A Publicação “NIST Special Publication 1270 | Towards a Standard for Identifying and Managing Bias in Artificial Intelligence” faz uma abordagem sobre os tipos de *Bias* existentes em Inteligência Artificial, quão impactantes eles podem ser e como é possível lidarmos a fim de mitigarmos esses vieses. O *Bias* (viés) é descrito como um efeito que priva um resultado estatístico de representatividade ao distorcê-lo sistematicamente, diferente de um erro aleatório, que pode distorcer em qualquer ocasião, mas é compensado na média. Como base principal, o artigo é estruturado em cima de três categorias dominantes que geram o AI bias (i) Sistêmico, (ii) Estatístico e Computacional e (iii) Humano.

(i) Sistêmico

Vieses sistêmicos resultam de procedimentos e práticas de instituições que operam de maneiras que resultam em certos grupos sociais sendo favorecidos e outros sendo desfavorecidos. Isso não necessariamente precisa ser o resultado de qualquer preconceito consciente ou discriminação, mas sim da maioria seguindo as regras ou normas existentes. Racismo e sexismo são os exemplos mais comuns. Esses vieses estão presentes nos conjuntos de dados usados em IA e as normas, práticas e processos institucionais em todo o ciclo de vida da IA e em culturas e sociedades mais amplas. Um exemplo dado na publicação é o caso Além da identidade pessoal, os rostos humanos codificam uma série de traços notáveis como expressão não-verbal, indicadores de atração e seleção sexual e emoção. A tecnologia de reconhecimento facial (FRT) é usada em muitos tipos de aplicações, incluindo identificação de gênero, que compara distâncias morfológicas entre rostos para classificar rostos humanos por gênero. O grau de dimorfismo sexual entre homens e mulheres parece variar com a idade e o grupo étnico. Como um consequência, a precisão da identificação de gênero FRT pode variar em relação ao idade e etnia.

(ii) Estatístico e Computacional

Vieses estatísticos computacionais decorrem de erros que resultam quando a amostra não é representativa da população. Um exemplo dado no documentário “Coded Bias” é o viés da Inteligência Artificial utilizada para fazer o reconhecimento facial de índivíduos, mas que se mostrou falha, uma vez que as amostras fornecidas para treinar o AI eram majoritariamente compostos de homens brancos.

(iii) Humano

Os vieses humanos refletem erros sistemáticos no pensamento humano com base em um número limitado de princípios heurísticos e predição de valores para operações de julgamento. Ainda, esse pode surgir devido ao desejo de agradar um grupo de pessoas. Um exemplo é a identidade visual no marketing, por exemplo, em que uma pessoa atraente e arrumada passa uma impressão de ser bem-sucedida.

Referências

https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Documentário Netflix – Coded Bias

https://research.aimultiple.com/ai-bias/

Site

https://www.visualcapitalist.com/18-cognitive-bias-examples-mental-mistakes/