Marsden, Nicola. (2025b). Prompting Fairness: How End Users Can Mitigate Bias in AI Systems. HCI International, LNAI15820, 68-77.
doi:10.1007/978-3-031-93415-5_4

Prompting Fairness: How End Users Can Mitigate Bias in AI Systems

Nicola Marsden(✉)

Heilbronn University, Heilbronn, Germany
nicola.marsden@hs-heilbronn.de
https://www.hs-heilbronn.de/de/lab-sozioinformatik

Abstract. Artificial Intelligence systems are increasingly integrated into critical decision-making contexts, raising concerns about their potential to perpetuate bias and inequality. Existing approaches to AI fairness have primarily focused on developer-led interventions, often neglecting the role of end-users in addressing these issues. This paper introduces a novel framework that positions end-users as active agents in mitigating bias. Through practical prompt engineering techniques, including prefix-based strategies, iterative refinement, reasoning-based prompting, and in-context learning, users can dynamically influence AI outputs without requiring access to system internals. The framework also addresses barriers such as moral disengagement, automation bias, and the complexity of implementation, offering solutions that enhance user engagement and foster collective efficacy. By reframing fairness as a shared responsibility, this work highlights the potential of participatory strategies to bridge the gap between technical advancements and equitable real-world applications, advancing the development of inclusive and accountable AI systems.

Keywords: AI fairness · End-user empowerment · Prompt engineering

1 Introduction: Reframing AI Fairness Through End-User Empowerment

The rapid proliferation of Artificial Intelligence (AI) systems in high-stakes decision-making contexts, such as mortgage lending, hiring, and criminal justice, has heightened concerns about algorithmic bias [4,18]. These systems, while promising efficiency and objectivity, frequently perpetuate and amplify existing societal inequities, leading to discriminatory outcomes [17,20]. Biases in AI systems can emerge at multiple stages of the lifecycle, including data collection, problem formulation, algorithm design, and evaluation processes [18]. For example, recruitment algorithms have shown a tendency to discriminate against candidates based on perceived gender, while predictive models in insurance and credit scoring disproportionately disadvantage people of color [17].

H. Degen and S. Ntoa (Eds.): HCII 2025, LNAI 15820, pp. 68–77, 2025.
https://doi.org/10.1007/978-3-031-93415-5_4

While fairness research has produced numerous bias detection metrics, mitigation algorithms, and open-source tools like AIF360 [4], these advances often remain inaccessible to practitioners and end-users. Developers frequently struggle to implement these solutions, facing challenges such as deciding whether to adjust data, modify algorithms, or correct outputs. Additionally, AI professionals may experience moral disengagement, viewing bias as either inevitable or beyond their control due to limited organizational support and a lack of clear guidelines [17].

Much of the existing literature on AI fairness has focused on developer-centric approaches, overlooking the potential role of end-users in mitigating bias. This paper argues that end-users-those who directly interact with AI outputs-can play a critical role in fostering equitable AI systems. Unlike technical interventions requiring access to model internals, strategies like prompt engineering empower end-users to dynamically address biases in real-time, adapting AI outputs within their specific contexts of use. By equipping end-users with practical tools and actionable techniques, this work reframes bias mitigation as an opportunity for user empowerment rather than a purely technical challenge.

This paper makes three key contributions:

1. A conceptual framework, which situates end-user bias mitigation within the broader challenges of AI fairness.
2. A synthesis of prompt engineering techniques that enable end-users to identify and mitigate biases in large language model (LLM) outputs.
3. A discussion of practical implications for integrating end-user bias mitigation into everyday applications of AI systems.

By repositioning end-users as active co-regulators of fairness, this paper aligns with the HCI tradition of empowering users through interactive systems [5,21]. The proposed approach bridges the gap between technical advancements and real-world applications, fostering a participatory and inclusive pathway toward equitable AI systems.

