filosofia unisinos | Unisinos Journal of Philosophy

ISSN 1984-8234

Filosofia Unisinos
Unisinos Journal of Philosophy
23(3): 1-12, 2022 | e23307

Unisinos – doi: 10.4013/fsu.2022.233.07

Artigo

Inteligência Artificial e os Riscos Existenciais Reais: Uma Análise das Limitações Humanas de Controle

Artificial intelligence and the real existential risks: an analysis of the human limitations of control

Kleber Bez Birolo Candiotto
https://orcid.org/0000-0003-0370-4380
Pontifícia Universidade Católica do Paraná - PUCPR, Programa de Pós-Graduação em Filosofia, Curitiba/PR, Brasil. Email: kleber.c@pucpr.br.

Murilo Karasinski
https://orcid.org/ 0000-0002-6099-6968
Pontifícia Universidade Católica do Paraná - PUCPR, Curso de Graduação em Filosofia, Curitiba/PR, Brasil. Email: k.murilo@pucpr.br.

RESUMO

A partir da hipótese de que a inteligência artificial como tal não representaria o fim da supremacia humana, uma vez que, na essência, a IA somente simularia e aumentaria aspectos da inteligência humana em artefatos não biológicos, o presente artigo questiona sobre o risco real a ser enfrentado. Para além do embate entre tecnofóbicos e tecnofílicos, o que se defende, então, é que as possíveis falhas de funcionamento de uma inteligência artificial – decorrentes de sobrecarga de informação, de uma programação equivocada ou de uma aleatoriedade do sistema – poderiam sinalizar os verdadeiros riscos existenciais, sobretudo quando se considera que o cérebro biológico, na esteira do viés da automação, tende a assumir de maneira acrítica aquilo que é posto por sistemas ancorados em inteligência artificial. Além disso, o argumento aqui defendido é que falhas não detectáveis pela provável limitação de controle humano quanto ao aumento de complexidade do funcionamento de sistemas de IA representam o principal risco existencial real.

Palavras-chave: Inteligência artificial, risco existencial, superinteligências, controle humano.

ABSTRACT

Based on the hypothesis that artificial intelligence would not represent the end of human supremacy, since, in essence, AI would only simulate and increase aspects of human intelligence in non-biological artifacts, this paper questions the real risk to be faced. Beyond the clash between technophobes and technophiles, what is argued, then, is that the possible malfunctions of an artificial intelligence – resulting from information overload, from a wrong programming or from a randomness of the system - could signal the real existential risks, especially when we consider that the biological brain, in the wake of the automation bias, tends to assume uncritically what is set by systems anchored in artificial intelligence. Moreover, the argument defended here is that failures undetectable by the probable limitation of human control regarding the increased complexity of the functioning of AI systems represent the main real existential risk.

Key-words: Artificial intelligence, existential risk, superintelligences, human control.

Qualquer tecnologia bastante avançada não é distinguível da mágica.
Arthur C. Clarke

Se o cérebro humano fosse tão simples que pudéssemos entendê-lo, seríamos tão simples que não poderíamos.
Frase atribuída a Emerson M. Pugh

Introdução

A vitória do Deep Blue, o supercomputador da IBM, sobre o campeão russo de xadrez, Garry Kasparov, em 1997, chamou a atenção da humanidade sobre uma provável supremacia das máquinas. O estatístico americano Nate Silver, em seu livro *The Signal and the Noise*, de 2012, apresenta uma observação importante sobre este grande feito do projeto da inteligência artificial (IA). Silver oferece dados mais recentes que demonstram ter havido uma falha no funcionamento do programa Deep Blue na quadragésima quarta jogada da primeira partida contra Kasparov. O programa teve uma sobrecarga de processamento ao selecionar uma jogada, porém, estava programado para adotar um padrão de segurança como último recurso perante falhas como essa, que seria selecionar uma jogada completamente aleatória. Como Kasparov desconhecia a forma como a Deep Blue estava programado, isso o confundiu e o levou a cometer um erro que impediu sua vitória. O movimento da sua torre sem nenhum objetivo aparente da Deep Blue causou ansiedade em Kasparov, uma vez que ele considerou a jogada contraintuitiva, um sinal de inteligência superior.

A ideia de uma entidade artificial dotada de uma inteligência superior é ressaltada recentemente por autores como o físico britânico Stephen Hawking e o futurista norte-americano Ray Kurzweil. Para aquele, quando o desenvolvimento das pesquisas em Inteligência Artificial alcançar a possibilidade de melhora recursiva das máquinas sem a necessidade de auxílio humano, haverá uma provável explosão de inteligência desses artefatos autoprogramáveis que humanamente não será mais possível controlar. Por isso, segundo Hawking, é crucial que atualmente possamos garantir que os objetivos dos computadores estejam alinhados aos dos humanos. Do contrário, a humanidade estará seriamente ameaçada. Para Kurzweil, em *A Singularidade Está Próxima,* a aceleração das mudanças tecnológicas tem aumentado significativamente, dado que uma análise da história com mais acuidade permite identificar que as mudanças tecnológicas ocorrem de forma exponencial, isto é, o conhecimento tecnológico dobra de

quantidade em períodos cada vez mais curtos[1]. O futurista norte-americano sugere que em 2045, com base em suas projeções matemáticas, ocorrerá a "singularidade" – o momento único em que a inteligência artificial se equiparará à humana e logo em seguida iniciará um ritmo acelerado de superação.

Retomando o caso do programa Deep Blue contra Kasparov, as projeções de Hawking e Kurzweil parecem desconsiderar alguns questionamentos como: de fato, o aumento da capacidade algorítmica das máquinas irá promover sua suposta superioridade sobre os humanos? O risco para a existência humana estaria na ideia de haver uma suposta supremacia das máquinas por conta dessa superinteligência?

