MC886

TP3: Ética em IA

Prof. Sandra Avila

Bruno Benitez de Carvalho - 167920
Eduardo Barros Innarelli - 170161

O que é ética em Inteligência Artificial?

A Comissão Europeia, através de seu grupo de peritos de alto nível sobre inteligência artificial (GPAN IA), publicou o documento *Orientações Éticas para uma IA de Confiança* [1], em que define ética em IA da seguinte forma: “A ética da IA é um subdomínio da ética aplicada que incide nas questões éticas suscitadas pelo desenvolvimento, pela implantação e pela utilização da inteligência artificial. A sua preocupação principal é identificar a forma como a IA pode melhorar ou suscitar preocupações para a vida das pessoas, quer em termos de qualidade de vida, quer de autonomia e liberdade humana necessárias para uma sociedade democrática.”

Tomando esta definição, é importante ressaltar os conceitos nela presentes. A ética é o estudo do conjunto de valores morais de um grupo, e a ética aplicada é o ramo que se dedica à análise de questões morais particulares na vida privada e pública [2]. Portanto, entende-se que a ética em IA é o estudo do impacto que os sistemas que se utilizam de IA trazem ou podem trazer para a sociedade humana, levando em conta o que se considera como moralmente aceitável ou inaceitável. O seu principal objetivo é identificar e antever preocupações que podem surgir devido aos sistemas que utilizam IA.

Pode-se encontrar muitas motivações para tal estudo. Por exemplo, o professor de ciência da computação do MIT, Joseph Weizenbaum, em 1976 alertou que a tecnologia em IA não deveria substituir pessoas em cargos como representantes de serviço ao consumidor, terapeutas, enfermagem, soldados, polícia e oficiais de polícia. O seu principal ponto sustentador para tal opinião é que as pessoas a cargo dessas posições precisam demonstrar empatia, algo que não é possível através de algoritmos.

Casos como o de Virginia Eubanks, que teve a requisição e serviços por parte e seu plano de saúde recusados por um programa de computador, porque ela tinha “indicadores comuns de fraude” [3], só reforçam a necessidade de um estudo profundo na área, e de que mudanças já precisam ser feitas.

Uma notícia atual

A notícia escolhida é intitulada *“Wrongfully Accused by an Algorithm”* (Injustamente Acusado por um Algoritmo) [4]. Foi publicada no dia 24 de Junho de 2020 no NY Times, e escrita pela jornalista especializada em tecnologia Kashmir Hill. A autora relata um caso ocorrido em Janeiro de 2020, no estado de Michigan dos EUA, em que um algoritmo de reconhecimento facial acusou erroneamente o cidadão afro-americano Robert Julian-Borchak Williams de um crime que não cometera.

O caso foi reconhecido como a primeira prisão sucedida de uma falha em um sistema de reconhecimento facial nos EUA. Tais algoritmos são utilizados pela polícia há

mais de duas décadas em território norte-americano, e não surpreende que a primeira vítima de um erro dessa natureza seja um homem negro. Estudos recentes do MIT e do NIST mostraram que a acurácia desses sistemas é maior para homens brancos do que para outras demografias, o que é típico de algoritmos enviesados, treinados com dados pouco diversos.

Esse acontecimento levantou o debate sobre o uso do reconhecimento facial no campo jurídico. A advogada Clare Garvie defende testes mais rigorosos de acurácia e enviesamento para os sistemas antes desses serem aplicados, visto que a própria empresa responsável pelo algoritmo em questão admite que essas métricas não eram formalmente calculadas.

Ainda assim, nunca um sistema de reconhecimento facial vai ser 100% preciso. Por isso, provedores de tecnologia e oficiais da lei enfatizam que uma predição realizada por um algoritmo deve servir de pista, e não de evidência. O fato de confiarem cegamente no resultado retornado pelo algoritmo a ponto de levarem Williams a prisão denuncia o baixo preparo das instituições policiais para lidar com tecnologias desse tipo.