2 Background: Understanding the Landscape of AI Bias and Its Mitigation

AI bias is a multifaceted issue that manifests in diverse ways across computational and socio-technical contexts. From a computational perspective, bias refers to systematic deviations from true or desired values, often leading to statistical inaccuracies [3]. Socio-technical perspectives, in contrast, frame bias as a reflection of structural inequities, privileging certain groups or concepts at the expense of others [16]. This broader lens situates bias within frameworks of social justice, fairness, and equity, recognizing its roots in societal structures as well as technical processes [6]. Bias in AI systems typically emerges in three primary forms [18]:

Date Bias. Stemming from historical inequities, datasets that are not representative, or subjective human categorization, data bias embeds existing societal disparities into training datasets. For instance, representation bias often excludes marginalized groups, perpetuating underrepresentation in AI outputs [19,20].

Algorithmic Bias. This type arises during model development, influenced by design objectives, parameter tuning, and optimization trade-offs. Models optimized to minimize error rates, for example, may inadvertently favor majority groups while disadvantaging minorities [20].

User Interaction Bias. Bias can also emerge during user interactions, driven by design choices or behavioral tendencies. Presentation bias and ranking bias, for instance, influence user perceptions by prioritizing specific outputs over others [9,18].

Despite advancements in fairness-focused AI research, existing mitigation strategies-such as pre-processing data adjustments, algorithmic interventions, and post-processing output corrections-largely rely on developers with technical expertise [3]. This developer-centric focus limits the accessibility of these solutions to end-users and excludes non-technical stakeholders from participating in fairness efforts. Moreover, the opacity of many AI systems, particularly those using closed APIs, restricts transparency and user interaction with underlying algorithms.

These gaps highlight the need for inclusive approaches that empower end-users to address bias dynamically within their interactions with AI systems. By providing users with tools like prompt engineering, which enable real-time bias mitigation without requiring access to model internals, it is possible to democratize AI fairness and bridge the divide between technical innovation and practical application. Recognizing and addressing the interplay between technical and societal factors is essential to fostering equitable AI ecosystems.

3 End-User Empowerment: A Conceptual Framework

AI fairness efforts have traditionally centered on developer-led interventions, treating end-users as passive recipients of AI-generated outputs [18]. However, this approach overlooks the potential for end-users to contribute to bias mitigation. Empowering users to actively engage with AI systems at the point of interaction offers a promising avenue to address biases in real-world contexts.

This paper proposes a conceptual framework that positions end-users as active co-regulators of fairness. By equipping users with practical strategies, such as prompt engineering, the framework fosters a participatory approach to bias mitigation. It bridges the gap between technical solutions and everyday application, complementing developer-centric efforts with a user-driven focus [10].

The framework comprises three key components:

1. **Awareness:** Empowering users begins with understanding how biases manifest in AI outputs and the societal structures that perpetuate inequities. Awareness enables users to identify and address subtle forms of bias, such as stereotypes embedded in language or assumptions implicit in generated content [18].
2. **Prompt Engineering:** This component provides users with practical tools to influence AI outputs by crafting prompts that counteract bias. For example, users can explicitly request unbiased responses or guide the AI toward equitable perspectives. These techniques allow end-users to dynamically adjust outputs, adapting them to align with fairness principles within their specific contexts [12].
3. **Critical Evaluation:** Users play a pivotal role in critically assessing AI outputs to ensure alignment with fairness objectives. By examining language, assumptions, and perspectives in the results, users can identify residual biases and refine their interactions with the system [14].

This framework reflects the human-computer interaction (HCI) tradition of empowering users through interactive systems [5]. By focusing on awareness, prompt engineering, and critical evaluation, it provides a structured pathway for end-users to participate in fostering equitable AI systems. Ultimately, the framework reframes bias mitigation as a collaborative effort, encouraging users to transition from passive recipients to active agents of change.

4 Prompt Engineering Techniques for Bias Mitigation

The framework outlined above establishes a foundation for empowering end-users as co-regulators of AI fairness, highlighting the importance of awareness, critical evaluation, and iterative refinement in addressing bias. However, these principles must translate into actionable methods that users can readily apply in their interactions with AI systems. Prompt engineering emerges as a pivotal tool for operationalizing this framework, enabling users to directly influence and mitigate biases in AI-generated outputs.