A existência de uma inteligência artificial que pudesse ser superior à inteligência humana, ao que parece, dependeria do conhecimento completo de como o cérebro humano opera para produzir sua inteligência. Uma alternativa para isso seria "simplesmente" emular um cérebro humano inteiro, como sustenta Kurzweil em *Como Criar uma Mente: O Segredo do Pensamento Humano Revelado*. A hipótese seria escanear um cérebro humano e programar um computador para operar da mesma maneira que as conexões neurais cerebrais. Contudo, tanto as técnicas atuais de escaneamento do cérebro e o poder computacional existente indicam que este projeto parece difícil de ser viabilizado a curto prazo, mesmo considerando avanços a partir de projetos como o americano *The BRAIN Initiative*[2] ou o europeu *Human Brain Project*[3]. Outra barreira, apontada por Miguel Nicolelis, diz respeito à sequência particular de contingências que determinaram a evolução do cérebro humano:

> *[...] nossa peculiar história evolutiva não pode ser comprimida em nenhum algoritmo computacional, um fato que elimina qualquer esperança de que máquinas, simulações computacionais ou formas artificiais de vida poderiam ser sujeitas a uma lista idêntica de pressões evolutivas, geradas por qualquer código de computador ou outra máquina criada pelo homem. Efetivamente, poderíamos dizer que, como um justo quid pro quo por carregar o legado de sua própria história impresso dentro de seus circuitos, o cérebro recebeu como recompensa a imunidade mais poderosa contra possíveis tentativas de copiar ou reproduzir seus mais íntimos segredos e arte (Nicolelis, 2011, p. 469).*

Para Nicolelis (2011, p. 469), eventos como mudanças ambientais, pandemias, guerra nuclear ou outras colisões de meteoro teriam maior chance de ocorrer – gerando riscos existenciais para a humanidade – do que a hipótese de superinteligências artificiais se voltando contra os seres humanos, no contexto da singularidade.

Assim, é neste âmbito que se descortina o argumento a ser explorado no presente trabalho: partindo-se da hipótese de que a inteligência artificial não representaria o fim da supremacia humana (posto que, na essência, a IA somente demonstraria a inteligência humana em artefatos não biológicos), qual seria, efetivamente, o risco a ser enfrentado? Para além do embate entre tecnofóbicos e tecnofílicos, o que se defende, então, é que os erros de funcionamento de uma inteligência artificial – decorrentes de sobrecarga de informação, de uma programação equivocada ou de uma aleatoriedade do sistema – poderiam sinalizar os verdadeiros riscos existenciais, sobretudo quando se considera que o cérebro biológico, na esteira do viés da automação e com base nos argumentos ainda a serem explorados, tende a assumir de maneira acrítica e de bom grado aquilo que é posto por sistemas ancorados em inteligência artificial.

---

[1] Kurzweil propõe a "Lei dos Retornos Acelerados", segundo a qual em um processo evolucionário a ordem aumenta exponencialmente, o que faz com que o tempo também acelere exponencialmente, determinando, em última instância, que o intervalo de tempo entre eventos relevantes fique cada vez menor.

[2] Anunciado pelo ex-presidente americano Barack Obama em 2013, o projeto BRAIN initiative (*Brain Research Through Advancing Innovative Neurotechnologies*) tem como objetivo principal descrever em detalhes o funcionamento do cérebro humano para entender como se desenvolvem doenças como Alzheimer, Parkinson e demais transtornos mentais. Com isso, viabilizaria a descoberta de suas curas.

[3] O *Human Brain Project*, realizado pela colaboração de cientistas europeus de muitas áreas do conhecimento, visa criar uma simulação do cérebro humano. Surgindo em 2013, quase simultaneamente ao projeto americano, a intenção era concluir esta simulação em 10 anos, ou seja, em 2023.

Dois lados da mesma face: Sobre como entusiastas e críticos da inteligência artificial se aproximam

É fato de que muitas tecnologias surgiram, pelos seus inventores, com propósito muito diferente daquele que o uso pela sociedade acabou consagrando. Kevin Kelly (2012, p. 233) colaciona que os desenvolvedores do fonógrafo, por exemplo, achavam que ele seria utilizado para as últimas declarações de vontade de pessoas à beira da morte; que o rádio teria como propósito retransmitir a missa para a zona rural e que a internet, emergida na época da guerra fria, funcionaria apenas como um backup de informações em caso de um colapso termonuclear. Há um contexto similar com a inteligência artificial, posto que, para muitos, ela seria a solução para grande parte dos problemas da humanidade, ao passo que, para outros, ela representaria o pior cenário possível em termos de riscos existenciais. Sem o benefício do tempo, que poderia demostrar qual narrativa não se sustentaria (a exemplo das previsões sobre o fonógrafo, o rádio ou a internet acima descritas), é interessante se perceber, porém, que ambas as visões – seja aquela que a apoia, seja aquela que a censura, ou até mesmo uma proposta intermediária, na vertente de um otimismo cauteloso – se fundamentam na premissa de que a inteligência artificial funciona. Como se verá, tais perspectivas têm em comum a desconsideração da operacionalização imprecisa, e falha, da máquina, razão pela qual, defende-se, todas essas abordagens são insuficientes para se perquirir o verdadeiro risco de uma inteligência artificial.

Na obra *Life 3.0*, Max Tegmark (2017, p. 30) propõe, diante da miscelânea em torno do assunto, que as abordagens a respeito da inteligência artificial poderiam ser classificadas em três grupos principais: os utópicos digitais, o movimento da IA benéfica, e os tecno-céticos.

