

I. Mirza, C. Huang, I. Vasista, R. Patil, A. Akalin, S. O'Brien, and K. Zhu (2025). MALIBU Benchmark: Multi-Agent LLM Implicit Bias Uncovered. *arXiv preprint* arXiv:2507.01019.

T.-N. Nguyen, L. Luo, T.-T. Vu, and D. Phung (2025). The Social Cost of Intelligence: Emergence, Propagation, and Amplification of Stereotypical Bias in Multi-Agent Systems. *arXiv preprint* arXiv:2510.10943.

T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, and K. Crawford (2021). Datasheets for Datasets. *Communications of the ACM* 64, 12, 86–92.

S. Phelps and R. Ranson (2023). Of Models and Tin Men: A Behavioural Economics Study of Principal-Agent Problems in AI Alignment Using Large-Language Models. *arXiv preprint* arXiv:2307.11137.

F. Piccialli, V. Mele, D. Cerciello, S. Sarwar, D. Chiaro, and P. Qi (2025). AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0. *Expert Systems with Applications* 291, 128404.

K. Li, J. Gao, and D. Wang (2026). Aligned Agents, Biased Swarm: Measuring Bias Amplification in Multi-Agent Systems. *arXiv preprint* arXiv:2604.08963.

R. Raskar, P. Chari, J. Zinky, M. Lambe, J. Grogan, S. Wang, R. Ranjan, R. Singhal, S. Gupta, R. Lincourt, R. Bala, A. Joshi, A. Singh, A. Chopra, D. Stripelis, B. B., S. Kumar, and M. Gorskikh (2025). Beyond DNS: Unlocking the Internet of AI Agents via the NANDA Index and Verified AgentFacts. *arXiv preprint* arXiv:2507.14263.

M. Raeis and S. J. Golestani (2021). Distributed Fair Scheduling for Information Exchange in Multi-Agent Systems. In *Proceedings of the 31st International Conference on Automated Planning and Scheduling (ICAPS 2021)*. AAAI Press.

A. Singh, A. Ehtesham, R. Raskar, M. Lambe, P. Chari, J. Grogan, A. Singh, and S. Kumar (2025). A Survey of AI Agent Registry Solutions. *arXiv preprint* arXiv:2508.03095.

A. Singh et al. (2024). A perspective on decentralizing AI. *MIT Media Lab* Technical Report.

D. Shi and K. Joo (2025). Sybil-Resistant Service Discovery for Agent Economies. *arXiv preprint* arXiv:2510.27554.

R. Song (2023). LinkDID: A Privacy-Preserving, Sybil-Resistant and Key-Recoverable Decentralized Identity Scheme. *arXiv preprint* arXiv:2307.14679.

A. Vaziry, S. Garzon, and A. Küpper (2025). Towards Multi-Agent Economies: Enhancing the A2A Protocol with Ledger-Anchored Identities and x402 Micropayments for AI Agents. *arXiv preprint* arXiv:2507.19550.

S. Wang, R. Raskar, M. Lambe, P. Chari, R. Singhal, S. Gupta, R. Ranjan, and K. Huang (2025). Using the NANDA Index Architecture in Practice: An Enterprise Perspective. *arXiv preprint* arXiv:2508.03101.

Y. Wang, S. Guo, Y. Pan, Z. Su, F. Chen, T. Luan, P. Li, J. Kang, and D. Niyato (2025). Internet of Agents: Fundamentals, Applications, and Challenges. *arXiv preprint* arXiv:2505.07176.

K. Yang and C. Zhai (2025). Ten Principles of AI Agent Economics. *arXiv preprint* arXiv:2505.20273.

Y. Yang, Y. Wen, J. Wang, and W. Zhang (2025). Agent Exchange: Shaping the Future of AI Agent Economics. *arXiv preprint* arXiv:2507.03904.

J. Zinky, H. Seshadri, M. Lambe, P. Chari, and R. Raskar (2025). NANDA Adaptive Resolver: Architecture for Dynamic Resolution of AI Agent Names. *arXiv preprint* arXiv:2508.03113.