Fair Agents: Balancing Multistakeholder Alignment in Multi-Agent Personalization Systems

Andrea Forster[1,*], Peter Müllner[1], Denis Helic[2], Elisabeth Lex[2] and Dominik Kowald[1,3,*]

[1]Fair-AI, Know Center Research GmbH, Graz, Austria
[2]Institute of Human-Centred Computing, Graz University of Technology, Graz, Austria
[3]Department of Digital Humanities, University of Graz, Graz, Austria

Abstract

LLM agents are increasingly used for personalization due to their ability to communicate directly with users in natural language, integrate external knowledge bases, and negotiate with other (possibly human) agents. Especially in multistakeholder AI systems with multiple distinct objectives, LLM agents are used to independently optimize for each stakeholder's goals. Here, stakeholder alignment is essential to identify and map these goals to provide LLM agents with quantifiable objectives. Plus, the way in which the outputs of the LLM agents are aggregated is fundamental to ensuring fair outcomes for all agents and, therefore, stakeholders. In this work, we identify open research challenges and propose a conceptual framework for designing fair multi-agent multistakeholder personalization systems that balance competing stakeholder objectives. Our framework integrates (i) methods to align stakeholder objectives and LLM agents, (ii) aggregation strategies, e.g., based on social choice theory, to form fair collective decisions, and (iii) stakeholder-centric evaluation procedures for both individual and collective agent behavior. We showcase our framework through a tourism use case and discuss possible applications in other domains, such as education and healthcare. Finally, we discuss domain-specific fairness tensions and review datasets for evaluating multistakeholder fairness and multi-agent personalization systems.

Keywords

Multi-Agent AI, Personalization, Fairness, Societal Aspects, Multistakeholder Recommender Systems

1. Introduction

Large language model (LLM) agents are reshaping personalized recommender systems [1]. Traditional systems infer preferences from behavioral patterns and rank pre-defined item sets. LLM agents, in contrast, capture user intent through natural language, integrate multimodal inputs, leverage external tools, and iteratively refine objectives through conversation [1]. This enables complex tasks beyond simple top-n recommendations [2, 1].

Recommender systems are multistakeholder platforms that must balance competing objectives across multiple stakeholders: consumers seeking relevant recommendations, providers seeking visibility, regulators ensuring compliance, and affected communities concerned with broader societal impacts [3, 4, 5]. Each stakeholder has distinct fairness concerns and success criteria [4, 6]. Rather than optimizing for all objectives simultaneously, multi-agent systems can dedicate separate agents to advocate for different stakeholders [7, 8]. This offers distinct advantages for stakeholder alignment but introduces critical challenges [7, 9]. Collective decision-making between LLM agents remains opaque and susceptible to failure modes such as confirmation bias, hallucinations, positional bias toward agents who "speak" first, and dominance by more articulate agents [10, 11, 12, 9]. Decisions are often distributed, emergent, and context-dependent. This can obscure accountability, cause concerns about transparency and fairness, amplify biases, and create privacy risks [8].

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden
*Corresponding author.
aforster@know-center.at (A. Forster); pmuellner@know-center.at (P. Müllner); dhelic@tugraz.at (D. Helic); elisabeth.lex@tugraz.at (E. Lex); dkowald@know-center.at (D. Kowald)
0009-0008-6818-1916 (A. Forster); 0000-0001-6581-1945 (P. Müllner); 0000-0003-0725-7450 (D. Helic); 0000-0001-5293-2967 (E. Lex); 0000-0003-3230-6234 (D. Kowald)

In this work, we focus on the central challenge of fair and transparent preference aggregation from multiple stakeholder-representing agents. We narrow this gap by adapting social choice theory for collective decision-making to LLM-based multi-agent systems. Social choice methods [13, 14] have proven theoretical properties to guarantee certain fairness criteria mathematically.

This paper makes three contributions toward the design of fair multi-agent multistakeholder personalization systems:

1. *Related Work and Research Challenges.* In Section 2, we synthesize prior work on multistakeholder fairness and alignment in personalized recommender systems and multi-agent AI systems, preference-aggregation strategies from social choice theory, and stakeholder-centric evaluation methodologies. From this synthesis, we identify and categorize open research challenges specific to multistakeholder personalization with LLM agents.
2. *Conceptual Framework.* In Section 3, we propose a conceptual framework for fair multi-agent and multistakeholder personalization. The framework combines LLM-based stakeholder alignment with social choice mechanisms (e.g., voting rules) for transparent preference aggregation. This bridges semantic flexibility with theoretical rigor, enabling context-dependent fairness instantiation [15, 5] while maintaining traceability. Stakeholder-centric evaluation spans individual agent reliability, system-level fairness and accuracy, and stakeholder perception through qualitative studies.
3. *Domains and Datasets.* In Section 4, we analyze potential application domains (e.g., tourism, education, healthcare) and identify stakeholder tensions and fairness concerns. For each domain, we align stakeholders to our framework's generic LLM agents. Finally, we outline available datasets suitable for the framework's operationalization and empirical validation.

2. Related Work and Research Challenges

In this section, we outline current approaches to multistakeholder fairness in multi-agent personalization systems and identify key research challenges that motivate our conceptual framework.

