January 15, 2025: Lecture Notes - Motivation: Potential Harms of Data-driven Systems

Marzieh Hosseini, Nastaran Darabi

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Instructor: Abolfazl Asudeh; www.cs.uic.edu/∼ asudeh/teaching/archive/cs516spring25/

1 Introduction: The Dual Nature of Computer Science

Computer science, rooted in the foundational work of pioneers like Alan Turing and John von Neumann, has profoundly shaped modern society over the last century. Its innovations have touched nearly every aspect of life, revolutionizing healthcare, scientific research, and economies, while even influencing judiciary systems. The advent of the Internet, sophisticated algorithms, and diverse practical applications are testaments to the transformative power of this discipline. Computer science has undeniably improved communication, raised living standards, and enabled data-driven decision-making across various domains.

However, alongside these remarkable achievements, significant challenges arise, particularly concerning ethical considerations and unintended negative consequences of its outputs. While motivations for pursuing computer science are diverse—ranging from personal fulfillment and economic opportunities to societal betterment—it is critical to address the inherent flaws and biases in data-driven systems. Several prominent examples illustrate these issues:

- **Google Image-Tagging Service (2015):** Despite achieving high accuracy in general image labeling, Google's service notoriously mislabeled individuals due to a lack of diversity in its training data. This incident resulted in reputational damage and the eventual discontinuation of the service, emphasizing the need for representative datasets.
- **Nikon Camera Eye-Tracking Feature:** Nikon's eye-tracking technology, designed to prevent capturing photos with closed eyes, exhibited bias against individuals of Asian descent. This failure highlighted limitations in data diversity, revealing how technologies can encode societal biases.
- **HP Webcam Face-Tracking:** HP webcams with face-tracking capabilities performed poorly for darker-skinned users due to homogeneity in training data, showcasing the importance of inclusivity in algorithm design.

These examples, along with broader concerns about algorithmic bias, are explored in Cathy O'Neil's *Weapons of Math Destruction*, which examines how data-driven systems can perpetuate societal inequalities in areas like policing, online advertising, and social systems. Subsequent sections will delve into examples of such "weapons of math destruction," focusing on online advertising and recommendation systems, and examine algorithmic fairness through the COMPAS recidivism score case study.

2 Influence Maximization and Bias in Online Advertising

Social networks can be modeled as graphs, with individuals represented as nodes and their connections as edges. **Influence maximization** involves selecting a subset of individuals (seed nodes) to maximize the spread of information, such as advertisements, across the network. Algorithms for influence maximization often prioritize high-degree nodes (those with many connections) under the assumption they have the greatest influence.

However, this approach can introduce biases:

1. **Unequal Opportunities for Influencers:** Algorithms that prioritize degree centrality may favor prominent groups, marginalizing less-connected communities and creating unequal opportunities within the influencer landscape.
2. **Exclusion of Loosely Connected Users:** Algorithms may fail to reach individuals with fewer connections, often representing diverse or marginalized groups, leading to disparities in access to information and benefits.

While influence maximization aims to optimize reach, its reliance on network structure risks reinforcing existing inequalities in social networks.

3 Echo Chambers and Bias in Recommendation Systems

Recommendation systems, prevalent in social networks like Twitter, curate content based on user preferences, social connections, and interaction history. Common factors include:

- **User History:** Past interactions inform individual preferences.
- **Connections:** Content consumed by a user's network shapes recommendations.
- **Similarity:** Users with shared characteristics receive similar recommendations.
- **Geography:** Location influences locally relevant recommendations.

While these systems enhance user engagement, they can create **echo chambers** by reinforcing users' existing views and limiting exposure to diverse perspectives. Examples include:

- **Political Polarization:** Social media platforms can amplify political echo chambers, isolating users from opposing viewpoints and exacerbating polarization.
- **Spread of Misinformation:** In tightly-knit communities, misinformation spreads rapidly, unchecked by diverse perspectives.

Thus, recommendation systems risk amplifying biases, contributing to societal polarization and misinformation.

