Case Study

Case 1: Biased Hiring Tool

The bias in Amazon's hiring tool stemmed from multiple interlinked causes beyond just historical data imbalance. The root issue was training data representation bias, as the model learned from resumes submitted over a decade, predominantly by men, embedding male-dominated language patterns and career trajectories as proxies for success. Proxy feature leakage further compounded this problem, where female-coded terms such as "women's chess club captain" became negative signals, as these terms deviated from patterns associated with past hires. Additionally, there was a model objective misalignment, since it optimized for replicating previous hiring decisions rather than identifying the most qualified candidates irrespective of gender, thus reinforcing systemic bias. Finally, had it been deployed, it risked creating a feedback loop, further entrenching gender disparities in future data and decisions.

To address these issues, three technically feasible fixes can be proposed. A data-centric approach would involve counterfactual data augmentation by introducing synthetic or real examples that reverse gender-coded terms or ensure balanced representation of female candidates with similar qualifications, though care is needed to maintain realism and avoid noise. A model-centric approach would apply fairness-constrained learning or adversarial debiasing, training the model to penalize gender-correlated representations; however, this requires careful tuning to avoid unacceptable trade-offs in predictive performance. Lastly, an evaluation-centric approach would implement post-processing fairness corrections, such as adjusting model outputs to equalize selection rates across genders through methods like Reject Option Classification, though this only mitigates output disparities without correcting internal model bias.

To evaluate these fixes effectively, concrete fairness metrics should include demographic parity to check for equal selection rates between genders, equal opportunity to ensure equally qualified male and female candidates have similar chances of selection, and disparate impact ratio to assess the proportional fairness of outcomes in line with regulatory thresholds. Together, these targeted strategies and metrics ensure AI hiring systems are not only technically sound but also ethically robust and compliant with fairness mandates in real-world deployment.

Case 2: Facial Recognition in Policing

Facial recognition in policing carries multifaceted ethical risks beyond wrongful arrests. Firstly, higher misidentification rates for minorities, especially Black and darker-skinned individuals, risk unjustified stops, detentions, and criminal records, as highlighted in studies such as the Gender Shades audit showing drastically lower accuracy for darker female faces. Secondly, disproportionate surveillance arises as these systems are often deployed in already over-policed communities, amplifying historical inequities and embedding systemic biases into technological infrastructures. Thirdly, chilling effects on civil liberties occur when pervasive facial surveillance deters individuals from exercising freedoms of assembly, protest, or religious

gathering, knowing their identities are constantly tracked. For example, in Hong Kong, protesters used masks to avoid facial recognition tracking, underscoring fears of political retribution. Finally, reinforcement of systemic discrimination occurs as these biased datasets and deployment patterns create a self-fulfilling cycle: more surveillance leads to more recorded offences in certain demographics, feeding back into predictive policing tools and justifying further targeted surveillance, thus exacerbating social injustice.

To deploy facial recognition systems responsibly in policing, robust, multi-layered policies are essential. Technically, implement strict accuracy and bias audits disaggregated by race, gender, and age before any operational use, with thresholds for deployment halting if disparities exceed acceptable levels. Governance-wise, establish independent oversight bodies with legal authority to review deployment decisions, assess proportionality, and halt use if ethical standards are violated. Transparency mechanisms should include clear public disclosures about where, when, and why facial recognition is used, ensuring affected communities are informed. Public engagement requires inclusive consultations, especially with marginalized groups disproportionately impacted, to assess societal acceptability and integrate community perspectives into deployment decisions. To ensure accountability and redress, law enforcement agencies must implement clear protocols enabling individuals to contest wrongful identifications, including immediate notification upon a facial recognition match, accessible appeal channels, and mandatory human verification before action is taken. Additionally, policies should define strict use-case limitations, banning real-time surveillance in public spaces unless under court-approved, narrowly defined, and time-bound circumstances. Together, these measures ensure that facial recognition, if deployed, aligns with democratic values, respects civil liberties, and mitigates technological amplification of systemic biases.

Ethical AI Report

This report documents a fairness audit conducted on the COMPAS recidivism risk score dataset using AI Fairness 360, as part of my commitment to deploying AI systems that are technically robust and ethically sound. Analysis revealed significant racial bias, with African-American defendants experiencing disproportionately higher false positive rates compared to Caucasian defendants. This means African-American individuals were more likely to be incorrectly classified as high-risk despite not reoffending, highlighting algorithmic unfairness that results in harmful misclassifications against a specific demographic group and exacerbates societal inequalities.

Such bias has profound ethical and societal implications within the criminal justice context. Inaccurate high-risk classifications can lead to harsher bail conditions, prolonged pretrial detention, or denial of parole for African-American defendants. Beyond individual harms, this undermines public trust in AI-assisted decision-making, perpetuates systemic racial disparities, and reinforces structural injustices within legal systems that are intended to uphold equitable treatment under the law.

To ensure this and future AI projects adhere to ethical AI principles, I adopt a structured approach. At the design phase, I proactively assess potential harms, misuse scenarios, and disproportionate impacts on vulnerable groups. For COMPAS, this involved examining how historical judicial biases might propagate through predictive models, leading to disparate

treatment outcomes. Throughout development, I operationalize fairness metrics such as equal opportunity difference and disparate impact ratio, applying pre-processing methods like Reweighing, in-processing interventions such as Adversarial Debiasing, and post-processing adjustments like Reject Option Classification. Transparency is ensured through clear model documentation, explainable outputs that enable stakeholders to understand decisions, and communication of model limitations.

Post-deployment, I establish continuous monitoring dashboards to track fairness and accuracy, define retraining triggers upon performance degradation, and designate ethical oversight roles to maintain accountability. In addition, I prioritize stakeholder engagement by involving domain experts, impacted communities, and legal professionals to co-design fairness objectives, validate outcomes, and embed feedback mechanisms for reporting misclassifications or harms, enabling iterative model improvements informed by real-world experiences.

To remediate the observed bias, I recommend data-level interventions such as fairness-aware sampling to address historical representation disparities, model-level interventions like adversarial debiasing to enforce parity in outcomes, and post-processing methods to adjust decision thresholds and reduce disparate impacts without requiring full model retraining. By systematically integrating ethical risk assessments, fairness-enhancing interventions, transparency mechanisms, continuous monitoring, and stakeholder engagement, I ensure AI systems I design and deploy remain fair, accountable, and socially responsible, thereby promoting equitable outcomes in all high-stakes domains where they operate.