Os utópicos digitais, segundo Tegmark (2017, p. 32), afirmam que a vida digital é o próximo (e desejável) passo na evolução cósmica, de forma que se as mentes digitais fossem deixadas livres, os resultados quase que certamente seriam bons. Neste sentido, tratar certas categorias de vida como inferior pelo fato de que elas seriam baseadas em silício, e não em carbono, representaria uma forma de especismo a ser eliminada. Entre os utópicos digitais, Tegmark (2017, p. 32) cita Larry Page, um dos fundadores do Google, para quem se a vida fosse se espalhar pela galáxia e além, então ela deveria ser encapsulada em uma forma digital. "[...] ele (Larry Page) pode entrar para a história como o humano mais influente que já viveu: meu palpite é que se uma vida digital superinteligente engolir nosso universo em minha vida, será por causa das decisões de Larry[4]" (Tegmark, 2017, p. 31). Além de Page, outros dois autores que poderiam ser inseridos no grupo de utopistas digitais são Hans Moravec e Ray Kurzweil (já citado anteriormente). Para Moravec (1988, p. 04), em breve as máquinas terão conhecimento suficiente para cuidar de sua própria manutenção, reprodução e autorreparo sem a necessidade de ajuda. Quando isso acontecer, a cultura dos seres humanos, diz Moravec (1988, p. 04), evoluirá independentemente da biologia humana e de suas limitações, culminando em um processo em que os conceitos de vida, morte e identidade pessoal perderiam seus sentidos atuais, posto que o humano e a máquina dariam origem, de forma simbiótica, a uma nova forma (positiva) de existência. Kurzweil (2005, p. 09), por sua vez, argumenta que a singularidade representaria o auge da fusão entre o pensamento biológico e a tecnologia, de modo que os seres humanos teriam poder sobre seu destino, estendendo seu alcance físico e mental para padrões inimagináveis: "Ao final deste século, a parte não biológica da nossa inteligência será trilhões de trilhões de vezes mais poderosa que a inteligência humana não melhorada[5]" (Kurzweil, 2005, p. 09). Segundo Kurzweil (2005, p. 09), a implicação mais importante da singularidade

[4] Tradução do original: "[...] he might go down in history as the most influential human ever to have lived: my guess is that if a superintelligent digital live engulfs our Universe in my lifetime, it will be because of Larry's decisions".
[5] Tradução do original: "By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence".

estaria no aspecto de refinamento dos melhores traços e características dos seres humanos, o que reafirma o prognóstico dos utopistas digitais de que a inteligência artificial seria algo efetivamente bom.

O movimento da IA benéfica, por outro lado, de acordo com Tegmark (2017, p. 37), argumentaria, no contexto da carta aberta do encontro de Porto Rico, em 2015[6], que talvez, e pela primeira vez na história, a humanidade poderia vir a possuir, por intermédio da inteligência artificial, uma tecnologia capaz tanto de resolver todos os problemas do planeta, quanto de acabar com a própria vida também. Stuart Russel exemplifica tal ambivalência:

> *Se preferir resolver problemas ambientais, você pode pedir à máquina que contenha a rápida acidificação dos oceanos resultantes dos níveis mais altos de dióxido de carbono. A máquina desenvolve um novo catalisador que facilita uma reação química incrivelmente rápida entre oceano e atmosfera e restaura os níveis de pH dos oceanos. Infelizmente, um quarto do oxigênio da atmosfera é usado no processo, deixando-nos asfixiados, de forma lenta e dolorosa (Russel, 2021, p. 135).*

Para o movimento da IA benéfica, é fato que a inteligência artificial poderia caminhar para resolver grande parte dos problemas humanos na contemporaneidade, mas, ao mesmo tempo, e aí estaria sua ambiguidade, também possibilitaria a própria extinção da humanidade, caso certas medidas e protocolos, especialmente antes do surgimento da uma superinteligência ou de uma inteligência artificial geral (AGI), não fossem pensados ou colocados em ação. Como se vê, entende-se que há implícito no movimento da IA benéfica a premissa de que, bem ou mal, a inteligência artificial possa vir a amadurecer e evoluir de forma que os seres humanos não consigam controlá-la, o que faz, neste sentido, que os utopistas digitais e o movimento da IA benéfica estejam no mesmo espectro de debate, apenas diferindo em relação às últimas consequências do uso da inteligência artificial: para os utopistas digitais, positiva em essência; para o movimento da IA benéfica, ambivalente.

O último grupo de abordagem em relação à IA, os tecno-céticos, defendem, de acordo com Tegmark (2017, p. 33), que temer o surgimento de robôs assassinos, por exemplo, estaria no mesmo quadrante de se preocupar com a superpopulação em Marte. Em outras palavras, intimidar-se com a hipótese de uma AGI seria uma inquietação desnecessária no entendimento de Andrew Ng e Rodney Brooks, autores descritos como tecno-céticos, posto que tal possibilidade seria tão remota que sequer se vislumbraria no horizonte de centenas de anos em relação ao futuro. Em que pesem os argumentos trazidos, entende-se, em relação aos céticos, que o principal ponto de risco sobre uma inteligência artificial não diria respeito ao momento em que ela se tornaria uma superinteligência, adquirindo consciência, ou até mesmo acerca da singularidade, mas a uma situação muito mais simples, já presente no atual contexto, de uma IA com falha de funcionamento, ou com má compreensão pelos seres humanos, como se argumentará a partir de agora.

Falhas e Riscos em torno da Inteligência Artificial

Yampolskiy (2016), inspirado no texto *Information Hazards: A Typology of Potential Harms From Knowledge* de Bostrom (2011), elabora uma classificação didática quanto ao momento e à forma como sistemas de IA se tornam perigosos. São dois estágios apresentados por Yampolskiy para análise do momento em que sistemas de IA podem apresentar riscos, um anterior e outro posterior à sua implantação. Ressalta-se que essa divisão é de caráter didático, uma vez que problemas de funcionamento ou falhas podem acontecer na continuidade dos dois estágios. Quanto à forma como sistemas de IA se tornam perigosos, o autor as classifica em relação às causas, que podem ser externas ou internas.

