Towards Increasing Adolescents' Awareness and Understanding of AI Bias

AIDAN Z. FITZSIMONS, Northwestern University, USA
KRISTIN FASIANG, Northwestern University, USA
ELIZABETH M. GERBER, Northwestern University, USA
DURI LONG, Northwestern University, USA

Adolescents increasingly use large language models (LLMs) such as ChatGPT and Gemini for schoolwork and personal writing (e.g., job materials). Because LLMs can reproduce and amplify societal biases present in training data, biased outputs may shape young peoples' sense of themselves, narrowing perceived possibilities and reinforcing inequities. We present two complementary approaches to fostering youth understanding of AI bias: (1) in-the-moment supports that overlay LLM use to help users recognize and mitigate bias during real-time interactions, and (2) learning interventions that scaffold authentic investigations of bias and its structural origins.

ACM Reference Format:
Aidan Z. Fitzsimons, Kristin Fasiang, Elizabeth M. Gerber, and Duri Long. 2026. Towards Increasing Adolescents' Awareness and Understanding of AI Bias. In *Proceedings of the* ***Sociotechnical Imaginaries of Responsible Design: A Case for Mitigating Gender-based Online Harm Workshop*** *at CHI 2026, April 15, Barcelona, Spain.* ACM, New York, NY, USA, 3 pages.

1 Introduction

Adolescents are rapidly integrating large language models (LLMs) such as ChatGPT and Gemini into their academic, social, and professional lives. Recent reports indicate widespread adoption of generative AI tools among teens for schoolwork, exploration, and personal writing [6, 15]. However, LLMs are known to reproduce and amplify social biases embedded in their training data [3, 11, 13]. Because training corpora often include large-scale internet data scraped from biased sources [3], models may reproduce gendered assumptions in the narratives they generate. Prior work has demonstrated that LLMs can construct gendered struggle narratives in college essays, particularly centering narratives for marginalized users around overcoming hardship [10]. More broadly, management science research shows that men and women tend to present qualifications differently, and evaluators interpret identical qualifications differently depending on gender cues [4, 5]. If such patterns are reflected in LLM outputs, adolescents using these tools may unknowingly receive advice that aligns with stereotyped expectations.

This creates a sociotechnical risk: biased outputs may shape the stories young people tell about themselves, narrowing perceived possibilities or reinforcing inequities [14]. While significant research has examined bias mitigation at the model level, there remains an urgent need for user-facing interventions that (1) help adolescents recognize bias when interacting with AI systems and (2) equip them with deeper conceptual understanding of how bias emerges in AI systems.

In this position paper, we describe two complementary forms of intervention to increase adolescents' awareness and understanding of AI bias: *(1) in-the-moment interventions* that overlay LLM use and support bias detection and mitigation during real-time interactions, and *(2) learning interventions* that scaffold authentic investigations of bias and its structural origins. Together, these interventions operate at both the experiential and conceptual levels of AI literacy [12], fostering critical awareness while preserving adolescents' agency.

Authors' Contact Information: Aidan Z. Fitzsimons, Northwestern University, Evanston, Illinois, USA, aidan.fitzsimons@u.northwestern.edu; Kristin Fasiang, Northwestern University, Evanston, Illinois, USA, kristinfasiang2029@u.northwestern.edu; Elizabeth M. Gerber, Northwestern University, Evanston, Illinois, USA, egerber@northwestern.edu; Duri Long, Northwestern University, Evanston, Illinois, USA, duri@northwestern.edu.

2 In-the-Moment Interventions Overlaying LLM Use

We are designing in-the-moment interventions to support adolescents while they are actively interacting with LLMs. These interventions embed bias awareness directly into AI use in moments when adolescents turn to AI. Adolescents often rely on LLMs for drafting emails, cover letters, and career exploration materials. However, pilot findings suggest that adolescents are not attuned to when models shift outputs toward gendered stereotypes. Therefore, the first class of intervention focuses on helping adolescents notice bias. Using a design-based research approach [1, 8], collaborative design sessions with college students can identify what forms of explanation or visualization make bias legible. These interventions align with AI literacy competencies emphasizing critical evaluation of AI outputs [12].

Adolescents also need concrete, usable strategies to reduce biased outputs. A second class of intervention involves developing validated prompt templates and usage guidelines that reduce narrative-level bias in LLM responses. Through iterative cycles of auditing and participatory evaluation [2], adolescents can help identify patterns of narrative bias and co-create prompt strategies. We will systematically test these interventions to determine whether they decrease the incidence of biased narrative tropes. Such low-burden tools empower adolescents to exercise agency within existing platforms, rather than requiring deep technical knowledge of model internals.

