Ética em Inteligência Artificial

MO434/MC934: Deep Learning - 2s2020 - Trabalho 2

Mauricio Araujo, 184477 | Raysa Benatti, 176483

1 Introdução: O que é Ética em Inteligência Artificial?

O conceito de ética relaciona-se aos conceitos de moral e de justiça — embora não sejam sinônimos. Debruçar-se filosoficamente e em profundidade sobre suas definições é uma tarefa impraticável neste trabalho; focamos, então, em noções de ética aplicada à inteligência artificial trazidas pela literatura.

Para Cointe et al. [1], a ética é uma "disciplina normativa, prática e filosófica sobre como humanos devem agir e ser em relação a outrem"; ela "usa princípios éticos para conciliar moral, desejos e capacidades do agente" (tradução nossa). Os autores sistematizam três grandes noções de ética: **de virtude** (ética associada à presença de valores virtuosos), **deontológica** (ética associada ao respeito por obrigações e permissões) e **consequencialista** (ética associada à moralidade das consequências de uma decisão). Em inteligência artificial, os agentes seriam construídos com uma ética individual baseada em algum(ns) desses preceitos.

A inteligência artificial, desde sua estruturação como uma área do conhecimento, lida com questões existenciais sem limites bem definidos — o que a coloca em constante confronto com a ética [2]. É recorrente a discussão sobre o desafio e a necessidade de regulamentar quais preceitos éticos devem guiar a construção de sistemas computacionais, além da incorporação de noções de ética na educação sobre a área [2].

O Código de Ética e Conduta Profissional da *Association for Computing Machinery* [3] foi atualizado em 2018 para incluir, dentre outras mudanças, menção explícita à necessidade de que riscos potenciais de sistemas de aprendizado de máquina sejam cuidadosamente avaliados [4]. O documento cita, ainda, princípios éticos gerais que devem guiar a área. Privilegiam-se as noções de evitar danos e contribuir para a sociedade e o bem-estar humano, reconhecendo-se que, em computação, todos os indivíduos são *stakeholders*. Embora todas as dimensões de ética estejam presentes no documento, a ética consequencialista parece prevalecer — o que faz sentido em um código de conduta voltado à prática profissional.

A UNESCO, em documento sobre o tema [5], também destaca a noção de ética como balizadora para o interesse público. O estudo descreve potenciais implicações da inteligência artificial em diversos campos: educação, ciência, cultura, comunicação e informação, paz, África, gênero e meio ambiente; cumpre, assim, seu papel como organização internacional, reconhecendo a ubiquidade da área, explorando suas dimensões multiculturais e não subestimando o desafio de lidar com a questão. Entendendo que o campo de estudo da inteligência artificial não é neutro — como nenhum é —, compreende-se a importância de prever, evitar, mitigar e remediar seus impactos sobre a sociedade.

Aqui, privilegiam-se perspectivas de ética consequencialista e de virtude: segundo as recomendações do documento, sistemas éticos seriam aqueles conformes a princípios de direitos humanos, inclusão, prosperidade, autonomia humana, explicabilidade, transparência, instrução, responsabilidade, *accountability*, democracia, boa governança e sustentabilidade.

Entendemos que as definições acima constróem satisfatoriamente uma noção de ética em inteligência artificial, que, concluímos, está relacionada à compreensão multifatorial de suas possíveis implicações sobre diferentes elementos da sociedade.

2 Descrição de caso

Um dos desdobramentos do estudo e aplicação de princípios éticos à computação é o reconhecimento acerca de vieses indesejáveis embutidos em sistemas de inteligência artificial usados para tomada de decisão — seja pelo uso de dados enviesados para informar o modelo, seja pela inadequação do modelo em si, seja por demais fatores. Recorrentemente, pesquisadores e profissionais alertam que tais sistemas podem criar, reforçar ou aumentar vieses negativos sobre determinados atores.

Isso é particularmente grave quando atinge grupos histórica e sistematicamente desfavorecidos, na medida em que pode efetivamente impedir seu acesso a direitos; nesse processo, gera-se desigualdade,

fomenta-se exclusão e ameaça-se a diversidade cultural de uma sociedade [5]. Trata-se, enfim, de um exemplo canônico sobre a importância de alinhamento da inteligência artificial a princípios éticos, sob pena de agravo ao interesse público.

Publicada em fevereiro de 2020 pela CBS News, a matéria *Is artificial intelligence making racial profiling worse?* [6] ilustra um caso típico sobre a questão. A notícia descreve como se deu o processo de adoção, por parte do Departamento de Polícia de Los Angeles, do *PredPol* — um sistema de policiamento preditivo baseado em inteligência artificial. O sistema extrai padrões de dados históricos sobre a ocorrência de crimes na cidade para inferir em quais localidades haveria maior probabilidade de novos crimes acontecerem.

Ocorre que indivíduos racializados são desproporcionalmente mirados por agentes de segurança pública em comparação aos não-racializados — o que se imprime nos dados históricos. Trata-se de um clássico exemplo de viés histórico nos dados, conforme definido por Mehrabi et al. [7]. Além disso, a localização na cidade — atributo principal do modelo — pode ser um *proxy* para etnia, o que reforça essa discriminação. Baseando-se nesses dados, o *PredPol* tenderá a desproporcionalmente prever ações criminosas praticadas em localidades racializadas — reforçando, assim, o racismo do sistema de justiça.

É interessante notar, ainda, como tais soluções encontram legitimidade na retórica do *apelo* (acrítico) *à ciência* e à ideia de *políticas públicas baseadas em evidências*[1] — afinal, se "números não mentem" e "dados são objetivos", conclusões geradas algoritmicamente por tais dados (e decisões delas decorrentes) seriam objetivamente corretas e, portanto, incontestáveis. A inteligência artificial é frequentemente associada a essa narrativa, o que intensifica a armadilha de aplicá-la em desconsideração a princípios éticos.