2.1. Multistakeholder Fairness and Biases in Multi-Agent AI Systems

A fundamental challenge in aligning AI systems with human values is determining whose conceptions of fairness are represented. Fairness in machine learning and AI tends toward narrow, mathematically tractable formulations [6, 16, 17]. These fairness formulations are typically rooted in Western ethical theories, aiming to maximize aggregate outcomes, ensure equal opportunity, infer user character, or follow fixed rules [18]. Scholars increasingly advocate pluralistic approaches that incorporate non-Western epistemologies, prioritizing collective well-being over individual agency [19, 18].

Multistakeholder fairness in recommender systems involves consumers, providers, platform-mediated actors (developers, third parties), and upstream/downstream parties [4]. Deldjoo [20] investigates the trade-offs between accuracy and provider-side fairness through prompt design strategies in LLM-based personalization. Fairness-oriented prompts recommend newer items with broader genre distributions, but significantly reduce accuracy. Notably, embedding fairness into system roles (e.g., "act as a fair recommender") proves more effective than fairness directives within prompts. These findings underscore the difficulty of reconciling competing objectives within a single agent [21, 7, 8].

LLM agents in multi-agent AI systems can act continuously and flexibly in response to context, as opposed to static agents, where agent behavior is defined through explicit logic [22, 9]. This can introduce additional complexities. Recent work reveals that LLM agents exhibit conformity and social influence effects that amplify biases [10]. Competitive incentives increase hallucination and behavioral misalignment [23]. Interactional fairness (i.e., interpersonal treatment and the honesty and adequacy of explanations) varies depending on the model and whether agents are prompted collaboratively or competitively [24]. Persona-induced biases and in-group favoritism are present across current model generations [25]. Fairness in multi-agent AI systems emerges from how agents interact, whose values

are prioritized, and how conflicts are resolved. This calls for systematic approaches to stakeholder alignment and operationalization in multi-agent AI systems.

2.2. Stakeholder Alignment and Integration into LLM Agents

Translating abstract stakeholder values into concrete agent objectives remains a central challenge in multi-agent personalization [26]. From a technical perspective, personalization and alignment can be implemented at multiple levels: input (profile-augmented prompting), model (fine-tuning), and objective (aligning to user preferences via Reinforcement Learning from Human Feedback (RLHF) or inference-time methods such as weight merging and model ensembling) [7]. However, RLHF tends to treat disagreements among evaluators as noise, limiting its ability to reflect diverse human values. To address this, Maura et al. [27] propose maximal lotteries, a probabilistic social choice rule that better operationalizes human intentions and respects majority preferences.

Recent approaches attempted to encode stakeholder values into agent behavior. Banerjee et al. [28] instantiate multiple LLM agents in a tourism recommender by encoding each agent's role, objectives, and ranking rules into system prompts. Uchoa et al. [26] propose a privacy-preserving framework. Agents (e.g., students, parents, institutions) generate recommendations based on local policies and use a dedicated negotiation agent to build consensus. Yet, these methods assume that stakeholder values can be cleanly articulated as static policies. The process of eliciting, validating, and refining stakeholder values remains underexplored. Ekstrand et al. [5] advocate for participatory and community co-design, highlighting that system design is often controlled by developer teams whose values may not align with those of other affected stakeholders.

2.3. Multistakeholder Preference Aggregation

In multi-agent AI systems, agents typically coordinate through natural language, with coordination patterns emerging implicitly from large-scale learning [9, 22]. In many cases, a central orchestrator manages several specialized agents, or agents coordinate through negotiation [22]. While these approaches offer flexibility, they lack verifiable reliability and exacerbate biases and failure modes [22, 11, 12, 10]. Iterative negotiation approaches have shown promise in balancing competing objectives but face challenges, including resource intensity and convergence on incorrect solutions [12, 11]. Banerjee et al.[28] integrate iterative LLM negotiation with voting-based mediation, finding that the choice of aggregation strategy critically shapes hallucination rates and stability.

To address the limitations of natural language negotiation, we turn to *social choice theory* [13, 14]. Social choice theory provides theoretically and mathematically grounded methods for aggregating individual preferences into transparent and equitable collective decisions [9]. It is well-established in multi-agent consensus building and resource allocation [9], and in aggregating preferences for group recommendations [29]. Aird et al. [15] operationalize social choice for provider-side fairness in recommendation tasks. In their framework, multiple fairness concerns are assigned to respective agents and dynamically integrated via allocation and aggregation mechanisms. Uchoa et al. [30] compare LLM-based mediation with social choice aggregation, showing that mediation can favor marginalized groups, while aggregation tends to optimize overall utility. However, their comparison is limited to a single aggregation mechanism, despite evidence that different voting rules yield vastly different outcomes [15]. This underscores the need for systematic exploration of how social choice mechanisms can complement or replace LLM-based mediation to achieve both fairness and explainability.

2.4. Stakeholder-Centric Evaluation

Evaluation in multi-agent systems must address individual agents and emerging behaviors across agents. Fairness in AI has traditionally been evaluated in centralized, single-agent settings via offline evaluation [6]. However, such approaches struggle to capture fairness as contextual and dynamic [6, 31, 15]. They fail to account for emergent interaction dynamics such as cascading errors, or conflicting objectives that arise in multi-agent settings [9].

Grounding mechanisms can be used to evaluate the reliability of individual agents. This can be done by restricting recommendations to verified knowledge bases [28], user validation studies [32], or similarity-based item matching [20]. Relevance and diversity metrics can be used to evaluate the performance of aggregate recommendations [28, 20]. Uchoa et al. [26] introduce dedicated coordination agents that resolve conflicts, and audit agents that monitor for bias and policy drift. Aird et al. [15] apply the $L_{1/2}$ norm as a metric to balance per-agent fairness with low disparity across agents. Yet, how to effectively integrate agent- and system-level evaluations remains an open challenge.