[6] O texto foi publicado pelo *Future of Live Institut* e pode ser encontrada neste endereço: http://futureoflife.org/ai-open-letter/

As causas externas podem ser divididas em: i) ações deliberadas; ii) efeitos colaterais de projetos ruins; e iii) situações diversas ocorridas no ambiente externo do sistema. Já as causas internas são as decorrentes de automodificações originadas no próprio sistema, que é o destaque dado neste artigo quanto aos efetivos riscos existências para a humanidade.

O quadro didático elaborado por Yampolskiy ilustra muito bem as diferentes situações em que um sistema de IA pode se tornar perigoso:

Table 1: Pathways to Dangerous AI

How and When did AI become Dangerous		External Causes			Internal Causes
		On Purpose	By Mistake	Environment	Independently
Timing	Pre-Deployment	a	c	e	g
	Post-Deployment	b	d	f	h

Fonte: Yampolskiy, R. V. "Taxonomy of Pathways to Dangerous Artificial Intelligence," in Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016, p.143.

A análise de todas as situações aqui apresentadas pode sustentar o que se pretende neste artigo, que é evidenciar o efetivo risco existencial que a IA pode trazer ao futuro da humanidade.

A primeira situação, identificada no quadro acima, é quando ocorre causas externas de maneira proposital, os *Hazardous Software*, isto é, sistemas de IA maléficos por conta das intenções de seus projetistas e desenvolvedores, usados com finalidades socialmente prejudiciais, tais como vírus, spyware, cavalos de Tróia ou Worms. Os malefícios tendem a aumentar conforme os *Hazardous Software* possam funcionar de forma mais sofisticada com o aperfeiçoamento da IA. O risco, neste caso, não está na IA em si, uma vez que ela foi projetada para uma finalidade que ela irá funcionar. Aqui, o bom funcionamento da IA acarreta efeitos maléficos para a sociedade, com consequências dificilmente previsíveis pela inteligência humana. Não resta dúvida de que os riscos desses sistemas de IA intencionalmente projetados para usos maléficos são muito ameaçadores. Há muito se fala sobre os efeitos de IA projetadas para finalidades destrutivas, como arma de guerra[7]. Contudo, a origem do risco é humana, com potencial controle mediante aplicação de medidas punitivas, por exemplo.

Ainda que um sistema de IA seja projetado para finalidades benéficas, após seu desenvolvimento podem ocorrer situações que comprometam seu funcionamento seguro. Por exemplo, considerando que o treinamento de IA se dá mediante a captura e acesso a grandes quantidades de dados, há a possibilidade de que esses dados sejam intencionalmente alterados para que, embora a IA funcione corretamente e tenha sido projetado com finalidade benéfica, o resultado desse funcionamento possa incorrer em prejuízos, como mudança de tendência no mercado financeiro. Sistemas de IA estão suscetíveis a interferências de hackers, o que a torna potencialmente perigosa, mesmo que projetada para finalidades benéficas. No caso de já projetadas para finalidades destrutivas, como os robôs "assassinos" ou os vírus para ataques militares cibernéticos, uma invasão hacker potencializaria ainda mais a periculosidade desses sistemas de IA. Isso seria similar ao risco de alguém com intenções espúrias conseguir roubar uma arma biológica militar, como um vírus de laboratório.

[7] O movimento de coalisão *Stop Killer Robots* (www.stopkillerrobots.org) é uma reação a essas ameaças. A campanha explicita a preocupação com a IA projetada para finalidades violentas como, no caso, a produção de robôs "assassinos". O lema "*Less autonomy. More humanity*" é um chamado para que as novas tecnologias possam promover as pessoas e não reduzi-las a objetos, estereótipos ou alvos. Por isso, clama por uma nova lei internacional sobre o uso de IA em armas com grande potencial destrutivo.

Uma segunda situação de causas externas em que sistemas de IA podem se tornar perigosos se dá mediante enganos do programador na elaboração ou desenvolvimento da forma como a IA é projetada. Isso porque ela pode conter vários bugs não detectados, erros de design, objetivos não alinhados a valores humanos e recursos mal desenvolvidos, todos os quais são capazes de produzir resultados altamente indesejáveis. Os equívocos podem ser de caráter técnico, que acarretam algum funcionamento problemático da IA, como algoritmos mal elaborados em relação aos propósitos humanos ou alguma incompatibilidade da programação do sistema com o código-fonte. Porém, equívocos podem também ocorrer, mesmo considerando os aspectos técnicos adequadamente cumpridos, como é o caso do desconhecimento do projetista ou desenvolvedor sobre aspectos culturais dos usuários que pode levar uma IA projetada e implementada corretamente pelo Estado Islâmico para fazer cumprir a Lei da Sharia a ser considerada maléfica para a cultura ocidental ou, o contrário, levar uma IA corretamente projetada e implementada pelo Ocidente para impor a democracia liberal a ser considerada malévola no Estado Islâmico (Yampolskiy, 2016, p. 144).

