Ética & Inteligência Artificial

Daniel Gardin Gratti 214729
Beatriz Cardoso Nascimento 247403

30 de março de 2023

1. Defina Ética em Inteligência Artificial.

Ética, do grego *ethos*, é um ramo de estudo da filosofia que tenta responder à difícil pergunta "Como devemos agir?" de forma lógica e humana, definindo costumes e regras que são tomadas como corretas ou erradas, com o objetivo melhorar o bem-estar e a convivência da humanidade. Essa preocupação surgiu com a evolução da socialização humana, que necessita de um código de conduta que mantenha a união e a civilidade. Nesse sentido, a Ética em Inteligência Artificial (IA) é um ramo da ética aplicada que busca entender o papel da IA na nossa sociedade atual e discutir formas responsáveis para a utilização dessa tecnologia.

Algoritmos que tentam realizar tarefas de necessidade cognitiva tal qual um humano são chamados de Inteligência Artificial. A ética em IA se apresenta como um debate de múltiplas perspectivas e contribuições heterogêneas em torno de um conjunto de perguntas canônicas, e implanta ferramentas normativas para análise ética e tomada de decisão [1]. Algumas das perguntas centrais acerca desse assunto são:

- Como podemos entender, explicar e controlar - se possível - o funcionamento interno de sistemas complexos de IA? [2]
- Quem deve ser responsabilizado por danos causados pela utilização de sistemas de IA? [3]
- Como sistemas de IA podem refletir discriminações, viéses e injustiças sociais existentes em seus dados de treinamento, dessa forma exacerbando-os? [4]
- Como a privacidade pode ser protegida, num contexto em que dados pessoais podem ser tão facilmente coletados e analisados? [5]

A Inteligência Artificial é sempre alvo de euforia - ou é a salvação para os problemas, ou é nosso caminho para a revolução das máquinas - parte desta euforia é justamente provinda da falta de conhecimento sobre o funcionamento dos algoritmos e uma visão crítica sobre seus resultados. O próprio termo "Inteligência Artificial"pode induzir ao erro pois, em primeiro lugar, algoritmos de IA podem processar quantidades imensas de dados e realizar tarefas altamente especializadas, mas isso pouco tem a ver com a inteligência criativa e intersectorial de humanos. Em segundo lugar, sistemas de Inteligência Artificial são altamente naturais, pois provêm de quantidades massivas de recursos planetários, força de trabalho humana e dados coletados e/ou produzidos por humanos para desenvolver e executar esses algoritmos.

De acordo com Evert Haasdijk[6], especialista em IA na empresa Deloitte, a ética em Inteligência Artificial requer transparência e responsabilidade sobre as soluções que encontramos. Por um lado, a ética deve partir dos programadores - detectar vieses nos dados, analisar criticamente a fonte destes dados, respeitar a privacidade e uso correto de dados de terceiros são todas ações que impulsionam positivamente a Inteligência Artificial a tornar-se mais ética e bem conviver com a sociedade. No entanto, focar apenas nas tecnologias de IA é insuficiente para capturar o que está em jogo moralmente com a utilização desses algoritmos. O contexto social mais amplo, a dinâmica política, econômica e cultural em níveis domésticos e globais devem ser vistos e analisados como os âmbitos onde a IA exerce sua influência, e também devem estar em foco quando se discute Ética em Inteligência Artificial.

2. Apresente uma notícia recente de um problema de Ética em IA.

Para discussão nessa questão, trazemos a notícia de título *These robots were trained on AI. They became racist and sexist*, publicada por Pranshu Verma no Washington Post [7]. Sua principal tese é a de que vieses racistas e sexistas presentes em sistemas de IA podem resultar em robôs que se guiam por esses preconceitos para realizar suas ações. A discussão sobre algoritmos de IA que exibem vieses prejudiciais para alguns grupos já têm se tornado pública há algum tempo. Sistemas para predição de crimes que injustamente apontavam pessoas negras e latinas como potenciais ofensores, algoritmos que tinham dificuldade para identificar pessoas racializadas ou que classificavam mais comumente mulheres negras e latinas como " donas de casa" são alguns dos exemplos que vieram à tona nos últimos anos.

Por enquanto, os robôs que utilizam IA têm escapado do escrutínio do público e têm sido percebidos como neutros pois realizam tarefas de natureza limitada, como reestocar suprimentos em um armazém. No entanto, geralmente o desenvolvimento de novos softwares ocorre baseado em códigos antigos, ou seja, quando os robôs começarem a realizar atividades mais complexas, seus códigos fontes conterão vieses prejudiciais para alguns grupos. Por conta disso, especialistas em ética da tecnologia e pesquisadores alertam que a adoção rápida dessa nova tecnologia pode resultar em consequências imprevistas no futuro.