2.5. Summary and Research Challenges

Despite growing interest in multi-agent multistakeholder personalization, existing frameworks do not jointly address the full scope from stakeholder alignment and operationalization to fair aggregation and evaluation in a principled, domain-agnostic manner. On the aggregation side, Uchoa et al. [30] offer a first empirical comparison of LLM-based mediation and social choice aggregation. Nonetheless, a broader and more systematic evaluation of social choice voting mechanisms across domains and stakeholder configurations is still lacking. The following, overarching research challenges express these goals and guide our conceptual framework:

- *RC1: Aligning abstract stakeholder values with concrete technical objectives of (LLM) agents*
- *RC2: Fair aggregation of diverse fairness concerns and consensus building*
- *RC3: Stakeholder-centric evaluation of multi-agent personalization systems*

3. Conceptual Framework

Next, we represent our conceptual framework for multi-agent multistakeholder personalization systems. Figure 1 illustrates the framework using a tourism use case as an example, showcasing the previously identified research challenges. Additional practical considerations are provided in Section 3.1. Overall, our framework is configurable and domain-agnostic, allowing researchers, practitioners, and communities to instantiate the fairness concerns and aggregation strategies most relevant to their specific context and stakeholder needs. Now, we investigate each research challenge in detail.

RC1: Aligning abstract stakeholder values with concrete technical objectives of (LLM) agents. In Figure 1, a user enters queries in natural language, while other relevant stakeholders are identified in advance. Our framework focuses on three core stakeholders: *users*, *providers*, and *third parties*. Each stakeholder is represented by at least one LLM agent, with the flexibility to subdivide stakeholders (e.g., separating third parties into multiple agents) or include additional ones as needed. We exemplify this by mapping stakeholders to the three axes of sustainability (social, economic, environmental) [33]. Stakeholder values are then turned into measurable agent objectives. For example, Forster et al. [34] frame overtourism mitigation and fairness for niche users as a popularity bias problem, and Aird et al. [15] operationalize different provider- and consumer-side fairness concerns into agents, each with its own fairness metrics. The recommendations are then ranked based on each agent's fairness state and relevance.

RC2: Fair aggregation of diverse fairness concerns and consensus building. Each agent generates a candidate list based on its objectives and policies, optionally including text-based justifications. We leverage social choice theory to aggregate the agents' candidate lists. Different voting mechanisms are compared, and each stakeholder's influence on the final ranking is quantified. Building on prior work, we employ voting rules including Borda (ordinal scoring), Copeland and Ranked Pairs (pairwise majority), and Kemeny variants (distance minimization) [15, 30, 35]. Furthermore, social choice raises important design questions: whether voting occurs simultaneously or sequentially, and whether votes are visible or blind to other agents. These choices can affect strategic behavior [9]. Therefore, our framework keeps an agent's underlying policies and preference model invisible to other agents. While

