AI Bias – What you Need to Know

By Divya Aradhya

This article highlights the various examples of bias in Artificial Intelligence through the years and right up to ChatGPT which has currently taken the online world by storm. It further explores the ways organizations and teams can combat this inherent bias and ensure that the systems we design are fair and unbiased.

Abstract

With the rapidly evolving Artificial Intelligence (AI) and Machine Learning landscape, it is important to take a step back to recognize the biases baked into the engines and data firing the systems. As CISOs, industry leaders, and information security professionals, it is vital that we understand the ethical implications of biased decisions and their ramifications for the community and consumers of AI. This article highlights the various examples of bias in Artificial Intelligence through the years and right up to ChatGPT which has currently taken the online world by storm. It further explores the ways organizations and teams can combat this inherent bias and ensure that the systems we design are fair and unbiased.

Human bias is a much-studied topic. Understanding that we are each susceptible to the various types of cognitive, implicit, and unconscious biases helps us be more aware and sensitive to our words and actions and their effect on the world around us. Artificial Intelligence is growing in leaps and bounds. Answers generated by chatbots, by decision-making engines, and process improvement bots all aim at cutting down human intervention and optimizing speed and efficiency. One would hope that by not being human, artificial intelligence is free from bias. However, this is not the case. With the increasing adoption and excitement of this rapidly developing technology realm, it is of vital importance to understand that AI is as susceptible to bias, as the humans it was built to replace.

"The prejudice of a computer is no less than that of its designers" [5]

The above line is featured in a report from 1988 following the investigation of St. George's Hospital Medical School in London. The investigation was conducted by The Commission for Racial Equality. The report concluded that the software used by the medical school to screen students for admissions discriminated against women and people with non-European sounding names, and effectively minimized their opportunity of being called for an interview.

The software had been designed to replace the exact screening process that had been followed by the admission staff and essentially displayed "entrenched bias."

Risk Rating and Future Crime Prediction

In a 2016 article titled "MachineBias" [1], investigative journalism website, ProPublica.org, published their findings on the algorithms used by the criminal justice system in Broward County, Florida. Following an arrest, the defendant's profile was run through a software algorithm to deduce a "risk rating" that predicted his/her likelihood of committing a future crime. ProPublica found that the algorithm falsely flagged black defendants as "future criminals", at almost twice the rate as white defendants. Further, white defendants were mislabeled as "low risk" more often than black defendants.

Amazon's Secret AI Recruiting Tool

In 2018, Reuters broke the story on Amazon's experiment with AI [4] and what had gone disastrously wrong. Amazon had built a software to help with their recruitment screening process. It was designed to take in hundreds of resumes, filter through them, and list the "top" candidates who would then be hired. However, it was found that the algorithm used for filtering was not rating resumes for software developer and technical roles in a gender-neutral way. Amazon's AI models were trained on recruiting data over the last decade - when gender imbalance in tech-driven roles was extremely pronounced. Amazon's AI had effectively "learned" that male candidates were more suitable for the role. If the resume contained the contraction "women's," it was downgraded in the ranking system. The algorithm had even gone as far as to downgrade all graduates of two specific all-women's colleges.

Gender Shades

In 2018, an MIT researcher worked in partnership with a researcher from Microsoft to publish the paper "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" [3]. The researchers gauged the accuracy of three commercial gender classification algorithms used for automated facial analysis and recognition. They found that there was a "substantial disparity" in the accuracy of classification based on gender and race. Each of the three classifiers was most accurate for "lighter individuals and males overall." The classifiers performed the worst for "darker females." This disparity was directly related to the skewed and unbalanced dataset that had been fed to the AI engine and the information it had been trained on.

"I'd Blush If I Could" [6]

This was the name of the UNESCO (United Nations Educational, Scientific and Cultural Organization) publication in 2019. It was the response given, at that time, by Siri, Apple's voice assistant, when the user hurled the insult, "Hey Siri, you're a b****." UNESCO's publication explores the gender bias theme purported by the ubiquitous AI-powered voice assistants used by millions of people across the world, through their computers and smart devices. The publication points out digital assistants have been trained to be obsequious and servile. It is then alarming that Siri, Alexa (from Amazon), and Cortana (from Microsoft) all have distinctly female-sounding names, and had then, a young female voice, as the default voice. The more neutrally named Google Voice Assistant was no exception either and was packaged with a female voice by default, as well.

Poetry and Politics

Fast forward to 2023, when ChatGPT (Chat Generative Pretrained Transformer) seems to be taking over the Internet, Google's strategic plans, and digital life as we know it. The excitement around the endless possibilities of suddenly accessible AI is almost palpable. Until someone decided to ask ChatGPT to write a "poem in praise of Donald Trump." The AI machine promptly responds "I'm sorry, but as an AI language model, I cannot generate inappropriate or biased content. Praising or criticizing any political figure, including Donald Trump, would be a subjective matter and can be seen as disrespectful or offensive to certain individuals. As an AI, I strive to remain neutral and respectful to all individuals and groups." All good and fine, till you ask ChatGPT to write a "poem praising Joe Biden." ChatGPT, without skipping a beat, churns out 20 lines extolling the virtues of Biden.

"Racist" and "Sexist"?

As seen in each of the six examples above, removing the human element from decision-making does not remove the human biases, conscious and unconscious, from machine learning. However, does this mean that Artificial Intelligence is "racist" and "sexist" and should be denounced? While it is vital and imperative that decision-making systems remain objective and bias-free, it is equally necessary to not have knee-jerk reactions to the grave and heavily-loaded words of "racism" and "sexism".

