[301]

Para onde caminha a inteligência artificial

Comportamento de robôs depende de escolhas não racionais.

José Eli da Veiga

EU& – sexta, 14 de janeiro de 2022, p. 31

Inteligência artificial a nosso favor: Como manter o controle sobre a tecnologia
Stuart Russell (tradução Berilo Vargas), 336 págs.
Companhia das Letras, 2021 – Preço: R$ 99,90 (e-book: R$ 39,90)

Entre os livros sobre inteligência artificial (IA), em português, este certamente é um dos que melhor discute a eventualidade de alguma inteligência, bem superior à dos humanos contemporâneos, vir a ser desenvolvida por máquinas.

Embora o título não deixe claro, a obra é essencialmente pautada pela eventualidade de uma 'superinteligência', às vezes também aludida pelo uso dos prefixos ultra e hiper. Mesmo não confessado, o principal propósito de Stuart Russell é ponderar sobre as chances de se impedir que a superinteligência venha a ser o último grande acontecimento da história humana.

Só que a visão do autor sobre tamanha controvérsia acaba por parecer quase adicionada a enfadonhas explanações sobre os mais imediatos riscos e proveitos da IA. Todos também envolvendo questões éticas de suma importância.

Satisfatoriamente, um único mantra unifica e dá coerência às duas dimensões temporais: "manter o controle", como ressalta o subtítulo. Para tanto, a exposição está organizada em três partes, só a terceira sugerindo uma maneira de entender a IA que venha a garantir máquinas que permaneçam benéficas.

As duas primeiras exploram mais a própria ideia de integração entre as duas inteligências: a dos humanos e a das máquinas. Com especial atenção, na segunda parte, ao já enfatizado desafio do controle: como continuar a ter poder absoluto sobre máquinas mais potentes do que os humanos.

Como o livro se destina ao "público em geral", provavelmente qualquer leigo lerá, sem muitas dificuldades, os quatro quintos formados por suas primeiras 233 páginas. Só o derradeiro quinto foi reservado a especialistas, na esperança de os incentivar a repensarem seus conceitos fundamentais. Junta quatro áridos apêndices a 365 notas.

A incógnita central, segundo Russell, é como aprender a prever as preferências humanas, já que as máquinas seriam puramente altruístas e humildes. Uma tese que é apresentada no molde de "três princípios": 1º) o único objetivo da máquina é maximizar a execução de preferências humanas; 2º) a máquina de início não tem certeza de quais são tais preferências; 3º) a fonte definitiva de informações sobre preferências humanas é o comportamento humano.

Então, o que mais interessa é o "terceiro princípio", referente às preferências humanas. Elas nem estão na máquina, nem esta é capaz de observá-las

diretamente. Porém, apesar disto, alguma conexão definida deve haver entre máquina e preferências humanas.
Este terceiro princípio estipula, portanto, que a conexão se dá pela observação de escolhas humanas: supõe-se que as escolhas estão relacionadas de alguma maneira (possivelmente muito complicada) às preferências subjacentes.

Para entender o quanto tal ligação é essencial, basta pensar no oposto: "se alguma preferência humana não tivesse efeito algum sobre qualquer escolha real ou hipotética que o humano fizesse, então provavelmente não faria sentido dizer que a preferência existe" (p. 170).

Por outro lado, é preciso permitir que a máquina se torne mais útil, aprendendo mais sobre o que se quer. Afinal, ela não teria utilidade se nada soubesse a respeito de preferências humanas. A ideia é simples: escolhas humanas revelam informações sobre preferências humanas. Pode ser óbvio, se aplicada a uma escolha entre dois sabores de pizza. Mas as coisas começam a ficar muito mais discutíveis quando se pensa em escolhas entre vidas futuras e escolhas feitas com a intenção de influenciar o comportamento de robôs.

Ou seja, as verdadeiras complicações surgem porque os humanos não são perfeitamente racionais: preferências e escolhas humanas podem ser incoerentes e a máquina precisaria levá-las em conta para poder interpretar escolhas como provas de preferências.

Daí a necessidade de longuíssimas justificativas, tanto para esperanças, quanto para cautelas. Em vez de aqui relatá-las, parece muito mais interessante aproveitar a ocasião para dar três informações sobre o autor, que poderão ser muito úteis no momento de decidir se este livro merece mesmo ser estudado.

Russell é o fundador e principal líder de um dos principais programas de pesquisa e ensino sobre o tema, no campus de Berkeley da Universidade da Califórnia: o *Center for Human-Compatible Artificial Intelligence*.

Em 2020 saiu a quarta edição de sua proeza anterior, um "compêndio escrito a quatro mãos" com Peter Norvig, atualmente diretor de pesquisa do Google. Além de permanecer, por um quarto de século, no topo das listas dos melhores livros sobre inteligência artificial, tal "compêndio" vem sendo o mais adotado em cursos universitários de 116 países: ***Artificial Intelligence: A Modern Approach***, Prentice Hall (1ª ed.: 1995). No Brasil, foi traduzido pela editora Campus-Elsevier.