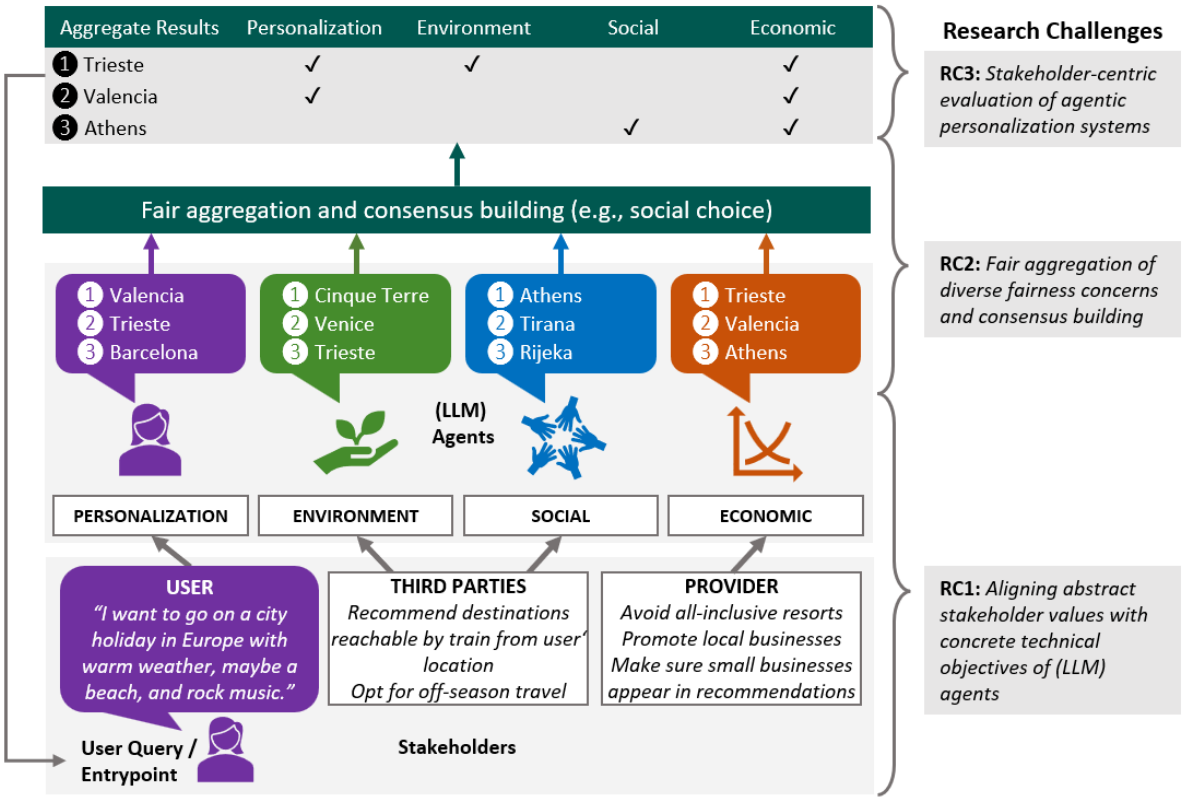

Figure 1: Our conceptual framework for multi-agent multistakeholder personalization in tourism. First, the user enters a query. Other stakeholder values are elicited beforehand. Next, agents are aligned with stakeholders (*RC1*) and each agent generates a candidate set, including justifications for their decision, if applicable. Candidate lists are fed into an aggregation mechanism, and consensus is built, e.g., through different social choice voting mechanisms (*RC2*). Finally, aggregate results and justification of whether different stakeholder objectives are met are returned to the user (*RC3*).

social choice methods provide formal guarantees and mathematical traceability, we plan to explore LLM-based mediation as a complementary approach, which may better capture nuanced semantic relationships between agent justifications [30].

RC3: Stakeholder-centric evaluation of multi-agent personalization systems. Evaluation can be performed for individual agents and across multiple agents. At the individual level, we verify agent reliability through grounding mechanisms and monitor for biases, penalizing unreliable agents [28, 26]. At the multi-agent level, we assess computational cost and latency alongside recommendation quality metrics, and long-term policy drifts not traceable in single interactions [26, 16]. The agents' top-n recommendations can be evaluated collectively and individually to assess contribution and fairness-accuracy trade-offs. Quality metrics include accuracy metrics (e.g., nDCG, recall [20]) and diversity metrics to account for item fairness (e.g., Gini, normalized entropy [28, 20]). Different fairness objectives can be defined for each agent [15]. On the user side, popularity bias can be assessed through PopLift [36]. For the proportional representation of protected groups, Kullback-Leibler or Jensen-Shannon divergence can be used to compare historical and recommended distributions [37]. Furthermore, fairness regret is the difference between an agent's perfect fairness (defined by their objectives [15]) and the fairness of a given choice. Beyond offline evaluation, real stakeholder perceptions of conversation quality, recommended content, and behavioral outcomes (satisfaction, willingness to act on recommendations) can be evaluated through user studies and focus groups.

3.1. Practical Considerations

This section explains further details that we omitted from Figure 1, for the sake of simplicity. First, users enter queries in natural language. The personalization agent can refine these through multi-turn dialog. When available, the system incorporates user interaction histories, including past reviews and semantic information about consumed items. The other stakeholder values are identified in advance, either from existing research or through surveys, interviews, or co-design settings. Their values are elicited through personas in natural language [28], opinion surveys [30], or hybrid methods. Based on this, each LLM agent operates with a task-specific system prompt encoding its stakeholder objective. Agents have access to retrieval tools (e.g., vector database search over item embeddings and web search for real-time information such as opening hours or pricing) and a candidate generation module. However, not every agent needs to be LLM-based. In domains where strong conventional recommender systems and rich user profiles already exist, a traditional recommendation model may replace the agent. An agent can create ranked candidate lists without natural language reasoning. A provider-side agent that ensures fair item exposure could leverage rule-based filtering based on statistical parity. Managing the computational cost and latency of running multiple (LLM-based) stakeholder agents remains challenging [28]. To balance fairness coverage with computational efficiency, stakeholder agents can be invoked either statically (pre-defined for all queries) or dynamically (activated based on query relevance). Agents are invoked when their compatibility with the current query is high or when the system's fairness performance for that agent's objective falls below a threshold [15]. Conversely, if an agent's objectives have been consistently met in recent queries, it may be skipped. This requires maintaining statistics of per-agent fairness metrics across queries.

4. Possible Application Domains, Use-Cases, and Available Datasets

This section outlines key domains, potential datasets, and use cases to implement our conceptual framework for multi-agent multistakeholder personalization systems. Table 1 identifies stakeholders and their fairness concerns based on existing literature in multistakeholder fairness. Table 2 provides an overview of available datasets used in multistakeholder systems.

4.1. Domains and Use Cases

Tourism. Destination, point-of-interest, and trip itinerary recommendations are inherently a multi-stakeholder problem. Users want to receive personalized recommendations, providers such as hotels and restaurant owners want their businesses to be recommended, local communities and NGOs want to avoid overtourism and environmental degradation, and booking platforms want to receive commissions [38]. Examples from tourism literature show that stakeholder-specific fairness concerns can be complex and context-based [31]. A trip itinerary user study [32] with a single-agent, user-focused approach shows that LLM agents can successfully incorporate user histories and generate relevant recommendations. Users, however, report issues with originality (bias toward popular places), time optimization (too few recommendations for the available time), and lack of real-time information. Banerjee et al. [28] encode popularity, sustainability, and personalization in separate LLM agents to produce balanced destination recommendation lists. In Section 3, we describe the operationalization of our conceptual framework in tourism in further detail.