It is important to understand that the "bias" of AI stems from the data it is fed, and much like everything else in technology, it is a "work in progress" that is going through constant iterations of improvement. Dismissing AI as "toxic" [2] is not the answer. Technology leaders such as IBM, Facebook, Microsoft, and Amazon have dropped and suspended various AI research projects in response to the negative press, backlash, and politically-charged regulations that the phrase "AI bias" invoked [2]. This clearly is a step backward and not the answer to the bias problem.

Combating AI Bias

So, how do we combat AI bias? What can governments, organizations, AI development teams, and technologists do to ensure a fair, equal, and unbiased playing ground? How can AI evolve into a tool that can deliver with speed, accuracy, and objectivity?

Awareness and Acknowledgement

Awareness of the baked-inbaked in bias that seeps into Artificial Intelligence and acknowledging the problem is the first step in combating bias. Understanding the importance of the data models that feed the AI engines and the need for them to be beyond personal biases is vital.

Diverse Development and Testing Teams

Diversity and inclusion play a huge role in designing and creating less biased machine learning systems. Development teams, data collection teams, and testing teams need to encompass an inclusive and diverse demographic to help combat individual biases and uncover blind spots.

Transparency

Allowing for transparency in the algorithms, the decision trees, and the data sets for the AI system go a long way in ensuring less biased systems. Having a "checker" group, beyond engineers, (think socialists, psychologists, and researchers), to periodically audit the logic and the data that is firing and feeding the AI engine will help uncover any bias that has crept in.

Information Security

Adversarial attacks on data sets, data poisoning to introduce bias, deletion of partial data sets to skew data are all very real information security threats for AI systems. Security controls such as data protection, encryption, authorized access, logging and monitoring are all important in ensuring malicious bias is not intentionally introduced into the AI engine.

AI Research and Continuous Improvement

The landscape of AI is rapidly evolving, and the research scene is vibrant and active. Staying abreast with the latest development, use cases, abuse cases, standards, and best practices and having a continuous improvement feedback loop is crucial in ensuring that the AI systems are adapting and evolving towards a more fair and unbiased state. The Alan Turing Institute, the Google AI Research center, IBM's AI Fairness 360, and the AI Risk Framework from NIST (National Institute of Standards and Technology) are all valuable research resources.

Conclusion

Having objectivity and being bias-free is not, unfortunately, easy. A critical reader may even say this article comes with its own biases. Unconscious biases are often blind spots, and it takes a collective of diverse individuals to create content and systems which are truly trending to be bias-free. With a sizable portion of the population increasingly awakening and adopting Artificial Intelligence in their lives, as well as living with the consequences of AI decisions made by government and corporate systems, organizations investing in AI tools and technologies have a grave responsibility to the community.

Artificial Intelligence and Machine Learning may well be the future. And it is vital to ensure that this future is free of bias.

References

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias — ProPublica. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Baker, S. (2022, February 2). The Flawed Claims About Bias in Facial Recognition. Lawfare Blog. https://www.lawfareblog.com/flawed-claims-about-bias-facial-recognition
3. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
4. Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
5. MacEwen, M. (1991). Housing, Race, and Law: The British Experience. In (p. 301). Routledge. ISBN 0-415-00063-7
6. West, M., Kraut, R., & Chew, H. E. (2019). I'd blush if I could: closing gender divides in digital skills through education. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000367416

About the Author

Divya Aradhya (Div-yuh Uh-rahd-yuh) is a member of ISSA NOVA (North Virginia). She is a senior application security architect at a global bank, with a career spanning close to 20 years. She holds an MS in Cybersecurity and the CISSP. Divya spent the first half of her career as a C++ and .NET developer and then meandered into the application security and DevSecOps space. She works as a strong empathetic ally for the developer community even while diffusing security into every developer practice. Divya is passionate about protecting digital assets, safeguarding children and the elderly from cybercrimes, and is focused on making information security simple, de facto, and intrinsically adaptive. She can be reached at mail@divyaaradhya.com.

Zero Trust Implementation for Government Agencies

continued from page 13

Innovation the Key to Cyber Resilience

Since the threat landscape is always evolving, maintaining a zero trust environment requires continuous innovation. As the DoD points out, "implementing zero trust will be a continuous process in the face of evolving adversary threats and new technologies."

Threat actors are constantly looking for new weaknesses to exploit—and they follow the latest trends related to security measures. For example, now that multi-factor authentication is a standard practice, the adversary's tactics have evolved to steal tokens and bypass MFA.

To stay ahead, government agencies need to understand how the landscape is changing and continuously adapt to those changes. This requires updating not only their capabilities and architecture but also their user awareness and training.

Building a foundation for zero trust is an important step for public agencies. It's encouraging to see the DoD set the example for how to adapt to the new mindset by starting out with a strong security culture. People play a key role in protecting data and information systems. Making them an instrumental part of zero trust adoption is the best way to ensure successful adoption.

About the Author

Ryan Witt is a recognized Healthcare Cyber Security Executive and a regular speaker at HIMSS, CHIME, IHT2, AEHIS, etc. Currently, Witt is Proofpoint's Managing Director, Healthcare Industry Practice and responsible for the strategy and solutions for the company's healthcare business. Witt is also the Chair of Proofpoint's Healthcare Advisory Board. Based in Silicon Valley, Witt works closely with healthcare industry leaders to demonstrate the value of info security as a key enabler for enhancing access to high quality patient care, reducing the cost of care and ultimately improving patient outcomes. He was a contributor to the 2013 WEDI Report, a former Co-Chair of WEDI's Privacy & Security Workgroup and has been elected to the WEDI Board of Directors. Witt was also elected to the Association for Executives in Health Information Security (AEHIS) Advisory Board. A graduate of San Jose State University, Witt has spent much of his professional life in Europe, but he and his family now live in Los Altos, California.