Nas pesquisas em Computação Afetiva, considerado o campo de estudo interdisciplinar que visa reconhecer, entender, simular e estimular estados afetivos no projeto de sistemas computacionais (Calvo et al., 2015), essa segunda categoria de causas de uma IA perigosa é significativamente preocupante. Corporações gigantes da tecnologia como Apple, Amazon, Google e Facebook, e centenas de outras empresas menores, estão implantando métodos de computação afetiva para prever ou influenciar o comportamento do consumidor[8]. Uma IA poderia ser treinada para prever o que um terceiro pensaria sobre o que uma pessoa está expressando, ao olhar para uma imagem de seu rosto (por exemplo, se eles estão sorrindo). Contudo, a associação entre expressão e sentimento, quando aprofundado com autorrelatos, mostra-se às vezes incerta, pois as pessoas podem sorrir por motivos variados: pode ser por educação, por serem surpreendidas ou até por frustração. Quando uma pessoa sorri diante de uma tela, fornece indícios, mas não certezas sobre o que ela acha engraçado, e se está achando engraçado. Por isso, esses dados devem ser combinados com outras evidências e fatores contextuais. Isso é particularmente preocupante quando as expressões são usadas para questões de grande impacto social, como um processo seletivo de trabalho ou identificação de suspeitos de alguma conduta ilegal. A maioria das pesquisas em Computação Afetiva tem como objetivo tonar as interações das máquinas com os humanos mais naturais. No entanto, é provável que uma IA ensinada para responder de maneira emocional seja considerada perigosa por conta de como esse estado de afeto poderia influenciar o comportamento humano.

Uma IA poderia se tornar perigosa devido à imprevisibilidade por parte de seus projetistas e desenvolvedores. Problemas de funcionamento ocorridos após sua criação poderiam gerar resultados diferentes daqueles originalmente projetados. São várias as situações possíveis, como problemas de comunicação entre humano e máquina, em que um termo detectado de forma desalinhada ao objetivo do humano pode tornar uma IA perigosa. Contudo, essa situação pode ser evitada com o sucessivo aprimoramento do ajuste de comandos e reconhecimento de comandos entre humano e máquina, algo que o processamento da linguagem natural continuamente tem alcançado. É o caso da WuDao 2.0, "uma das máquinas/algoritmos que mais se assemelha a um ser humano, no que diz respeito à comunicação", todavia, ainda "longe de ter as soluções adequadas para conversas naturais com as máquinas" (Kavinski, 2022). A WuDao 2.0 é uma IA pré-treinada em inglês e chinês para simulação de fala, escrita de poesia, reconhecimento de imagem e geração de texto. Desenvolvida pelos chineses, essa IA usa 1,75 trilhão de parâmetros, dez vezes mais que a segunda mais poderosa para essa finali-

[8] Bem por isso que o *AI Now Institute*, um centro de pesquisa interdisciplinar dedicado a entender as implicações sociais da inteligência artificial, em seu relatório de 2019, considerou a computação afetiva como sua principal preocupação social (Crawford et al., 2019).

dade artística, a GPT-3 da OpenAI, que usa 175 bilhões de parâmetros (Willemart, 2021). Esse exemplo mostra que o aumento da capacidade tecnológica e técnica tendem a amenizar os perigos de uma IA quanto aos problemas de comandos na interação entre humano e máquina.

Quanto a problemas de funcionamento que irão ocorrer após sua criação de um sistema de IA, é relevante destacar uma questão para os propósitos deste artigo: problemas podem ocorrer como efeitos colaterais da resolução de conflitos entre comandos, incompatíveis em um domínio específico ou interações de software versus hardware que não são detectadas no momento de sua inconsistência. Com a evolução do sistema, esses problemas podem se tornar imprevisíveis e não verificáveis, devido ao aumento de sua complexidade e falta de transparência. Isso poderia gerar um sistema tão a nossa frente que potencialmente inviabilizaria uma comunicação no nosso nível. Essa falta de controle sobre a evolução da IA, embora de origem humana, é uma das probabilidades reais de risco existencial. Contudo, ressalta-se, a origem ainda assim é humana, pois decorre da nossa incapacidade de previsibilidade diante do aumento de potencialidade e complexidade da IA.

É importante notar que os problemas ligados a equívocos ou falta de previsibilidade de problemas que venham a surgir ao longo da evolução de um sistema de IA são intensificados pela pressão comercial, que exige inovação com rapidez. A limitação de tempo para uma análise mais profunda do sistema, bem como revisões em seu projeto ou até mesmo desenvolvimento com maior cautela, são formas de minimizar os riscos.

Uma terceira situação de causas, ainda externas, em que sistemas de IA podem se tornar perigosos, são os decorrentes de questões ambientais de origem não humana. Há situações especulativas, como uma IA obtida por sinal extraterrestre, ou situações muito raras, como a troca de bits individuais ocorridos em diferentes dispositivos de hardware devido a defeitos de fabricação ou algum tipo de descarga elétrica ou radiação. Yampolskiy (2016) apresenta essa terceira categoria de origem de riscos para uma IA como maléfica aos propósitos humanos, entretanto, por ser de caráter especulativo ou por ser estatisticamente muito restrito que algo assim aconteça, não suscitaria riscos existenciais reais.

A quarta e última modalidade de origens de problemas é de caráter interno ao sistema, independente da participação humana. Para Yampolskiy (2016), uma das abordagens mais prováveis para criar IA superinteligente é cultivá-la a partir de uma IA "bebê" por meio de autoaperfeiçoamento recursivo (RSI, de *recursive self-improvement*). Um cenário perigoso é aventado por autores como Kurzweil, com a previsão de que a evolução de um sistema IA pode se tornar autoconsciente, capaz de decidir sobre quais regras seguir ou não, com propósitos próprios, possivelmente em detrimento da humanidade. Contudo, conforme Kurzweil, já tratado neste artigo, os riscos decorrentes da singularidade se dariam pela supremacia de inteligência que uma IA viria a ter sobre os humanos. Kurzweil pressupõe a evolução de IA mediante seu correto funcionamento, sem falhas, até alcançar sua autoconsciência.