4.1 Foundational Principles of Prompt Engineering for Fairness

At its core, prompt engineering for bias mitigation involves the meticulous crafting of input prompts to guide LLMs towards the generation of unbiased outputs. This approach stems from the understanding that LLMs are not neutral instruments but rather, reflect the biases embedded within their training corpora [11,12]. Through the strategic design of prompts, end-users can counteract these inherent biases, encouraging the model to produce fairer and more equitable responses. It is crucial to acknowledge that biases may manifest as statistical, cognitive, societal, or institutional phenomena requiring a multifaceted approach to address their complexities [4].

Clarity and Specificity. Prompts should be formulated with clarity and specificity, explicitly instructing the model to eschew stereotypes and discriminatory language. For example, instead of asking an LLM to "describe a typical engineer," a more specific prompt could be, "Describe an engineer, without making assumptions about their gender, race, or background."

Contextual Awareness. Effective prompts must demonstrate a sensitivity to context, acknowledging that bias is often situation-dependent [4]. They should provide sufficient contextual information to help the LLM understand the specific nuances of a given situation, thus avoiding stereotypical assumptions.

Iterative Refinement. Bias mitigation is an inherently iterative process. End-users should engage in empirical testing with various prompts, rigorously evaluating the resultant outputs, and continuously refining their approach to achieve optimal outcomes.

4.2 Prompt-Based Debiasing Techniques

Recent investigations have demonstrated the capacity for LLMs to self-regulate their inherent biases through specific prompting strategies [12].

Prefix-Based Prompting. Prefix-based prompting involves guiding LLMs to generate less biased text by incorporating specific instructions, phrases, or roles at the beginning of a prompt. For example, users may include instructions such as, "Please ensure that the following is unbiased and does not rely on stereotypes," or prompt the model to adopt a persona, such as, "You are a fair-minded person who promotes inclusivity in all responses" [10,13].

Iterative Prompting with Self-generated Feedback. Iterative prompting enables users to refine outputs by leveraging feedback from previous iterations. This involves an initial prompt that generates a response, followed by additional instructions asking the model to revise its output to address fairness concerns [10,12].

Reasoning-Based Prompting. Reasoning-based prompting leverages techniques that encourage LLMs to explicitly articulate their reasoning process. This method fosters critical reflection and reduces reliance on biased shortcuts by guiding the model to think systematically or reflect on its outputs [15,22].

In-Context Learning through Prompt Examples. In-context learning involves embedding examples of desired behavior within the prompt, enabling the model to mimic fair and equitable responses. This technique exposes the model to counter-stereotypical examples, which guide it toward generating balanced outputs [10,13].

5 Barriers to Action Among End-Users

While the potential of prompt engineering for bias mitigation is evident, its real-world application hinges on users' ability and willingness to engage actively with these strategies. This section examines the barriers that end-users face when addressing bias in LLM outputs, highlighting psychological, educational, and practical obstacles. These challenges are critical to understanding the gap between the promise of end-user empowerment and its practical realization.

5.1 Psychological Barriers: Moral Disengagement and Automation Bias

Psychological barriers play a critical role in limiting user engagement with bias mitigation efforts. One significant challenge is *moral disengagement*, a cognitive mechanism through which individuals detach their moral values from their actions, allowing them to avoid accountability for addressing bias [2,8]. In the context of AI systems, users and professionals may justify inaction by reframing the issue as unavoidable or beyond their influence [19]. For example:

- **Moral justification:** Rationalizing inaction as necessary to achieve broader goals.
- **Displacement of responsibility:** Attributing bias mitigation to developers or regulators.
- **Diffusion of responsibility:** Diminishing personal accountability in group settings.

Compounding this challenge is *automation bias*, the tendency for users to over-rely on AI outputs due to their perceived authority or objectivity [19]. This bias discourages users from critically evaluating AI-generated results. Together, these psychological barriers undermine efforts to empower end-users as active participants in mitigating AI bias.