4 Algorithmic Bias and Feedback Loops: Perpetuating Inequality

Algorithmic bias is not limited to isolated incidents; it is a systemic issue that permeates numerous applications, including sensitive areas like predictive policing, automated hiring processes, and credit scoring. Understanding how bias is not just present but actively amplified requires examining the concept of the **bias feedback loop**. This cyclical process elucidates how seemingly neutral algorithmic systems can inadvertently perpetuate and even worsen existing societal inequalities. The feedback loop can be broken down into the following stages:

1. **Biased World State: The Seed of Inequality:** The cycle often originates from pre-existing societal biases and inequalities that are deeply embedded within our social structures, historical contexts, and cultural norms. These biases can manifest in various forms, including racial prejudice, gender stereotypes, socioeconomic disparities, and other forms of systemic discrimination. This biased world state is not a neutral starting point but rather a landscape already unevenly shaped by historical and ongoing injustices. For example, historical discriminatory housing policies can lead to segregated neighborhoods, which in turn can influence crime rates and policing patterns.
2. **Biased Data Collection: Reflecting and Amplifying Prejudices:** Data, often praised for its objectivity and factual nature, is actually shaped by the collection process and the context from which it originates. When data is gathered from a biased environment, it inevitably mirrors and can even amplify these existing biases. This amplification can occur through various mechanisms. **Selection bias** arises when the data collection process itself

systematically excludes or underrepresents certain groups. For instance, if crime data is primarily collected in specific neighborhoods due to targeted policing, it will overrepresent crime in those areas and underrepresent it elsewhere, regardless of the actual distribution of criminal activity. **Measurement bias** occurs when the way data is measured or recorded is systematically skewed for certain groups. For example, facial recognition technology has been shown to be less accurate for individuals with darker skin tones due to datasets being mainly composed of lighter-skinned faces, leading to biased performance in real-world applications like surveillance and security.

3. **Biased Model Development: Learning and Perpetuating Discrimination:** Machine learning models are trained on data to identify patterns and make predictions. When these models are trained on biased data, they inevitably learn and internalize the biases present in that data. The algorithms are designed to optimize patterns within the training data. However, if the data is skewed or discriminatory, the resulting model can perpetuate and even amplify these biases in its predictions and decision-making. For example, a hiring algorithm trained on historical data that reflects gender imbalances in certain professions may inadvertently favor male applicants over equally qualified female candidates. This happens because the historical data disproportionately associates maleness with success in those roles. The model itself is not malicious; it simply mirrors and codifies the biases it encountered during training.

4. **Biased Actions and Interventions: Reinforcing the Cycle of Inequality:** The predictions generated by biased models have real-world consequences; they often inform important actions and interventions. When these biased predictions are acted upon, they can create feedback loops that reinforce and amplify the initial biases. For example, in the context of loan applications, if a biased model predicts a higher risk of default for individuals from certain demographic groups—based on historical data that reflects past discriminatory lending practices—denying loans to these individuals based on the model's prediction will further restrict their economic opportunities. This lack of access to capital can perpetuate socioeconomic disparities, potentially validating the model's initial biased prediction in what becomes a self-fulfilling prophecy. Thus, the actions taken based on biased algorithms contribute to a cycle where inequality is not only maintained but actively intensified over time.

Predictive policing serves as a concerning example of this bias feedback loop in action. Algorithms that are trained on historical crime data often reflect biased policing practices, such as targeting specific neighborhoods or demographics. These algorithms are used to predict future crime hotspots, which then influences how police resources are deployed. As a result, there is an increased police presence and surveillance in areas identified as high-risk by the algorithm.

This heightened police presence can lead to a higher number of arrests in these areas, regardless of whether the actual crime rate has increased. Instead, the rise in arrests may simply be due to increased scrutiny and detection of crime. The newly generated arrest data further reinforces the algorithm's initial prediction, perpetuating the notion that these areas are high-crime zones.

This creates a self-reinforcing cycle that not only leads to a biased understanding of crime but also disproportionately focuses law enforcement attention on certain communities. This dynamic can erode trust between law enforcement and these communities, exacerbate existing social tensions, and perpetuate discriminatory policing practices.

5 Case Study: The COMPAS Recidivism Score and Algorithmic Fairness Dilemmas

The **COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)** algorithm, employed to predict the risk of recidivism among criminal defendants, has become a central case study in the debate surrounding algorithmic fairness. A landmark 2016 investigation by ProPublica brought to light significant racial disparities in COMPAS scores, igniting a widespread discussion about the ethical implications of using algorithmic risk assessments in the justice system.