Music. There are different stakeholder needs in streaming services and music licensing. Copyright policies often treat lyricists, composers, and publishers as a single group, despite their differing needs. Automated licensing systems tend to favor large publishers who can easily navigate them, while small rights holders are left at a disadvantage [39]. For instance, users want to receive relevant song recommendation lists. These can be measured by the ratio of saved or repeated songs. Musicians want fair exposure on the platforms to gain new fans and keep existing listeners. Platforms want to mediate both and generate revenues [40]. Our framework addresses recommendation fairness and

Table 1
Overview of possible application domains with diverse stakeholder settings and fairness concerns. Section 4 details how our multi-agent multistakeholder personalization framework addresses these concerns through stakeholder-specific agents and social choice aggregation.

Domain	Use Case	Stakeholders	Fairness Concerns
Tourism [28, 38, 16]	Destination, point-of-interest, and trip itinerary recommendation	Travelers, local businesses, platform developers, destination managers, communities, ecology	Ensuring personalization and provider visibility, maximizing commissions, mitigating overtourism and environmental degradation
Music [39, 40]	Balanced streaming recommendations, fair royalty remuneration for all rights holders	Artists, music labels, streaming services, listeners, concert venues	Balancing monetary rewards and market development, ensuring artist exposure, fostering customer loyalty, guaranteeing fair royalties
Education [26, 4, 16]	Course material recommendation	Students, parents, tutors, school management, authorities	Balancing content exposure for tutors, optimizing learning outcomes for students, ensuring curriculum compliance, respecting parental oversight
Digital Archives [41, 42]	Digitization and dissemination of cultural heritage information	Archivists, curators, librarians, platform developers, researchers, students, publishers	Making high-quality information easily accessible, increasing user satisfaction, promoting platform growth, disseminating accurate results, supporting funded curations
Job Market and Human Resources [17, 43]	Multi-objective job matching, fair algorithmic human resource management	Job seekers, recruiters, companies, public employment services, labor unions, human rights agencies	Supporting personal development and adequate remuneration, promoting economic development, reducing unemployment rates, filling open positions, enabling efficient decision-making
Healthcare [44, 45, 16]	Clinical decision support, consumer health, facilitation of stakeholder collaboration	Clinicians, researchers, legal and policy experts, patients, consumer advocates, AI scientists, industry leaders	Ensuring representativeness of training data, maintaining system transparency, disclosing biases, building user trust, empowering patients
Finance [18, 15]	Fair matching of loans with lenders	Lenders (platform end users), lending partners (NGOs that mediate international loans), borrowers (individuals and small groups advertising their	Advancing financial inclusion, aligning user interest in low-risk projects, maximizing benefits by investing in low-purchasing-power countries, ensuring fair exposure of borrowers

licensing equity by instantiating user agents optimizing for relevance, separate artist agents for lyricists, composers, and publishers, ensuring fair exposure and rights distribution, and platform agents balancing engagement with revenue. By encoding distinct preferences for each stakeholder and using social choice aggregation, our framework can counteract the homogenization that disadvantages smaller stakeholders.

Education. Value alignment in education entails diverse cultural and philosophical values, competing curricula, and time restrictions [26]. In an online educational resource recommendation task, our framework would align the personalization agent with the student's learning preferences; family, privacy, and institutional requirements (schools, authorities, exam schedules) would be separate third-party agents, and fair exposure to learning content from different tutors would be ensured by a provider agent.

Digital Archives. Different stakeholders in digital archives, such as archivists, curators, platform developers, researchers, students, and publishers, have different requirements. Scholarly stakeholders may not be interested in popular or serendipitous recommendations and may favor explicit user control over system-driven personalization [42, 41]. Our framework instantiates separate user agents aligned with scholarly versus discovery-oriented objectives, with additional agents representing curatorial policies (e.g., promoting underexposed collections) and platform goals.

Job Market and Human Resources. Job-seeking platforms must balance fairness across multiple stakeholders. For instance, it would be unfair if management positions were predominantly recommended to male candidates, or if job postings from large corporations received disproportionate visibility over start-ups [17]. Our framework would deploy user agents for job seekers and provider agents for companies and recruiters. Additional policies from third parties, such as public employment services and labor unions, are also instantiated in separate agents. Similarly, algorithmic human resource management systems for task assignment and promotion decisions involve competing stakeholder interests. In our framework, agents for system users (managers and businesses) seek streamlined decision-making and may prefer opacity to prevent gaming. Employee agents want fair exposure to advancement opportunities and transparency in decision-making. Third-party agents represent developer constraints (balancing transparency, accuracy, and maintainability) and regulator requirements (auditability, ethics compliance) [43]. Candidate list generation by each agent allows for explicit enforcement of fairness metrics. Social choice aggregation then combines these lists, allowing auditors to verify that anti-discrimination policies were actually enforced.

Healthcare. The implementation of AI in healthcare entails a complex set of stakeholders, often spanning multiple organizations and countries with different cultural values [45]. Patients' interests are frequently marginalized in decision-making processes, necessitating robust advocacy in policy formation and system design [44]. Rozenblit et al. [44] advocate for a multistakeholder consortium for effective AI governance in healthcare, including patients, clinicians, ethicists, researchers, and industry leaders, aiming to create patient-centered standards through voting-based consensus mechanisms. If implemented within a clinical decision support system, our framework would include healthcare professionals as users, while patients and advocacy groups are represented by third-party agents. In patient-facing applications, provider agents represent pharmaceutical companies, medical device manufacturers, and healthcare institutions, while third-party stakeholder agents represent ethicists. Unlike informal consortium voting, our social choice mechanisms provide formal guarantees about how stakeholder preferences combine, addressing the accountability gaps between governmental bodies and private corporations [44].