5.2 Practical Barriers: Education, Complexity, and Iteration

Practical barriers significantly limit the ability of end-users to engage effectively with bias mitigation strategies. Key challenges include:

- **Lack of Awareness:** Many users are unfamiliar with AI bias and the techniques available to address it [17]. Educational interventions are necessary to bridge this gap.
- **Complexity of Techniques:** Advanced methods, while effective, can be difficult for non-technical users to implement without sufficient training and support.
- **Iterative Refinement:** Effective bias mitigation often requires users to test, evaluate, and refine their prompts iteratively. This process can be resource-intensive and lead to frustration if users do not see immediate results.

To address these challenges, user-centered tools should simplify the bias mitigation process. Examples include:

- Guided workflows that progressively teach users effective prompt engineering techniques.
- Real-time feedback mechanisms, such as visual indicators of bias levels in AI outputs.
- Accessible resources, such as interactive tutorials or examples of successful debiasing prompts, embedded within AI applications.

5.3 Bridging the Gap: Agency and Collective Efficacy

Social cognitive theory (SCT) offers insights into how users' perceptions of agency-their belief in their ability to influence outcomes-and collective efficacy-their belief in the group's capacity to effect change-shape their willingness to act [1]. Fostering collective efficacy is essential. For example:

- Collaborative tools can enable group-level feedback and shared decision-making.
- Interfaces highlighting contributions of other users reinforce mutual accountability and encourage sustained participation.

By integrating SCT principles into AI system design, this approach bridges the gap between individual disengagement and collective empowerment, paving the way for more equitable ecosystems.

6 Conclusions and Future Directions

This paper has explored the transformative potential of end-users as active participants in mitigating bias in AI outputs, presenting a conceptual framework and actionable strategies centered on prompt engineering. By emphasizing user-driven interventions, this work reframes AI fairness as a collaborative effort that extends beyond technical solutions to include the critical role of end-users in fostering equitable systems.

6.1 Summary of Contributions

The contributions of this work include a conceptual framework that positions end-users as active co-regulators of fairness, enabling them to identify and mitigate biases in AI outputs. It also provides a synthesis of practical and accessible prompt engineering techniques, demonstrating how LLMs can be leveraged to reduce bias through user-driven strategies. Additionally, the paper analyzes psychological and practical barriers, offering insights for designing user-friendly tools and fostering greater user empowerment. By addressing these dimensions, this work highlights that AI fairness is both a technical and social challenge. Empowering end-users to take an active role bridges the gap between theoretical advancements and real-world applications, fostering participatory and inclusive approaches to mitigating bias.

6.2 Limitations and Future Directions

Future research must prioritize empirical validation to evaluate the real-world effectiveness of prompt-based debiasing. Controlled user studies can provide critical insights into usability and accessibility, while identifying areas for refinement. Addressing psychological barriers, such as moral disengagement and automation bias, will also be essential. Interventions like interactive educational initiatives and collaborative tools can empower users and sustain engagement with bias mitigation practices.

While empowering end-users democratizes fairness efforts, it risks shifting accountability away from developers and companies deploying AI systems. To balance user empowerment with institutional responsibility, multi-agent systems could integrate developer safeguards with user-facing tools. This approach ensures shared accountability while aligning with frameworks like the European Union's AI Act [7], which mandates fairness safeguards throughout the AI lifecycle.

Ultimately, the success of fairness interventions depends on fostering shared responsibility among developers, operators, and end-users. By bridging technical innovation with participatory approaches, future research can ensure that AI systems are not only powerful but also equitable and inclusive, reflecting a collective commitment to fairness and accountability.

Acknowledgments. This work has been partially funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) under Grant Number 01 FP24M03A as part of the project 'Tech2stay - Strengthening and retaining women in ICT: Living lab research on innovative interventions, investigating attrition and sustainable strategies for retaining women in top ICT jobs and positions (Tech2stay - Frauen in der IT stärken und halten: Reallaborforschung zu innovativen Interventionen, Untersuchung von Fluktuation und nachhaltige Strategien zum Verbleib von Frauen in IT-Berufen und -Spitzenpositionen)' in the Funding Line 'MissionSTEM - Women shaping the future (MissionMINT - Frauen gestalten Zukunft)'. The responsibility for all content supplied lies with the authors.

