Brookings Institution Study

One aspect of the Brookings Institution study, *“Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,”* that particularly stood out to me was the emphasis on transparency as a key strategy for mitigating algorithmic bias. The study outlines the importance of ensuring that algorithms, particularly those deployed in critical sectors like finance, healthcare, and criminal justice, are auditable and interpretable. I appreciated how the authors advocated for systems that can be examined not only by their developers but also by external stakeholders such as regulators and independent researchers. This focus on transparency is compelling because it acknowledges the real-world consequences of opaque algorithmic decision-making and supports the idea that accountability should be built into the design of these systems from the outset.

I strongly agree with the study’s argument that transparency is essential for fostering public trust and enabling meaningful oversight. Algorithms that operate as “black boxes” can reinforce existing inequalities without offering affected individuals any recourse. However, I felt the study could have pushed further on how to achieve meaningful transparency in practice. Simply making source code or model weights available does not necessarily make a system understandable to non-experts or even to other developers. More attention should have been given to strategies for communicating algorithmic logic to diverse audiences, including policymakers and the general public. For example, visualizations, plain language summaries, or standardized documentation protocols could be discussed more thoroughly as tools to bridge this gap.

One limitation of the study is that while it acknowledges the role of regulatory frameworks in enforcing transparency, it stops short of outlining specific legislative or institutional mechanisms that could support this goal. There is little discussion of how current laws may fall short or how international examples might inform U.S. policy. Additionally, the study could benefit from a deeper exploration of the tension between transparency and intellectual property rights, especially in the private sector. Balancing proprietary interests with the need for openness is a complex but crucial issue that deserves more nuanced treatment. Overall, while the emphasis on transparency is well-placed, the study would be strengthened by a more detailed roadmap for how this principle can be implemented in real-world systems.

Potential Uses of Facial Recognition

Given the current imperfect but improving state of facial recognition software, where (if at all) do you think it should and/or should not be used?

○

Considering the present state of facial recognition software, imperfect but improving, I believe it should be very limited and carefully regulated, especially when it is used in high-stakes applications such as law enforcement. For instance, facial recognition used to identify individuals sought for police questioning creates significant possibilities for misidentification, especially among people of color, as various studies have documented larger error rates for non-white subjects. It would thus lead to wrongful detentions or arrests and heighten current issues of bias within the criminal justice system. Till the technology becomes demonstrably reliable across all demographic groups, its use in these contexts should be either paused or heavily restricted to prevent harm.

This is in contrast to more benign uses, such as professors taking attendance in large lecture halls, which would be more acceptable if there were clear protocols of consent and protection of privacy. In these settings, the consequences of a misidentification are much lower, and students could have the option to opt out if they're uncomfortable. Even so, institutions should remain transparent about how the data is stored, who has access to it, and how long it's retained. Without these safeguards, even low-risk applications could contribute to a larger erosion of privacy and individual autonomy.

Suppose facial recognition improved to a point where identifications are nearly flawless. Then, where (if at all) do you think it should and/or should not be used?

If facial recognition technology reached a state of near-perfect accuracy, then the range of its potential applications would expand, but this again would not warrant its use without limits. Even in a flawless status, deploying the technology for government identification of suspected potential terrorists raises ethical and civil liberties concerns. Who defines "suspected"? What mechanisms are in place to prevent abuse or profiling? Notice that even accuracy would not negate the possibility of governmental overreach and surveillance creep. Near-flawless systems may be justifiable in highly specific, narrowly defined contexts, such as airport security checkpoints, if coupled with strong oversight, transparency, and opt-out options. In the end, technological accuracy cannot override the need for ethical scrutiny, accountability, and respect for individual rights.

Algorithmic Hiring

If you are applying for a job where there will be many applicants and an initial screening process to narrow the search down to a much smaller number who will be considered seriously, and the choices are to have that screening done solely by an algorithm or solely by humans, which would you prefer and why? Please give at least two reasons.

If I were applying to a job with a high volume of applicants, I would much rather the initial screening be done by humans and not just an algorithm. One key reason is that algorithms, while efficient, often rely on narrow criteria that can overlook important nuances in a candidate's background. For example, an algorithm might filter resumes based on keywords or gaps in employment, potentially excluding qualified applicants who bring valuable but unconventional experiences. A human reviewer is more likely to interpret context, such as a career change or time taken off for caregiving, with empathy and understanding that an algorithm cannot replicate.

A second reason is that such systems might contain embedded bias. While algorithms might be appreciatively improved by training on large data sets, the fact is that they can still pick up existing biases, particularly in training data representing historical inequalities in hiring practices. This leads to a systemic exclusion of underrepresented groups, even if the algorithm appears neutral on its face. A human, especially one trained in bias-free hiring, may have a greater ability to discern such bias in candidates and adjust accordingly during the screening process. Although humans are also fallible with regard to biases, there is at least the possibility of reflecting on one's own biases and changing course, which is not as easily built into an inflexible algorithmic model.

Do you feel there always needs to be a human involved in any hiring process, or are there cases where you feel the entire process is best conducted via algorithms? Please illustrate your response with at least one example.

I don't believe every part of the hiring process requires human involvement. For positions that are either not very specialized or with very standardized requirements, like temporary data entry positions or customer service, automated systems may be suitable to handle the whole process, at least if such an algorithm is designed to be fair and transparent. For example, a company hiring dozens of seasonal warehouse employees might use an algorithm to screen candidates based on basic qualifications (availability, location, and legal working status) and then automated interviews, even the scheduling of onboarding. In those situations, logistics are more important than deep candidate evaluation, and efficiency can outweigh the need for personal judgment.

In most professional or long-term positions, a hybrid approach that balances algorithmic assistance with human oversight will most likely constitute the most ethical and effective strategy.