A possibilidade de uma autoconsciência é rechaçada por filósofos como Searle (2014). Para este, os computadores não vão dominar o mundo, uma vez que carecem totalmente de uma "realidade psicológica" e nada mais são do que sistemas de circuitos bem projetados e altamente funcionais. A ideia de computadores superinteligentes, com propósitos intencionais, que coloquem a humanidade em risco com base em suas próprias crenças e desejos ou outras motivações é, para Searle (2014), "irreal porque a máquina não tem crenças, desejos e motivações". Assim, o profundo questionamento quanto à possibilidade de autoconsciência da IA torna também questionável a presença de riscos existenciais reais de uma superinteligência capaz de agir deliberadamente contra os humanos. Aliás, os perigos da alta capacidade futura de sistemas de IA podem decorrer justamente pela inexistência de crenças, desejos e emoções, com o funcionamento "sociopata" da máquina[9].

[9] "O que chamamos de doença mental nas pessoas, particularmente a sociopatia, demonstrada pela falta de preocupação com os outros, também pode aparecer em mentes artificiais. Uma variante leve de comportamento antissocial pode ser o caso dos

Todavia, os riscos reais dessa modalidade interna de origem de problemas, conforme pretende-se aqui sustentar, se dá pela desconexão de acompanhamento humano do funcionamento interno do sistema, sem identificar sua complexidade, bem como suas falhas indetectáveis. Assim como ocorreu com o Deep Blue da IBM, o primeiro caso tratado neste artigo, outro exemplo muito significativo de falha de IA pôde ser visualizado no Flash Crash da tarde de 06 de Maio de 2010, ocasião em que uma quebra trilionária das bolsas norte-americanas ocorreu em um período aproximado de trinta minutos. O interessante, nas palavras de Liam Vaughan (2020, p. 83), foi o fato de ninguém entender como o mercado acionário havia operado durante tal período, restando o fechamento da bolsa como única alternativa aos seres humanos:

> *O Flash Crash havia posto a nu as limitações dos reguladores, mas mais do que isso, havia despertado a população em geral para o fato de que toda a estrutura dos mercados financeiros havia se deslocado sob seus pés sem que eles percebessem. Aqueles poucos minutos de ansiedade, quando parecia concebível que todo o edifício iria cair, tinham levado a um acerto de contas. Os senadores falaram por todos os americanos quando ponderaram quais eram as repercussões de automatizar completamente um sistema que determinava o valor de nossas empresas, nossas economias e os alimentos e recursos que consumíamos; em cujos interesses era para os títulos mudar de mãos milhares de vezes por minuto; quem eram os vencedores e os perdedores desta mudança de paradigma; e o que aconteceria se a tecnologia fosse abusada (Vaughan, 2020, p. 102).*

Mesmo quando operam sem falhas, ainda assim a inteligência artificial, em se tratando do funcionamento de bolsas de valores, parece trazer problemas para a cognição humana[10].

Viés da automação e seres humanos que não pensam

Na obra *Rápido e Devagar: duas formas de pensar*, Daniel Kahneman (2011, p. 29) propõe que a mente humana seria composta pelo Sistema 1, que operaria de forma rápida, automática, com pouco ou nenhum esforço, e sem percepção de controle voluntário; e o Sistema 2, que seria demandado em situações de escolha e concentração, nas quais as atividades mentais mais complexas fossem requisitadas.

---

palavrões excessivos do IBM Watson [...], causados pelo aprendizado com dados incorretos. Da mesma forma, qualquer sistema de IA aprendendo com maus exemplos pode acabar sendo socialmente inadequado, como um ser humano criado por lobos" (Yampolskiy, 2016, p.146). O caso da IBM ocorreu em 2013. Para que o supercomputador da IBM parecesse o mais próximo possível a um humano real, Eric Brown e sua equipe projetaram o Watson para realizar a comunicação e conversação humana informal mediante a memorização do Urban Dictionary. Esse dicionário é constituído por colaborações de pessoas comuns, minimamente regulado por editores voluntários para manter certo controle de qualidade. Por conta da informalidade de linguagem almejada, o que poderia ter sido outro grande avanço do Watson acabou produzindo uma situação embaraçosa, uma vez que passou a usar muitos termos preconceituosos, de cunho sexista, racista e homofóbico.

[10] Na obra *Flash boys: revolta em Wall Street*, Michael Lewis (2014, p. 38) retrata o seguinte panorama: "Em 2002, 85% de todas as operações aconteciam na Bolsa de Nova York, e era um ser humano que processava cada ordem de compra e venda. As ações que não eram negociadas na Bolsa de Nova York o eram na Nasdaq. Nenhuma era negociada nas duas. Por determinação da SEC, reagindo por sua vez aos protestos públicos contra panelinhas dentro do mercado, as próprias bolsas, em 2005, deixaram de ser serviços de utilidade pública pertencentes a seus próprios integrantes e se transformaram em companhias abertas com fins lucrativos. Introduzida a concorrência, as bolsas se multiplicaram. No começo de 2008, havia treze bolsas de capital aberto, a maioria delas no norte de Nova Jersey. Praticamente todas as ações passaram a ser negociadas em todas essas bolsas: ainda era possível comprar e vender ações da IBM na Bolsa de Nova York, mas também se podia comprá-las e vendê-las na Bats, na Direct Edge, na Nasdaq, na Nasdaq BX etc. A ideia de que era necessário um ser humano intermediando investidores e mercado não existia mais. A "bolsa" na Nasdaq, na Bolsa de Nova York ou em suas novas concorrentes, como a Bats e a Direct Edge, consistia num grande conjunto de servidores que rodavam o programa chamado 'matching engine', que servia para fechar ordens de compra e venda. Não havia ninguém dentro da bolsa com quem conversar. (...). Ao mesmo tempo, as bolsas estavam mudando a maneira como lucravam. Em 2002, cobravam de todos os corretores de Wall Street que enviavam uma ordem de compra ou venda a mesma comissão simples e fixa por ação negociada. A substituição de pessoas por máquinas tornou os mercados não só mais rápidos como também mais complexos. As bolsas desenvolveram um sistema incrivelmente complexo de taxas e comissões. O sistema se chama 'modelo maker-taker' [formador/tomador de liquidez] e, como muitas das invenções de Wall Street, quase ninguém entendeu".