No caso que analisamos, por exemplo, as próprias declarações da instituição policial em resposta às críticas lançam mão dessa retórica. Afirma-se que o algoritmo "mira o crime, não pessoas" e cita-se uma referência acadêmica para argumentar que o *PredPol* não introduz novos vieses na tomada de decisões, sendo equivalente às melhores práticas de policiamento usadas até então. Ignora-se, assim, que: (a) a formulação de políticas públicas deve ser multidimensional e sempre passa por fatores sociais, escolhas políticas e disputas de narrativa, não havendo como ser um processo neutro ainda que informado por parâmetros objetivos; (b) mesmo (supostamente) sem novos vieses, os vieses preexistentes nos dados utilizados são suficientes para caracterizar como antiética a adoção desse tipo de sistema.

3 Um caminho possível

Para mitigar ou solucionar problemas como o descrito na seção anterior, diversas intervenções sobre o processo de desenvolvimento, aplicação e avaliação de sistemas baseados em inteligência artificial foram propostas na literatura [7]. Destacamos, aqui, o caminho proposto por Sloane et al. [8].

Diante do reconhecimento de que sistemas podem embutir vieses contra determinados grupos, é comum que se sugira maior participação de membros desses grupos nos processos de construção desses sistemas. Ocorre que a participação, por si só, não é suficiente, como explicam os autores — além de ser, muitas vezes, explorada sob uma lógica que não deixa de reforçar estruturas de poder. A tendência de limitar-se a aumentar a diversidade dos atores do processo via participação é criticada como *participation washing* [9].

A solução proposta, então, é aprofundar o significado da participação em todas as suas dimensões, tornando-a uma ferramenta efetiva para ajudar a construir sistemas mais éticos. Tais dimensões são as seguintes:

- **Participação via trabalho**: Refere-se à participação para execução de tarefas operacionais do processo (e.g. coleta de dados, rotulagem, interação com o sistema etc.);
- **Participação via consulta**: Refere-se à participação como consultor/a em qualquer etapa do processo;
- **Participação como justiça**: Refere-se à parceria de longo prazo com diversos atores interessados no processo.

Segundo o estudo, desafios associados à participação como um fim em si mesmo incluem: ausência de reconhecimento ou compensação por trabalhos relevantes prestados no processo; integração pobre com participantes; inviabilidade de custos; participação performativa; contexto desfavorável.

Para superar esses desafios, os autores propõem a incorporação, **desde a concepção do projeto**, dos seguintes eixos à ideia de participação:

[1] O desenvolvimento dessa discussão demanda um aprofundamento que está além do escopo deste trabalho, mas julgamos importante introduzi-la.

- **Reconhecimento da participação**: trabalhos operacionais relevantes devem ser reconhecidos e compensados adequadamente — sempre com transparência;
- **Atenção ao contexto**: Se é inviável escalar o resultado com a adequada participação em todas as etapas, seu escopo deve ser limitado ao contexto para o qual ela foi considerada;
- **Participação genuína e de longo prazo**: Transparência e compartilhamento de conhecimento genuíno entre os atores devem fazer parte de todo o processo, com manutenção e articulação constantes — e previsão de alocamento de recursos para tal.

O trabalho propõe, por fim, que o desenvolvimento de um banco de dados acessível de precedentes sobre a questão facilitaria o processo para pesquisadores e profissionais. A ideia seria registrar falhas de design de participação, associadas a outros atributos de interesse, para que a comunidade pudesse aprender com erros passados.

Referências

[1] Nicolas Cointe, Grégory Bonnet e Olivier Boissier. "Ethical Judgment of Agents' Behaviors in Multi-Agent Systems". Em: *AAMAS '16: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. May. International Foundation for Autonomous Agents e Multiagent Systems, 2016, pp. 1106–1114. ISBN: 9781450342391. URL: https://dl.acm.org/doi/10.5555/2936924.2937086.

[2] C. Dianne Martin e Toma Taylor Makoundou. "Ethics by Design in AI". Em: *ACM Inroads* 8.4 (2017), pp. 35–37. DOI: https://doi.org/10.1145/3148541.

[3] Association for Computing Machinery. *ACM Code of Ethics and Professional Conduct*. 2018. URL: https://www.acm.org/code-of-ethics.

[4] InfoQ. *Why Should We Care about Technology Ethics? The Updated ACM Code of Ethics*. 2019. URL: https://www.infoq.com/articles/acm-code-ethics.

[5] World Commission on the Ethics of Scientific Knowledge e Technology of UNESCO. *Preliminary study on the Ethics of Artificial Intelligence*. 2019. URL: https://unesdoc.unesco.org/ark:/48223/pf0000367823.

[6] Taylor Mooney e Grace Baek. *Is artificial intelligence making racial profiling worse?* CBS News. 2020. URL: https://www.cbsnews.com/news/artificial-intelligence-racial-profiling-2-0-cbsn-originals-documentary/.

[7] Ninareh Mehrabi et al. *A survey on bias and fairness in machine learning*. Marina del Rey, 2019. arXiv: 1908.09635.

[8] Mona Sloane et al. "Participation is not a Design Fix for Machine Learning". Em: *Proceedings of the 37th International Conference on Machine Learning*. Citações Google Scholar: 6. Vienna, 2020. URL: https://arxiv.org/abs/2007.02423.

[9] Mona Sloane. *Participation-washing could be the next dangerous fad in machine learning*. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/08/25/1007589/participation-washing-ai-trends-opinion-machine-learning.