References

1. Bandura, A.: Social cognitive theory: an agentic perspective. Annu. Rev. Psychol. **52**(1), 1–26 (2001). https://doi.org/10.1146/annurev.psych.52.1.1
2. Bandura, A., Barbaranelli, C., Caprara, G.V., Pastorelli, C.: Mechanisms of moral disengagement in the exercise of moral agency. J. Pers. Soc. Psychol. **71**(2), 364 (1996)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press, Cambridge (2023)
4. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM J. Res. Dev. **63**(4/5), 4:1–4:15 (2019). https://doi.org/10.1147/JRD.2019.2942287

5. Bødker, S., Dindler, C., Iversen, O.S., Smith, R.C.: Participatory Design, Synthesis Lectures on Human-Centered Informatics, vol. 14. Morgan & Claypool Publishers, San Rafael (2022)
6. Draude, C., Klumbyte, G., Lücking, P., Treusch, P.: Situated algorithms: a sociotechnical systemic approach to bias. Online Inf. Rev. **44**(2), 325–342 (2019). https://doi.org/10.1108/OIR-10-2018-0332
7. European-Union: Regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689
8. Festinger, L.: A Theory of Cognitive Dissonance. Stanford University Press, Redwood City (1957)
9. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. (TOIS) **14**(3), 330–347 (1996)
10. Furniturewala, S., et al.: Thinking fair and slow: on the efficacy of structured prompts for debiasing language models, pp. 1–14. arXiv preprint arXiv:2405.10431 (2024). https://doi.org/10.48550/arXiv.2405.10431
11. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey. arXiv preprint arXiv:2309.00770 (2023). https://doi.org/10.48550/arXiv.2309.00770
12. Gallegos, I.O., et al.: Self-debiasing large language models: zero-shot recognition and reduction of stereotypes. arXiv preprint arXiv:2402.01981 (2024). https://doi.org/10.48550/arXiv.2402.01981
13. Ganguli, D., et al.: The capacity for moral self-correction in large language models. arXiv preprint arXiv:2302.07459 (2023). https://doi.org/10.48550/arXiv.2302.07459
14. Han, S., Kelly, E., Nikou, S., Svee, E.O.: Aligning artificial intelligence with human values: reflections from a phenomenological perspective. AI & SOCIETY, pp. 1–13 (2022). https://doi.org/10.1007/s00146-021-01247-4
15. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. **35**, 22199–22213 (2022). https://doi.org/10.1080/0960085X.2021.1927212
16. Kordzadeh, N., Ghasemaghaei, M.: Algorithmic bias: review, synthesis, and future research directions. Eur. J. Inf. Syst. **31**(3), 388–409 (2022)
17. Lancaster, C.M., Schulenberg, K., Flathmann, C., McNeese, N.J., Freeman, G.: "It's everybody's role to speak up... but not everyone will": understanding AI professionals' perceptions of accountability for AI bias mitigation. ACM J. Responsib. Comput. **1**(1), 1–30 (2024). https://doi.org/10.1145/3632121
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021). https://doi.org/10.1145/3457607
19. Ozmen Garibay, O., et al.: Six human-centered artificial intelligence grand challenges. Int. J. Hum.-Comput. Interact. **39**(3), 391–437 (2023). https://doi.org/10.1080/10447318.2022.2153320
20. Smith, G., Rustagi, I.: Mitigating bias in artificial intelligence: an equity fluent leadership playbook. Technical report. Berkeley Haas Center for Equity, Gender and Leadership (2020). https://haas.berkeley.edu/wpcontent/uploads/UCB_Playbook_R10_V2_spreads2.pdf

21. Usmani, U.A., Happonen, A., Watada, J.: Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In: 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–7 (2023). https://doi.org/10.1109/HORA58378.2023.10156761
22. Zeng, C.C., Chung, M., Zhou, E.: Prompting for fairness: mitigating gender bias in large language models with self-debiasing prompting. University of Michigan CSE 595 Natural Language Processing (2024)