3 Educational Resources for Deeper Understanding of AI Bias

We have also designed structured learning experiences to scaffold adolescent understanding of bias through engagement with underlying model processes and authentic AI auditing and bias mitigation processes. These learning experiences can involve *simulations of underlying model behaviors* that help learners "see" from the perspective of AI. Large Language Madlibs is an unplugged activity that engages young people in simulating how a LLM uses probability to make decisions about the next word to display by engaging in a process of dice rolling and coin flipping to generate a sentence [9]. Probabilities encoded in the activity design prompt learners to observe and reflect on potential gender bias [9].

Learning experiences can also *scaffold the process of auditing real-world models* to enable learners to observe, reflect on, and mitigate bias. For example, BiasViz is a learning platform that engages youth in an authentic process of auditing an AI model by predicting, documenting, and quantifying biases that arise in LLM outputs [7]. Learners engage in red-teaming practices to craft prompts to elicit biased outputs, quantify whether observed biases are replicated across a variety of outputs, and reflect on the underlying structural causes of bias. In future versions, we plan to explore approaches for scaffolding bias mitigation practices by engaging learners in augmenting training data or fine-tuning models. This will enable young learners to exercise increased agency when interacting with LLMs, increasing their awareness of AI biases and also their ability to modify AI to better reflect their values.

4 Discussion and Implications

Addressing AI bias in adolescence requires moving beyond solely technical mitigation strategies. Because identity formation and AI interaction are intertwined during this developmental period, interventions must operate at both experiential and conceptual levels. In-the-moment overlays foster situated awareness, helping adolescents detect bias during consequential writing tasks. Prompt toolkits provide practical means of resistance. Complementary educational resources (to be developed) can scaffold deeper understanding of training data, sociotechnical systems, and the structural origins of bias. Future work should evaluate these interventions longitudinally, examining not only bias detection but also impacts on adolescents' narrative agency, career exploration breadth, and perceptions of AI systems. Ultimately, increasing adolescents' awareness and understanding of AI bias is not only a matter of fairness in outputs—it is a matter of protecting young people's developing sense of possibility.

References

[1] Sasha Barab and Kurt Squire. 2004. Design-Based Research: Putting a Stake in the Ground. *Journal of the Learning Sciences* 13 (Jan. 2004), 1–14. https://doi.org/10.1207/s15327809jls1301_1

[2] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300773

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[4] Elizabeth Lauren Campbell and Oliver Hahl. 2022. He's Overqualified, She's Highly Committed: Qualification Signals and Gendered Assumptions About Job Candidate Commitment. *Organization Science* 33, 6 (Nov. 2022), 2451–2476. https://doi.org/10.1287/orsc.2021.1550

[5] Katherine B. Coffman, Manuela R. Collis, and Leena Kulkarni. 2024. Whether to Apply. *Management Science* 70, 7 (July 2024), 4649–4669. https://doi.org/10.1287/mnsc.2023.4907

[6] Common Sense Media Group. 2024. *The Dawn of the AI Era: Teens, Parents, and the Adoption of Generative AI at Home and School.* Technical Report.

[7] Hasti Darabipourshiraz, Maalvika Bhat, and Duri Long. 2025. Introducing AI Without Computers: Hands-On Literacy and Ethical Sense-Making for Young Learners. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3706599.3719876

[8] Matthew W. Easterday, Daniel G. Rees Lewis, and Elizabeth M. Gerber. 2018. The logic of design research. *Learning: Research and Practice* 4, 2 (July 2018), 131–160. https://doi.org/10.1080/23735082.2017.1286367 _eprint: https://doi.org/10.1080/23735082.2017.1286367.

[9] Kristin Fasiang and Duri Long. 2025. Large Language MadLibs. In *Proceedings of the Fifteenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

[10] Aidan Z. Fitzsimons, Elizabeth M. Gerber, and Duri Long. 2025. AI constructs gendered struggle narratives: Implications for self-concept and systems design.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2290–2301. https://doi.org/10.1145/3715275.3732156

[11] Paula Hall and Debbie Ellis. 2023. A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review* 47, 7 (March 2023), 1264–1279. https://doi.org/10.1108/OIR-08-2021-0452

[12] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[13] Sinead O'Connor and Helen Liu. 2024. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY* 39, 4 (Aug. 2024), 2045–2057. https://doi.org/10.1007/s00146-023-01675-4

[14] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. https://doi.org/10.1145/3600211.3604673

[15] Mátyás Turós, Róbert Nagy, and Zoltán Szűts. 2025. What percentage of secondary school students do their homework with the help of artificial intelligence? - A survey of attitudes towards artificial intelligence. *Computers and Education: Artificial Intelligence* 8 (June 2025), 100394. https://doi.org/10.1016/j.caeai.2025.100394