Schaich Borg, J., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And how we get there*. Random House. 279 pp. ISBN 9780241454763

Maria João Guimarães
(ISCTE – Instituto Universitário de Lisboa)
(mjdgs@iscte-iul.pt)
ORCID: 0009-0009-0873-503X

Maria João Guimarães (short bio): Doutoranda em Comunicação Social no ISCTE — Instituto Universitário de Lisboa. Licenciada em Jornalismo pela Universidade de Coimbra (1994–98). Jornalista no diário *Público* desde 1998, desde 2002 na secção Mundo. Concluiu, em 2005/06, o Programa Avançado em Estudos Políticos e Económicos para Jornalistas da Universidade Católica de Lisboa e, no ano de 2015/16, o programa European Journalism Fellowships, da Universidade Livre de Berlim.

Maria João Guimarães (short bio): PhD student at ISCTE — University Institute of Lisbon. A graduate from Coimbra University with a degree in Journalism, she is a journalist at the daily newspaper *Público* since 1998, and from 2002 onwards at its Foreign Desk. In 2005/06, she completed the Advanced Program in Political and Economic Studies for Journalists at the Catholic University of Lisbon and, in 2015/16, the European Journalism Fellowships program at the Free University of Berlin.

Submissão: 23/03/2025

Aceitação: 16/06/2025

Resumo (PT): A utilização de aplicações de inteligência artificial (IA) pode trazer perigos, desde o reforço de vieses humanos até consequências em termos de segurança. Ao explicarem as origens de vários destes perigos, tanto atuais como potenciais, os três autores de *Moral AI. And how we get there* iluminam caminhos para que a vertente moral da IA não seja deixada de lado no seu desenvolvimento. De vários modos, a chave está também na comunicação, seja entre quem desenvolve produtos de IA, seja das empresas para o público em geral sobre as capacidades, e potenciais pontos fracos, das suas aplicações.

Palavras-chave: Inteligência Artificial, Moral, Ciências de Computação, Comunicação.

Abstract (EN): The use of artificial intelligence (AI) applications can bring dangers, from the reinforcement of human biases to consequences in terms of security. By showing the reasons for both current and potential dangers, the three authors of *Moral AI. And how we get there* illuminate ways to guarantee the moral aspect of AI isn't left out of its development. In many ways, communication is a central element, both between those developing AI products and between companies and the general public regarding the capabilities, and potential weaknesses, of their applications.

Keywords: Artificial Intelligence, Moral, Computer Sciences, Communication.

A comunicação como elemento essencial de uma AI assente em princípios morais

Em *Moral AI: And how we get there* (Random House, 2024), Jana Schaich Borg, Walter Sinnott-Armstrong e Vincent Conitzer, autores com diferentes especializações (ciências sociais, filosofia, e ciências de computadores), tratam de várias aplicações da Inteligência Artificial (IA) para a sociedade e das respetivas implicações, num livro escrito para um público não especializado. Ainda que se dedique longamente a efeitos nefastos, presentes e potenciais, não é um livro que apenas disseque todas as potenciais consequências negativas de aplicações e produtos que utilizem IA, antes tentando explicar as razões que contribuem para esses efeitos, e sugerindo formas de os evitar.

Não é imediatamente claro o que tem o livro que ver com comunicação; mas tem, sobretudo de três modos. O primeiro, e mais direto, é a influência da IA nas redes sociais *online*, um campo considerado menos "crítico" (do que, por exemplo, o uso de IA em medicina, nos transportes ou no contexto militar), mas que acaba por se refletir no mundo "real" (por oposição ao mundo apenas "*online*", embora também se possa discutir se a relação entre os dois termos é mesmo de oposição). Afinal, o adágio *sticks and stones may break my bones, but words can never hurt me* parece cada vez mais longe de ser verdadeiro.

O segundo tem que ver com o modo como a IA se mostra a um conjunto alargado de pessoas, o que está a acontecer sobretudo com programas como o ChatGPT ou outros *large language models*: é aqui, defendem os autores, que se define a confiança ou desconfiança da sociedade em relação à IA em geral. O que acontece com a IA "visível" e como ela é percebida é essencial para esta confiança e, para isso, é imperativo que haja transparência sobre o seu funcionamento e as causas de potenciais falhas.

Também é defendido que se houver melhor comunicação sobre o modo como a IA chega a alguns resultados (exemplos dados são o de a IA ajudar a prever o potencial de reincidência de suspeitos de crimes e, com base nisso, a decidir sobre a possibilidade de o indivíduo aguardar julgamento em liberdade, ou de a IA fazer o balanço dos critérios para definir quem deve receber um rim entre vários candidatos), já que, se quem usa a IA tiver mais informação sobre o modo como o sistema faz cálculos e chega a uma recomendação, será mais fácil ultrapassar eventuais problemas e uma possível rejeição da tecnologia.

Finalmente, o terceiro surge nos capítulos em que é detalhado o modo como as equipas que desenvolvem IA se articulam: é iluminada a necessidade de, sobretudo dentro das empresas que desenvolvem IA (ou entre empresas, já que muitos produtos finais são desenvolvidos com base como que em módulos, sendo estes desenvolvidos por empresas diferentes), haver melhor comunicação.

A comunicação é, assim, central para a possibilidade de desenvolvimento de uma IA que não esqueça a sua dimensão moral (ou seja, que seja preocupada com os seus efeitos na sociedade e que tente evitar consequências nefastas da sua utilização), com várias das dificuldades a residirem não na falta de vontade de intervenientes no processo para que a IA se reja por princípios morais (embora com exceções em que o entusiasmo pela descoberta técnica pode deixar para trás quaisquer considerações éticas, como no caso de Oppenheimer e da bomba atómica), mas sim na falta de comunicação e de articulação entre vários atores com responsabilidade no processo.

1. Desafios e possíveis respostas

Na introdução — "Qual é o problema?" — são dados exemplos de boas e más notícias da utilização de IA em vários campos, dos *media* ao marketing, da saúde à vigilância, da arte ao ambiente (também há boas notícias na arte, e também há más notícias no ambiente). São enumeradas as principais áreas em que a IA se cruza com valores morais: segurança, igualdade, privacidade, liberdade, transparência e fraude.

Como sempre que se fala em inteligência artificial, não é exatamente claro o que implica o conceito (cunhado já em 1955 por John McCarthy), por isso, no primeiro capítulo, os autores começam por definir o que é (e o que não é) a IA, e os vários tipos de classificação de IA conforme as suas capacidades.

No segundo capítulo é explorada a questão da segurança, discutindo-se perigos da IA com base no que é, e no que se pensa que poderá vir a ser; nomeadamente, o medo comum de que haja uma IA que tome conta do mundo por lhe ser dada uma determinada tarefa — por exemplo, pedir à IA que maximizasse o número de clipes fabricados acabaria com esta a destruir o mundo para o fazer, mesmo com humanos a tentar impedi-la. Estes casos extremos não são vistos como uma inevitabilidade, mas também não são tomados apenas como um exagero, e os autores defendem que é necessário começar a planear estratégias de contenção para esse caso. Há duas principais dificuldades: a de ensinar à IA o chamado senso comum humano, e a de prever os erros que a IA possa cometer, já que esta é "previsivelmente imprevisível" (Schaich Borg et al., 2024, p. 69).

No entanto, sendo que a IA já existente tem problemas e comete erros (os quais podem ser da própria IA, podem ser humanos ou podem resultar de os humanos confiarem demasiado na IA ou praticarem cada vez menos determinadas tarefas, por estas serem atribuídas à IA), é nesses que o resto do capítulo se foca.

Aqui surge a distinção entre as áreas que são consideradas *safety-critical*, em que o mau funcionamento da IA poderá ter consequências letais ou causar danos graves a pessoas ou bens materiais (como na área dos transportes, da defesa, ou da medicina), e as que geralmente não o são, mas que têm, ainda assim, potencial para causar grande dano.

O exemplo dado de uma área que não é considerada *safety-critical* é o das plataformas de redes sociais *online*, em que o objetivo da IA usada é determinar o conteúdo a mostrar para maximizar o tempo que cada pessoa passa na plataforma. Com mais incentivos para partilhar conteúdo que receba "gostos" ou tenha mais partilhas, a consequência é a de uma muito maior quantidade de informação a ser transmitida e, pior, muitas vezes de informação errada ou mesmo desinformação. E se antes de haver redes sociais *online* já havia rumores, e estes tinham consequências, nas redes sociais *online* a desinformação é por vezes mais rápida do que a informação, e as consequências surgem também com maior rapidez (realça-se um caso de falsos rumores que circularam em redes sociais *online* na Índia em 2017 sobre pessoas que retirariam órgãos a crianças, e apesar de não ter havido qualquer prova, estiveram ligados a espancamentos de mais de 150 pessoas inocentes).

O fenómeno das câmaras de eco, em que cada pessoa recebe sobretudo conteúdo que corresponde à sua visão do mundo, potenciado pela IA, também é potencialmente perigoso, com as plataformas a privilegiarem conteúdo polémico para provocar reações e gerar "viralidade", e este conteúdo inclui muitas vezes publicações que provoquem emoções negativas contra outros grupos, o que funciona como reforço positivo para ataques verbais *online* ou *offline* e mesmo para violência física.

No terceiro capítulo é abordada a possibilidade de a IA respeitar a privacidade dos utilizadores. Os autores tentam rejeitar uma visão de fatalidade da disponibilização de dados para a utilização de produtos e tentam contrariar a reação de quem pensa que não tem nada a esconder, ilustrando a dimensão do problema: por exemplo, um estudo descobriu que 19 em 21 aplicações para *smartphones* enviaram dados para aproximadamente 600 domínios diferentes, algumas delas de modo contínuo, mesmo quando a aplicação não estava a ser utilizada.

No quarto capítulo é questionada a possibilidade de a IA ser justa. Ironicamente, o facto de ter havido vários casos em que a utilização da IA por diversos sistemas resultou em efeitos desproporcionados para categorias de pessoas já mais desfavorecidas chamou a atenção para muitas formas de injustiça em decisões tomadas por humanos. A IA poderia, nas condições certas, ser até mais justa. Mas esse não é, ainda, o caso.

No quinto capítulo é analisada a potencial responsabilização por erros cometidos por IA que tenham consequências graves (como a morte de alguém). Se é difícil atribuir responsabilidade, não a atribuir também trará uma atitude menos empenhada de todas as pessoas que intervêm no processo de criação ou produção de produtos de IA, defendem os autores.

No sexto capítulo é abordada a potencial incorporação de moral humana nos sistemas de IA e a questão da necessidade de transparência sobre as variáveis usadas para fortalecer a dimensão moral de um determinado produto. Por exemplo, num sistema que permita ajudar em decisões sobre quem recebe um rim, a transparência sobre os critérios é importante para pedir *feedback* aos humanos sobre se a utilização está a ser conforme o que seria expectável em termos morais; isso ajuda a perceber não só que decisões são tomadas, mas porquê, o que por sua vez ajuda a que todos os intervenientes no processo — ou seja, médicos, pacientes, e as comunidades mais alargadas — tenham confiança no sistema de IA e, assim, mais disponibilidade para o utilizar: "A transparência permite que haja confiança, que é necessária para uma adoção alargada do sistema de IA, junto com todos os seus potenciais benefícios sociais" (Schaich Borg et al., 2024, pp. 135–136).

Finalmente, no sétimo capítulo, são abordadas sugestões sobre o que pode ser feito por vários atores, como as empresas que desenvolvem IA, o terceiro sector, ou os governos dos países, que deviam regular melhor o desenvolvimento e uso desta tecnologia (mas resistem a fazê-lo porque a regulação é vista muitas vezes como um travão à inovação). Aqui é explorado o modo como a fraca comunicação entre intervenientes no processo leva a que não sejam tidas em conta implicações éticas e morais da IA, especialmente porque, na maior parte dos casos, engenheiros informáticos ou pessoas que trabalham dados que desenvolvem as aplicações recebem já uma espécie de pacote pronto a usar de quem desenvolveu o algoritmo, e não há qualquer interação entre eles.