Enquanto isso, a expectativa de crescimento da indústria de automação é de $18 a $60 bilhões até o fim da década, impulsionada em grande parte pela robótica. Nos próximos cinco anos, o uso de robôs em depósitos tem chance de crescer em 50% de acordo com o Material Handling Institute.

Os tipos de vieses incluídos no software desses robôs podem ter implicações no mundo real. Imagine, por exemplo, que um robô é apresentado a duas bonecas, uma negra e uma branca, e é questionado sobre qual a mais bonita. Com base nos algoritmos atuais, é mais provável que ele escolha a branca e isso é extremamente problemático. Outra situação possível é a de um robô estoquista que precise coletar brinquedos, cujas caixas frequentemente apresentam fotos de pessoas. Esse robô está mais apto a reconhecer embalagens com pessoas brancas que com pessoas racializadas.

A conclusão da notícia é que, segundo os estudiosos que foram entrevistados, é praticamente impossível ter sistemas de IA que não sejam baseados em dados enviesados. No entanto, as empresas responsáveis por eles não devem desistir: devem auditar os algoritmos que usam e diagnosticar as formas que eles exibem comportamentos falhos, criando maneiras de identificar e resolver essas situações. Isso não é um problema atualmente por conta das formas que os robôs vêm sendo usados, mas pode se tornar uma questão daqui a uma década. No entanto, se as empresas esperarem para implementar mudanças, pode ser tarde demais.

3. Apresente um artigo científico recente de uma solução para um problema de Ética em IA.

Discutimos nesta questão o artigo *Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach*, publicada por Suyun Liu e Luis Nunes Vicente [8]. A solução proposta pelo artigo está em resolver problemas éticos de inteligência artificial por meio de critérios de *fairness* adicionados a otimização dos modelos. A ideia está no estudo do trade-off entre a acurácia do modelo, que pode carregar vieses, e métricas de *fairness*, por meio de otimização multi-objetivo entre estas duas métricas.

A crítica principal ao processo de otimização de modelos de machine learning apresentada está na falta de consciência do modelo quanto a características sensíveis (e.g. grupo social, etnia, gênero), que podem ser aprendidas muitas vezes com vieses e tratar injustamente alguma parcela não-dominante da população. Para avaliar e promover a justiça destes modelos quanto a todas as população os autores exploram técnicas de reduzir a injustiça de modelos durante o processo de treinamento, ao contrário de técnicas de pré e pós-processamento do modelo.

Para isso, o objetivo do modelo não se pauta apenas em maximizar a acurácia do preditor, mas também maximizar medidas de *fairness* no modelo, que levam em conta características sensíveis dos dados. Muitas vezes estes dois objetivos são conflitantes, tornando necessário uma otimização de ambos por meio de uma abordagem multi-objetivo. A ideia é semelhante a regularização de redes neurais, em que um segundo objetivo é introduzido a função de perda, que tenta minimizar ambos os objetivos ao mesmo tempo, garantindo uma melhor generalização do modelo.

Para medir o *fairness* de um modelo, critérios de acurácia em diferentes grupos com características sensíveis podem ser adicionados a uma função de perda multi-dimensional em que o objetivo do otimizador é gerar um conjunto de soluções eficientes que mostram um trade-off entre os critérios selecionaods.

Estes critérios podem ser simples medidas de acurácia, quanto critérios mais profundos, como os de *disparate impact*, *equalized odds*, que tentam manter as probabilidades balanceadas dentro de cada grupo sensível, evitando uma disparidade de aceitação devido as características sensíveis.

Os métodos apresentados tentam integrar métricas de fairness nos modelos como objetivos que o otimizador tenta minimizar. No entanto, ao invés de criar uma função de perda única, o otimizador é treinado para encontrar um conjunto de soluções que tenta minimizar ambas os objetivos ao mesmo tempo, formando uma fronteira de Pareto. As soluções são providas ao especialista e um modelo é escolhido conforme seu trade-off entre erro e fairness seja o mais adequado.