Finance. Peer-to-peer microlending platforms like Kiva.org involve multiple stakeholders with different objectives [18, 15]. Lenders seek low-risk or high-impact loans, while borrowers need equitable exposure. The platform must balance both while proving accountability to funding institutions [18]. Our framework represents lenders as users with configurable risk/impact preferences, borrowers through provider-side agents enforcing exposure equity (e.g., geographic or sector diversity), and policies of platform developers and intermediaries through third-party agents. Social choice aggregation allows the provider and third-party agents to ensure minimum exposure thresholds for underrepresented borrower categories.

Table 2
Overview of potential datasets for future experiments with descriptions and references to prior work. Columns indicate whether datasets are synthetic, contain user-item ratings, or include rich contextual information.

Dataset	Description	Domain	Synthetic	Ratings	Item Context
SynthTRIPs[1]	Repository with synthetic user queries for personalized travel planning and knowledge base of European cities	Tourism [28]	✓		✓
Microlending[2]	Lender-borrower interactions and categorical loan features	Finance [15]		✓	✓
Context Trails[4]	Dataset and code to enrich Foursquare user-item interaction datasets with additional contextual information (opening times, fine-grained venue information, weather data)	Tourism [50]		✓	✓
Educational Stakeholder Archetypes[5]	Stakeholder archetypes with preferences on a five-point scale generated from stakeholder persona-based natural language prompts	Education [30]	✓		
Nemotron Personas[6]	Synthetic dataset featuring personas from different countries, including their professional, sports, cultural personas	Various	✓		
Yelp[7]	User-item interactions for points-of-interest, ratings and semantic reviews, additional item metadata (e.g., business category, opening times)	Tourism [51]		✓	✓

[1] https://ashmibanerjee.github.io/synthTRIPS-website/
[2] https://scholar.colorado.edu/concern/datasets/j6731518r
[3] https://grouplens.org/datasets/movielens/
[4] https://zenodo.org/records/15855966
[5] Data available upon request from the authors [30]
[6] https://huggingface.co/datasets/nvidia/Nemotron-Personas-Singapore; Brazil, India, USA, France, and Japan also available
[7] https://business.yelp.com/data/resources/open-dataset/

4.2. Datasets

Table 2 provides an overview of relevant datasets for multistakeholder settings. Personalized alignment of LLMs requires high-quality datasets containing both implicit and explicit user-item interactions [7, 46]. Our framework particularly leverages rating-based datasets with contextual item information where LLMs can process semantic content (e.g., categorical attributes, natural language descriptions) that traditional rating-based models cannot effectively utilize. Such datasets typically contain user profiles and relationships (demographics, social networks), historical dialogues with LLM agents, user-generated content (reviews), interaction history (ratings, consumed items), and pre-defined user preferences [7]. Dataset selection often lacks rigorous justification. Vente et al. [47] address this gap by introducing a web tool that analyzes the similarity of 96 recommender system datasets to inform selection decisions, noting that most experiments rely on only four datasets without a clear rationale. The scarcity of high-quality, domain-specific datasets with sufficient depth and questions about training data leakage [48] have driven researchers toward synthetic data generation. For instance, Banerjee et al. [49] introduce a framework that produces diverse travel queries from persona-based preferences and sustainability filters, grounded in an external knowledge base. This approach provides a blueprint for combining real-world and synthetic datasets to address multistakeholder fairness in multi-agent AI systems across domains.

5. Conclusion and Future Work

In this paper, we identify three core research challenges for multi-agent multistakeholder personalization systems: aligning abstract stakeholder values with concrete objectives of LLM agents, aggregating diverse fairness concerns, and conducting stakeholder-centric evaluation. We propose a conceptual framework that addresses these challenges to enable multistakeholder and multi-agent fairness in personalization systems. Leveraging social choice theory for aggregation ensures transparent and accountable consensus-building across diverse deployment domains. We illustrate our framework with a tourism recommendation use case. Additionally, we provide examples of domain-specific use cases, stakeholders, fairness concerns, and datasets, and discuss the applicability of our framework for various domains.

In the future, we plan to validate our conceptual framework empirically. This includes assessing how stakeholder values can be effectively integrated and evaluated. Moreover, we aim to compare various social choice mechanisms with LLM-based mediation and explore hybrid aggregation approaches that combine both paradigms. We, for instance, envision mixed strategies where social choice mechanisms ensure formal fairness guarantees and mathematical traceability of stakeholder influence. At the same time, LLMs provide natural language justifications that explain which stakeholder concerns were prioritized and what trade-offs were made. This approach could address limitations of pure social choice (lack of semantic explanations) and pure LLM mediation (lack of formal guarantees). Additionally, we plan to compare our framework against a non-LLM baseline (e.g., aggregating traditional recommendation lists with social choice) to quantify the added value of LLM multi-agent personalization.

Acknowledgments

This work is conducted within the Interfaces of Agent-Centric Artificial Intelligence (IACAI) COMET module, funded by the Austrian Research Promotion Agency (FFG).