Essa notícia também reforça o debate acerca de se a IA está colaborando para uma sociedade mais justa e unida, ou para a segregação. Os EUA são um país historicamente conhecidos pelo problema da segregação racial. O fato de os algoritmos serem mais ineficientes em reconhecimento facial com pessoas de pele negra, e continuarem a ser usados é extremamente alarmante. O caso do Sr. Williams é o primeiro a ser tornado público, mas pode não ser o primeiro que aconteceu. A humilhação que ele sofreu diante de sua esposa e filhas, e a forma como os policiais os trataram provavelmente causaram danos que carregarão por toda a vida. Mais uma vez, cabe a reflexão sobre a seriedade que deve ser atribuída ao tomar a decisão sobre utilizar IA em áreas tão sensíveis.

Artigo científico

O artigo científico a ser apresentado é *“Transparent Interpretation with Knockouts”* [5]. Seu autor é Xing Han, da University of Texas at Austin. Foi publicado inicialmente no arXiv.org em 1 de novembro de 2020, sofrendo uma revisão em 4 de novembro de 2020. Por se tratar de um trabalho muito recente, ele ainda não possui citações do Google Acadêmico.

O artigo traz à atenção o problema da transparência de modelos de Machine Learning, uma vez que eles são utilizados, muitas vezes, por pessoas que não tem um conhecimento profundo no assunto. Por isso, torna-se necessário um método de explicação para tais usuários finais de como e porquê o modelo que elas utilizam tomou determinada decisão. A solução proposta pelo autor baseia-se em identificar a(s) instância(s) do conjunto de treinamento que são responsáveis por tal decisão. Isto significa que se a(s) instância(s) do conjunto de treinamento fossem retiradas, a decisão do modelo seria diferente.

Adicionalmente, o autor utilizou o método desenvolvido para resolver outro problema de ética em IA, o de identificação de “dados contaminados” (Data Poisoning Detection). Contaminação de dados é um tipo de ataque a um modelo de Machine Learning, em que dados muito parecidos aos já presentes no conjunto de treinamento são adicionados nele mas com os identificadores invertidos. A solução do autor se mostrou eficiente em detectar tais dados contaminados também.

A solução apresentada é a mesma, independente do modelo utilizado responsável pela predição que procura-se justificar, e pode ser resumida pelo seguinte algoritmo:

Algorithm 1 Proposed method for finding a set of supports

Given dataset $\mathcal{D} = \{z_i := (x_i, y_i)\}_{i=1}^{n}$, the test point x_k and an empty set $\mathcal{B}$.
Initialize $\mathcal{D}' = \mathcal{D}$.
repeat
  Optimize unconstrained classifier $\hat{\pi}$ using equation (1) on $\mathcal{D}'$;
  Optimize constrained classifier $\hat{\pi}_c$ using equation (2) on $\mathcal{D}'$;
  Select data instance by $\hat{i} = \arg\min_i(\delta(z_i, \hat{\pi}_c) - \delta(z_i, \hat{\pi}))$;
  Add data instance $z_{\hat{i}}$ to set $\mathcal{B}$.
  Remove data instance $z_{\hat{i}}$ from $\mathcal{D}'$: $\mathcal{D}' := \mathcal{D}'/z_{\hat{i}}$.
until the difference between $L(\hat{\pi}_c)$ and $L(\hat{\pi})$ is statistically insignificant.

Figura 01: Algoritmo para detecção de instâncias de treinamento responsáveis por uma decisão. As equações e notações são referenciadas no artigo.

O autor ainda apresenta otimizações em seu algoritmo, além de aplicá-lo em um modelo real utilizado em decisões de linha de crédito.

O grupo conclui que esta é uma solução válida para um problema importante na área de Machine Learning e IA, uma vez que estas tecnologias estão sendo cada vez mais usadas, e cada vez em áreas mais delicadas. Eventualmente, estes modelos podem falhar, e é de suma importância saber porque houve tal falha, tendo a possibilidade de corrigi-la o mais breve possível, além de dar explicações satisfatórias para usuários finais, que não necessariamente são peritos no assunto de IA e Machine Learning.

Links e Referências

[1] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
(Documento em .pdf disponível em português. Citação na página 11, parágrafo 32)
[2] https://iep.utm.edu/ethics/
[3] https://metalurgicos.org.br/noticias/algoritmos-e-desigualdade/
[4] https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html
[5] https://arxiv.org/abs/2011.00639