Máquinas Preconceituosas: Machine Learning e Ciências Sociais

O avanço das técnicas de aprendizagem de máquina e ciência de dados tem aflorado com vários aplicativos e algoritmos utilizados para reconhecimento de faces, casamento de padrões para construção de perfis de usuários, entre outras coisas. Os algoritmos estão cada vez mais interpretando comportamentos e dados sobre as pessoas. Porém, por mexer com pessoas, esses algoritmos acabam tendendo a ficar preconceituosas. Esses “bugs” de discriminação tem vários motivos, às vezes nos dados coletados para treinamento, ou por rotulação feita após a coleta. Isso pode levar a vários problemas éticos, de discriminação e acabar tendo repercussões muito negativas.

FALAR SOBRE COMO ESSES “BUGS” SÃO CAUSADOS

Em uma análise inicial sobre os danos causados por essa discriminação divide em dois principais tipos de danos causados: os danos de alocação e os danos de representação.

Os danos de alocação são as mais comuns associações com discriminação. O dano alocativo é quando um sistema aloca ou não aloca certo grupo para um oportunidade ou recurso. Por exemplo, aceitar ou não uma proposta de empréstimo em um banco para diferentes grupos de pessoas.

Os danos de alocação podem ser divididos em 5 tipos: Estereotipação, Reconhecimento, Difamação, Sub-representação e Exoneração.

A estereotipação acontece quando um sistema assume que um certo grupo sempre terá o resultado que a maioria terá. Como por exemplo: O google translate traduz palavras de línguas que não tem gênero com o gênero mais utilizado.

O reconhecimento acontece quando um grupo é apagado ou simplesmente invisível para um certo sistema. Como por exemplo um bug do Google Photos que colocava pessoas negras não como faces e sim como animais. Ou mesmo o reconhecimento de faces de câmeras Nikon que não reconhecia pessoas de tons de pele mais escuros.

A difamação acontece quando um sistema denigre certos grupos com rótulos, ou ofensas culturais. O exemplo citado anteriormente do Google Photos também é em parte difamação, assim como também acontece em sistema de auto sugestões ao sugerir coisas como “judeus devem”.

A sub-representação e a exoneração acontece quando grupos são removidos ou sub-representados em uma área do sistema, o que leva a má interpretação dos dados pelo sistema. Como, por exemplo, em uma pesquisa de imagens por CEOs no google a maioria dos resultados, na primeira página, eram homens brancos, tendo somente uma mulher.

Já os dano de representação são quando um sistema reforça as subordinações de certos grupos fazendo uma alocação má distribuída de acordo com raça, gênero, religião, etc.

APRESENTAR SOLUÇÕES!!!!!!!

* trabalho de sociologos e data scientists para definição de labels
* e para estudos exploratórios de novos comportamentos a partir dos dados.
* mais cuidado na analise e divisão dos grupos de minorias para análise mais a fundo desses grupos.
* Confirmation bias is the process of seeking or favoring information that confirms one’s preconceived hypotheses, while ignoring or downplaying contradictory evidence.
* First, it’s important to be aware of concepts like implicit bias and to be willing to challenge negative stereotypes about minority groups, even when this requires additional effort. Second, it we want others to use our methods fairly and responsibly, and without our hands-on involvement, we need a community-wide increase in commitment to the public understanding of science.