# Graduate Research Plan Stament

The rapid proliferation of artificial intelligence (AI) across various aspects of society holds the promise of transformative benefits. However, this unchecked growth of AI has brought to the forefront a pressing concern—bias and fairness within AI systems. As AI infiltrates diverse domains, it has been responsible for alarming instances of discrimination, stereotyping, and exacerbated societal inequalities. This research proposal aims to address the pervasive issue of bias in AI by developing standardized strategies for mitigation. My objective is not only to acknowledge the problem but to actively contribute to the creation of ethical and fair AI systems that can be trusted across various domains. Given the real-world consequences of AI's unchecked biases, this research is both timely and necessary.

One significant instance of AI bias can be observed in algorithmic hiring practices. AI-driven hiring algorithms have demonstrated biases against certain demographic groups, such as gender. Notably, Amazon’s recruitment AI system was found to favor male applicants over female ones (Dastin, 2018), highlighting the technology’s potential to perpetuate gender disparities in employment. Another concerning case relates to facial recognition technology; when employed without proper safeguards, it can exhibit racial bias. Research has revealed that some commercial facial recognition systems are less accurate in identifying individuals with darker skin tones, potentially leading to misidentification and unjust consequences (Buolamwini & Gebru, 2018). Moreover, biases have infiltrated the criminal justice system, where AI algorithms are used for predicting recidivism rates. These algorithms have demonstrated biases against minority groups (Larson et al., 2016), raising concerns about unequal treatment and unjust sentencing, further exacerbating societal inequalities. These examples underscore the pressing need to comprehensively address AI bias.

In my pursuit of addressing bias in AI systems comprehensively, my research approach consists of several key phases. Phase 1 will be focused on the development of methodologies to standardize bias assessment within AI datasets and algorithms. Phase 2 will shift the research focus toward devising versatile bias mitigation strategies employing various techniques to systematically reduce bias in AI domains. Phase 3 will incorporate cross-domain validation through collaborations between industry to bridge the gap between my theoretical research and real world applications.

**Phase 1:** Initially, I will develop advanced methodologies dedicated to assessing and quantifying bias within AI datasets and algorithms. Drawing inspiration from the Fairness, Accountability, and Transparency (FAT) models (Barocas et al., 2019), my aim is to expand and refine these models, ultimately creating a standardized framework for bias assessment. This framework will serve as a crucial foundation for my research, enabling a thorough understanding of the various facets of bias that can exist within AI systems.

**Phase 2:** Once bias assessment methodologies are established, my research will pivot towards the development and evaluation of standardized bias mitigation strategies. These strategies will be versatile and adaptable, with the goal of mitigating bias across a wide array of AI domains. Leveraging techniques such as re-sampling (adjusting the sample distribution to balance underrepresented groups), re-weighting (assigning different weights to instances in the dataset), and adversarial training (employing adversarial networks to make models robust against biases), I intend to systematically reduce the impact of bias within AI systems. My objective is to equip AI practitioners and developers with a toolkit of effective strategies for addressing bias.

**Phase 3:** To ensure the real-world applicability and effectiveness of my standardized framework and bias mitigation strategies, my research will incorporate cross-domain validation. Collaboration with industry partners, including technology companies and AI developers, will enable me to conduct comprehensive testing and validation across various AI applications. By engaging in practical, industry-driven validation efforts, my research aims to bridge the gap between theory and real-world implementation, ultimately fostering the development of fair and ethical AI systems (Larson et al., 2019).

The outcomes of my research into mitigating bias in AI systems hold significant potential for both the field of AI and society at large. By developing standardized methodologies for bias assessment and mitigation, my research seeks to pave the way for fairer and more ethical AI applications across diverse domains. One of the primary societal impacts of this work is the potential to reduce harmful and discriminatory outcomes associated with biased AI algorithms. In domains such as criminal justice, finance, and healthcare, where AI systems play a pivotal role in decision-making, the reduction of bias can lead to fairer outcomes and increased trust in AI technologies.

Furthermore, my research will contribute to advancing the broader understanding of AI ethics, promoting transparency, and fostering accountability in AI development by actively engaging with academic, industry, and policymaking stakeholders. This collaborative approach will facilitate the adoption of ethical AI practices and the development of guidelines that ensure AI systems are designed with fairness and inclusivity in mind. Ultimately, the societal impact of this research extends to creating a more equitable and inclusive technological landscape, where AI systems are not only technically proficient but also built to benefit all individuals and communities, regardless of their background or characteristics. This transformation will reinforce the societal trust and confidence necessary for AI's responsible integration across various domains.

The successful completion of this research endeavor requires access to a spectrum of resources spanning data, computing infrastructure, and collaborative networks. Access to large and diverse datasets is crucial for developing comprehensive bias assessment methodologies and evaluating bias mitigation strategies across various AI domains. Additionally, substantial computational resources are essential for the implementation and testing of complex AI algorithms and models. Collaborative partnerships with industry stakeholders, including tech companies and AI developers, are indispensable for real-world validation and practical application of the research findings.

**References**.

1. Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women." Reuters.

2. Buolamwini, J., & Gebru, T. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification." Proceedings of Machine Learning Research, 81.

3. Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica