1 Motivation

Over the past decade, artificial intelligence (AI) has moved from a niche area of computer science to a transformative force across industries and public life. Yet, some of the most pressing societal challenges remain unresolved: algorithmic fairness, demographic bias, and privacy. My research addresses these three ethical dimensions of AI, with a particular emphasis on *embedding-based representations*—the vectors at the heart of social recommender systems, large language models (LLMs), and natural language processing (NLP). I believe that tackling these problems will be central to building AI that is not only powerful but also trustworthy, inclusive, and safe.

Why fairness, bias, and privacy? These issues directly determine whether AI systems empower or marginalize people. For example, social networks increasingly mediate political debate, professional opportunities, and cultural visibility. Recommendation algorithms, however, tend to amplify popular nodes, reinforcing inequalities and diminishing the voice of structural minorities [6]. Similarly, LLMs now guide product discovery, hiring, and education, but our studies reveal that they encode and reproduce gender and racial stereotypes in subtle, implicit ways [8, 3]. Privacy is equally urgent: the explosion of sensitive textual data in law, healthcare, and governance demands mechanisms that protect individuals without rendering text unusable. Our recent work on *CluSanT* demonstrated that it is possible to balance semantic coherence with strong differential privacy guarantees [1].

Where are we with responsible AI today? Considerable progress has been made, but important gaps remain. Fairness research has produced diverse algorithms, yet most optimize abstract metrics (statistical parity, exposure fairness) without ensuring the *long-term visibility of minority communities*. Bias audits in LLMs are flourishing, but they often focus on benchmark datasets rather than real-world deployment contexts like consumer recommendations. Privacy in NLP is typically achieved either at the cost of text quality or through representations that humans cannot interpret. Meanwhile, policymakers and the public are demanding AI systems that meet high standards of transparency, accountability, and compliance with frameworks like the EU AI Act, AIDA, and the NIST AI RMF.

This combination of **high societal impact and technical difficulty** makes fairness, bias, and privacy a fertile area for research. Large language models offer new opportunities to evaluate privacy, probe bias, and generate counterfactuals for fairness, but also concentrate hidden associations and power dynamics. My agenda is to both understand these risks and design algorithms that enhance the inclusivity, accountability, and usability of AI systems.

2 Five-Year Plan

Over the next five years, my research will pursue a unified agenda for **embedding-centric responsible AI**. At its core lies a simple but powerful observation: the vectors that drive modern AI systems—embeddings of users, tokens, and documents—are not

neutral. They encode visibility, bias, and confidentiality. To design AI that is trustworthy, we must learn to reshape these representations so that fairness, inclusivity, and privacy are not afterthoughts, but *built into the foundation of the models themselves.*

- **Fairness as a dynamic property of networks.** Current recommender systems optimize short-term accuracy while amplifying structural inequalities. The next frontier is to design algorithms that treat fairness not as a static constraint, but as a *temporal dynamic* of evolving networks. My vision is to create fairness-aware recommenders that can sustain minority visibility across time, scale, and shifting communities. Recommender systems continue to show gaps in exposure and diversity, amplifying popular content while marginalizing minority users [10]. Counterfactual methods help separate genuine preferences from spurious effects of sensitive attributes [11], yet the field still lacks a coherent theory of long-term fairness. My work will establish this foundation.

- **Bias in LLMs as systemic rather than local.** Audits of large language models have shown that stereotypes persist across embeddings, prompts, and downstream tasks. But bias in LLMs is not merely a problem of offensive outputs—it is a systemic issue that shapes how knowledge is represented and retrieved. I will develop benchmarks and mitigation methods that treat bias as a *structural property of embedding spaces*, not only of text samples. This will require combining computational audits [9] with fairness identification frameworks [4], and extending them into multilingual and domain-specific settings such as law and healthcare. The goal is to move from surface-level "debiasing" toward systemic rebalancing of knowledge in generative systems.

- **Privacy as coherence-preserving transformation.** Most privacy mechanisms in NLP degrade fluency or utility. My approach is to treat privacy not as deletion, but as *representation transformation*: designing embedding-level operations that remove sensitive information while preserving semantics and readability. Building on our work with CluSanT, I envision frameworks that integrate privacy into the generative loop, yielding outputs that remain legally and socially intelligible. Recent papers such as PAPILLON [5] and RemoteRAG [2] signal a shift toward privacy-conscious deployments; my contribution will be to supply the theoretical and algorithmic backbone that makes these approaches broadly usable.

- **Towards a unified theory of ethical embeddings.** Fairness, bias, and privacy have been studied in parallel, but they share a common foundation: all three are properties of how embeddings encode people and knowledge. My long-term goal is to articulate a unified framework of *ethical embeddings.* This means building mathematical models that can measure how embeddings distribute visibility, encode stereotypes, and leak sensitive attributes—and then designing algorithms that can reshape these spaces under normative constraints. Just as statistical learning theory defined the limits of generalization, I aim to help define the theoretical limits of ethical AI.

3 Past Research

I have been working on fairness, bias, and privacy in AI throughout my doctoral research and early publications, with a consistent focus on how embedding-based representations shape the inclusivity and trustworthiness of modern AI systems. This body of work provides the foundation for my five-year plan:

- **Fairness in social recommender systems.** My research introduced *MinWalk*, a fairness-aware link recommendation algorithm that improves the long-term visibility of structural minority communities in evolving networks [7]. Unlike existing methods that optimize exposure metrics at a single snapshot, MinWalk explicitly models fairness as a dynamic property of the network. Extensive experiments on real-world graphs showed that it achieves a more stable balance between diversity and accuracy. This line of work demonstrates how fairness interventions can be designed to counteract systemic amplification of popularity and opens the path toward temporal theories of algorithmic fairness.

- **Bias detection in large language models.** I have analyzed how stereotypes are encoded in modern embeddings from OpenAI, Google, Microsoft, Cohere, and BGE. Using metrics such as SC-WEAT, clustering, and distributional divergence, I showed that gender and racial associations persist at scale, often surfacing in downstream tasks like consumer product recommendations [8, 3]. These studies introduced multi-method audit pipelines—including Marked Words, SVM classifiers, and Jensen–Shannon divergence—to capture implicit bias in generative outputs. This work highlights how bias in LLMs is not incidental but systemic, and provides methodological tools to evaluate and mitigate such risks.

- **Privacy-preserving NLP with semantic coherence.** In response to the challenge of sanitizing sensitive text without destroying readability, I developed *CluSanT*, a framework that combines token clustering with controlled replacement mechanisms [1]. Unlike prior approaches that sacrificed fluency for privacy, CluSanT balances both dimensions, producing legally and semantically intelligible text while achieving stronger privacy guarantees. Evaluation on a legal benchmark showed consistent improvements in grammar, coherence, and semantic similarity over baselines. This research lays the foundation for embedding-level privacy methods that can integrate seamlessly into generative systems.

References

[1] Ahmed Musa Awon, Yun Lu, Shera Potka, and Alex Thomo. Clusant: Differentially private and semantically coherent text sanitization. In *NAACL*, pages 3676–3693, 2025.

[2] Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. Remoterag: A privacy-preserving LLM cloud RAG service. In *Findings of the Assoc. for Computational Linguistics: ACL*, pages 3820–3837. ACL, 2025.

[3] Poomrapee Chuthamsatid, Shera Potka, and Alex Thomo. Word embedding bias in large language models. In *I-SPAN*, pages 267–282. Springer, 2025.

[4] Wei Liu, Baisong Liu, Jiangcheng Qin, Xueyuan Zhang, Weiming Huang, and Yangyang Wang. Fairness identification of large language models in recommendation. *Sci. Rep.*, 15(1):5516, 2025.

[5] Stephen Meisenbacher, Alexandra Klymenko, and Florian Matthes. Papillon: Privacy preservation from internet-based and local models. In *Proc. NAACL-HLT*, pages 3112–3128. ACL, 2025.

[6] Shera Potka, Isla Li, Jason Kepler, and Alex Thomo. Enhancing structural minority visibility in link recommendations. In *MEDES*, pages 55–68. Springer, 2024.

[7] Shera Potka and Alex Thomo. Community structure and coherence in digital humanities works. In *IISA*, pages 1–8. IEEE, 2023.

[8] Ke Xu, Shera Potka, and Alex Thomo. Gender and race bias in consumer product recommendations by large language models. In *AINA*, pages 245–258. Springer, 2025.

[9] Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Proc. ACL*. ACL, 2025. to appear.

[10] Yunqi Zhao. Fairness and diversity in recommender systems: A survey. *ACM Trans. Recomm. Syst.*, 3(2):1–38, 2025.

[11] Ziwei Zhu, Huaxiu Yao, Jinyang Li, Peizhao Zhao, Yaliang Li, Meng Jiang, and Suhang Wang. Path-specific counterfactual fairness for recommender systems. In *Proc. ICLR*, 2023.