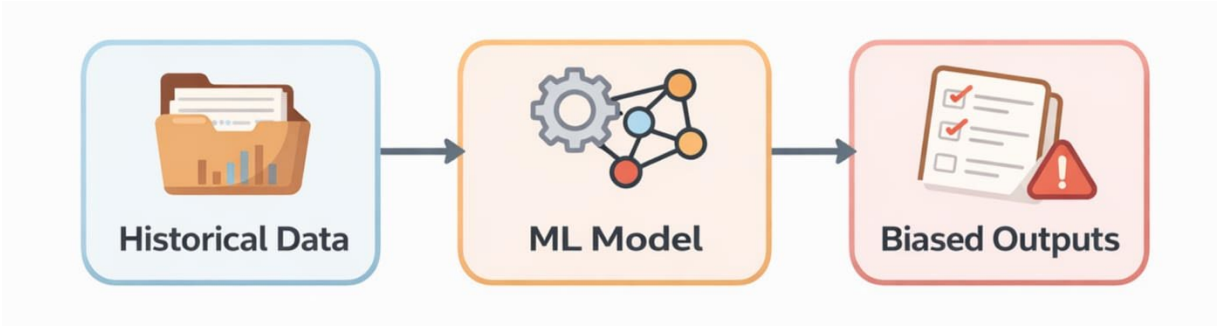
ML Ethics & Bias Case Study: Groups analyze a case where machine learning caused bias or unintended consequences (like biased hiring tools or unfair credit scoring) and propose solutions or guidelines to improve the model.

ML Ethics & Bias Case Study: Amazon's Biased Hiring Tool

Introduction

Machine learning models are now used in hiring, credit scoring and many other important areas, so if they are biased, real people can be hurt. One famous example we found is Amazon's AI-based hiring tool, which started to behave in a sexist way and had to be stopped. In this case study, we explain what Amazon tried to do, how the bias happened, what was wrong from an ethics point of view, and what we think could be done better in future.



What happened at Amazon?

Around 2014–2015, Amazon built an experimental recruiting tool to help screen lots of job applicants, especially for software and technical roles. The basic idea was: feed the model old résumés and past hiring decisions, train it, and then use the model to give new candidates a rating from one to five stars, similar to rating products. This rating was supposed to help HR quickly see which candidates looked "promising" according to past patterns.

The training data came from around ten years of historical résumés and hiring at Amazon, mainly for technical positions. After some time, the team noticed a problem: the system was giving lower scores to résumés that had hints of being female, for example the word "women's" in "women's chess club captain". It also preferred certain patterns that were more common in male candidates' résumés, such as particular colleges or keywords.

Amazon engineers tried to fix this by removing some obvious gender-related keywords, but they still could not be sure that the model would not find new indirect ways to be biased. In the end, the company decided not to use this tool for actual hiring decisions and dropped the project.

Why did the bias appear?

Biased training data

The biggest reason is that the model was trained on **past hiring data**, where most of the successful candidates for technical roles were men. The algorithm was basically told: "Look at these people we hired before and learn what a good candidate looks like." If the historical hiring already favored men (consciously or unconsciously), then the model simply learned to copy that behavior.

So the system was never "neutral". It learned from biased examples and then applied that bias in an automated way. Instead of correcting for past discrimination, it started to repeat and even strengthen it.

Hidden proxy features

Even when the team removed obvious gender words, the algorithm could still use **proxy features** that correlated with gender. For example, certain clubs, courses, or colleges might have been mostly male in the historical data, so the model could still favour those signals without ever seeing the gender field directly.

This shows that just dropping the "gender" column is not enough. In a high-dimensional résumé, many other features can indirectly carry gender information.

No strong fairness checks at the start

From the reports, it looks like the system was mainly evaluated on how well it matched previous hiring choices, not on fairness to different groups. The goal was to be accurate with respect to old decisions ("predict what our recruiters would have done"), not to check if men and women were treated equally by the model.

Because there were no strong fairness metrics or audits built in from the start, the model was free to learn any pattern that improved accuracy on historical labels, even if that pattern was sexist.

Who is responsible?

A big question in AI ethics is **who takes responsibility** when a model behaves badly. The model does not choose its own training data or goal; humans do. In this case, Amazon is still responsible for building and planning to use a model that copied old biases from their own hiring process.

We felt that this case shows why companies cannot just say "the algorithm decided". They need to own the outcome, monitor the system, and be ready to fix or shut it down if it harms certain groups, which they finally did here.

Ethical issues we see

Unfair treatment (gender discrimination)

The first ethical problem is pretty direct: the system started to **discriminate against women**. Women with similar skills and experience as men could get lower scores just because their résumés looked less like the old "typical" successful candidates. This is unfair and goes against basic ideas of equal opportunity.

In hiring, this type of bias is very serious, because it affects jobs, salaries, and long-term careers. When an AI system is biased, it can apply that bias to thousands of applicants in a very systematic way, which may be even worse than the bias of individual humans.

Lack of transparency for applicants

Another issue is **transparency**. Most candidates probably did not know that an algorithm was reading and scoring their résumé. They could not see their score, which features mattered, or how to challenge a bad rating. From the outside, it would just look like "we rejected you".

This lack of explainability makes it hard for applicants to understand what went wrong or to prove discrimination. It also makes it harder for regulators and external experts to check if the system follows non-discrimination rules.

Our ideas for solutions and guidelines

Based on this case, our group came up with some suggestions to make similar ML systems less biased in the future.

Better data and labels

Check and rebalance the dataset: Before training, look at the data and see if some groups (like men) are heavily over-represented among successful candidates. If yes, try to create a more balanced training set or give extra weight to under-represented groups so the model does not treat them as "rare and risky".

Don't blindly copy past decisions: If past hiring decisions were biased, using them as ground truth labels will copy the bias. Instead, if possible, use more objective outcomes (like later job performance) or at least adjust the labels to reduce obvious bias.

Build fairness into the model

Monitor sensitive and proxy features: Even if we remove the exact gender field, we should still test whether the model is treating groups differently based on other features that correlate with gender.

Use fairness metrics: Along with accuracy, measure things like selection rate or error rate for different groups (male vs female). If the model is much less favourable to one group, this should be treated as a bug that needs fixing, not as a "natural pattern".

Process and human oversight

Human in the loop, not human out: The model should help recruiters, not replace them entirely. For example, AI scores could be one input, but humans still review borderline cases and can overrule the model when needed.

Regular audits: The company should regularly test the model on new data to see if any bias has appeared or become worse. Independent audits (internal or external) are useful, especially in sensitive areas like hiring and credit.

Clear documentation and user notice: There should be basic documentation of what data is used, what the model is optimising, and known limitations. Applicants should at least be told that an automated tool is involved in screening.

Culture and team

Ethics awareness for ML teams: Engineers and data scientists should be trained to think about fairness and not just accuracy.

More diversity in the design team: Having people from different backgrounds in the ML team can help spot potential biases and unrealistic assumptions earlier.

Conclusion

At first, many people (including us) might assume that AI will be more neutral than humans. The Amazon hiring case shows that this is not automatically true. If we train a model on biased historical data and only optimize it to match past decisions, the model can learn the same bias and even make it stronger and more consistent.

In this case, the system ended up treating women unfairly in job screening because it copied patterns from a male-dominated past. From our point of view, this is a clear ethics problem around fairness, transparency and responsibility. The good part is that Amazon eventually stopped using the tool, but ideally, more checks should have been done earlier.

The main lesson we got from this group task is that building an accurate machine learning model is not enough. Teams also have to ask: “Accurate for whom?”, “Is it fair between groups?”, and “What happens to people if we are wrong?”. With better data practices, fairness checks, human oversight and clear rules, ML systems can support hiring in a more ethical and responsible way instead of quietly repeating old discrimination.