RESENHA

Inteligência artificial, valores e incerteza: um problema insolúvel para a regulação?

CHRISTIAN, B. (2020) *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company.

Resenha por
Felipe Leitão Valadares Roquete
felipe.roquete@fgv.edu.br

Bacharel em Direito (UFMG), Mestre em Ciência Política (UnB) e Doutorando em Direito da Regulação (FGV-RJ). Desde 2004, é servidor público federal da carreira de Especialista em Políticas Públicas e Gestão Governamental. Atualmente, é Coordenador-Geral de Análise Antitruste da Superintendência-Geral do Conselho Administrativo de Defesa Econômica. As opiniões do presente artigo não representam a posição institucional do Cade.

V. 2 / N. 2 / DEZEMBRO DE 2021
PÁGINAS 150 A 154

RESENHA POR
FELIPE LEITÃO VALADARES ROQUETE

É preciso coragem, antes de tudo. Pois não são triviais os desafios que se colocam para quem se propõe a escrever sobre temas técnicos e complexos para um público mais amplo, não especializado. E Brian Christian consegue, em seu livro *The Alignment Problem*, equilibrar-se na linha tênue que separa a abordagem simples da simplória, a concisão da superficialidade e, principalmente, a clareza do didatismo rasteiro.

Quatro anos de pesquisa calcada em extensa e cuidadosa bibliografia e quase cem entrevistas depois, ao abordar os diversos momentos da fascinante — e, não raras vezes, angustiante — caminhada do desenvolvimento da inteligência artificial (IA), o autor nos apresenta casos concretos (nos quais impera a angústia), bem como relata estórias sobre os principais personagens que contribuíram para a evolução do campo de pesquisa (aqui prevalece o fascínio).

Dividido em três capítulos, o livro aborda, com a mesma qualidade técnica, os sistemas de inteligência artificial já em uso e como seus vieses estão sendo tratados (capítulo intitulado "Profecia"), os desafios e as potencialidades dos sistemas atualmente em desenvolvimento ("Ação") e, finalmente, quais estratégias os pesquisadores estão utilizando para garantir o uso seguro e a obtenção de resultados acurados ("Normatividade").

Os casos apresentados partem do prosaico (por exemplo, as primeiras redes neurais comercialmente bem-sucedidas, utilizadas para ler os códigos postais manuscritos em cartas e os valores de cheques depositados em caixas automáticos), chegando ao trágico (o uso de algoritmos de classificação de imagens que são incapazes de identificar certos grupos sociais e que, quando implementados, implicaram situações de discriminação de gênero e raça), passando pelo anedótico (como ocorre nos sistemas que passaram regularmente a enganar seus desenvolvedores, em uma espécie de "jeitinho algorítmico", atingindo os objetivos postos por intermédio de atalhos nada funcionais).

Nesse percurso, o autor tece um diálogo entre as pesquisas acadêmicas aplicadas desenvolvidas nos últimos 80 anos e as diversas iniciativas que implementaram algoritmos de aprendizagem supervisionada, aprendizagem não supervisionada e *reinforcement learning* no mercado privado e no setor público, com objetivo não de demonstrar um enciclopedismo estéril, mas de esclarecer os pontos-chave dos métodos em benefício dos leitores: aqui, a transdisciplinaridade que está na origem daquelas pesquisas é traduzida de forma a explicitar tanto os fundamentos das técnicas de IA quanto os elementos que lhes são subjacentes e, usualmente, negligenciados.

E tais elementos subjacentes representam o cerne da análise empreendida por Brian Christian, pois envolvem valores e normas que podem ser afetados potencialmente — e, em diversos exemplos descritos no livro, efetivamente o são — pelo uso de IA. Trata-se, portanto, de buscar estratégias para nos assegurarmos que os modelos de inteligência artificial capturem nossos valores e nossas normas, entendam o que quisemos dizer e pretendemos fazer e, acima de tudo, o que queremos, quais os nossos reais objetivos com o uso de tais ferramentas: tal seria o "problema do alinhamento" (de um lado, as tecnologias de uso pervasivo e de constituição opaca e, de outro, os valores e as normas que estão no cerne da convivência humana), que emergiu como das mais centrais e urgentes questões científicas da atualidade. E prevenir tal desalinhamento, tal divergência, significa evitar resultados catastróficos.

Isso porque sistemas de inteligência artificial são constituídos por modelos que possuem uma característica peculiar: sua utilização continuada e sem revisão pode reforçar e potencializar problemas preexistentes — como vieses e tratamentos discriminatórios —, pois sua estrutura é sujeita a efeitos de *feedback loops*. Em outras palavras, partindo de uma realidade na qual existam iniquidades que estejam, como usualmente

estão, refletidas nas bases de dados que serão utilizadas por aqueles modelos, a implementação de sistemas de IA produz resultados que refletem os vieses originários e, como tais resultados retornarão ao sistema, agora como insumo para futuras análises e decisões, cristalizam-se aquelas iniquidades, em um círculo nada virtuoso de reiteração e confirmação de vieses. O modelo de IA, portanto, não apenas mimetiza, mas também, em longo prazo, muda a realidade na qual é utilizado, pois consolida e expande as situações de iniquidade preexistentes.

Mas tal diagnóstico não é unívoco, ou melhor, não resulta em uma única estratégia preferencial para lidar com o problema do alinhamento. Assim, seria possível identificar duas comunidades que predominariam entre pesquisadores, desenvolvedores e usuários na área de IA: de um lado, aquela que se preocuparia mais com os riscos éticos atualmente vivenciados, advindos do uso de tais tecnologias (por exemplo, a baixa acurácia de sistemas de reconhecimento facial ou os vieses de sistemas utilizados para implementação de políticas públicas) e, de outro, aquela que estaria mais focada nos perigos futuros, que surgiriam quando os sistemas de inteligência artificial se tornassem onipresentes e responsáveis por decisões que afetassem nossas relações privadas, nossa vida em sociedade e nossa relação com o Estado.

