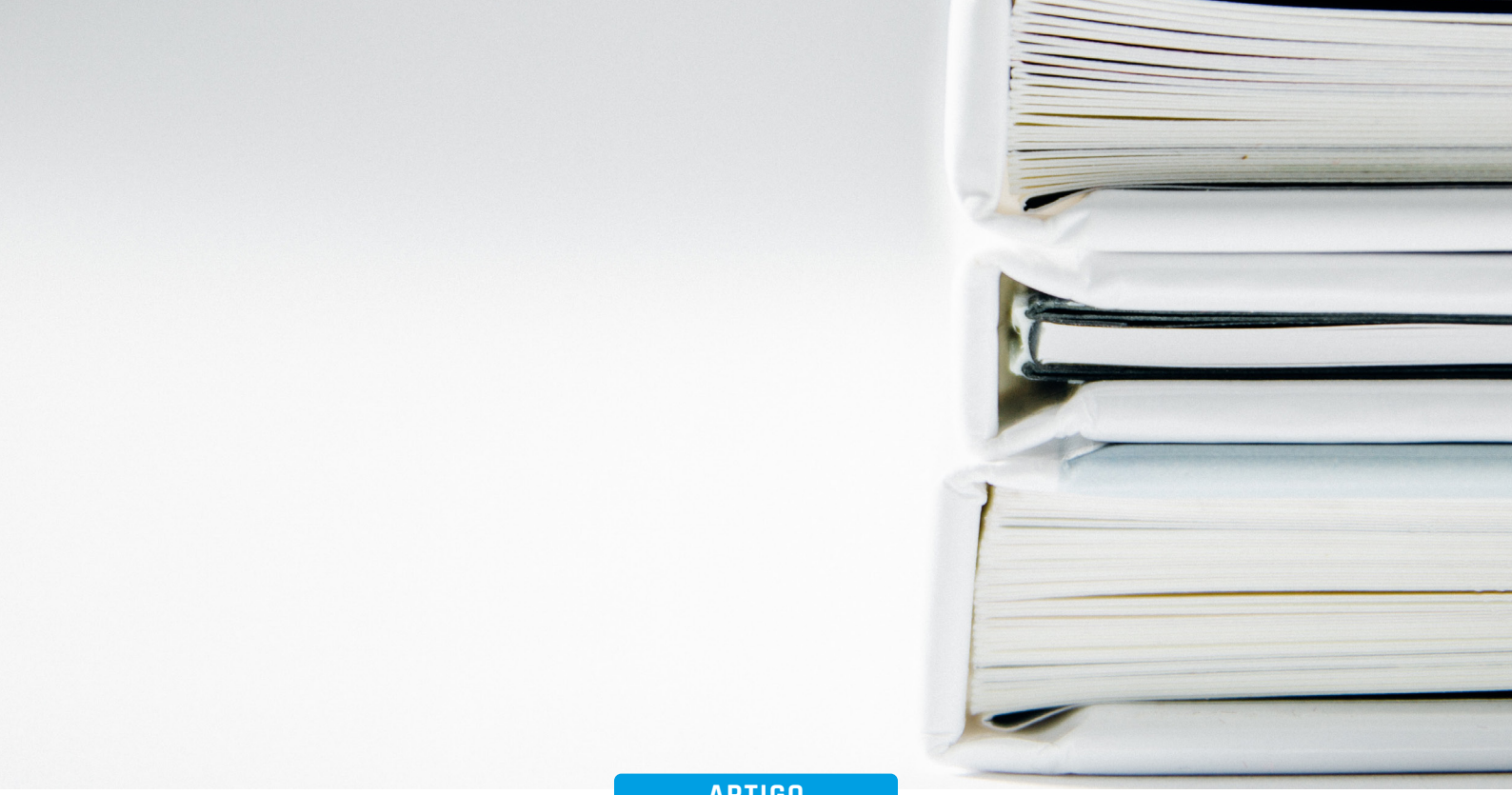

ARTIGO

INTELIGÊNCIA ÉTICA

POR
Flávio S. Corrêa da Silva e Nina S. T. Hirata
fcs@ime.usp.br e nina@ime.usp.br

Inteligência Artificial (IA)

Duas linhas de trabalho têm caracterizado, genericamente, as atividades na área de IA, com a atenção primordial da comunidade científica e de engenharia oscilando periodicamente entre elas [3]: (1) IA simbólica (IA-S), em que padrões de raciocínio são caracterizados utilizando sistemas lógicos; o alinhamento entre padrões de raciocínio empiricamente observados e modelos formais produzidos com base em sistemas lógicos é produzido com base em argumentação e exemplos suficientemente convincentes, e (2) IA adaptativa (IA-A), em que técnicas para identificação de padrões são aproximadas assintoticamente por funções matemáticas pertencentes a uma classe específica, fazendo uso de métodos iterativos denominados de Aprendizado de Máquina.

Estas duas linhas de trabalho são complementares. A IA-S produz sistemas que, por design, mas apenas teoricamente, podem garantir transparência, equidade na geração de soluções e explicabilidade. Sua escalabilidade para problemas complexos e de grande porte, bem como a robustez da validação de seu alinhamento com rela-

ção aos problemas resolvidos, entretanto, sempre foram pontos fracos relevantes. A IA-A, por outro lado, apresenta ótima escalabilidade e alinhamento estatisticamente mensurável com relação aos problemas que pretende resolver, mas se mostra resistente à caracterização rigorosa de métricas e métodos para garantia de transparência, equidade e explicabilidade.

Big Data

O cenário atual de grande volume de dados propulsionou um grande avanço da IA-A, especialmente *deep learning*, despertando interesses crescentes por tecnologias viabilizadas por esse avanço: sistemas de tradução, assistentes virtuais, veículos autônomos, sistemas de recomendação e reconhecimento biométrico etc. No entanto, vários desses sistemas têm apresentado falhas de natureza ética [6] como, por exemplo, desempenhos discriminatórios ou estereotipados, desfavoráveis a algum subgrupo particular, o que na prática pode perpetuar e ampliar os vieses presentes na sociedade. Estas falhas vêm suscitando amplas discussões acerca dos aspectos éticos associados à IA.

Um olhar sobre o sistema ético

Um sistema ético é um conjunto de normas e valores capazes de guiar pessoas em direção a uma vida que valha a pena viver [7]. Este conceito acompanha as civilizações, especialmente ocidentais, desde a antiguidade grega. Como corolário, também herdado dos pensadores gregos, temos que sistemas éticos buscam diferenciar o bem do mal, para que indivíduos e sistemas sociais possam se afastar do mal e se dirigir ao bem, com base na premissa de isto conduzir à vida que valha a pena viver. Os sistemas éticos têm sido classificados em três categorias: sistemas baseados em (1) virtudes, em que o bem resulta de cultivar e colocar em prática atributos virtuosos como generosidade, altruísmo, compaixão etc.; (2) deveres, em que o bem resulta da obediência a leis, regras e normas que conduzam a uma sociedade justa, equitativa e capaz de garantir direitos individuais; e (3) bom balanceamento entre consequências de ações, em que o bem resulta de garantir que cada ato de cada indivíduo, ainda que possa produzir efeitos negativos, produza efeitos positivos capazes de compensá-los.

Ética no desenvolvimento de sistemas inteligentes

A ética baseada em consequências de ações é, dentre estas três categorias, a mais simples de considerar no design de sistemas inteligentes: ao projetar um sistema inteligente para a resolução de problemas, são estabelecidas ações que este sistema será capaz de efetuar. Os possíveis efeitos negativos destas ações podem ser considerados, para que sejam implementadas possíveis ações corretivas, capazes de evitar estes efeitos ou mitigar suas consequências - por exemplo, regras podem ser inseridas para solucionar situações imprevistas ("botão de pânico"), ou variáveis descritivas são revisadas com o propósito explícito de impedir que informação sigilosa possa ser inferida (por exemplo, pela exclusão de variáveis sensíveis).

Desse modo, raramente são incluídos como requisitos na especificação de um sistema os possíveis efeitos positivos das

ações prescritas para o sistema, assumindo que seus requisitos funcionais sejam suficientes para caracterizar os efeitos positivos esperados da operação do sistema. O foco, portanto, na maioria das ocasiões em que requisitos éticos são considerados no design de sistemas inteligentes apoiados em princípios de consequências de ações, tem sido em evitar malefícios, em vez de garantir benefícios.

A ética baseada em deveres também tem sido tratada explicitamente no design de sistemas inteligentes, fundamentada na explicitação de mecanismos de normas e valores e sua formalização através de sistemas deônticos formais. Estes sistemas são implementados, na grande maioria dos casos, como restrições e condições que delimitam e guiam o comportamento dos sistemas. A inclusão de mecanismos deônticos, portanto, frequentemente foca em evitar infrações e, portanto, em evitar malefícios.

Recentemente, a ética baseada em virtudes tem sido considerada com maior atenção [1,5]. Esta categoria de sistemas éticos se mostra mais desafiadora para implementação em sistemas inteligentes porque, diferentemente das categorias anteriores, requer a busca de benefícios a partir de mecanismos internos de resolução de problemas, além da garantia de evitar malefícios a partir do atendimento a mecanismos externos de controle de comportamento[2]. Por exemplo, e considerando a terminologia proposta por Etzioni [4], ações podem ser programadas seguindo a linha simbólica ou inferidas seguindo a linha adaptativa, considerando não apenas uma visão hedonista que priorize a satisfação imediata, mas também uma perspectiva que leve a benefícios comunitários e de auto-realização, com base em ações afirmativas.

Assegurando a ética

Para tal, é necessário que requisitos técnicos específicos possam ser formulados e o alinhamento de sistemas com estes requisitos seja mensurável. Neste sentido, cumpre destacar que instituições com alta reputação internacional, como o IEEE, têm empreendido esforços na construção de normas e recomendações técnicas relacionadas, especificamente, à garantia de comportamento ético de sistemas baseados em IA[1].

Além de estabelecer normas e recomendações técnicas, é preciso desenvolver mecanismos para promover e garantir seu cumprimento. Dentre as iniciativas em direção à promoção da ética, podemos citar métodos algorítmicos que visam garantir sistemas sem comportamento discriminatório, por exemplo, pela eliminação de variáveis sensíveis na inferência de um modelo ou pelo emprego de métricas para quantificar imparcialidade.

Observa-se também uma maior aproximação entre as abordagens simbólica e adaptativa da IA, assim como entre as abordagens gerativas e discriminativas em Aprendizado de Máquina, avançando em direção a uma maior transparência e explicabilidade. Nota-se também o surgimento de empresas especializadas em auditar sistemas inteligentes[2] representando um reforço fiscalizador. Outra iniciativa inte-

1 https://standards.ieee.org/industry-connections/ec/autonomous-systems/, https://standards.ieee.org/ieee/7010/7718/, https://globalcxi.org/
2 https://www.ethicalintelligence.co/

ressante são sistemas de software livre[3] que auxiliam o processo de auditoria ou explicação de modelos computacionais.

Uma vez que os desenvolvedores, auditores e usuários desses sistemas são os indivíduos, a sensibilização e conscientização têm também um importante papel. Algumas iniciativas nesta direção são a inclusão de disciplinas sobre ética em grades curriculares, a formação de times multidisciplinares e a ampliação de diversidade em TI. Finalmente, para garantir que essas normas sejam seguidas, será importante estabelecer regras de regulação e fiscalização jurídicas, mas mais do que isso é importante desenvolver uma cultura ética na sociedade, com participação do setor privado, governos, universidades, organizações independentes (ONGs, Fundações etc.), e indivíduos. O volume de debates e iniciativas em curso indica que estamos caminhando nessa direção.

3 https://fairlearn.org/ e https://shap.readthedocs.io/en/latest/index.html

Referências

1. Bilal A., Wingreen S. and Sharma R. Virtue ethics as a solution to the privacy paradox and trust in emerging technologies. Proceedings of the 3rd International Conference on Information Science and System, p. 224-228, 2020.
2. Correa da Silva F.S. Towards Positive Artificial Intelligence. In: Baldoni M., Bandini S. (eds) AIxIA 2020 – Advances in Artificial Intelligence. AIxIA 2020. Lecture Notes in Computer Science, vol 12414. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-77091-4_22
3. Correa da Silva F.S. and Sawhney S. Good design can go a long way to support ethical behaviour of intelligent systems in healthcare. Submitted, 2022.
4. Etzioni A. Happiness is the wrong metric: A liberal communitarian response to populism. Springer Nature, 2018.
5. Gamez P., Shank D. B., Arnold C. and North M. Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. AI & SOCIETY, 35(4), 795-809, 2020.
6. UNESCO World Commission on the Ethics of Scientific Knowledge and Technology. Preliminary study on the Ethics of Artificial Intelligence. UNESCO, 2019.
7. Vallor S. An introduction to data ethics. Course module, Santa Clara, CA: Markkula Center for Applied Ethics, 2018.

FLÁVIO S. CORRÊA DA SILVA é Professor Associado de Ciência da Computação na Universidade de São Paulo. Atua na área de Inteligência Artificial. Sua pesquisa tem se focado em desenvolvimentos teóricos da Inteligência Artificial, bem como aplicações em saúde, desenvolvimento urbano e regional, bem-estar, atenção a idosos e promoção de valores éticos.

NINA S. T. HIRATA é Professora Associada do Instituto de Matemática e Estatística da Universidade de São Paulo. Seus interesses de pesquisa estão concentrados na área de aprendizado de máquina, cobrindo desde fundamentos, algoritmos até aplicações, especialmente no processamento e análise de imagens.