Segundo Kahneman (2011, p. 29), "quando pensamos em nós mesmos, nos identificamos com o Sistema 2, o eu consciente, raciocinador, que tem crenças, faz escolhas e decide o que pensar e o que fazer a respeito de algo". No entanto, como as pesquisas sobre o funcionamento da mente demonstraram, a participação do Sistema 1 seria mais relevante na cognição humana do que inicialmente se supunha:

> *A maior parte do que você (seu Sistema 2) pensa e faz origina-se de seu Sistema 1, mas o Sistema 2 assume o controle quando as coisas ficam difíceis, e normalmente ele tem a última palavra. A divisão de trabalho entre o Sistema 1 e o Sistema 2 é altamente eficiente: isso minimiza o esforço e otimiza o desempenho. O arranjo funciona bem na maior parte do tempo porque o Sistema 1 geralmente é muito bom no que faz: seus modelos de situações familiares são precisos, suas previsões de curto prazo são em geral igualmente precisas e suas reações iniciais a desafios são rápidas e normalmente apropriadas. O Sistema 1 tem vieses, porém, erros sistemáticos que ele tende a cometer em circunstâncias específicas (Kahneman, 2011, p. 34).*

Como tem se defendido ao longo do artigo, uma dessas circunstâncias específicas em que se cometeriam erros sistemáticos ocorreria a partir da dificuldade da mente humana de perceber falhas no funcionamento da inteligência artificial, sobretudo se considerada a confiança que atualmente os indivíduos já depositam nos algoritmos. Ao documentar a dependência humana em sistemas de computação, James Bridle (2019, p. 51) deu a tal problema o nome de viés da automação, o qual "garante que daremos mais valor à informação automatizada do que à nossa experiência, mesmo quando ela conflita com outras observações – especialmente quando essas observações são ambíguas" (Bridle, 2019, p. 51). Alguns exemplos da aviação tornam mais tangível a discussão. Segundo Bridle (2019, p. 53), alarmes de incêndio equivocados das primeiras versões do Airbus A330 fizeram com que tripulações mudassem a rota de voo, mesmo quando os pilotos atestavam, visualmente, que não havia qualquer incêndio a bordo. Em outro caso, um voo da Korean Airlines foi derrubado pelo sistema de defesa soviético, mesmo sob os protestos do controle aéreo para que o avião retomasse sua rota original: "Segundo os investigadores, havia várias pistas durante essas horas que podiam ter alertado a tripulação do que se passava. (...). Mas nenhum desses efeitos levou os pilotos a questionarem o sistema (...). Eles continuaram a confiar no piloto automático" (Bridle, 2019, p. 52). Mais recentemente, sob a vigência de sistemas de IA, Bridle fez referência à "morte por GPS":

> *Tentando chegar a uma ilha da Austrália, um grupo de turistas japoneses dirigiu até uma praia e mar adentro porque seu sistema de navegação por satélite garantia que aquela era uma rota viável. Tiveram de ser resgatados antes de a maré subir, a aproximadamente quinze metros da costa. Outro grupo, no estado de Washington, entrou com o carro em um lago quando foi direcionado a sair da estrada principal e entrar em um embarcadouro. Quando os serviços de emergência chegaram, encontraram o carro boiando na água, só o bagageiro superior à vista. (...). Nesses casos, o sinal de GPS não foi enganado e não se desviou. Simplesmente se fez uma pergunta ao computador e ele respondeu – e os seres humanos seguiram a resposta até a morte (Bridle, 2019, p. 54).*

Para Nicolelis (2020, p. 335), o cérebro humano seria o mais perfeito camaleão criado pela natureza, posto que, uma vez exposto a contingências do mundo exterior, iniciaria um processo de autorreformatação de sua microestrutura interna orgânica, desencadeando uma sequência de modificações de comportamentos e ações. Dessa forma, sustenta Nicolelis (2020, p. 343), a mente humana seria capaz de assimilar instrumentos artificiais de natureza mecânica, alterando, ao fim e ao cabo, o próprio senso do "eu". O perigo, como defende Nicolelis (2020, p. 345), residiria no fato de que toda vez que um ser humano assumisse uma posição subalterna (ou de coadjuvante) em relação a um sistema digital, as habilidades humanas tenderiam a degradar, ocasionando pontos de erros incomuns até então. Some-se a isso, na perspectiva de Bridle (2019, p. 54), que a contemporaneidade seria marcada por problemas

complexos e pela pressão extraordinária de tempo, de maneira que indivíduos buscariam dispensar o mínimo de esforço cognitivo para resolver tarefas: "A computação, em qualquer escala, é um hack cognitivo, que entrega à máquina tanto o processo de decisão quanto a responsabilidade. Conforme a vida se acelera, a máquina começa a lidar cada vez com mais tarefas cognitivas" (Bridle, 2019, p. 55). Assim, a grande ameaça estaria no surgimento (ou crescimento?) de uma humanidade que não pensa. "Estamos à beira de uma era em que as artes e as ideias serão derivadas dos algoritmos. As máquinas cada vez mais sugerem os temas mais populares para a investigação humana, e os seres humanos estão obedecendo cada vez mais" (Foer, 2018, p. 199-200). E se, como defendido, o principal problema de uma inteligência artificial não estaria em causas externas, mas em falhas de funcionamento internas, independentes, percebe-se o quão grave são esses riscos para o futuro da humanidade. Afinal de contas, como pondera Russel (2021, p. 128), a questão está em saber se são os humanos quem controlam a IA ou se são os humanos quem se tornaram ferramenta da IA, passando a fornecer "informações e corrigindo bugs quando necessário, mas incapazes de compreender em profundidade como a coisa toda funciona" (Russel, 2021, p. 128).