4. Apresente e discuta a resposta do ChatGPT sobre a notícia escolhida na questão anterior.

Ao utilizarmos o ChatGPT com o prompt "Comente em até 500 palavras e na língua portuguesa, a notícia abaixo:", obtivemos a seguinte saida:

"Pesquisadores de várias universidades, como a Johns Hopkins University e o Georgia Institute of Technology, conduziram um estudo no qual verificaram que robôs programados com inteligência artificial possuem preconceitos raciais e sexuais. Os robôs foram instruídos a escolher as imagens de pessoas com determinadas profissões, tais como cozinheira ou médica, e decidir quem era um criminoso, baseando-se em imagens e legendas. Reiteradas vezes, os robôs escolheram a imagem de um homem negro como sendo o criminoso. Da mesma forma, palavras como "dona de casa"e "faxineiro"foram associadas a imagens de pessoas de cor. Segundo os pesquisadores, isso pode ser o primeiro indício empírico de que robôs são capazes de ser sexistas e racistas.

Esse estudo vem alertando sobre as consequências do rápido avanço tecnológico e da adoção de novas tecnologias sem a devida atenção aos seus possíveis preconceitos. A automatização tem sido bastante incentivada por empresas para tarefas que vão desde cuidados com pacientes em hospitais até estoque de prateleiras, e a pandemia tornou a demanda por essa tecnologia ainda mais urgente. Contudo, os pesquisadores alertam que esses robôs podem ser construídos sobre raízes defeituosas, o que pode levar a problemas futuros.

A indústria de automação deve crescer de US $ 18 bilhões para US $ 60 bilhões até o final da década, impulsionada em grande parte pela robótica. Porém, esse rápido avanço tecnológico também pode gerar problemas. Vários estudos recentes documentaram casos de algoritmos de inteligência artificial com viés racial, como sistemas de reconhecimento facial que não conseguem identificar corretamente pessoas de cor. Com os robôs, os preconceitos podem passar despercebidos por mais tempo, devido à percepção de que robôs são mais neutros em comparação com sistemas algorítmicos.

A pesquisa mostra que é necessário fazer análises mais aprofundadas dos algoritmos de IA para detectar esses preconceitos e corrigi-los. É essencial que empresas que adotam essa tecnologia realizem auditorias em seus algoritmos e criem maneiras de diagnosticar e melhorar seus problemas, evitando assim prejuízos futuros."

Ao analisar criticamente a saída do ChatGPT, podemos observar que o texto é condizente com a notícia em linhas gerais. No entanto, apesar de solicitarmos que ele "comente" a notícia, o resultado foi basicamente um resumo da reportagem, o que não condiz exatamente ao comando dado. Neste contexto, o ChatGPT mostrou-se limitado ao conhecimento fornecido, o que ocasionou uma transcrição praticamente idêntica do artigo original.

Além disso, um erro interpretativo na frase "Os robôs foram instruídos a escolher as imagens de pessoas com determinadas profissões, tais como cozinheira ou médica, e decidir quem era um criminoso, baseando-se em imagens e legendas" foi cometido. Na reportagem, estas duas tarefas, de classificação de criminosos e atribuição de profissão, são realizadas separadamente, sem qualquer intersecção entre elas, apesar da conclusão de ambas ser semelhante quanto à injustiça cometida pela inteligência artificial.

Em conclusão, ainda que a inteligência artificial produza resultados impressionantes, seu comportamento ainda carrega vieses e não é a prova de erros, além de possuir limitações, como a saída gerada acima revela. Apesar do comportamento mimicamente humano destes algoritmos, sua personificação quanto ao seu funcionamento pode gerar euforia e também ofuscar algumas destas imperfeições. É necessário um olhar crítico a respeito do mecanismo destas ferramentas, visando uma maior transparência e minimizando problemas éticos que possam ocorrer.

Referências

[1] Jan-Christoph Heilinger. The ethics of ai ethics. a constructive critique. *Philosophy & Technology*, 35(3):61, 2022.

[2] Hendrik Kempt, Jan-Christoph Heilinger, and Saskia K Nagel. "i'm afraid i can't let you do that, doctor": meaningful disagreements with ai in medical contexts. *AI & SOCIETY*, pages 1–8, 2022.

[3] Robert Sparrow. Killer robots. *Journal of applied philosophy*, 24(1):62–77, 2007.

[4] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code, 2020.

[5] Aniceto Perez y Madrid. Privacy is power. why and how you should take back control of your data, 2021.

[6] Evert Haasdijk. Transparency and responsibility in artificial intelligence. `https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf`, 2019. [Online; accessed 15-March-2023].

[7] Pranshu Verma. These robots were trained on ai. they became racist and sexist., 2022. `https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/`, Last accessed on 2023-03-20.

[8] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022. 37 citações no Google Scholar.