Tratam-se, portanto, de visões que estão, respectivamente, calcadas em avaliações baseadas nos conceitos de risco e de incerteza. Frank H. Knight (2021) elaborou tal definição que, em linhas gerais, diferencia realidades nas quais predomina o *risco* — em que seria possível calcular a probabilidade *a priori* da ocorrência de determinados fatos, por meio da indução com base na experiência e na avaliação empírica — de outras nas quais prevalece a *incerteza*, ou seja, quando a análise de riscos tradicional pode ser inadequada para tratar riscos não quantificáveis, dada a inexistência de bases válidas para classificar novos eventos, implicando a necessidade de realizar estimativas e, assim, de conviver com o erro.

Dessa forma, aqueles que privilegiam o enfrentamento dos riscos éticos dos sistemas de IA já em funcionamento supõem que o mundo é composto por elementos que, sob certas condições, sempre se comportarão de determinada maneira, o que, no limite, tornaria sempre possível calcular a probabilidade de ocorrência de determinados fenômenos.

Já o segundo grupo, que centra suas preocupações nos impactos de longo prazo do uso de sistemas de IA, reconhece a impossibilidade de calcular a probabilidade de ocorrência de eventos futuros, pois envolveriam riscos desconhecidos. Nesse caso, vigeria a ignorância quanto a problemas que sequer sabemos que existirão, que ainda nem sabemos formular ou para os quais talvez nem tenhamos vocabulário para exprimir.

Brian Christian nos apresenta, portanto, um cenário que baliza eventuais propostas de regulação de sistemas de inteligência artificial. Explica-se.

Se se parte da compreensão de que os efeitos deletérios do uso de sistemas de inteligência artificial para automação de tomada de decisão, tanto no setor privado quanto no setor público, são passíveis de estimação prévia, propostas de regulação estarão preparadas para lidar com problemas semelhantes aos que já foram identificados anteriormente: em outras palavras, conseguirão manejar ferramentas baseadas em riscos, forjando instrumentos regulatórios capazes de evitar que os mesmos erros se repitam no futuro. A governança regulatória, portanto, analisará o presente por intermédio de um olhar retrospectivo, no qual a incerteza não encontra guarida.

Mas se, como afirmou Frank Rosenblatt, referindo-se à possível utilidade do primeiro sistema de redes neurais, "o uso segue a invenção", seria inviável antecipar quais os possíveis usos e, consequentemente, os resultados em termos

de desalinhamento entre valores e tecnologia, do uso pervasivo de sistemas de IA. Quando o que queremos (e o que não queremos) é difícil de determinar direta e completamente, estamos diante de uma situação de ignorância e, portanto, instrumentos regulatórios precisam lidar com ambientes de incerteza. Aqui, a governança deve estar preparada para, com base em um olhar prospectivo, manejar instrumentos regulatórios que sejam *future proof* — conforme Rehman, Ryan e Efatmaneshnik (2017) —, ou seja, sustentáveis, resilientes e que se adaptem às mudanças complexas inerentes ao desenvolvimento de sistemas de inteligência artificial.

Ainda que o ponto de partida tenha sido delineado precisamente por Brian Christian — em uma tomada de decisão na qual há múltiplas e complexas etapas (como o processo de geração dos dados, a construção das bases de treinamento, a definição dos parâmetros do modelo, os testes e a sua implementação), abre-se espaço para que se crie uma cadeia de vieses, potencializada ao longo dos diversos elos do sistema de IA — tal tarefa, em que pese seu caráter incontornável, não esgota os desafios que nos são apresentados.

Como, exatamente, uma eventual regulação deve buscar traduzir — em termos computacionais — os princípios, direitos e ideais articulados pela legislação que proíbe tratamento discriminatório? Quais os instrumentos regulatórios adequados para garantir, simultaneamente, de um lado, ambiente viabilizador da inovação e do desenvolvimento tecnológico e, de outro, estrutura flexível que permita a identificação tempestiva de resultados deletérios em sistemas de IA que sequer foram ainda desenvolvidos? Enfim, como a regulação dever abordar a incerteza, inerente à evolução da inteligência artificial?

Brian Christian deu um passo inicial significativo, pois seu livro traz um panorama abrangente e compreensivo do desenvolvimento da IA, que funciona como um nivelador conceitual e técnico. E, ausente esse balizador inicial, a tarefa seguinte — a análise quanto à utilidade, viabilidade e/ou necessidade de criar um arcabouço regulatório para sistemas de inteligência artificial — torna-se excessivamente onerosa para todos aqueles que desejam enfrentar o debate acerca da regulação de novas tecnologias.

Ao fim do percurso, restaria-nos perguntar se o problema do alinhamento, trazido pela inteligência artificial, não estaria apenas espelhando questões históricas cuja solução, no caso brasileiro, nos desafia há séculos. Nesse caso, a tecnologia estaria apenas ampliando — ou potencializando — uma imagem iníqua (distorcida, mas ainda assim um reflexo de nossa sociedade) que nos aterroriza, ou deveria nos aterrorizar a todos, desde tempos imemoriais.

Referências bibliográficas

Knight, F. H. (2021). *Risk, Uncertainty and Profit*. Las Vegas: Pantiano Classics.

REHMAN, O. U., RYAN, M. J. & EFATMANESHNIK, M. (2017). Future Proofing Process. *INCOSE International Symposium*, 27(1), p. 921–934. Acessado em 24 de novembro, disponível em https://www.researchgate.net/publication/319407223_Future_Proofing_Process.