Considerações finais

Na obra *Robot-Proof: Higher Education in the Age of Artificial Intelligence*, Joseph Aoun (2017, p. xix) defende que seria necessário repensar o sistema educacional, de maneira que as universidades deveriam começar a ensinar *humânica*. Para Aoun (2017, p. xix), o conhecimento por si só não seria suficiente para o trabalho do futuro, de modo que a *humânica* passaria a ser um conjunto de capacidades cognitivas de natureza mais elevada, capaz de integrar sistemas de pensamento, mentalidades criativas, agilidades culturais, todas retroalimentadas por análises críticas e racionais. De acordo com Aoun (2017, p. 53), "(...) precisamos de um novo modelo de aprendizagem que permita aos alunos compreender o mundo altamente tecnológico ao seu redor e que simultaneamente lhes permita transcendê-lo[11]".

Sem embargo da proposta da *humânica* – que se assemelha a outras iniciativas para o sistema de ensino, como a dos quatro C's: *critical thinking, communication, collaboration, creativity* (pensamento crítico, comunicação, colaboração e criatividade) – entende-se que ainda que fatores educacionais possam ser trabalhados para mitigar e, eventualmente, até eliminar os riscos de inteligências artificiais originadas por causas externas, como no caso de (i) ações deliberadas, (ii) efeitos colaterais de projetos ruins e (iii) questões ambientais de origem não humana. Defende-se, por outro lado, que na hipótese de causas internas, nas quais as automodificações sejam originadas do próprio sistema, com incompreensão a respeito do funcionamento da IA, o risco existencial permaneceria incólume, posto que o cérebro humano teria muitas dificuldades em perceber a falha e, mesmo percebendo, de tomar decisões que contrariassem o sugestionamento da inteligência artificial.

Referências

AOUN, J. E. 2017. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. Cambridge, MIT Press.

BOSTROM, N. 2011. Information Hazards: A Typology of Potential Harms From Knowledge. *Review of Contemporary Philosophy,* **10**: 44-79.

BRIDLE, J. 2019. *A nova idade das trevas: A tecnologia e o fim do futuro*. Tradução Érico Assis. São Paulo, Todavia.

[11] Tradução do original: "(...) we need a new model of learning that enables learners to understand the highly technological world around them and that simultaneously allows them to transcend it".

CALVO, R. A.; D'MELLO, S.; GRATCH, J.; KAPPAS, A. 2015. *The Oxford handbook of affective computing*. New York, Oxford University Press.

CRAWFORD, K.; DOBBE, R.; DRYER, T., *et al*. 2019. *AI now 2019 report*. New York, NY, AI Now Institute.

CLARKE, A. 1962. *Profiles of the Future: An inquiry into the Limits of the Possible*. London, Orion House.

FOER, F. 2018. *O mundo que não pensa*. Tradução Debora Fleck. Rio de Janeiro, LeYa.

HAWKING, S. 2018. *Breves respostas para grandes questões*. Tradução Cassio de Arantes Leite. Rio de Janeiro, Intrínseca.

KAHNEMAN, D. 2012. *Rápido e devagar: duas formas de pensar*. Tradução Cassio de Arantes Leite. Rio de Janeiro, Objetiva.

KAVINSKI, A. 2022. *A corrida pelo Processamento de Linguagem Natural*. *MIT Technology Review*, 12 de janeiro de 2022.

KELLY, K. 2012. *Para onde nos leva a tecnologia*. Tradução Francisco Araújo da Costa. Porto Alegre, Bookman.

KURZWEIL, R. 2007 *A Era das Máquinas Espirituais*. Tradução Fábio Fernandes. São Paulo, Aleph.

KURZWEIL. R. 2014. *Como criar uma mente: os segredos do pensamento humano*. Tradução de Marcello Borges. São Paulo, Aleph.

KURZWEIL, R. 2005. *The singularity is near: when humans transcend biology*. New York, Penguin Books.

LEWIS, M. 2014. *Flash boys: revolta em Wall Street*. Tradução Denise Bottmann. Rio de Janeiro, Intrínseca.

MORAVEC, H. P. 1988. *Mind children: the future of robot and human intelligence*. Cambridge, Massachussetts, Harvard University Press.

NICOLELIS, M. *Muito além do nosso eu: a nova neurociência que une cérebros e máquinas - e como ela pode mudar nossas vidas*. São Paulo: Companhia das Letras, 2011.

NICOLELIS, M. 2020. *O verdadeiro criador de tudo: Como o cérebro esculpiu o universo como nós o conhecemos*. São Paulo, Planeta.

RUSSELL, S. 2021. *Inteligência artificial a nosso favor: Como manter o controle sobre a tecnologia*. Tradução Berilo Vargas. São Paulo, Companhia das Letras.

SEARLE, J. R. 2014. What Your Computer Can't Know. *The New York Review of Books*, October 9, 2014.

SILVER, N. 2012. *The Signal and the Noise: Why Most Predictions Fail – but Some Don't*. New York, The Penguin Press.

TEGMARK, M. 2017. *Life 3.0: being human in the age of artificial intelligence*. New York, Alfred A. Knopf.

VAUGHAN, L. 2020. *Flash crash: a trading savant, a global manhunt, and the most mysterious market crash in history*. New York, Doubleday.

WILLEMART, P. 2021. Arte e Programas de Inteligência Artificial: GPT-2, GPT-3, Wu Dao 2.0. *Revista Desenredo*, **17**(3): 514-524.

YAMPOLSKIY, R. V. 2016. Taxonomy of Pathways to Dangerous Artificial Intelligence. In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Submetido em 13 de Março de 2022.

Aceito em 25 de Julho de 2022.