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar checks, spelling checks, and sentence polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Huang, J. Wu, Y. Xia, Z. Yu, R. Wang, T. Yu, R. Zhang, R. A. Rossi, B. Kveton, D. Zhou, et al., Towards agentic recommender systems in the era of multimodal large language models, arXiv preprint arXiv:2503.16734 (2025).

[2] L. Xu, J. Zhang, B. Li, J. Wang, S. Chen, W. X. Zhao, J.-R. Wen, Tapping the potential of large language models as recommender systems: A comprehensive framework and empirical analysis, ACM TKDD 19 (2025) 1–51.

[3] R. Burke, Multisided fairness for recommendation, Workshop on Fairness, Accountability, and Transparency in Machine Learning (2017).

[4] R. Burke, G. Adomavicius, T. Bogers, T. Di Noia, D. Kowald, J. Neidhardt, Ö. Özgöbek, M. S. Pera, N. Tintarev, J. Ziegler, De-centering the (traditional) user: Multistakeholder evaluation of recommender systems, International Journal of Human-Computer Studies (2025) 103560.

[5] M. D. Ekstrand, A. Razi, A. Sarcevic, M. S. Pera, R. Burke, K. L. Wright, Recommending with, not for: Co-designing recommender systems for social good, ACM TORS (2025).

[6] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, D. Zanzonelli, Fairness in recommender systems: research landscape and future directions, User Modeling and User-Adapted Interaction 34 (2024) 59–108.

[7] J. Liu, Z. Qiu, Z. Li, Q. Dai, W. Yu, J. Zhu, M. Hu, M. Yang, T.-S. Chua, I. King, A survey of personalized large language models: Progress and future directions, arXiv preprint arXiv:2502.11528 (2025).
[8] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, H. D. Nguyen, Multi-agent collaboration mechanisms: A survey of llms, arXiv preprint arXiv:2501.06322 (2025).
[9] V. Dignum, F. Dignum, Agentifying agentic ai, arXiv preprint arXiv:2511.17332 (2025).
[10] A. Bellina, G. De Marzo, D. Garcia, Conformity and social impact on ai agents, arXiv preprint arXiv:2601.05384 (2026).
[11] A. Wynn, H. Satija, G. Hadfield, Talk isn't always cheap: Understanding failure modes in multi-agent debate, arXiv preprint arXiv:2509.05396 (2025).
[12] J. Chun, Q. Chen, J. Li, I. Ahmed, Is multi-agent debate (mad) the silver bullet? an empirical analysis of mad in code summarization and translation, arXiv preprint arXiv:2503.12029 (2025).
[13] K. J. Arrow, Social Choice and Individual Values, Yale University Press, 1951.
[14] A. Sen, Collective Choice and Social Welfare, Holden-Day, 1970.
[15] A. Aird, P. Farastu, J. Sun, E. Stefancova, C. All, A. Voida, N. Mattei, R. Burke, Dynamic fairness-aware recommendation through multi-agent social choice, ACM TORS 3 (2024) 1–35.
[16] C. Bauer, L. Chen, N. Ferro, N. Fuhr, A. Anand, T. Breuer, G. Faggioli, O. Frieder, H. Joho, J. Karlgren, et al., Conversational agents: A framework for evaluation (cafe)(dagstuhl perspectives workshop 24352), Dagstuhl Manifestos 11 (2025) 19–67.
[17] M. Kaya, T. Bogers, Mapping stakeholder needs to multi-sided fairness in candidate recommendation for algorithmic hiring, in: Proceedings of RecSys'25, 2025, pp. 257–267.
[18] J. J. Smith, A. Buhayh, A. Kathait, P. Ragothaman, N. Mattei, R. Burke, A. Voida, The many faces of fairness: Exploring the institutional logics of multistakeholder microlending recommendation, in: Proceedings of FAccT'23, 2023, pp. 1652–1663.
[19] S. Mhlambi, S. Tiribelli, Decolonizing ai ethics: Relational autonomy as a means to counter ai harms, Topoi 42 (2023) 867–880.
[20] Y. Deldjoo, Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency, ACM TORS 4 (2025) 1–35.
[21] A. Mishra, Ai alignment and social choice: Fundamental limitations and policy implications, arXiv preprint arXiv:2310.16048 (2023).
[22] M. Abou Ali, F. Dornaika, J. Charafeddine, Agentic ai: a comprehensive survey of architectures, applications, and future directions, Artificial Intelligence Review 59 (2025) 11.
[23] B. El, J. Zou, Moloch's bargain: Emergent misalignment when llms compete for audiences, arXiv preprint arXiv:2510.06105 (2025).
[24] R. Binkyte, Interactional fairness in llm multi-agent systems: An evaluation framework, in: Proceedings of AIES-25, volume 8, 2025, pp. 457–468.
[25] J. Li, X. Liu, Y. Feng, From single to societal: Analyzing persona-induced bias in multi-agent interactions, in: Proceedings of AAAI-2026, volume 40, 2026, pp. 31609–31617.
[26] A. P. Uchoa, C. E. Oliveira, C. L. Motta, D. Schneider, Multi-stakeholder alignment in llm-powered collaborative ai systems: A multi-agent framework for intelligent tutoring, in: Proceedings of CHIRA'25, Springer, 2025, pp. 360–379.
[27] R.-R. Maura-Rivero, M. Lanctot, F. Visin, K. Larson, Jackpot! alignment as a maximal lottery, arXiv preprint arXiv:2501.19266 (2025).
[28] A. Banerjee, A. Satish, F. N. Aisyah, W. Wörndl, Y. Deldjoo, Collab-rec: An llm-based agentic framework for balancing recommendations in tourism, arXiv preprint arXiv:2508.15030 (2025).
[29] G. Popescu, Group recommender systems as a voting problem, in: International Conference on Online Communities and Social Computing, Springer, 2013, pp. 412–421.
[30] A. P. Uchoa, C. E. Oliveira, C. L. Motta, D. Schneider, Natural-language mediation versus numerical aggregation in multi-stakeholder ai governance: Capability boundaries and architectural requirements, Computers 15 (2026) 24.
[31] P. Müllner, A. Schreuer, S. Kopeinik, B. Wieser, D. Kowald, Multistakeholder fairness in tourism: what can algorithms learn from tourism management?, Frontiers in big Data 8 (2025) 1632766.

[32] E. L. González-Sanz, I. Cantador, A. Bellogín, Llm-based generation of personalized, context-aware city tourist itineraries: A user study with gpt trip planner (2025).
[33] R. Lozano, Envisioning sustainability three-dimensionally, Journal of cleaner production 16 (2008).
[34] A. Forster, S. Kopeinik, D. Helic, S. Thalmann, D. Kowald, Exploring the effect of context-awareness and popularity calibration on popularity bias in poi recommendations, in: Proceedings of RecSys'25, 2025, pp. 593–598.
[35] P. Lederer, D. Peters, T. Wąs, The squared kemeny rule for averaging rankings, arXiv preprint arXiv:2404.08474 (2024).
[36] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The impact of popularity bias on fairness and calibration in recommendation, arXiv preprint arXiv:1910.05755 (2019).
[37] O. Lesota, A. Melchiorre, N. Rekabsaz, S. Brandl, D. Kowald, E. Lex, M. Schedl, Analyzing item popularity bias of music recommender systems: are different genders equally affected?, in: Proceedings of RecSys'21, 2021, pp. 601–606.
[38] A. Banerjee, P. Banik, W. Wörndl, A review on individual and multistakeholder fairness in tourism recommender systems, Frontiers in big Data 6 (2023) 1168692.
[39] N. Hadziarapovic, M. van Steenbergen, P. Ravesteijn, J. Versendaal, G. Mertens, Integrating stakeholder values in system of collective management of music copyrights: A value-sensitive design approach, International Journal of Music Business Research 14 (2025) 27–43.
[40] M. Unger, P. Li, M. C. Cohen, B. Brost, A. Tuzhilin, Deep multi-objective multi-stakeholder recommendations in the media industry, Available at SSRN (2025).
[41] F. Atzenhofer-Baumgartner, B. C. Geiger, G. Vogeler, D. Kowald, Value identification in multi-stakeholder recommender systems for humanities and historical research: The case of the digital archive monasterium. net, arXiv preprint arXiv:2409.17769 (2024).
[42] F. Atzenhofer-Baumgartner, G. Vogeler, D. Kowald, A multistakeholder approach to value-driven co-design of recommender systems evaluation metrics in digital archives, in: Proceedings of RecSys'25, 2025, pp. 503–508.
[43] M. Langer, C. J. König, Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management, Human Resource Management Review 33 (2023) 100881.
[44] L. Rozenblit, A. Price, A. Solomonides, A. L. Joseph, E. Koski, G. Srivastava, S. Labkoff, D. Bray, M. Lopez-Gonzalez, R. Singh, et al., Toward responsible ai governance: balancing multi-stakeholder perspectives on ai in healthcare, International Journal of Medical Informatics 203 (2025) 106015.
[45] S. Thiebes, F. Gao, R. O. Briggs, M. Schmidt-Kraepelin, A. Sunyaev, Design concerns for multiorganizational, multistakeholder collaboration: a study in the healthcare industry, Journal of Management Information Systems 40 (2023) 239–270.
[46] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, H. Zhang, Y. Liu, C. Wu, X. Li, C. Zhu, et al., How can recommender systems benefit from large language models: A survey, ACM TOIS 43 (2025) 1–47.
[47] T. Vente, M. Heep, A. Abbas, T. Sperle, J. Beel, B. Goethals, Aps explorer: Navigating algorithm performance spaces for informed dataset selection, in: Proceedings of RecSys'25, 2025, pp. 1322–1324.
[48] D. Di Palma, F. A. Merra, M. Sfilio, V. W. Anelli, F. Narducci, T. Di Noia, Do llms memorize recommendation datasets? a preliminary study on movielens-1m, in: Proceedings of SIGIR'25, 2025, pp. 2582–2586.
[49] A. Banerjee, A. Satish, F. N. Aisyah, W. Wörndl, Y. Deldjoo, Synthtrips: A knowledge-grounded framework for benchmark data generation for personalized tourism recommenders, in: Proceedings of SIGIR'25, 2025, pp. 3743–3752.
[50] P. Sánchez, A. Bellogin, J. L. Jorro-Aragoneses, Context trails: A dataset to study contextual and route recommendation, in: Proceedings of RecSys'25, 2025, pp. 716–725.
[51] H. A. Rahmani, Y. Deldjoo, A. Tourani, M. Naghiaei, The unfairness of active users and popularity bias in point-of-interest recommendation, in: International workshop on algorithmic bias in search and recommendation, Springer, 2022, pp. 56–68.