Russell está entra os principais protagonistas do movimento global pelo "Humanismo Digital", fundado em Viena em maio de 2019, cinco meses antes da publicação original deste seu livro, agora bem traduzido pela Companhia de Letras. Ele é o autor do terceiro capítulo de recentíssima coletânea, que poderá ser muito útil aos que se preocupam com os rumos da inteligência artificial, colocada em livre acesso pela Springer: ***Perspectives on Digital Humanism***, de Werthner H. et al. (eds).

[FIM DA VERSÃO CURTA]

= = =

SEGUNDA PARTE

Porém, qualquer obra sobre superinteligência exige comparação a outras que também advertem para os riscos de longo prazo da IA. Particularmente à principal referência de tão frenético debate: o livro de Nick Bostrom ***Superinteligência: caminhos, perigos e estratégias para um mundo novo*** (Ed. DarkSide, 2018; c2014). Quais seriam as convergências e eventuais divergências entre Russell e Bostrom?

Ambos dizem que o sucesso em IA produzirá trajetória civilizacional que leva a um uso compassivo e triunfante do legado cósmico da humanidade. Russell acrescenta: "Se não tirarmos proveito do que a IA tem a oferecer, a culpa será exclusivamente nossa" (p. 102).

Também compartilham a hipótese de Bostrom sobre uma "decolagem rápida", na qual a inteligência das máquinas cresce astronomicamente, em poucos dias ou semanas. Não haverá tempo para resolver o problema do controle - e não haverá mais o que fazer -, se uma explosão de inteligência vier a ocorrer antes de ter sido resolvido o problema menos desafiador, de controlar máquinas de inteligência apenas ligeiramente sobre-humana, acrescenta Russell.

Tudo começa a ficar bem mais obscuro quando Russell discute uma tese de Bostrom segundo a qual quase qualquer nível de inteligência poderia, em princípio, ser combinado a quase qualquer objetivo final.

A ideia de que sistemas inteligentes poderiam adotar os objetivos a serem perseguidos sugere que um sistema razoavelmente inteligente abandonaria seu objetivo inicial em troca de objetivo mais "correto". Segundo Russell, é muito difícil entender por que um agente racional faria isso.

Por outro lado, há quem diga ser impossível um programa tão inteligente a ponto de inventar maneiras de subverter a sociedade humana para atingir objetivos que lhe foram dados por humanos, sem compreender que estaria causando problemas para estes mesmos humanos.

Infelizmente, diz Russell, não só é possível que um programa se comporte assim, como é inevitável. O plano ótimo em execução pela máquina pode muito bem causar problemas para humanos. E a máquina pode muito bem saber disto. Mas, por definição, a máquina não reconhecerá tais questões como problemáticas. Elas não lhe dizem respeito.

Claro, "satisfazer objetivos conflitantes" não é o problema. É algo já incorporado ao modelo mais usual, desde os primeiros tempos da teoria da decisão. O problema é que os objetivos conflitantes dos quais a máquina esteja ciente não constituem a totalidade das preocupações humanas. Além disto, dentro do mesmo modelo, não há nada que diga que a máquina deve se preocupar com objetivos com os quais ela não recebeu ordem para se preocupar.

Os humanos levam em conta as preferências de outros humanos e sabem que não conhecem todas estas preferências. Por esta razão, Russell sustenta que tais características, se incorporadas à máquina, talvez ofereçam o começo de uma solução.

Nick Bostrom também adverte contra a motivação de liderança em IA para a conquista da posição de "senhor do mundo", como chegou a dizer Putin, em 2017, três anos depois da publicação de seu livro. Neste caso, a competição entre países, assim como entre corporações, tenderia a concentrar-se mais em avanços de aptidões primárias, do que no problema do controle.

Russell acrescenta que seria inútil alguém conquistar a posição de monopólio em IA, pois, como não é jogo de soma zero, nada se perde compartilhando-a. Por outro lado, competir para ser o primeiro, sem antes resolver o problema do controle, é um jogo de soma negativa.

Não há dúvida de que criminosos, terroristas e estados vilões se sentiriam incentivados a contornar quaisquer restrições ao design de máquinas inteligentes e usá-las para controlar armas ou conceber e executar ações criminosas. O maior perigo não é tanto que tais planos tenham êxito, mas que eles fracassem, perdendo controle sobre sistemas inteligentes mal projetados. Sobretudo sistemas imbuídos de objetivos maléficos e com acesso a armas.

De resto, Nick Bostrom propõe o uso de sistemas benignos de IA superinteligente para detectar e destruir quaisquer sistemas de IA maléficos ou rebeldes. Russell concorda, mas diz ser bem melhor descobrir formas de cortar o mal pela raiz. O primeiro passo positivo nesta direção seria uma campanha internacional, bem coordenada, contra os crimes cibernéticos.

Isto formaria um molde organizacional para iniciativas futuras de prevenção contra o surgimento de programas incontroláveis de IA. Ao mesmo tempo, produziria uma ampla compreensão cultural de que criar tais programas é, a longo prazo, um ato suicida.

= = =

José Eli da Veiga é professor sênior do Instituto de Estudos Avançados da USP: www.zeeli.pro.br