- **ProPublica's Argument: Demographic Disparity and the Pursuit of Justice:** ProPublica's investigation centered on the finding that black defendants were significantly

more likely than white defendants to be incorrectly classified as "high-risk" by COMPAS, meaning they were predicted to re-offend but did not. Conversely, white defendants who did re-offend were sometimes incorrectly labeled as "low-risk." This disparity, ProPublica argued, pointed to a fundamental racial bias within the algorithm. Their argument strongly advocated for **demographic parity**, also known as statistical parity or disparate impact. From ProPublica's perspective, fairness in this context demands that the probability of receiving a "high-risk" classification should be roughly equal across different racial groups. They contended that disproportionately labeling black defendants as high-risk, even when inaccurate, has profound and unjust consequences, potentially leading to harsher sentences, denial of bail, and other detrimental outcomes that disproportionately affect minority communities already facing systemic disadvantages within the justice system. ProPublica's stance emphasizes that true justice requires algorithms to avoid perpetuating or exacerbating existing racial inequalities, and demographic parity serves as a crucial metric to assess and mitigate such disparate impacts.

- **Northpoint's Defense: Accuracy Parity and Predictive Validity:** Northpoint (now Equivant), the developers of COMPAS, countered ProPublica's claims by emphasizing **accuracy parity**. They argued that the algorithm was fair because it demonstrated comparable predictive accuracy across racial groups. Their core claim was that if COMPAS predicted someone to be "high-risk," the probability of that prediction being correct (i.e., the individual actually re-offending) was roughly the same for both black and white defendants. Northpoint's defense prioritized the algorithm's overall predictive validity and consistency across demographics. Their rationale for focusing on accuracy parity stems from the perspective that a risk assessment tool's primary function is to accurately predict risk, and fairness is achieved when this predictive power is applied equally effectively across all groups. From this viewpoint, if the algorithm is equally good at identifying high-risk individuals regardless of race, it fulfills its intended purpose fairly, even if the distribution of risk scores differs between groups.

- **Wisconsin Supreme Court's Decision: Equalized Odds and Error Rate Equity:** The Wisconsin Supreme Court, in a legal challenge to the use of COMPAS, ultimately sided with a perspective that emphasized **equalized odds**. The court focused on the algorithm's error rates, specifically false positives (incorrectly predicting high-risk) and false negatives (incorrectly predicting low-risk). Their decision hinged on the finding that COMPAS exhibited roughly equal false positive rates and false negative rates across racial groups. In other words, the probability of COMPAS mistakenly labeling a non-re-offender as high-risk was similar for both black and white defendants, and similarly, the probability of mistakenly labeling a re-offender as low-risk was also similar across racial groups. The court concluded that because the algorithm's error rates were roughly equitable, it met a reasonable standard of fairness and could continue to be used in judicial decision-making. The Wisconsin Supreme Court's emphasis on equalized odds reflects a concern for ensuring that the algorithm's mistakes, both in terms of false positives and false negatives, are distributed equitably across different demographic groups, minimizing the potential for disproportionate harm from algorithmic errors.

These differing fairness metrics highlight the inherent complexities in defining fairness in algorithmic systems. The COMPAS case starkly illustrates that there is no single, universally agreed-upon definition of fairness. Choosing to prioritize one metric often necessitates accepting trade-offs and potential compromises in relation to others. For example, aiming for perfect demographic parity might require sacrificing some degree of predictive accuracy or potentially leading to disparities in false positive or false negative rates. Conversely, optimizing for accuracy parity may result in unequal distributions of risk scores across groups, potentially violating demographic parity concerns. The selection of the most appropriate fairness metric is not a purely technical decision but rather a deeply ethical and value-laden choice that must consider the specific context of application, the potential societal impacts, and the competing values of accuracy, equity, and justice. The COMPAS dilemma underscores that algorithmic fairness is not a monolithic concept but a multifaceted challenge requiring careful consideration of different fairness principles and their implications.

6 Conclusion: Navigating the Complexities of Algorithmic Fairness

Computer science's potential must be balanced with ethical accountability. Cases like COMPAS highlight the trade-offs and complexities of fairness metrics, emphasizing the need for transparency, robust methodologies, and interdisciplinary collaboration. Addressing these challenges is essential for fostering equitable and just outcomes in a data-driven society.