Article Discussion

Google Researcher Timnit Gebru Said She Was Fired. Then Came the Backlash” (The New York Times, 2020)

What really struck a chord in this article was the way in which it framed the tension between corporate interests and the integrity of academia, particularly in the domain of research on ethical AI. Timnit Gebru, a highly regarded AI researcher at Google, was either fired or forced to resign after submitting a research paper that critically examined the bias and environmental impact of large language models, technologies that Google profits from. What really struck me most was how a company that publicly touts diversity and ethical AI appears to have shut down internal criticism once it clashed with business goals. The dismissal of Gebru mobilized significant backlash within the tech community, shining a light on how even top researchers can be censored when their work threatens the dominant narrative.

The key ethical issue here has to do with the suppression of dissent and lack of academic freedom in corporate research environments. When corporations fund research that has the potential to affect their bottom line, a conflict of interest, of sorts, is created. From an ethical perspective, with principles of transparency and accountability in full view, AI research should be open to critical inquiry, even when the findings make people uncomfortable. The Gebru case underlines how marginalized voices in tech, particularly Black women, are exceptionally more susceptible to retaliation and raises broader questions about whose voices are valued in shaping the future of AI.

Google's Ideological Echo Chamber" by James Damore (2017 Memo)

A controversial internal memo by former Google engineer James Damore argues that biological differences between men and women may explain the reason for some gender disparities in tech. The most intriguing thing to me in this article was how it framed the debate about diversity and inclusion as a threat to intellectual freedom. Damore said that Google had created an "ideological echo chamber" where dissenting views on diversity were not tolerated. His memo sparked heated debates both inside and outside the company about where the line falls between free expression and discriminatory speech in the workplace.

But the central ethical issue here is the abuse of scientific reasoning to justify inequality in the workplace. While open discussion and diversity of perspective are important, ethical frames such as justice and respect for persons remind us that freedom of speech does not license apparent rationalizations of systemic discrimination. In fact, despite apparent appeals to data, arguments like Damore's suppress another potent form of underrepresentation: social and structural barriers. For many, particularly women and underrepresented minorities, the memo created a hostile work environment and raised pressing questions about how companies balance free speech with the obligation to preserve an inclusive, respectful workplace.

Article Discussion

The article *"The Problem With Superintelligent AI"* by philosopher Nick Bostrom, published in *The New York Times* (October 31, 2019), explores the existential risks posed by the future development of artificial general intelligence (AGI) and superintelligent systems. Bostrom argues that while today's AI systems are still narrow and specialized, we must begin preparing for the possibility of AGI surpassing human intelligence. He warns that once AI systems are able to recursively improve themselves, they could quickly become uncontrollable and potentially act in ways misaligned with human values. The central concern is that we may not have sufficient mechanisms in place to ensure that these systems act safely, ethically, or in humanity's best interests once they exceed our cognitive capacities.

I found the article thought-provoking but somewhat one-sided in its framing. While I agree with Bostrom's point that proactive research into AI safety and ethics is essential, I felt the focus on hypothetical existential threats overshadows more immediate and concrete ethical challenges in AI. Issues like algorithmic bias, data privacy, surveillance, and the social implications of automation are already affecting millions of people today. Focusing too heavily on "doomsday scenarios" can divert attention and resources from the tangible harms that marginalized communities are already experiencing due to current AI systems. In short, I don't disagree with the long-term concerns, but I believe the article would have been more balanced if it acknowledged the pressing, real-world problems we face right now.

To improve the article, I think Bostrom could have connected his long-term concerns to the current ethical landscape in AI. Bridging the gap between present-day harms and future risks would make the argument more relevant to a broader audience. For example, he might have explored how flawed or biased systems today are a sign of what can go wrong if we don't get AI design and oversight right from the start. This connection could ground his philosophical insights in lived realities, helping readers understand why both short- and long-term thinking are critical in shaping AI's future responsibly.

Generative AI Concerns

Of the various ethical concerns that have been raised about generative artificial intelligence, which two concern you most, and why?

One of the most alarming ethical issues of generative AI is the spread of misinformation and the potential for large-scale manipulation. These systems, such as those used to create deepfakes or synthetic news content, can produce convincing but completely fabricated information at scale and speed. This poses serious risks to public trust, democratic processes, and social cohesion. As tools become increasingly available, malicious actors will leverage them to impersonate public figures, fabricate news events, or create viral content to spread false narratives. The power of AI to convincingly emulate reality makes the line separating the truthful from the fabricated increasingly difficult for an average person to distinguish.

Another major concern is the potential of generative AI to reinforce and amplify societal biases, especially in areas such as hiring, law enforcement, or education where AI is integrated into decision-making systems. Because these models are trained on large datasets drawn from the internet, they can easily absorb and replicate the biases found in the data they were trained on.

Thus, such systems may promote stereotypes, marginalize some groups, or espouse discriminatory premises, especially when these tools are used without careful control. Of course, bias in AI isn't anything new, but this challenge is much harder to detect and overcome given the scale and complexity of generative models.

Select one of those issues and answer: what do you feel should be done about this issue to make it less of a concern?

About the misinformation problem, I think regulation, transparency, and digital literacy can help mitigate its harm. Governments and tech companies should work together to establish certain basic labeling requirements on AI-generated content, such as watermarks or metadata that indicate if a piece of media was generated by a machine. Platforms should be responsible for identifying and removing malicious synthetic content in a timely manner. But equally important is educating people to think critically about digital content. Media literacy should be taught in schools, and awareness campaigns for adults should be run. By doing so, we will create societal resilience against AI-generated misinformation. This problem will not be overcome by technology alone; it will take a broad, collective effort to safeguard truth and trust in the digital age.