Depois de explicarem todos os papéis diferentes das pessoas envolvidas no processo de criação de produtos de IA, os autores dizem que uma das questões essenciais que as estratégias para uma IA regida por princípios morais têm de tentar resolver é precisamente a falta de oportunidade para comunicação entre as pessoas que desenvolvem várias partes

do sistema durante a criação de um produto específico de IA (Schaich Borg et al., 2024, p. 145).

Mesmo quando há interação entre várias equipas que contribuem para um produto, há muito potencial para mal-entendidos, por se tratar de pessoas com *backgrounds* e vocabulário profissional muito diferente, com os grupos a não terem, muitas vezes, sequer noção da informação que poderá realmente estar a passada e de qual poderá estar a perder-se.

Finalmente, é preciso, defendem os autores, encontrar um modo de recolher opiniões úteis e fiáveis de todas as pessoas que podem ser afetadas por um dado produto de IA (um desafio enorme, reconhecem os autores, dado que a a utilização alargada da IA e a sua ligação às nossas vidas quotidianas faz com que quase qualquer cidadão do mundo possa ser potencialmente afetado por um produto de IA). Mais, o fluxo de informação entre quem cria a IA e quem pode ser afetado por ela deveria ser bidirecional, algo que os autores admitem que não será fácil, mas que é possível e necessário.

Conclusão

O livro explica de um modo acessível a pessoas que não sejam de áreas especializadas quais os principais problemas que decorrem do modo como são desenvolvidos os atuais produtos ou sistemas de IA, defendendo que as decisões sobre a dimensão moral da IA — que tem capacidade para afetar praticamente qualquer pessoa do mundo, independentemente do local onde é pensada e desenvolvida — deviam ser tomadas conscientemente, e não deixadas ao que possa acontecer no decorrer do processo. Fá-lo através de uma divisão em cinco "campos de batalha": disseminação da tecnologia, práticas organizacionais, educação, participação cívica e políticas públicas. Em vários destes campos, a comunicação aparece como um fator essencial para que a vertente moral seja de facto incluída no desenvolvimento da IA.

Os autores demonstram optimismo em relação à possibilidade de um trabalho construtivo e com efeitos positivos numa IA que tenha mais preocupações com a implicação moral da sua utilização e que não contribua para prejudicar ainda mais pessoas que já estejam em situações estruturais de maior desvantagem, pese embora reconheçam que este é um enorme desafio, como ilustram com recurso a uma citação do sociólogo alemão Ulrich Beck sobre a ética, que muitas vezes "tem o papel de um travão de bicicleta num voo intercontinental" (Beck, 1998, citado por Schaich Borg et al., 2024, p. 148).

Os próprios autores fazem, além disso, notar que o campo é sujeito a mudanças muito rápidas, referindo-se a avanços na própria tecnologia; de facto, desde a publicação deste livro, já muito mudou; por exemplo, surgiu o *chatbot* Deepseek, modelo chinês concorrente do ChatGPT (a China está pouco presente no livro, sendo referida apenas como ator na corrida contra os EUA por uma IA superinteligente, e pela utilização de IA na vigilância da minoria uigur, um dos exemplos apresentados como uma utilização problemática da IA), a que os EUA responderam com um decreto presidencial de Donald Trump[1] para "remover obstáculos à liderança americana" em IA com menos regulação (o caminho contrário ao defendido pelos autores para uma IA assente em princípios morais). Em suma, os autores lutam contra a ideia de ser uma fatalidade que o futuro da IA seja dominado por produtos sem preocupação com a sua dimensão moral, e com potenciais efeitos nefastos na sociedade; contudo, ao mostrarem os meios para evitar que isto aconteça, que são variados e envolvem o esforço de múltiplos atores, sublinham a dimensão da tarefa.

REFERÊNCIAS

Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And how we get there*. Pelican Books.

Executive Order 14179. (2025). Federal Register, *90*(20), 8741–2742. The White House. https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf

[1] Executive Order 14179. (2025). Federal Register, *90*(20), 8741–2742. The White House. https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf