2018 JAN-JUN
Nº 17

TECCOGS

revista digital de tecnologias cognitivas

EXPEDIENTE

TECCOGS – Revista Digital de Tecnologias Cognitivas, nº 17, Jan-Jun 2018, ISSN: 1984-3585
Programa de Pós-graduação em Tecnologias da Inteligência e Design Digital (TIDD) | PUC-SP

SUMÁRIO

RESENHA

Yuval Noah Harari, *Homo Deus:* uma breve história do amanhã

Por Rodrigo Petrônio

EDITORIAL

Dora Kaufman[1]

Esta 17ª edição da TECCOGS é dedicada ao tema da Inteligência Artificial (IA), mais precisamente aos avanços recentes e seus impactos na sociedade, no mercado, nas empresas e nos indivíduos. Diversas expressões têm sido utilizadas por diferentes comunidades – reconhecimento de padrões, modelagem estatística, mineração de dados, descoberta de conhecimento, análise preditiva, ciência de dados, sistemas adaptativos, sistemas de auto-organização, e outros – e alguns até o denominam simplesmente Inteligência Artificial. Independente do nome e da funcionalidade, a IA permeia atualmente quase todas as atividades do planeta, facilitando a vida do século XXI e, simultaneamente, colocando novos desafios. A chamada *weak* IA tem foco no desenvolvimento de algoritmos e técnicas para solucionar determinados problemas, executar uma tarefa específica. A *strong* IA, apesar de ainda ser considerada ficção, é tema de debates em eventos internacionais, acadêmicos e não acadêmicos, e de vasta bibliografia pelos seus potenciais impactos na humanidade.

As tecnologias de IA são "disruptivas" em praticamente todos os domínios. Emergem inúmeras questões, parte delas são abordadas ao longo dos sete artigos que compõem essa edição de autoria de pesquisadores com múltiplas formações, afiliações e repertórios. Nosso propósito é contribuir minimamente para que os pesquisadores, particularmente de ciências sociais e humanas, adquiram um entendimento básico que possibilite se situar e atuar na "sociedade de dados".

A visibilidade conquistada pela Inteligência Artificial desde os anos 2006-2010 decorre dos estrondosos resultados obtidos com base no processo denominado de aprendizagem profunda (*deep learning*), que permite às máquinas aprenderem a partir de exemplos. A técnica (como alguns a consideram) transformou os dados (*big data*),

[1] Pós-doutoranda pelo TIDD/PUC-SP, pós-doutora pela COPPE-UFRJ, doutora ECA-USP com período na Université Paris – Sorbonne IV. Pesquisadora visitante e palestrante no Computer Science Department, Courant Institute of Mathematical Sciences, NYU (2009, 2010), e no Alexander von Humboldt Institute for Internet and Society, Berlim (2015). Pesquisadora do Atopos ECA/USP (desde 2011), participa do Grupo de Estudos em Inteligência Artificial do Instituto de Estudos Avançados da USP.
E-mail: kaufman1955@gmail.com.

gerados exponencialmente em nossas interações digitais, em informações úteis e procedimentos que vão desde diagnósticos médicos até simples recomendações de filmes e músicas. Dada sua relevância, o Dossiê é inteiramente dedicado a descrever noções básicas de aprendizagem profunda agregando contribuições de autores relevantes.

A entrevista dessa edição é com Davi Geiger, professor associado em Ciência da Computação e Ciência Neural do Instituto Courant da Universidade de Nova York, PhD pelo Instituto de Tecnologia de Massachusetts (MIT) e há 35 anos dedicado ao tema da IA, com experiências no desenvolvimento de dispositivos e soluções práticas, como um recente sistema de automação para o varejo. O professor Geiger nos contempla com sua definição sobre IA, suas percepções sobre os recentes avanços e questões éticas envolvidas.

O primeiro artigo, "Inteligência Artificial: uma utopia, uma distopia", é do professor titular da Poli USP e PhD pela Carnegie Mellon University, Fabio Gagliardi Cozman, que enumera um conjunto de "distopias" e "utopias" associado a evolução da IA finalizando com um cenário denominado por ele de "utopia realista" em contraponto à "distopia realista". Como pesquisador das ciências exatas, o professor Fabio nos contempla com uma visão singular.

O segundo artigo, "O protagonismo dos algoritmos de Inteligência Artificial: observações sobre a sociedade de dados", é de autoria da pesquisadora dos impactos sociais da IA, Dora Kaufman, editora dessa edição, doutora pela ECA/USP, pós-doutora pela COPPE/UFRJ e pós-doutoranda no programa de Tecnologias da Inteligência e Design Digital (TIDD) da PUC-SP. O artigo descreve e analisa alguns dos impactos da chamada "sociedade de dados" na perspectiva dos indivíduos e na perspectiva dos mercados e das empresas.

O terceiro artigo, "Efeitos das construções mecânicas de significados: questões semióticas entre a possibilidade, o dever e o direito a explicações em IA", é fruto de um projeto de pesquisa da IBM Research–Brazil. Os seus autores são Joel Luis Carbonera, doutor em Ciência da Computação na Universidade Federal do Rio Grande do Sul, Bernardo Gonçalves, doutor pelo Laboratório Nacional de Computação Científica (LNCC) com pós-doutorado na Universidade de Michigan–Ann Arbor e atualmente

doutorando em Filosofia da Ciência/USP, e Clarisse de Souza, professora titular do Departamento de Informática PUC-Rio e doutora em Linguística Aplicada.

O quarto artigo, "Uma cartografia comum aproximando Inteligência Artificial, Filosofia e Psicologia", é de coautoria dos professores titulares da UNINTER, Luciano Frontino de Medeiros, doutor pela Universidade Federal de Santa Catarina, Marilene Garcia, com pós-doutorado pelo TIDD, e Alvino Moser, doutor em Filosofia e Ética pela Université Catholique de Louvain/Bélgica. O artigo relaciona quatro problemas epistemológicos a partir de uma perspectiva onde a Inteligência Artificial compartilha um domínio de conhecimento comum com as áreas da Filosofia e da Psicologia.

O quinto artigo, "Interação, indistinguibilidade e alteridade na Inteligência Artificial", é de autoria de João Cortese, doutor em co-tutela na Université Paris 7 e Filosofia da USP, aborda a questão ética por meio do tema da eficiência a partir da constatação de que a tecnologia da computação moderna, para ter competência sobre uma tarefa, não necessita ter compreensão sobre ela. Para o autor, pensar sobre uma inteligência artificial é a outra face de se pensar sobre a inteligência humana.

O sexto artigo, "O menosprezado debate sobre o artificial em IA", é de autoria de Orlando Lima Pimentel, mestrando em filosofia da ciência pela USP e colaborador da Associação Filosófica Scientiae Studia, e explora o papel da artificialidade presente no estudo da Inteligência Artificial (IA) debatendo os principais sentidos do termo "artificial" e como não é adequado associá-los à inteligência, com referências a Turing e Babbage.

O sétimo artigo, "Pode uma máquina desejar", de autoria de Midierson Maia, doutor e mestre em Ciência da Computação pela USP, com base na pergunta de Alan Turing – *Can machines think*? – coloca a questão sobre a capacidade das máquinas desejarem, considerando a relação entre pensamento e linguagem com referências em Decartes, Lacan e Bishop.

O professor titular da FAAP e pós-doutorando no TIDD, escritor e filósofo Rodrigo Petrônio apresenta uma resenha sobre a obra *Homo Deus*: uma breve história do amanhã do escritor e historiador israelense Yuval Hoah Harari. Autor dos best-sellers *Sapiens: Uma Breve História da Humanidade* e *Homo Deus*, Harari tornou-se um

conferencista internacional e uma referência nos esforços de compreender as transformações em curso na sociedade e as perspectivas futuras da humanidade.

Desejamos a todos uma boa leitura!

entrevista

Entrevista com Davi Geiger[1]

Dora Kaufman[2]

Dora Kaufman (DK): John McCarthy, que cunhou o termo em 1955, define Inteligência Artificial como "a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes". Russell e Norvig, autores do livro de referência nas universidades americanas, definem como "o estudo e concepção de agentes inteligentes, onde um agente inteligente é um sistema que percebe seu ambiente e realiza ações que maximizam suas chances de sucesso". Qual a sua definição de IA?

Davi Geiger (DG): Gosto da definição do McCarthy, mas falta a definição do que é "inteligência". Na definição do Russell e Norvig, faltam definições do que é "sucesso", do que é "perceber um ambiente", do que é "ação". Então eu adiciono à ideia do John McCarthy a definição de inteligência como sendo todas as funcionalidades do cérebro. O cérebro é algo que sabemos o que é e, em princípio, podemos saber quais são suas funcionalidades. Então acho que assim se completa uma definição: Inteligência Artificial é a ciência e a engenharia de criar máquinas que tenham funções exercidas pelo cérebro dos animais.

DK: A partir de meados dos anos 2010, os algoritmos de IA estão cada vez mais presentes e interferindo no nosso cotidiano. Quais os avanços recentes que justificam esse protagonismo?

DG: Vejo três componentes. Um é o avanço tecnológico no processamento dos computadores: hoje usamos GPUs (Graphics Processing Units) que executam o processamento paralelo (como também ocorre no cérebro), tornando o computador

[1] Professor Associado em Ciência da Computação no Courant Institute of Mathematical Sciences, New York University. Professor Assistente (1997-2000) e Visiting Assistant Professor (1994-1997) no Courant/NYU. Visiting Scientist no Institute of Mathematics, Cambridge University (1993). Research Scientist, Siemens Corporate Research, Princeton (1990-93). PhD no Physics Department and Artificial Intelligence Laboratory, Massachusetts Institute of Technology - MIT. Prêmio da National Science Foundation Career - NSF Career Award (1998). Co-fundador da DeepMagic (IA e IoT/mobile commerce/varejo) e consultor desde 2002 da N-hega (sistemas para indústrias). E-mail: dg1@nyu.edu.

[2] Ver Editorial, p. 5.

mil vezes mais rápido para o processamento de imagens, por exemplo. O segundo é o avanço em *storage* data e *cloud computing*. Com *storage* na *cloud*, por exemplo, o *imagenet* é um data set para estudos com 100.000 categorias (ou *labels*) e cada categoria possui 1000 imagens, i.e., são 1.000.000 imagens categorizadas. Com *cloud*, o software é compartilhado e, a partir daí, multiplicado muito rapidamente. E o terceiro é o avanço científico em *machine learning* para extrair informação com tais computadores. As técnicas básicas de *machine learning* foram elaboradas nos anos 60-80, mas apenas com os avanços tecnológicos mencionados foi possível observar uma performance que surpreendeu a todos.

DK: A mídia e a sociedade em geral tendem a responsabilizar os algoritmos pela invasão de privacidade e diversos outros comportamentos não éticos advindos da manipulação dos dados digitais (*big data*). Você poderia nos explicar o que são exatamente os algoritmos e por que a culpa não é deles?

DG: A questão da privacidade é muito delicada e importante. Não é apenas uma questão de identificar algoritmos, pois o problema inclui o acesso a dados, e o uso e a distribuição de tal dados. Sim, a combinação e correlação de dados de fontes diferentes produzindo novas informações pessoais são feitas por algoritmos, que obtêm correlações estatísticas; logo, os algoritmos são simples técnicas estatísticas. Até onde essas atividades são positivas para oferecer melhores serviços individualizados, e até onde essas atividades invadem e limitam a liberdade dos indivíduos? Essas são discussões que sempre existiram quando novidades são introduzidas na vida de todos.

Por exemplo, na questão médica de seguro, pode se obter a informação de um indivíduo de um hospital e também de outro site como o registro da carteira de motorista. Cruzando as informações obtêm-se conclusões sobre o indivíduo, como a idade e problemas médicos. Esses sites sozinhos foram criados com proteções a informações individuais, mas explorando todas informações possíveis de outros sites, pode-se tirar informações de um indivíduo as quais não estão nos acordos sociais existentes. Isso pode alterar seguros de saúde, ajudar a aumentar preconceitos. Não

sou um expert nisso, mas vamos precisar identificar os problemas, trazer o legislativo para criar novas leis, e o público em geral deve participar do processo boicotando as empresas que apoiam atividades não éticas.

DK: Pesquisadores, ativistas, celebridades e até mesmo empresários à frente de empresas de tecnologia, alertam para as ameaças e riscos da IA. Um dos temores é o desconhecimento de como exatamente funcionam os processos de aprendizado das máquinas com base em grande quantidade de dados e em tentativa-e-erro. Quais suas impressões desses riscos? Eles preocupam você? É legítima a preocupação da sociedade?

DG: É verdade que não sabemos (e talvez nunca venhamos a saber) exatamente o que é aprendido por máquinas de aprendizado, mas podemos ver o efeito do aprendizado. Mas eu pergunto: isso é diferente do que acontece nos humanos? Alguém sabe "como" e "o que" as pessoas aprendem? Tudo que observamos é como nós funcionamos em várias situações. No caso de máquinas, é um pouco mais controlado já que podemos copiar o software que foi aprendido em outra máquina, e assim duas máquinas são mais iguais do que duas pessoas. Além disso, o software não muda ao longo do tempo (a não ser que se queira modificá-lo), enquanto pessoas mudam seus comportamentos. Nesse sentido existe mais controle em máquinas. O método de verificar se a máquina, ou uma pessoa, aprendeu algo é por testes, tanto quanto deveríamos testar uma pessoa. Esses testes devem ser muito bem elaborados, e assim reduzir os riscos. Acredito que desse modo podemos controlar melhor os erros de uma máquina do que em humanos. Por exemplo, os carros autônomos, sem motoristas, estão rodando nas ruas e sendo testados e também comparados à performance dos humanos. Em paralelo, claro, os cientistas estão cada vez entendendo melhor como funcionam essas máquinas. Em princípio é mais fácil vir a entender essas máquinas do que os humanos porque sabemos o software dessas máquinas, podemos modificá-las, i.e., temos mais controle sobre todo o processo. Mas sim, essa compreensão é de interesse dos cientistas e é ainda limitada, assim como entender como nós – humanos – funcionamos continua de interesse de todos e é ainda mais

limitada, a compreensão. Concluindo, a preocupação é legítima, assim como é legítimo dizer que as máquinas podem funcionar melhor do que os humanos, desde que os testes demonstrem o fato.

DK: Ainda sobre os processos de aprendizado dos sistemas de IA, estudiosos alertam sobre o risco de perpetuar vieses e preconceitos humanos, e a impossibilidade de verificação, ou seja, diagnosticar e corrigir erros do processo com precisão. Como você enfrenta esses desafios? Eles são reais? Há como evitá-los no estágio de desenvolvimento atual?

DG: Todos os erros são possíveis, e por que não esses do preconceito? O chamado “bias". A razão dos preconceitos é ignorância, falta de conhecimento. No caso de máquinas de aprendizado, se os dados fornecidos são selecionados por humanos com algum bias, então a máquina vai aprender esse bias. Por exemplo, se um sistema de reconhecimento de faces nunca foi exposto a fotos de pessoas orientais, talvez não as reconheça. Seria isso preconceito? Acho que sim, uma forma de ignorância (não conhecer um tipo de pessoa que existe). Mas se o sistema for refeito e imagens dessas pessoas forem apresentadas durante o aprendizado, esse sistema irá então também reconhecer orientais. Acho válido lembrar que as máquinas aprendem a partir de informações que nós – humanos – provemos e assim preconceitos podem ser perpetuados ou corrigidos.

DK: Os avanços recentes da IA e as perspectivas para um futuro próximo requerem novos arcabouços legais e regulatórios. Observamos iniciativas de agências governamentais, particularmente nos EUA e Europa que, contudo, esbarram em dois desafios: o conhecimento limitado do tema pelos legisladores e a velocidade das transformações em curso. Você está de acordo com esses dois aparentes obstáculos? A sociedade tem como controlar os impactos negativos da IA?

DG: Concordo com a preocupação. Acho que sua pergunta é um bom diagnóstico de um problema. Acho que a solução é continuar trazendo o assunto para o

legislativo. Não sei mais o que se pode fazer. É difícil antecipar todos os problemas, todos os impactos negativos de IA, que certamente virão, e então é necessário procurar modos mais dinâmicos de mudanças. Então concordo com a preocupação, o modo de controlar é com o sistema legislativo existente, tentando chamar atenção para sua importância.

DK: Sabemos que não há consenso entre os experts sobre o futuro da Inteligência Artificial. As pesquisas apontam ser alta a probabilidade da "superinteligência", como definida por Nick Bostrom – "um intelecto que excede em muito o desempenho cognitivo dos seres humanos em praticamente todos os domínios de interesse" –, ser criada ainda no século XXI. Qual o futuro que você delineia? Faz sentido investir no desenvolvimento de uma inteligência autônoma (sem controle humano)?

DG: A história do homem é de criação, da curiosidade sendo o motor da ciência, do desenvolvimento. A bomba atômica mostrou que somos capazes de usar tal desenvolvimento científico de modo destrutivo. Ao mesmo tempo, talvez por conta da bomba atômica, hoje, países poderosos não entram em guerra direta. Será que essa curiosidade humana, de entender o seu próprio cérebro e criar uma inteligência artificial vai ser reprimida? Ou vamos continuar tal pesquisa e também investir em encontrar métodos de controle dos aspectos destrutivos? Eu prefiro acreditar no segundo *approach*.

DK: Uma ampla variedade de serviços está usando algoritmos de IA para estabelecer e direcionar padrões (*pattern-oriented matching*) como as plataformas de música *Spotify* e *Apple Music* ou para recomendar produtos como *Netflix* e *Amazon*. Contudo, esses processos ainda não contemplam todas as dimensões das preferências disponíveis dos usuários. Você está de acordo? Qual a perspectiva dessa evolução em termos de eficiência e *timing*?

DG: Não sou *expert* em quais dimensões os algoritmos não incluem. Eu sei que todas essas empresas estão contratando PhDs em Inteligência Artificial, o que sinaliza que querem melhorar o que já existe. Em princípio, como disse no início desta entrevista, com mais data *available*, com mais poder de computação, todos os sistemas de AI vão melhorar e mais dimensões podem ser incluídas. No aspecto de melhora dos algoritmos existe igualmente muita atividade. As conferências sobre o *state of the art* teórico em AI também estão crescendo muito. Uma boa e nova ideia quando publicada tem boas chances de se tornar obsoleta em alguns meses, já que existe um exército de pesquisadores que vai usá-la e modificá-la, gerando novas ideias. Então a perspectiva é muito otimista, já que estamos melhorando em todos os aspectos, com mais data, mais *computing power*, e melhores algoritmos e teorias. O sucesso de AI dentro dessas empresas vem dos últimos dez anos, e cresceu muito nos últimos cinco anos, então acredito que estamos no início desse crescimento de AI, quer dizer então que, nos próximos dez anos, o crescimento será exponential.

dossiê

Deep learning: a Inteligência Artificial que domina a vida do século XXI

Dora Kaufman[1]

Latanya Sweeney, ex-chefe de tecnologia da Comissão Federal de Comércio dos EUA e atualmente professora da Universidade de Harvard, foi informada por uma colega que o Google AdSense associava seu nome a anúncios sugerindo sua prisão. Intrigada, ela digitou o nome de outro de seus colegas, Adam Tanner, e o anúncio da mesma empresa surgiu sem a sugestão de prisão. Testando nomes racialmente associados, Sweeney encontrou discriminação estatisticamente significativa, sendo que um nome estereotipado como de negro era 25% mais propenso a receber um anúncio de registro de detenção – claramente um viés do sistema de busca ao reproduzir os preconceitos raciais da sociedade.[2]

A participação de pesquisadores chineses nos principais congressos mundiais de Inteligência Artificial (IA) cresceu de 10% em 2012 para 23% em 2017, enquanto de americanos no mesmo período caiu de 41% para 34% (AGRAWAL; GANS; GOLDFARB, 2018). A tendência é a China assumir a liderança na pesquisa e na aplicação comercial de IA, consequência de vários fatores, dentre eles a ausência de proteção à privacidade dos dados de seus cidadãos, significando uma vantagem comparativa vis-a-vis de regiões, como a Europa, com regulamentação e cultura rigorosas.

Os dois eventos, aparentemente não relacionados, têm em comum o processo denominado de *deep learning* (aprendizagem profunda), capaz de transformar grandes volumes de dados em informação útil. Tarefas tradicionalmente desempenhadas pelos seres humanos (reconhecimento visual, tomada de decisão, reconhecimento de voz, tradução) e outras que superam a capacidade humana (manipular e processar grandes bases de dados, *big data*), estão sendo executadas por máquinas inteligentes. Aprendizagem profunda é sobre previsão, e permeia grande parte das atividades do século XXI. Quando digitamos uma consulta ao Google, é ele que seleciona a resposta personalizada e os anúncios apropriados ao perfil do usuário, bem como traduz um

[1] Ver Editorial, p. 5.
[2] L. Sweeney, "Discrimination in online ad delivery", Communications of the ACM 56, no. 5 (2013), p. 44-54, <https://dataprivacylab.org/projects/onlineads/>.

texto de outro idioma, assim como filtra os e-mails não solicitados (*spam*). A Amazon e o Netflix recomendam livros e filmes pelo mesmo processo, do mesmo modo o Facebook usa aprendizado de máquina para decidir quais atualizações mostrar e o Twitter faz o mesmo para os tweets. Quando acessamos um computador, em qualquer de seus formatos, provavelmente estamos acessando concomitantemente um processo de *deep learning*. "O aprendizado de máquina faz inferências a partir de dados. E quanto mais dados eles têm, melhor elas ficam. Agora não precisamos programar computadores, eles se programam" (DOMINGOS, 2015, p. xi). Cada interação acessa dois níveis: "O primeiro é conseguir o que você quer. O segundo nível, e no longo prazo o mais importante, é ensinar o computador sobre você. Quanto mais você ensina, melhor ele pode servir – ou manipular você" (ibid., p. 264).

Os volumes de dados gerados atualmente inviabilizam o uso da tradicional programação computacional (com regras definidas a priori). A vantagem dos sistemas de aprendizado é que eles próprios estabelecem os algoritmos, i.e., adaptam-se automaticamente aos requisitos da tarefa. "Para muitas aplicações não fomos capazes de criar algoritmos apropriados apesar das décadas de pesquisa que começaram nos anos 1950. […] O aprendizado de máquinas agora é a força motriz da Inteligência Artificial":

> Especialmente nos últimos vinte anos ou mais, as pessoas começaram cada vez mais a se perguntar o que poderiam fazer com todos esses dados. Com esta pergunta, toda a direção da computação é revertida. Antes, os dados eram o que os programas processavam e cuspiam – os dados eram passivos. Com esta pergunta, os dados começaram a conduzir a operação; não são mais os programadores, mas os dados em si que definem o que fazer a seguir. (ALPAYDIN, 2016, p. x-xiii)

A grande quantidade de dados não é o único fator restritivo. No reconhecimento de imagem facial, por exemplo, os seres humanos têm certa facilidade, mas não conseguem explicá-lo (conhecimento tácito), o que não permite programar o computador. Ao analisar diferentes imagens de rosto de uma pessoa, um programa de aprendizado captura o padrão específico para essa pessoa e, em seguida, verifica esse padrão em uma dada imagem (ibid.). Uma das aplicações que mais tem surpreendido é a tradução interlingual automática.

Existem muitas expressões sendo utilizadas por diferentes comunidades – reconhecimento de padrões, modelagem estatística, mineração de dados, descoberta de conhecimento, análise preditiva, ciência de dados, sistemas adaptativos, sistemas de auto-organização, e outros – e alguns até o denominam simplesmente Inteligência Artificial (DOMINGOS, 2015). Independente do nome e da funcionalidade, o foco é no desenvolvimento de algoritmos e técnicas para solucionar determinados problemas, executar uma tarefa específica. A *strong* IA ainda é ficção.

A título de introdução

Cunhado em 1955, por John McCarthy, o termo "Inteligência Artificial" deu início a um campo de conhecimento associado à linguagem e à inteligência, ao raciocínio, à aprendizagem e à resolução de problemas (RUSSELL; NORVIG, 2009[3]). A IA propicia a simbiose entre o humano e a máquina ao acoplar sistemas inteligentes artificiais ao corpo humano (prótese cerebral, braço biônico, células artificiais, joelho inteligente e similares), e a interação entre o homem e a máquina como duas "espécies" distintas conectadas (homem-aplicativos, homem-algoritmos de IA). Tema de pesquisa em diversas áreas – Computação, Linguística, Filosofia, Matemática, Neurociência, entre outras –, a diversidade de subcampos e atividades, pesquisas e experimentações, dificulta descrever o estado da arte atual da IA. Os estágios de desenvolvimento bem como as expectativas variam entre os campos e suas aplicações, que incluem os veículos autônomos, reconhecimento de voz, games, robótica, tradução de linguagem natural, diagnósticos médicos, assim por diante. Atualmente, os sistemas inteligentes estão em todas as áreas de conhecimento.

Existem inúmeras definições de Inteligência Artificial, reflexo das especificidades intrínsecas a cada campo. Russell e Norvig (2009, p. 1-5) listam oito delas agrupadas em duas dimensões – as relativas a processos mentais e raciocínio e as relativas a comportamento – contudo, duas definições generalistas servem ao nosso propósito. Conforme a primeira de John McCarthy, Inteligência Artificial "é a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador

[3] Publicado originalmente em 1994 e seguido de várias edições, adotado nas universidades americanas como o livro de referência sobre IA.

inteligentes".[4] A segunda, de Russell e Norvig, define IA como o estudo de "agentes inteligentes" capazes de "perceber seu meio ambiente e de realizar ações" com a expectativa de "selecionar uma ação, que maximize seu desempenho" (2009, p. viii, 37).

Em 1959, Arthur Lee Samuel, pioneiro norte-americano no campo de jogos de computador e Inteligência Artificial, cunhou o termo "machine learning" (ML) (enquanto funcionário da IBM), inaugurando um subcampo da IA cuja finalidade é prover os computadores da capacidade de aprender sem serem programados. Evoluindo a partir do estudo do reconhecimento de padrões e da teoria de aprendizagem computacional na IA, o *machine learning* explora o estudo e a construção de algoritmos que, seguindo instruções, fazem previsões ou tomam decisões baseadas em dados – modelos elaborados a partir de entradas de amostras. Originada na estatística, em que migrar de observações particulares a descrições gerais é chamado de "inferência", a aprendizagem é chamada de "estimativa", e a classificação é chamada "análise discriminante" (ALPAYDIN, 2016). O aprendizado de máquina é empregado em uma variedade de tarefas de computação, nas quais programar os algoritmos é difícil ou inviável. Esses modelos analíticos permitem que pesquisadores, cientistas de dados, engenheiros e analistas produzam decisões e resultados confiáveis e replicáveis, e revelem ideias ocultas em relacionamentos históricos e tendências de dados.

> Usamos o aprendizado de máquina quando acreditamos que existe uma relação entre observações de interesse, mas não se sabe exatamente como. Porque não sabemos sua forma exata, não podemos simplesmente seguir em frente e anotar o programa de computador. Portanto, nossa abordagem é coletar dados de observações de exemplo e analisá-lo para descobrir o relacionamento. (ALPAYDIN, 2016, p. 29)

Na década de 1980, inspirados no cérebro humano, cientistas da computação criaram um subcampo da ML propondo um processo de aprendizado com base nas redes neurais, com resultados mais concretos nesta década.[5] O pioneiro foi Geoffrey Hinton, com a ideia de "neural networks" em artigo publicado na revista Nature de

[4] "Q. What is artificial intelligence? A. It is the science and engineering of making intelligent machines, especially intelligent computer programs" Disponível em: <http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html>. Acesso em 29 de jan. 2018.
[5] "O inicio da década de 1980 trouxe esperança de que os engenheiros pudessem programar cuidadosamente sistemas especialistas para replicar domínios habilidosos como diagnósticos médicos, mas estes eram caros de se desenvolver, complicados e incapazes de lidar com a miríade de exceções e possibilidades, levando ao que ficou conhecido como o 'Inverno da IA'" (AGRAWAL; GANS; GOLDFARB, 2018, p. 32).

1986.[6] O avanço não ocorreu acidentalmente. "Ganhamos capacidade de construir hardware paralelo conectando milhares de processadores, e as redes neurais artificiais despertaram interesse como uma possível teoria para distribuir cálculos em um grande número de unidades de processamento, todas em paralelo" (ibid., p. 28). O processo de aprendizagem profunda começou a florescer na década de 1990, com foco em problemas solucionáveis de natureza prática, relacionados a uma tarefa concreta. "As redes profundas ainda funcionam em domínios relativamente restritos, mas estamos vendo resultados mais impressionantes todos os dias à medida que as redes aumentam e são treinadas com mais dados" (ALPAYDIN, 2016, p. 109). O treinamento consiste em mostrar exemplos e ajustar gradualmente os parâmetros da rede até obter os resultados requeridos, denominado "aprendizagem supervisionada": são fornecidos os resultados desejados (*output*) e, por "tentativa e erro" chega-se ao resultado – meta.

> Uma vez que temos dados – e hoje em dia temos dados "grandes" – uma computação suficiente disponível – e agora temos centros de dados com milhares de processadores – apenas esperamos e deixamos o algoritmo de aprendizagem descobrir tudo o que é necessário por si só. [...] Descobrir essas representações abstratas é útil não só para a previsão, mas também porque a abstração permite uma melhor descrição e compreensão do problema. (ALPAYDIN , 2016, p. 108)

A abordagem é chamada de "retropropagação" (*back propagation*), aprendendo por exemplos. Muitos problemas foram transformados de problemas algorítmicos (Quais são as características de um gato?) em problemas de previsão (Essa imagem é similar a uma imagem que já vi antes?) (AGRAWAL; GANS; GOLDFARB, 2018).

A rede geralmente tem entre 10 e 30 camadas empilhados de neurônios artificiais. Num reconhecimento de imagem, por exemplo, a primeira camada procura bordas ou cantos; as camadas intermediárias interpretam as características básicas para procurar formas ou componentes gerais; e as últimas camadas envolvem interpretações completas. Na identificação de fotos nas redes sociais, a máquina percebe padrões e "aprende" a identificar rostos, tal como alguém que olha o álbum de fotos de uma família desconhecida e, depois de uma série de fotos, reconhece o

[6] In: Rummerlhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. Learning representations by back-propagating errors. *Nature*, Vol. 323, p. 533-536. October, 1986. Disponível em: <https://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf>. Acesso em: 29 de jan. 2018.

fotografado. O reconhecimento de voz, que junto com a visão computacional está entre as aplicações mais bem-sucedidas, já permite a comunicação entre humanos e máquinas, mesmo que ainda precária (Siri, Alexa, Google Now). Na cognição houve, igualmente, importantes avanços.

É importante notar que as máquinas inteligentes não reproduzem o funcionamento do cérebro, cuja complexidade ainda é relativamente pouco entendida, inviabilizando qualquer tentativa nessa direção. É mais correto dizer que a construção dessas máquinas é inspirada no cérebro humano. O cérebro é composto de neurônios, que por sua vez são formados por detritos que se conectam por meio de sinapses: cada vez que os detritos dos neurônios se encontram provocam uma sinapse (conexão). Essa configuração é denominada "redes neurais" em que, por analogia, o equivalente aos neurônios no computador são as unidades, ou seja, cada unidade do computador equivale a um neurônio no cérebro humano. Se temos 100 "sinapses" num computador, significa que temos 100 informações chegando e se conectando. As novas unidades, localizadas numa nova camada, recebem as informações, processam e "cospem" o *output* para as unidades de uma nova camada.

No processo de visão, por exemplo, a retina, um sensor de luz, representa a primeira camada. A retina é impactada por feixes de luz, que são as primeiras informações originadas no exterior. O mesmo se passa no ouvido com relação ao som, no olfato do nariz com relação ao cheiro, e no tato da pele com relação a sensibilidade. São informações elétricas e químicas, posteriormente enviadas para o cérebro. O aparelho perceptivo da visão é o único dos sentidos em que a primeira camada contém neurônios (logo, já é "cérebro"). Não por coincidência é o mais sofisticado, correspondendo a 1/3 do cérebro, ou seja, esta parcela do cérebro é dedicada à visão (a segunda atividade predominante no cérebro são os movimentos). A luz inicialmente encontra o censor da retina, que é a primeira camada, em seguida segue para uma nova camada, neste caso localizada na parte de trás do cérebro chamada de V1, continua se deslocando entre várias camadas, até retomar para a parte frontal do cérebro (*vision path way*). O cérebro tem dez áreas, e cada área cerca de 140 milhões de neurônios. O computador criado pela Microsoft há cerca de dois anos, considerado o mais avançado atualmente na tarefa de reconhecimento de imagem – Image res.Net –, tem 152

camadas, ou seja, as unidades vão se conectando e transmitindo informação a outras unidades ao longo de 152 camadas.

Fazendo um paralelo entre a visão humana e a câmara fotográfica, a nossa retina corresponde ao sensor de imagem da câmara. Em ambos, o que desencadeia o processo é a incidência de luz. O *input* da luz se transforma num número. Como isso é possível? A luz é composta de fótons, então importa calcular quantos fótons "caíram" na retina por unidade de tempo. Simplificando, o que permite diferenciar um objeto de outro é o número de fótons que sensibiliza a retina. Se todos os *inputs* viram números, temos um conjunto de números na primeira camada. Os processos no cérebro e os nas máquinas são semelhantes. Cada unidade que corresponde ao neurônio humano tem a decisão sobre o que será enviado à outra camada (ou não enviado). O que sai de uma camada não é necessariamente igual ao que entrou da camada anterior, significando certo grau de autonomia em relação ao operador.

Cada unidade recebe informações (*inputs*) de muitas unidades da camada anterior. No estado de evolução atual da IA, o operador humano arbitra o número de camadas. No futuro, existe forte indicação neste sentido, as máquinas vão construir outras máquinas inteligentes (sem arbitragem humana). O que define uma máquina inteligente são dois componentes: o valor de cada conexão e a arquitetura, traduzido no número de camadas.

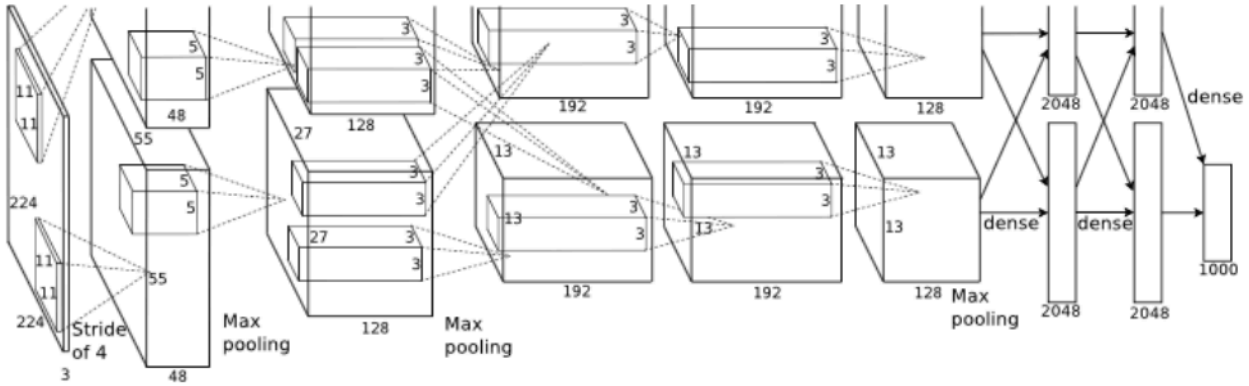

Figura 1. Arquitetura típica – com delimitação de responsabilidades – CPUs.
Fonte: ImageNet classification with deep convolutional neural networks, p. 5.
Disponível em: <https://www.nvidia.cn/content/tesla/pdf/machine-learning/imagenet-classification-with-deep-convolutional-nn.pdf>. Acesso em: 16 abr. 2018.

A figura 1 ilustra uma arquitetura típica, com a delimitação de responsabilidades entre duas CPUs (*Central Processing Unit*).[7]

Aprendizagem profunda no século XXI

O avanço do processo de aprendizagem profunda a partir dos anos 2006 a 2010 com a obtenção de resultados explícitos, deve-se fundamentalmente a três fatores: (a) crescente disponibilidade de grande quantidade de dados (*big data*), (b) maior capacidade computacional e (c) evolução dos algoritmos. Vejamos dois desses componentes.[8]

Big data

Em fevereiro de 2008, o *Centers for Disease Control and Prevention* (CDC) identificaram um crescimento de casos de gripe no leste dos EUA; na ocasião, o Google declarou ter detectado um aumento nas consultas sobre os sintomas da gripe duas semanas antes do lançamento do relatório. A partir dessa experiência, sua unidade filantrópica criou um sistema de alerta, o "Google Flu Trends".[9] Anteriormente ao aparecimento do vírus H1N1, pesquisadores do Google publicaram um artigo na revista Nature, ignorado pelas autoridades, sobre a capacidade de previsão da propagação da gripe de inverno, com base nos dados gerados em sua plataforma. A metodologia basicamente estabelecia correlações entre a frequência de certas consultas e a disseminação da gripe ao longo do tempo e espaço, identificando regiões específicas em tempo real. Esse evento influenciou uma mudança na mentalidade sobre o uso de dados (MAYER-SCHÖNBERGER; CUKIER, 2013). "*Big data* refere-se a coisas que se pode fazer em grande escala para extrair novos insights ou criar novas formas de valor, mudando os mercados, as organizações, a relação entre cidadãos e governos e muito mais" (ibid., p. 6).

Mayer-Schönberger e Cukier indicam como desdobramentos a capacidade de analisar grandes quantidades de dados sobre um tópico específico, e não mais se ater a amostras; a disposição em adotar a desordem do mundo real dos dados, deixando de

[7] Esse conteúdo foi transmitido diretamente à autora pelo pesquisador em Computer Science Davi Geiger da NYU.
[8] A capacidade computacional adentra em um campo de conhecimento particular, do domínio das ciências exatas/ tecnologias.
[9] Artigo do NYT. Disponível em: <https://www.nytimes.com/2008/11/12/technology/internet/12flu.html>. Acesso em: 31 de mai. 2018.

privilegiar a exatidão; e o respeito crescente por correlações em vez de causalidades. Trata-se de sacrificar a exatidão para ter acesso à tendência geral. Agrawal, Gans e Goldfarb (2018) destacam três funções desempenhadas pelos dados: (a) primeiro temos os dados de entrada (*input*), que alimentam os algoritmos e são utilizados no processo de previsão; (b) segundo, os dados de treinamento (*training data*), usados para aperfeiçoar os algoritmos; e (c) terceiro, temos os dados de feedback com a função de melhorar o desempenho dos algoritmos com base na experiência dos usuários (ibid., p. 43).

Qualquer interação com tecnologias digitais deixa "rastros", alguns voluntários como as publicações nas redes sociais – Facebook, Twitter e Instagram –, e outros involuntários, como as informações armazenadas nos bancos de dados digitais na compra com cartão de crédito, na movimentação bancária *online*, no acesso aos programas de fidelidade, no vale-transporte, nas comunicações por telefonia móvel, e inúmeras outras ações presentes em nossa rotina. Esses podem ser usados pelas plataformas originais ou "reusados" por terceiros, ou combinados pela fusão de conjunto de dados, com as mais variadas finalidades, e são responsáveis por inúmeros benefícios da sociedade do século XXI (e, igualmente, por inúmeras ameaças). "O entusiasmo pela 'internet das coisas' – incorporando chips, sensores e módulos de comunicação aos objetos do cotidiano – é, em parte, relacionado à rede, mas também sobre digitalizar a informação de tudo que nos rodeia" (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 96). Por meio deles é possível revelar uma infinidade de questões relacionadas à população, desde quais grupos são mais suscetíveis a determinadas doenças até qual é o perfil do cidadão propenso a honrar um empréstimo bancário, até segmentar os consumidores em perfis. Pentland (2015) defende que o *big data* oferece a chance de ver a sociedade em toda a sua complexidade; para ele, uma vez desenvolvida uma visualização mais precisa dos padrões de vida humana, podemos esperar compreender a sociedade de forma mais adequada à nossa rede complexa e interligada de seres humanos e tecnologia.

A importância dos dados cresce gradativamente.[10] Em 2012, a operadora espanhola Telefónica criou uma empresa separada – Telefónica Digital Insights – com o

[10] Ver sobre "Capitalismo de Dados" no artigo de Dora Kaufman desta edição.

propósito de comercializar dados anônimos e agregados de localização de assinantes para varejistas e outros. Uma divisão do cartão de crédito Mastercard – Mastercard Advisors – agrega e analisa cerca de 65 bilhões de transações de 1,5 bilhões de titulares de cartão em 210 países procurando identificar tendências de negócios e consumo para, em seguida, comercializar a informação à terceiros (MAYER-SCHÖNBERGER; CUKIER, 2013).

Algoritmos

"Algoritmos estão em toda parte. Dominam o mercado de ações, compõem música, dirigem carros, escrevem artigos de notícias e autênticas provas matemáticas – e seus poderes de autoria criativa estão apenas começando a tomar forma" (FINN, 2017, p. 15). Como pondera Domingos (2015), atualmente se todos os algoritmos parassem de funcionar, seria o fim do mundo. Algoritmo é um conjunto de instruções matemáticas, uma sequência de tarefas para alcançar um resultado esperado em um tempo limitado. Os algoritmos antecedem os computadores – o termo remonta ao século IX ligado ao matemático al-Khwãrizmi, cujo livro ensinava técnicas matemáticas a serem equacionadas manualmente. "Algorismus" era originalmente o processo de calcular numerais hindo-arábicos (FINN, 2017). Ed Finn define um algoritmo como "qualquer conjunto de instruções matemáticas para manipular dados ou raciocínios através de um problema" (ibid., p. 17). Para Ethem Alpaydin, "um algoritmo é uma sequência de instruções que são realizadas para transformar a entrada (*input*) na saída (*output*)" (2016, p. 14). Brian Christian e Tom Griffiths (2016) extrapolam o conceito para além do âmbito da Matemática: "Quando você cozinha pão a partir de uma receita, você está seguindo um algoritmo, o mesmo quando você tricota uma peça com base num determinado padrão. […] Algoritmo faz parte da tecnologia humana desde a Idade da Pedra" (2016, p. 4). A ideia de associar algoritmo à receita de culinária, contudo, é contestada por Domingos (2015) para quem a receita não especifica exatamente a ordem e as etapas – quanto de açúcar, por exemplo, está contido em uma colherada. "Se quiséssemos programar um robô de cozinha para fazer um bolo, teríamos que dizer como reconhecer o açúcar do vídeo, como pegar uma colher e assim por diante. Portanto, uma receita culinária está muito longe de um algoritmo" (ibid., p.

3). O algoritmo requer instruções precisas e não ambíguas, o suficiente para serem executadas por um computador. "Algoritmos são um padrão exato. Costuma-se dizer que você realmente não entende algo até que possa expressá-lo como um algoritmo" (ibid., p. 4).

Os algoritmos têm sua própria complexidade, que Domingos (2015) agrupa em (a) complexidade do espaço, traduzida no número de bits de informação que precisa armazenar na memória do computador, (b) complexidade do tempo, traduzida no tempo necessário para "rodar", i.e., quantas etapas à percorrer até produzir resultados, e (c) complexidade humana, traduzida nos limites do cérebro humano, inclusive para identificar os erros e corrigi-los.

Como mencionado anteriormente, a aprendizagem profunda é sobre "previsão" com base em correlações, e é sobre reduzir significativamente o custo de previsão. O objetivo não é identificar causalidades entre distintos fenômenos ou simples eventos, mas descobrir padrões e correlações que geram *insights*. "Antes do *big data*, nossa análise geralmente se limitava a testar um pequeno número de hipóteses que definíamos bem antes de coletar os dados. Quanto deixamos os dados falarem, podemos fazer conexões que nunca imaginamos que existissem" (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 14).

Os algoritmos de aprendizado não funcionam da mesma forma, e suas diferenças impactam os resultados e, por vezes, o próprio modelo de negócio. Domingos (2015) compara os modelos de recomendação da Amazon e do Netflix: "Se cada um tivesse guiado você através de uma livraria física, tentando determinar o que é 'certo para você', a Amazon estaria propensa a levá-lo até às prateleiras que você frequentou anteriormente; a Netflix levaria você a seções estranhas da loja" (ibid., p. xvi). O algoritmo da Netflix tem uma compreensão mais profunda das preferências do usuário, e explica-se porque seu modelo depende de expandir a demanda para filmes e vídeos da "cauda longa", que custam menos do que os *blockbusters*.[11] Pelo modelo de negócio da Amazon, a concentração nas mesmas preferências é positiva porque facilita a logística. Além disso, como clientes, as pesquisas indicam que estamos mais dispostos

[11] O valor da assinatura seria deficitário se os usuários escolhessem apenas filmes e videos "blockbusters".

a ter uma chance em um item estranho num modelo de assinatura do que se tivermos que pagar por esse item individualmente (ibid.).

A sociedade dos dados e dos algoritmos

Os algoritmos de IA já estão presentes no nosso cotidiano. Parte do sucesso da Netflix, por exemplo, está em seu sistema de personalização, em que algoritmos analisam as preferências do usuário e de grupos de usuários com preferências semelhantes e, com base nelas, sugere filmes e séries. Acessamos sistemas inteligentes para programar o itinerário com o Waze, pesquisar no Google e receber do Spotify recomendações de músicas. A Siri da Apple, o Google Now e a Cortana da Microsoft, são assistentes pessoais digitais inteligentes que nos ajudam a localizar informações úteis com acesso por meio de voz. Existe uma multiplicidade de algoritmos de IA permeando as interações nas redes sociais, dentre eles os algoritmos do Feed de Notícias do Facebook.

Na sua operação diária, a Amazon captura grandes volumes de dados incluindo não apenas os livros que os usuários compram, mas quais os livros que eles só olham e por quanto tempo. Inicialmente, as recomendações derivaram de processar semelhanças entre clientes baseadas em amostras com resultados limitados; a partir de 1998, seu sistema de recomendação passou a buscar associações entre os próprios produtos[12] utilizando todos os dados com resultados mais precisos. "Quando a Amazon realizou um teste comparando a venda produzida por editores humanos com a venda produzida por conteúdo gerado por computador, os resultados nem chegaram perto. O material derivado e dados gerou muito mais vendas" (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 51).

O crescimento exponencial dos dados inviabiliza a programação tradicional, remetendo inevitavelmente às técnicas de aprendizado de máquinas. A Amazon não pode codificar os gostos do conjunto de seus clientes em um programa de computador, assim como o Facebook desconhece como escrever um programa para identificar as melhores atualizações no Feed de Notícias. A Netflix pode ter cem mil títulos de DVD em estoque, mas se os clientes não souberem como encontrar suas

[12] Na ocasião, a Amazon registrou uma patente sobre "filtragem colaborativa item-a-item", como a técnica é conhecida (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 51).

preferências, o padrão será escolher os hits. A "cauda longa" só decola com os algoritmos de aprendizado (recomendação assertiva). Os algoritmos de aprendizado são os casamenteiros: eles encontram produtores e consumidores um para o outro com o melhor dos dois mundos: a diversidade de opções e o baixo custo da grande escala, com o toque da personalização associado aos pequenos. Um dos efeitos é a concentração de mercado: "Quem tem mais clientes acumula a maior parte dos dados, aprende os melhores modelos, conquista os novos clientes e assim por diante, em um círculo virtuoso" (ibid., p. 12).

Cada um de nós é, simultaneamente, um gerador e um consumidor de dados. "Queremos ter produtos e serviços especializados. Queremos que nossas necessidades sejam compreendidas e nossos interesses sejam previstos" (ALPAYDIN, 2016, p. 16). O paradoxo é que, ao mesmo tempo, queremos preservar a privacidade de nossos dados. O desafio colocado é encontrar um equilíbrio entre a abertura de dados, inclusive como enfrentamento da concentração de mercado, e a transparência sobre o uso dos dados.

Referências

AGRAWAL, Ajay; GANS, Joshua; GOLDFARB, Avi. *Prediction machines*: the simple economics of Artificial Intelligence. Boston, MA: Harvard Business Review Press, 2018.

ALPAYDIN, Ethem. *Machine learning*. Cambridge, MA: MIT Press, 2016.

CHRISTIAN, Brian; GRIFFITHS, Tom. *Algorithms to live by:* the computer science of human decisions. New York, NY: Henry Holt, 2017.

DOMINGOS, Pedro. *The master algorithm*: how the quest for the ultimate learning machine will remake our world. New York, NY: Basic Books, 2018.

FINN, Ed. *What algorithms want*: imagination in the age of computing. Cambridge, MA: MIT Press, 2017.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big data*: a revolution that will transform how we live, work, and think. New York, NY: Houghton Mifflin Harcourt, 2013.

RUSSELL, Sutart J.; NORVIG, Peter. *Artificial Intelligence*: a modern approach, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2009.

PENTLAND, Alex. *Social physics*: how social networks can make us smarter. New York, NY: Penguin, 2015.

artigos

Inteligência Artificial: uma utopia, uma distopia

Fabio Gagliardi Cozman[1]

Resumo: A busca por inteligências artificiais parece ter atingido um ponto de inflexão no qual várias tecnologias há muito prometidas têm se tornado realidade. Em particular, o uso intensivo de grandes bases de dados tem levado ao desenvolvimento de sistemas de reconhecimento de imagens, compreensão de linguagem natural e tomada de decisão, cujo desempenho chega a igualar, e em alguns casos superar, o desempenho humano. Esses avanços tecnológicos têm gerado reações de otimismo e pessimismo. Algumas opiniões são entusiasticamente positivas, entendendo que o futuro trará riquezas imensuráveis a todos; outras vislumbram o fim da espécie humana. Entre esses extremos pode-se encontrar um leque matizado de posições. Este artigo procura capturar essas posições por meio de um conjunto de "distopias" e "utopias". Ao final, chega-se a uma "utopia realista", que apresenta um objetivo plausível para a sociedade atual, em contraponto a uma "distopia realista" que representa um modelo verossímil mas que deve ser evitado.

Palavras-chave: Inteligência Artificial. Aprendizado de máquina. Mercado de trabalho.

Abstract: The search for artificial intelligences seems to have reached a point where many long-promised technologies have become reality. In particular, the heavy use of large databases has led to the development of systems for image recognition, natural language understanding, and decision making, whose performance often reaches, and sometimes surpasses, human performance. This technological advance has produced optimistic and pessimistic reactions. Some opinions are enthusiastically positive in believing that the future will bring enormous riches to everyone, while others foresee the end of humanity. Between these extremes, one can find a varied set of positions. This paper tries to capture these positions using a set of "dystopias" and "utopias". In the end we reach a "realistic utopia" that serves as a plausible goal for society, in contrast to a "realistic dystopia" that represents a credible model that must be avoided.

Keywords: Artificial Intelligence. Machine Learning. Job market.

Inteligência artificial: um verão escaldante, mas com possíveis trovoadas

Embora a busca por uma Inteligência Artificial seduza a mente humana há séculos (NILSSON, 2009), foi só em 1950 que Alan Turing apresentou a primeira formalização do conceito de uma "máquina que pensa" (TURING, 1950). Poucos anos

[1] Professor Titular da Poli - USP/Departamento de Engenharia Mecatrônica, PhD Carnegie Mellon University, Livre-docência pela USP. Atualmente é coordenador da comissão de Inteligência Artificial da SBC, Associate Editor do Int. Journal on Approximate Reasoning, membro do Editorial Board do Artificial Intelligence Journal, e Associate Editor do Journal of Artificial Intelligence Research. E-mail: fgcozman@usp.br.

depois ocorreu o primeiro encontro de pesquisadores interessados no tema em que foi cunhado o termo “Inteligência Artificial” (MCCARTHY, 1955).

Desde então, a ideia de uma “Inteligência Artificial” tem fascinado a sociedade. Esse interesse se reflete em referências culturais, por exemplo, na literatura e no cinema e também em discussões científicas e econômicas.

Grandes promessas fazem parte da história da Inteligência Artificial. Durante os seus anos iniciais, a área foi cercada de previsões otimistas e em alguns casos mirabolantes. O entusiasmo sofreu um duro golpe em 1974, quando um relatório encomendado pelo parlamento britânico indicou que a área falhava na prática. Durante vários anos subsequentes o financiamento da área foi reduzido em todo o mundo, um período que hoje é conhecido como “inverno da IA”. Mas o pessimismo não durou para sempre: durante os anos oitenta houve um novo momento de euforia, marcado pelo desenvolvimento de “sistemas especialistas” que teriam condições de reproduzir as regras usadas por especialistas na solução de problemas específicos. E, de novo, a área enfrentou um longo inverno durante o final dos anos 80 e começo dos anos 90. Olhando para esse período, pode-se ver que muitas técnicas fundamentais da área estavam em gestação, mas ainda não haviam maturado ao ponto de resolver problemas práticos significativos. O inverno foi longo, e bastante frio.

Hoje a área vive um verão escaldante. Algumas das maiores empresas do mundo, como Google e Facebook, baseiam-se em métodos de Inteligência Artificial, como entendimento de textos e reconhecimento de faces. Grande parte das empresas anuncia estratégias que incluem técnicas de Inteligência Artificial para processar montanhas de dados acumuladas por toda parte. A sociedade é bombardeada com notícias sobre avanços da tecnologia e de seus impactos.

Como ocorreu essa mudança climática? Em primeiro lugar, houve exponencial aumento na capacidade de processamento de computadores, em particular computadores com processamento paralelo. E, além do aumento de capacidade, computadores e equipamentos como câmeras de vídeo tiveram enormes reduções de preço. Em segundo lugar, a área se beneficiou do acúmulo exponencial de dados na sociedade e indústria em função da redução de custos de sensores e da melhoria das redes de comunicação.

Mas, além desses dois pontos práticos importantes, houve também a maturação de técnicas computacionais ao longo de duas décadas de esforços. Houve grande avanço em algoritmos de aprendizado, que extraem padrões de grandes bases de dados, bem como algoritmos de otimização aplicados a problemas de planejamento. Além disso, houve significativa maturação de técnicas de modelagem baseadas em lógica e probabilidades. Isso levou a um grande progresso na geração automática de diagnósticos e de planos, e na compreensão de imagens e de linguagem natural. A área conseguiu finalmente concretizar algumas das promessas feitas no seu início, logrando realizar computacionalmente atividades que de fato parecem inteligentes.

Esse progresso naturalmente gera reações, algumas otimistas, outras pessimistas. Os otimistas creem, em resumo, que o aumento de produtividade causado por máquinas inteligentes será extraordinário, e o consequente aumento de riqueza mais que compensará alguns desconfortos causados por erros eventuais de serviçais automáticos. Os pessimistas creem que o aumento de produtividade virá, mas acompanhado de tantos problemas que o balanço geral será negativo: alguns julgam que os problemas passam pela própria destruição da humanidade, enquanto outros preveem uma inaceitável perda de privacidade e de controle, além de um aumento da desigualdade entre seres humanos. Previsões sobre o futuro variam e são de difícil classificação.

Para tentar compreender melhor como se dá a discussão sobre Inteligência Artificial, este artigo procura dividir as várias opiniões em “modelos de futuro”, alguns utópicos, outros distópicos. Um futuro utópico, mesmo que improvável, oferece um objetivo a perseguir. Um futuro distópico, mesmo que pouco verossímil, descreve circunstâncias a evitar, suscitando debate sobre como a sociedade deve se organizar hoje para desviar-se de um desastre amanhã.

O artigo passa por várias utopias e distopias de baixa probabilidade; ao final, o artigo procura descrever uma “utopia realista” e uma “distopia realista”. Ou seja, o artigo procura encontrar um modelo de futuro otimista e que pode de fato ser atingido se houver esforço para tal, e um modelo de futuro pessimista que pode de fato ocorrer se não houver preocupação nem prudência.

Distopias Catastróficas

Existem muitas distopias totalmente pessimistas que se baseiam na evolução da Inteligência Artificial. Esse é um tema que tem recebido grande atenção na literatura e no cinema, e que está profundamente aninhado no imaginário popular. Em linhas gerais, pode-se identificar duas diferentes distopias dessa natureza.

A primeira distopia se baseia na noção de que computadores podem ter uma ação fundamentalmente maléfica devido a falhas de operação ou devido a uma maldade intrínseca (o "desejo de dominar o mundo"). O cinema tem sido pródigo em explorar esse tema há muito tempo; provavelmente, um dos exemplos mais populares está no filme *Terminator* e suas sequências. Por anos, essa foi a distopia mais comum relacionada à Inteligência Artificial.

A segunda distopia se baseia na noção hoje designada por superinteligência artificial (BOSTRON, 2014). Nessa visão, os computadores terão uma melhoria exponencial em sua capacidade cognitiva, em um certo momento superando em muito o ser humano. Nesse ponto o ser humano terá se tornado, para a superinteligência artificial, um ser inútil e inconveniente – como um enxame de mosquitos que apenas perturbam. A superinteligência poderá então decidir destruir os seres humanos, baseado em argumentos estritamente racionais; por exemplo, para evitar divisão de recursos energéticos. A distopia envolve a destruição da humanidade ou sua redução a uma situação de servidão completa (talvez uma servidão na qual os seres humanos remanescentes estejam subjugados e inconscientes da sua real situação).

Embora tais distopias catastróficas mereçam atenção, esse artigo não as discutirá em profundidade. Essas distopias estão hoje muito mais próximas de experimentos mentais sobre a sociedade humana do que de reais possibilidades de futuro. A Inteligência Artificial hoje em desenvolvimento não tem condições de planejamento em escala global nem de negociação sofisticada; tampouco tem condições de assumir aparatos físicos que possam realmente oferecer dano a populações de grande porte.

Mas não é pela sua improbabilidade que essas distopias são aqui deixadas em segundo plano. O ponto aqui é que, para que distopias catastróficas possam acontecer,

distopias mais sutis muito provavelmente terão sido antes atingidas. Vale a pena entender o que seriam essas distopias mais sutis.

Distopias: um guia para o pessimista

Considere agora que uma distopia catastrófica não ocorre, portanto a humanidade continua sobrevivendo e interagindo com suas máquinas. Quais ameaças são contempladas por uma visão pessimista da Inteligência Artificial?

As ameaças parecem se dividir em dois grupos.

Em primeiro lugar, há ameaças que podem ser resumidas como "perda do controle". A mais simples é relacionada à perda de privacidade: máquinas que usam dados sigilosos sem autorização. Outra ameaça é que decisões discriminatórias sejam tomadas por algoritmos sem que haja controle social: já existem preocupações sobre decisões judiciais tomadas com base em dados enviesados (OSOBA; WELSER IV, 2017). Uma outra ameaça é a perda de entendimento, por parte de seres humanos, das razões pelas quais decisões automáticas são tomadas. O ser humano se transformaria em uma minúscula engrenagem em uma sociedade kafkiana em que decisões são tomadas com base em algoritmos que ninguém consegue entender. Aqui há uma ameaça ao próprio debate democrático na sociedade: como é possível debater assuntos de interesse geral se ninguém sabe realmente como o sistema funciona? Uma sociedade em que todas as ações são mediadas por dispositivos artificiais pode se tornar uma sociedade em que relacionamentos profundos são raros ou impossíveis. Finalmente, máquinas podem falhar, falhas podem ter efeitos tão mais nefastos quanto mais responsáveis forem as máquinas. Uma falha pode disparar mísseis erradamente ou pode causar acidentes aéreos de grande porte; uma sequência de falhas pode causar uma guerra ou uma epidemia. E uma sociedade controlada por máquinas pode ser muito mais vulnerável a ataques perpetrados por humanos mal-intencionados que podem danificar elementos centrais do processo decisório.

Em certa medida, todas as ameaças citadas no parágrafo anterior já podem ser observadas em alguma medida na sociedade atual: já somos submetidos a "sistemas" de controle bancário com pouca transparência, já perdemos alguma privacidade, já temos que lidar com decisões automáticas que não são explicadas e já nos

defrontamos com “falhas não humanas”. O pessimista pode se sentir plenamente justificado ao raciocinar que a perda de controle já em curso só pode se acentuar, em certo momento tornando a vida insuportável.

Em segundo lugar, há ameaças relacionadas ao mercado de trabalho. Existe muita discussão sobre quais profissões serão afetadas pelo aumento de capacidade cognitiva de computadores (FREY; OSBORNE, 2013). Algumas profissões parecem ter pouco futuro: funcionários de empresas de propaganda por telefone terão menos relevância no momento em que essas propagandas forem realizadas de forma eletrônica. Mas ninguém sabe como será o mercado de trabalho no futuro, e essa insegurança gera angústia: cada trabalhador se pergunta se sua profissão está em risco.

Mas quais são exatamente as ameaças relacionadas ao mercado de trabalho?

A substituição de postos de trabalho por máquinas pode gerar dramas pessoais momentâneos, mas se a sociedade tiver condições de reeducar pessoas, em princípio será possível atingir um patamar mais produtivo em que novos empregos florescerão. A simples necessidade de modificar a força de trabalho atual, levando-a a um patamar mais refinado, talvez a um custo alto, não parece ser elemento suficiente para uma distopia. Porém, outras consequências da automação de empregos parecem estar na raiz de várias distopias. Uma possível consequência é que a turbulência no mercado de trabalho nunca acabe: que o ser humano esteja sempre submetido a um processo humilhante de obsolescência, passando de profissão em profissão, perseguido por máquinas cada vez mais capazes. Outra possível consequência é que ocorra uma concentração cada vez maior da renda, com máquinas controladas por uma minoria da população – enquanto a maior parte da humanidade estaria competindo arduamente com as máquinas, um grupo pequeno poderia auferir lucros extraordinários. Ou poderia haver uma crescente desigualdade no mercado de trabalho: seres humanos com certas habilidades de difícil automação se distanciariam cada vez mais de trabalhadores com habilidades facilmente reproduzidas artificialmente (BRYNJOLFSSON; McAFEE, 2014). Uma terceira possível consequência é a captura dos empregos com maior capacidade cognitiva por agentes artificiais, relegando o ser humano a tarefas repetitivas e de baixa remuneração (FREEMAN, 2016). Nessa última visão, o problema

não é necessariamente que faltarão empregos: novos tipos de trabalho poderão ser criados, mas a vantagem competitiva dos seres humanos sobre as máquinas desaparecerá e elas controlarão as atividades mais nobres.

O pessimista parece ter amplo espectro de opções. O que diz o otimista?

A grande utopia

A Grande Utopia da Inteligência Artificial (GUIA) é, em essência, simples. Máquinas realizam todo o trabalho repetitivo com extraordinária precisão e produtividade, enquanto seres humanos gastam algum tempo controlando o processo produtivo e democraticamente debatendo como regular a sociedade, trabalhando de forma prazerosa e dedicando tempo considerável à saúde e bem-estar.

Na GUIA a infraestrutura artificial opera de forma clara e ética, seguindo diretrizes produzidas pela sociedade humana em um processo democrático amplo. Nessa utopia, máquinas oferecem diagnósticos médicos a todos os cidadãos, e médicos humanos especializados se ocupam apenas de novos casos e também da interação mais pessoal com pacientes; máquinas oferecem serviços legais simples, enquanto advogados humanos se ocupam de negociações mais complexas e da concepção de novos instrumentos jurídicos; e assim por diante. Essa sociedade utópica também dará grande assistência ao envelhecimento da população: assistentes artificiais estarão disponíveis de forma ininterrupta, até para conversar e confortar. Um melhor planejamento urbano, acoplado a um melhor planejamento individual sobre opções de transporte, pode reduzir níveis de poluição e evitar escassez de recursos; similares ganhos podem ser obtidos em todos os aspectos da vida humana.

Na GUIA o mercado de trabalho é completamente subvertido pela Inteligência Artificial, mas o resultado é bem-vindo: atividades não desejadas são ocupadas por máquinas que operam com grande eficiência, e seres humanos se ocupam de atividades de seu interesse.

Há possíveis variantes dessa utopia no que diz respeito à distribuição de riquezas e de empregos. Uma possível GUIA seria obtida com uma população humana altamente qualificada operando em um mercado de trabalho competitivo (mas talvez com regras sobre competição não-humana). Uma outra GUIA contemplaria mecanismos

de distribuição de renda, por exemplo, oferecendo programas de renda mínima para garantir que toda a população tenha benefícios relativamente equânimes.

Uma utopia "realista"

Embora a Grande Utopia da Inteligência Artificial sirva como um objetivo meritório, alguns de seus aspectos parecem excessivamente idealizados. Mesmo em um mundo com máquinas superinteligentes e dóceis poderá haver problemas com proteção de privacidade, falta de explicação, falhas – talvez causadas por seres humanos incompetentes ou mal-intencionados. Também poderá haver desigualdade, até mesmo disputa sobre os meios de produção. Parece razoável buscar um objetivo mais realista, e em particular uma utopia que sirva para um país como o Brasil.

Como seria essa utopia?

Em primeiro lugar, o uso intensivo de máquinas elevará enormemente a produtividade. Portanto a riqueza total gerada pela sociedade crescerá. Os ganhos de produtividade melhorarão a vida da sociedade em geral, aprimorando os serviços e produtos do setor privado, oferecendo mais recursos para o setor público. Produtos serão mais úteis e mais eficientes, reduzindo por exemplo a poluição. Os recursos públicos, administrados e fiscalizados com suporte de agentes artificiais, serão utilizados com maior eficiência e sob maior controle social.

Em segundo lugar, as ameaças relativas à perda de controle (violação de privacidade, decisões automáticas discriminatórias ou sem explicação) serão controladas por meio de legislação apropriada. Aqui se trata de encontrar legislação que proteja a sociedade sem impedir a inovação; proibições genéricas baseadas em medos abstratos só aumentarão a burocracia e reduzirão a produtividade. Uma boa legislação deve incentivar o progresso e evitar as ameaças.

Em terceiro lugar, e com uma boa dose de otimismo utópico, a sociedade encontrará diretrizes éticas que serão codificadas em máquinas. Em particular, máquinas receberão objetivos claros e éticos. Parece difícil que todos os dilemas éticos possam ser assim equacionados, mas será importante que os casos mais simples sejam codificados, deixando os casos complexos para debate entre os seres humanos. Da mesma forma, questões legais sobre atribuição de responsabilidade sobre

comportamento artificial serão em grande parte codificadas, deixando os casos complexos para análise por agentes humanos.

Em quarto lugar, o número de "falhas não humanas" será reduzido a um ponto tão baixo que não haverá questionamento quanto ao uso de máquinas nas mais variadas atividades. Afinal, o próprio ser humano comete falhas; o objetivo realista deve ser produzir dispositivos que apresentem muito menos falhas que os operadores humanos, e não produzir dispositivos que nunca falham.

Em resumo, a infraestrutura artificial da sociedade será onipresente mas transparente, justa e ética, praticamente sem falhas, suprindo os seres humanos de suas necessidades.

Em uma sociedade com essas características, com menor pressão do trabalho e melhor distribuição de riquezas, relacionamentos humanos de qualidade poderão ocorrer naturalmente. Mesmo o uso de "amigos artificiais" poderá ser útil para lidar com ocasionais episódios de solidão ou de doença. Seres humanos poderão dedicar mais tempo a esportes, artes, lazer.

Para completar essa utopia, é necessário imaginar como ocorreria de fato a distribuição de trabalho e riquezas. Essa é uma questão que tem suscitado as mais diversas propostas. Parece razoável supor que a maior produtividade na sociedade poderá levar a programas universais de educação e saúde. Além disso, parece razoável supor que o sistema de educação evolua de tal forma que os seres humanos estejam continuamente se aprimorando e se colocando à frente das máquinas em atividades fundamentais. Mas e se houver um contingente de pessoas sem colocação? Seria o caso de oferecer programas de renda mínima a todas essas pessoas? Ou isso será um incentivo para que os seres humanos deixem de trabalhar e se coloquem em uma posição subalterna às máquinas? Esse é um debate necessário para os próximos anos, à medida em que a tecnologia avançar. Na versão utópica, a sociedade conseguirá encontrar mecanismos de distribuição equânime de oportunidades que permitam a todos manter sua relevância e dignidade.

E uma distopia "realista"

Se a utopia realista da Inteligência Artificial serve como um objetivo a almejar, uma distopia realista pode servir como indicação sobre o que evitar. Na verdade, a simples continuidade de várias mazelas da sociedade já parece ser suficiente para gerar uma distopia realista bastante preocupante. Suponha que, ao tentar atingir a utopia realista, façamos alguns erros e, ao mesmo tempo, não consigamos nos libertar de nossos medos e preconceitos. O que pode acontecer, em particular pensando no Brasil?

Pode ocorrer uma resistência corporativa tremenda à Inteligência Artificial; leis podem ser promulgadas para impedir o avanço dessa tecnologia ou para criar tanta burocracia que o avanço se torne impossível. Nesse caso o país pode perder a oportunidade de empreender na área, no futuro tornando-se importador da tecnologia. Por exemplo, imagine que associações de médicos proíbam a análise automática de radiografias: isso impedirá a existência de empresas no Brasil com essa tecnologia, e em pouco tempo a população (grande parte da qual hoje não tem acesso a médicos) poderá enviar suas radiografias a outros países para análise. Outro exemplo: imagine uma legislação que, no afã de evitar comportamento não-ético, torne impossível qualquer teste de um dispositivo artificial que se comunique com seres humanos. Ou ainda se pode tentar controlar o mercado de trabalho, impedindo que a tecnologia floresça e que novas formas de ocupação sejam criadas – nesse caso também podemos nos ver em pouco tempo importando máquinas, bens e serviços de outros países, enquanto nossa força de trabalho míngua devido a sua falta de preparo.

É importante entender que a tecnologia de Inteligência Artificial é relativamente barata e depende apenas de boas ideias e boa formação, permitindo que empreendedores descubram novas maneiras de aumentar a qualidade de vida da população. O país não pode perder essa oportunidade por medo de enfrentar suas consequências.

De uma forma ou de outra, a produtividade geral deve aumentar com a implantação de máquinas inteligentes. Pode então ocorrer um crescimento da desigualdade social. Alguns poucos controlam a produção; alguns outros se dedicam a tarefas criativas; muitos se dedicam a tarefas menores. Alguns trabalham longas horas;

outros não têm o que fazer. Para evitar um colapso social, pode ser então instituída uma alta taxação redistributiva, gerenciada por um governo ineficiente e pouco transparente, em que decisões humanas e artificiais são igualmente incompreensíveis para a maioria da população. Nesse cenário, uns poucos se beneficiam muito do avanço tecnológico, enquanto a maioria vive uma situação ambígua: todos podem obter análises médicas de forma automática, e todos podem obter muitos outros benefícios da tecnologia onipresente, mas praticamente todos vivem com meios limitados, cercados por máquinas de alta produtividade. Todo esse cenário é realmente distópico e significa um aumento de problemas que já nos acompanham. Para evitá-lo, é preciso dar suporte à tecnologia e regrá-la adequadamente, além de conduzir um debate sério sobre a forma mais efetiva de distribuir riquezas geradas por inteligências artificiais.

Em resumo, é preciso evitar essa distopia realista a todo custo, mantendo todo o foco na utopia realista que é de fato possível.

Enviado: 1 março 2018
Aprovado: 31 abril 2018

Referências

BOSTRON, Nic. *Superintelligence*: paths, dangers, strategies. Oxford: Oxford University Press, 2014.

BRYNJOLFSSON, Erik; McAFEE, Andrew. *The second machine age*. New York, NY: Norton, 2014.

FREEMAN, Richard B. Who owns the robots rules the world, *Harvard Magazine*, vol. 118, nº. 5, MayJune, p. 37-39, 2016.

FREY, Carl Benedikt ; OSBORNE, Michael A. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, vol. 114, issue C, p. 254-280, 2017.

NILSSON, Nils. *The quest for artificial intelligence*: a history of ideas and achievements. Cambridge: Cambridge University Press, 2009.

MCCARTHY, J. MINSKY, M. L. ROCHESTER, N. SHANNON, C. E. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955. *AI Magazine*, vol. 27, nº 4. Disponível em: <https://aaai.org/ojs/index.php/aimagazine/article/view/1904/1802>. Acesso em: 3 jun, 2018.

OSOBA, Osonde. WELSER IV, William. *An intelligence in our image*: the risks of bias and errors in artificial intelligence. Santa Monica, CA: Rand, 2017.

TURING, Alan. Computing machinery and intelligence, *Mind*, vol. 59, issue 236, October 1950, p. 433-460.

O protagonismo dos algoritmos de Inteligência Artificial:

observações sobre a sociedade de dados

Dora Kaufman[1]

Resumo: A Inteligência Artificial (IA) está presente no nosso cotidiano: nos algoritmos de busca do Google, na recomendação de filmes e música do Netflix e Spotify, nas redes sociais, no Waze, nos assistentes pessoais, nos videogames, nos sistemas de vigilância e segurança, e mais em um conjunto de benesses que facilitam a vida do século XXI. Proliferam impactos negativos à serem compreendidos e equacionados. Multiplicam-se iniciativas de proteção com foco na transparência dos modelos matemáticos, e no uso dos dados. Em paralelo, as tecnologias de IA transformam a economia (migração do capitalismo financeiro para o capitalismo de dados), e as empresas (no mínimo, impactando custo e eficiência). Essas e outras questões serão abordadas no artigo.

Palavras-chave: Inteligência Artificial. Tecnologias. Sociedade de dados.

Abstract: Artificial Intelligence (AI) is present in our daily lives: Google's search algorithms, Netflix and Spotify music and movie recommendations, social networks, the Waze, personal assistants, video games, surveillance and security systems, and more in a set of blessings that make life easier for the 21st century. There are negative impacts to be understood and equated. Multiplying protection initiatives with focus on the transparency of mathematical models, and the use of data. In parallel, AI technologies transform the economy (migration from financial capitalism to data capitalism), and business (at least, impacting cost and efficiency). These and other issues will be addressed in the article.

Keywords: Artificial Intelligence. Technologies. Data society.

Introdução

Em meados de 2013, com a revelação do esquema de espionagem da NSA (*National Security Agency*), o livro "1984" de George Orwell (1949) teve um aumento de vendas de 7.000% na plataforma de comércio online Amazon, passando da posição 13.074 para 193 da lista de livros mais vendidos. Mais de quatro anos depois, em janeiro de 2017, uma semana após a posse do Presidente Donald Trump, o mesmo livro figurou como o mais vendido dentre todos os gêneros na Amazon. Longe de caracterizar fenômenos pontuais, o livro de Orwell esteve entre os 100 mais vendidos

[1] Ver Editorial, p. 5.

na plataforma entre os anos 2013-2016. Orwell concebeu uma distopia chamada "Oceania" em que a "realidade" é definida pelo governo. Nos conceitos atuais, e explica sua recente visibilidade, é uma reflexão sobre tecnologia versus privacidade e controle, temas que estão na pauta da sociedade do século XXI.

Observa-se uma sofisticação dos dispositivos de vigilância, e a mineração dos dados online pelas tecnologias de IA (redes sociais, cartão de crédito, exames e tratamentos médicos, movimentação bancária, GPS/Waze, pesquisas, reservas, e assim por diante). No âmbito corporativo, simultâneo a uma flexibilização nos modelos de gestão, consolidam-se novos sistemas de controle; a Amazon, por exemplo, referência em inovação, aprimora a cultura tradicional de comando-e-controle com base em dados (MAYER-SCHÖNBERGER; RAMGE, 2018).

O artigo se propõe a descrever e analisar alguns dos impactos da "sociedade de dados" na perspectiva dos indivíduos, e na perspectiva dos mercados e das empresas.

Na perspectiva dos indivíduos

A Inteligência Artificial está presente no nosso cotidiano:[2] nos algoritmos de busca do Google, na recomendação de filmes e música do Netflix e Spotify, nas redes sociais, no aplicativo Waze, nos assistentes pessoais, nos videogames, nos sistemas de vigilância e segurança, e mais em um conjunto de benesses que, efetivamente, têm o potencial de facilitar a vida do século XXI.

O marketing e a propaganda usam os algoritmos de IA para identificar os hábitos e preferências dos consumidores e produzir campanhas mais assertivas e segmentadas. O mesmo ocorre com as áreas comerciais: no setor imobiliário, os algoritmos permitem identificar se você foi designado para uma função em outra cidade e/ou contratado por uma empresa com escritório em outra cidade, acessar os locais e os tipos de moradia que você vem pesquisando na Internet, qual o tamanho de sua família e assim por diante, aumentando a chance de ofertas apropriadas de imóveis. O varejo físico incorpora as "vantagens" do varejo online por meio de dispositivos que permitem identificar por onde o cliente circulou nas lojas, a trajetória do seu olhar nas prateleiras, por quantas vezes e por quanto tempo. São os algoritmos

[2] Pelo menos de uma parte da população que tem acesso à internet e aos dispositivos digitais.

de IA que transformam em informação útil a imensidão de dados gerados pelas movimentações digitais ("rastros digitais").

Seus benefícios são inegáveis, e os indivíduos e a sociedade os reconhecem. Em paralelo, contudo, proliferam impactos negativos a serem compreendidos e equacionados. Dentre eles, destacam-se (a) o viés nos processos de decisão automatizados, (b) a invasão da privacidade e as novas formas de controle, e (c) a personalização dos acessos e pesquisas online.

Viés nos processos de decisão automatizados

Um sistema chamado COMPAS[3] (*Correctional Offender Management Profiling for Alternative Sanctions*) no Estado de Wisconsin e similares em outros estados americanos, baseados em algoritmos, determinam o grau de periculosidade de criminosos e consequentemente a pena do condenado. A intenção, segundo seus defensores, é tornar as decisões judiciais menos subjetivas. A metodologia de avaliação, criada por uma empresa privada comercial, vem sendo fortemente contestada. O modelo matemático FICO,[4] usado por agências de crédito como Experian/Serasa, Transunion e Equifax, avalia o risco de um indivíduo não quitar um empréstimo bancário (propensão à inadimplência) com base em seu histórico. Em 2013, a Comissão de Comércio Federal Americana informou que 5% dos clientes (cerca de dez milhões) tiveram um erro em um de seus relatórios de crédito, resultando em taxas maiores. Os sistemas de avaliação das agências apresentam resultados díspares: estudo de 500.000 arquivos indicou que 29% das agências de crédito tinham pontuações diferentes em pelo menos cinquenta pontos, implicando em custos mais altos de empréstimos ou financiamentos (PASQUALE, 2015).

As áreas de RH das empresas valem-se de pontuações de crédito nos processos de contratação, supondo que o mau crédito se correlaciona com o mau desempenho no trabalho, implicando numa espiral descendente (dificuldade em honrar empréstimos acarreta dificuldade de realocação profissional). Acessam igualmente o histórico médico dos candidatos recorrendo a um cada vez mais unificado banco de dados (*big data*). Vasconcelos, Cardonha e Gonçalves (2017) apontam três problemas na chamada

[3] Correctional Offender Management Profiling for Alternative Sanctions.
[4] FICO: disponível em: <http://www.fico.com/en/customers>. Acesso em: 18 mai. 2018.

"contratação algorítmica": (a) dados históricos do candidato podem não ser adequados para a finalidade de filtragem, (b) dados extraídos de redes sociais podem ser questionáveis do ponto de vista ético, e (c) substituição de vários tomadores humanos de decisões por um único algoritmo pode implicar em perda de diversidade. Os autores indicam diretrizes para mitigar esses efeitos: projetar um processo de supervisão para buscar explicitamente correlações fortes sobre atributos sensíveis, tentando desvendar o preconceito antes que o sistema seja implantado. O resultado, todavia, é sempre suscetível a preconceitos porque depende de dados gerados pelo homem, imputados diretamente ou por meio de processos de aprendizado.

Brynjolfsson e McAfee (2017) admitem que existem riscos na decisão automatizada, mas ponderam que "embora todos os riscos da IA sejam muito reais, o padrão de referência adequado não é a perfeição, mas sim a melhor alternativa possível. Afinal, nós humanos temos vieses, cometemos erros e temos problemas para explicar, de fato, como chegamos a determinada decisão" (ibid.). Por outra linha de raciocínio, pode-se argumentar que esses modelos são simples referências no processo de tomada de decisão. Ou ainda, que no estágio atual, em que as máquinas ainda dependem da supervisão humana, cabe a este inserir nas máquinas os parâmetros, ou seja, a responsabilidade sobre o processo. "Não seremos capazes de superar todos os preconceitos humanos e falhas de decisão; mesmo que os seres humanos escolham usar sistemas de aprendizado de máquinas inteligentes em mercados ricos em dados, essa escolha ainda será humana" (MAYER-SCHÖNBERGER; RAMGE, 2018, p. 14).

Para Cathy O'Neil (2016) os nossos valores e desejos, expressos nos dados que selecionamos, influenciam nossas escolhas, ou seja, os modelos são opiniões incorporadas em Matemática. "A questão, no entanto, é se eliminamos o viés humano ou simplesmente o camuflamos com tecnologias" (ibid., p. 25).

Invasão da privacidade e as novas formas de controle

A cultura e a prática de controle permeiam as empresas desde sua origem na Revolução Industrial, supondo-se uma correlação entre o grau de eficiência dos controles internos e o grau de eficiência da própria empresa. O exercício do controle, todavia, extrapola os relatórios e os sistemas criados com essa finalidade; a jornada de trabalho

constitui-se por si só em um poderoso domínio sobre os indivíduos. Foucault (2002, 2005, 2008) introduz esse debate ao analisar as transformações na sociedade com relação à utilização do tempo[5] e do espaço.[6] No final do século XX, concepções e metodologias de gestão introduziram novas dinâmicas,[7] promovendo ambientes corporativos mais flexíveis e engajados.

A partir de uma lógica distinta, contrariando a tendência das últimas décadas, as novas tecnologias propiciam controles sofisticados desde o processo de seleção e contratação, ao desligamento final do funcionário. Investigação do *New York Times*, em 2015, sobre as condições de trabalho nos escritórios da Amazon, constatou que os funcionários são responsabilizados por um conjunto de métricas sobre diferentes aspectos operacionais (descritos em cerca de cinquenta páginas), sendo solicitados semanalmente a explicar as ineficiências detectadas (MAYER-SCHÖNBERGER; RAMGE, 2018). Aparentemente, a política da Amazon está sendo adotada pelas empresas de tecnologia, em um processo denominado pela *The Economist taylorismo digital*. Complementando os controles internos, as empresas acessam as publicações de seus funcionários nas redes sociais e, em alguns casos, um conjunto mais amplo de dados (prontuário médico até viagens de férias).

No evento *Sustainable Brands*,[8] David O'Keefe, da telefonica Dynamic Insights, apresentou um produto derivado dos dados captados das linhas móveis (*mobile phone data*). Com o título "usando dados comuns globais e aprendizado de máquina para fornecer informações de relacionamento digital em multinacionais" (*using global communication data and machine learning to provide digital relationship insights in multinationals*), O'Keefe descreveu o "produto" em que, por meio dos dados dos celulares dos funcionários de uma empresa multinacional (quem ligou para quem, com que frequência, quanto tempo durou a ligação, etc.) é possível identificar as redes informais internas, importante elemento nas estratégias de gestão. Essas redes, mais do que as redes formais definidas nos organogramas, indicam as conexões de

[5] O tempo pensado aqui como um instrumento de dominação, de controle do próprio corpo do indivíduo. Manter a qualidade e a produtividade demanda eliminar os fatores que possam perturbar ou distrair o desempenho das funções, formando o que Foucault denominou de "tempo integralmente útil".
[6] A "arquitetura para vigiar" organiza-se um inusitado tipo de vigilância a partir de um controle intenso e contínuo ao longo de todo o processo de trabalho.
[7] Como, por exemplo, a arquitetura de "espaços abertos" nos escritórios em contraposição ao modelo anterior de salas ou divisões individuais ("baias de trabalho"). Num processo mais recente, surge o sistema de trabalho conhecido como *home-office* com os indivíduos recuperando, relativamente, parte do poder sobre seu tempo e espaço (Kaufman, 2017).
[8] Disponível em: <https://events.sustainablebrands.com/sb17saopaulo/pt/>. Acesso em: 18 mai. 2018.

influência e de poder nas empresas (além do tempo que cada funcionário "gasta" ao celular com assuntos externos ao trabalho). Parece ficção científica, mas é realidade e supera de longe as previsões de George Orwell no livro *1984*, publicado vários anos antes do termo Inteligência Artificial ter sido cunhado (1956).

A questão não se resume a ter ou não acesso a dados específicos, inclusive porque vários têm mecanismos de proteção. O risco maior está na combinação e correlação de dados originados em distintas fontes, que geram novos dados privados (correlações estatísticas) livres de supostos acordos de privacidade.

Personalização dos acessos e pesquisas online

A recente explosão de dados na Internet e Web substituiu a ideia de "liberdade" pela ideia de "relevância", na formação do fluxo de informações online;[9] o acesso à informação passou à ser personalizado. Sofisticados algoritmos de Inteligência Artificial individualizam as consultas ao Google, i.e., os resultados variam em função do perfil de quem está buscando a informação. Pariser (2011), ativista da Internet, alerta para o processo invisível de filtragem de conteúdo que, ao gerar resultados personalizados, nos coloca em contato com o que queremos ver e não com o que devemos ver, e que temos que assegurar o acesso não só ao que é relevante, mas também ao que é desconfortável, desafiador e outros pontos de vista. Pariser denuncia a falta de transparência.

A rede social Facebook utiliza algoritmos de IA no gerenciamento das publicações no feed de notícias de seus usuários. São disponibilizados diariamente cerca de 2 mil itens para cada usuário (mensagens, imagens, vídeos); dentre esse conjunto de informações, os algoritmos identificam e selecionam de 100 a 150 publicações com a intenção de facilitar a experiência do usuário.[10] Para processar com assertividade a seleção de "conteúdos relevantes" e estabelecer correlações, os algoritmos[11] precisam ter acesso a uma grande e diversificada quantidade de dados.[12] No processo os algoritmos interferem na mediação entre seus usuários.

[9] Um dos desafios é deliberar como e a quem cabe a função de "curadoria".
[10] A comunicação não é controlada inteiramente pelo Facebook, o usuário que pode acessar os recursos de ajustes disponibilizados pela plataforma.
[11] Algoritmo: conjunto de instruções matemáticas que serve para implementar estratégias e objetivos pré-definidos (referência: dossiê dessa publicação).

A principal crítica aos sistemas inteligentes é a formação de "bolhas", ou "câmara de ecos", ao promover a homogeneização das relações sociais, mantendo as pessoas em círculos sociais fechados formados por iguais. Guess, com base em um estudo[13] no qual é um dos autores, pondera que são exageradas as afirmações de prevalência desses fenômenos. Para ele, "a narrativa 'câmaras de eco' captura, no máximo, a experiência de uma minoria do público". Ele alerta, inclusive, que é difícil estudar o problema na medida em que os dados e os algoritmos das plataformas de mídias sociais são em sua maioria proprietários (ou seja, acesso limitado), e crê que há mais evidências de "câmaras de eco" na vida real do que online.

Na perspectiva dos mercado e das empresas

Ao longo da história, com distintas visões, as transações entre os agentes econômicos foram regidas pela oferta e demanda de produtos e serviços, reguladas pelo fator "preço": quanto maior a oferta dada uma demanda estável, os preços dos bens[14] tendem a reduzir, e quanto maior a demanda dada uma oferta estável, os preços dos bens tendem a aumentar. A mesma dinâmica pode ser observada pelo prisma da quantidade do bem: excesso de demanda gera escassez do bem e a tendência do preço de equilíbrio é subir, e, no sentido inverso, excesso de oferta gera abundância do bem e a tendência é o preço de equilíbrio cair. O pressuposto desses modelo é de um mercado eficiente e que os agentes tomam decisões racionais. Nas economias tradicionais, o fluxo de informações converge para o preço, mensurando as preferências do consumidor dentre diversas outras variáveis econômicas, i.e, o preço conecta oferta e demanda.

No regime denominado por Mayer-Schönberger e Ramge (2018) de *data capitalism*, o preço perde sua centralidade, os agentes utilizam os dados para identificar *better matches* explorando várias dimensões, em uma transição do capitalismo financeiro para o capitalismo de dados. No primeiro, a informação, difícil e cara, convergia para o "preço"; no segundo, a informação é múltipla, complexa, rápida e

[12] A seleção do feed de notícias extrapola a movimentação de um usuário individual, os algoritmos buscam similaridades com outros usuários no processo denominado "aprendizagem profunda" (*deep learning*) – referência: Dossiê dessa publicação).
[13] Avoiding The Echo Chamber About Echo Chambers: Why selective exposure to like-minded political news is less prevalent than you think, Knight Foundation. Disponível em: <https://medium.com/trust-media-and-democracy/avoiding-the-echo-chamber-about-echo-chambers-6e1f1a1a0f39>. Acesso em: 15 mai. 2018.
[14] Bens aqui entendidos como produtos e serviços.

barata. Os autores identificam três tecnologias-chave:[15] (a) linguagem padrão para comparar e compartilhar os dados sobre os bens e as preferências, (b) capacidade para identificar "matches" em várias dimensões e selecionar os parceiros e transações adequados, e (c) capturar e usar as preferências de maneira eficaz (assertividade). Para os autores, os dados estão substituindo o preço como elemento estrutural da relação produtor e consumidor, e a moeda como meio de pagamento.[16] Hoje, já pagamos vários serviços com dados (pesquisa no Google, benefícios do Facebook – relações sociais e plataforma de negócios) e, em breve, essa prerrogativa deve se estender às anuidades dos cartões de crédito, as taxas bancárias e aos custos da telefonia, setores que concentram grandes volumes de dados de seus clientes.

As grandes empresas de tecnologia – as FAANGS como denominadas pelo *Financial Times* (Facebook, Amazon, Apple, Netflix e Google) –, são a parte mais visível da economia de dados, mas não a única. A queda de receita na função "voz" pressiona as empresas de telecomunicações na busca por produtos alternativos e, aparentemente, a inovação disruptiva está no uso dos dados de seus usuários, particularmente na telefonia móvel. Os bancos, talvez o setor com mais acesso a dados privados, ainda não estão usando os dados de seus clientes em sua plena dimensão concentrados em reduzir os custos através da migração de plataformas físicas para digitais. Existem fortes indícios de que em breve, os bancos vão se reinventar como intermediários de informação,[17] preservando as funções de transferência e armazenamento de valor:[18]

> Embora, em teoria, os bancos devam se sentir muito à vontade trabalhando com muitos dados, porque eles coletam e operam uma grande variedade de dados financeiros detalhados de seus clientes há muitas décadas, eles não fizeram muito com os dados que possuem. Neste contexto, eles são ricos em dados, mas pobres em insights. (Ibid., p. 154)

[15] A geração e armazenamento de dados não é tão recente na economia, o novo é a capacidade de manipular esses dados transformando-os em valor (tecnologias de Inteligência Artificial).
[16] Os dados são o que os economistas chamam de "bem não rival", ou seja, os mesmos dados podem ser utilizados por múltiplos agentes, o que já se constitui numa vantagem sobre a moeda.
[17] As criptomoedas (moedas virtuais privadas) associadas com as tecnologias do blockchain (permite transferir propriedade de ativos e documentos), equacionando os problemas atuais (custo de energia, alta volatilidade, etc.) têm o potencial de reconfigurar o sistema financeiro tanto do ponto de vista dos bancos privados quanto dos bancos centrais (autoridade monetária).
[18] O processo passa por investimentos pesados dos bancos nas Fintechs (empresas do setor financeiro intensivas em tecnologia, termo deriva da junção de Finanças e Tecnologia) e alternativas relacionadas a sistemas de blockchain (redução de custos de transferencia, no mínimo). Na última década, o investimento global em FinTech cresceu mais de 15 vezes: de um total aproximado de US$ 6,8 bilhões em 2005 para US$ 107 bilhões em 2017, concentrado em soluções para pagamentos, transferências, processamentos, destacando-se a emergência recente do Blockchain. Do total investido, US$ 72,1 bilhões de dólares foram alocados no mercado americano (fonte: Darryl West, do HSBC, no evento RISE, Hong Kong, julho/2017).

Para que a economia de dados funcione é imprescindível rotular e categorizar a informação, ou seja, registrar digitalmente e detalhadamente as referências individuais de produtos e serviços. A falta de uma antologia reduz o número de transações pela limitação em encontrar um "match", i.e., a falta de filtros de identificação compromete a eficiência do mercado. A previsão de Mayer-Schönberger e Ramge (2018) é que os próprios dados vão impulsionar as antologias de dados.[19] Na origem da Amazon, em meados da década de 1990, ao perceber a impraticabilidade de lançar uma loja online tudo, Jeff Bezos, fundador e CEO, analisou uma lista de vinte possíveis categorias de produtos optando pelos livros: além de serem *commodities*, existiam três milhões de livros impressos em todo o mundo, e os catálogos sazonais dos editores tinham sido digitalizados (STONE, 2013).

Não é suficiente, contudo, a disponibilidade dos dados brutos; extrair as informações demanda um processo de correspondência que seja inteligente o suficiente para levar em conta as múltiplas dimensões de preferências e seu peso relativo. Plataformas como Spotify, Apple Music, Netflix e Amazon, utilizam-se de IA para combinar as preferências dos seus usuários e recomendar com mais precisão músicas, filmes, ou produtos em geral. Os algoritmos de IA viabilizam esses processos identificando padrões complexos embutidos nos dados, analisando o comportamento passado para prever o futuro, e criando estratégias para sensibilizar os clientes ideais.

Uma das consequências mais perversas do capitalismo de dados é a concentração do mercado,[20] supostamente derivada de três efeitos: escala, que reduz custos; rede ou "externalidade da rede", que expande adesão (quanto maior o número de usuários maior as novas adesões);[21] e feedback frequente, que aprimora o produto e gera ganhos de eficiência (MAYER-SCHÖNBERGER; RAMGE, 2018). Endossando a tese dos economistas Ariel Ezrachi e Maurice Stucke de que os sistemas de aprendizado de

[19] Alation, Corrigon e Expertmaker são algumas das start-ups com foco nesses processos.

[20] Google concentra cerca de 4 de 5 solicitações de pesquisa originadas em desktop e 9 de 10 solicitações originadas de dispositivos móveis, e seu similar Baidu tem 60% do mercado chinês de busca; Amazon tem mais de 40% das receitas de varejo online nos Estados Unidos. Facebook tem 2 bilhões de usuários no mundo, e a chinesa Tencent, proprietária do aplicativo WeChat (pagamento online e troca de mensagens instantânea), é a primeira empresa chinesa a superar os 500 bilhões de dólares em valor de mercado; Alibaba tem cerca de 51,3% de market-share na China,seu principal concorrente, Jingdong, tem 32,9%. Os nichos menores reproduzem padrão similar: O GoDaddy, maior registrador de nomes de domínio da Internet é 4 vezes maior do que seu concorrente, o WordPress domina os registros de blog, o Netflix governa *streaming* de filmes, o Instagram tem mais de 500 milhões de usuários ativos por dia contra 173 milhões de seu principal concorrente, Snapchat. Facebook e Google detêm mais de 60% do mercado de anúncios online (Fonte: McKinsey).

[21] Padrão de estruturação denominado por Barabási de *rich get richer* ("ricos ficam mais ricos"), segundo o qual as redes não são conectadas igualitariamente, mas, ao contrário, as que têm mais conexões, mais links, ampliam as oportunidades de gerar mais conexões (BARABÁSI, Albert-László. Linked: a nova ciência dos networks. São Paulo: Leopardo, 2009).

máquinas estejam minando a concorrência, Mayer-Schönberger e Ramge (ibid.) refutam a visão de que a solução passa unicamente pela abertura dos algoritmos.

> Os algoritmos, por si só, não são suficientes para permitir que pequenos competidores e novos concorrentes compitam com empresas estabelecidas, porque os algoritmos não são a matéria-prima […] os reguladores que desejam garantir mercados competitivos devem exigir o compartilhamento de dados. (Ibid., p. 168)

A vantagem comparativa estaria na posse dos dados, e não no conhecimento dos algoritmos. Se e quando os dados dos grandes participantes estiverem disponíveis para os concorrentes menores, a tendência será a inovação se disseminar com a posse dos dados deixando de ser uma barreira de entrada.

Empresas

Uma das principais diferenças entre o mercado e a empresa é a maneira como as decisões são tomadas: no mercado, descentralizada e compartilhada entre os participantes; na empresa, centralizada e investida em um número relativamente restrito de executivos. No ambiente corporativo, os fluxos de informação, as decisões e a estrutura comunicativa permanecem concentrados nos níveis mais altos de gerência. Em decorrência, Mayer-Schönberger e Ramge (2018) apontam dois potenciais benefícios da IA que não estão sendo apropriados pelas empresas: (a) a automação nos processos de decisão, função da relativa pouca geração de dados associados às decisões nas funções gerenciais, ou seja, ausência de dados suficientes para os processos de aprendizado dos sistemas inteligentes;[22] e (b) a inovação radical, função do fato de que as novas ideias não estão contidas nos dados, ou seja, os sistemas de IA não tem pontos de referência para aprender e propor.

Observa-se uma gradativa incorporação da IA na operação interna das empresas, aparentemente privilegiando, com a automação, os efeitos de eficiência e redução de custos dos processos. Raras são as experiências que ensejam transformações disruptivas nos modelos de negócio. Agrawal, Gans e Goldfarb (2018) atribuem as tecnologias de IA um papel também na redução dos custos de previsão,

[22] Essa restrição, em parte, retarda a substituição de humanos por máquinas nas funções cognitivas nas empresas, particularmente nos níveis mais altos da hierarquia.

desde a projeção de um inventário até o treinamento dos carros autônomos (previsão da ação humana dada a determinadas condições).

Em paralelo, apropriando-se das vantagens das novas tecnologias, emergem inéditos modelos de negócio alterando os critérios de competitividade e gerando novos líderes setoriais (Airbnb, Uber, Amazon, Alibaba, Google, Facebook, Netflix, dentre outros). Um desses modelos são as empresas – plataformas, organizações centradas em tecnologia, Tom Slee (2017) discute esse modelo como a mais importante transformação do capitalismo do século XXI.[23] O arcabouço regulatório será um dos fatores determinantes na sobrevivência ou na falência de alguns desses modelos.

Reflexões finais

René Descartes, fundador da filosofia moderna, defendeu a segregação dos mundos humano e animal considerando o primeiro como o único ser vivo capaz de pensar racionalmente (GUNKEL, 2012); com o conceito de "animal-máquina" (*bête-machine*), o filósofo relacionou os animais aos autômatos, identificando uma similaridade entre animais e máquinas. Com base em Descartes, Gunkel propõe pensar sobre às máquinas inteligentes a partir da ideia de agente que produz uma ação intencional (como os humanos e animais). Santaella (2018) aponta duas diferenciações entre humanos e animais: a linguagem[24] humana é evolutiva (transforma-se, adapta-se), e o humano é o único animal que fala. "A julgar por seus avanços recentes, restam poucas dúvidas acerca do fato de que, mais cedo ou mais tarde, a IA deverá abranger muitas das competências que até agora julgamos serem privilégios exclusivos dos humanos" (ibid., p. 2).

As incertezas sobre o futuro da IA coloca a questão filosófica se faz sentido investir no desenvolvimento de uma inteligência sem controle humano ou se é mais prudente abdicar de seus potenciais benefícios. Proliferam iniciativas de proteção, envolvendo pesquisadores, empresas, governos, agências regulatórias, particularmente na Europa. No âmbito do poder público, dois obstáculos têm o

[23] A extensão e complexidade dos novos modelos de negócios demandam um artigo exclusivo, a intenção é apenas incluir o tema nos impactos das tecnologias de IA no mercado/economia.
[24] "Linguagem" extrapola a linguagem verbal, extendendo-se à outras linguagens: visuais, sonoras, gráficas, notacionais, simbólicas, hipermídia, linguagem de máquina, linguagem de programação, linguagem algorítmica etc. (SANTAELLA, 2018).

potencial de comprometer os resultados: falta de conhecimento relativo dos reguladores sobre as novas tecnologias (do outro lado estão os maiores e mais bem pagos profissionais e poderosas empresas) e a acelerada evolução versus o desafio de manter a legislação atualizada. Com isso, aumenta a responsabilidade da sociedade, particularmente com o tema da "transparência" não aceitando como inevitável a opacidade dos modelos matemáticos. O desafio é que, segundo os especialistas, não sabemos exatamente como as máquinas aprendem, observamos apenas o efeito do aprendizado através de testes que, se bem elaborados, reduzem os riscos.

Quanto aos dados, temos o uso primário derivado de dados coletados nas fontes primárias, ou seja, dados coletados na origem como nas nossas movimentações no Google ou Facebook. E temos os dados secundários, ou reúso dos dados, que são os dados adquiridos por terceiros.[25] É viável supor que uma legislação apropriada possa exercer controle sobre as geradoras de dados para impedir o acesso de terceiros? Como definir o que seja "bom" uso e "mau" uso dos dados? E não são apenas as empresas de tecnologia, mas igualmente os cartões de crédito, os bancos, as farmácias, as seguradoras, os laboratórios médicos, em fim todos os setores que acumulam dados sensíveis na identificação de hábitos, comportamentos, características, perfis, por isso mesmo valiosos.

Para Pasquale (2015) a lei da informação protege muito mais do que a lei de privacidade pessoal, e denuncia dois movimentos opostos: "enquanto empresas poderosas, instituições financeiras e agências governamentais escondem suas ações por trás de acordos de não divulgação, 'métodos proprietários' e regras de mordaça, nossas próprias vidas são livros cada vez mais abertos" (ibid., p. 3). Ele observa que as empresas coletam cada vez mais dados sobre seus usuários, mas combatem as regulamentações que permitiriam a esses mesmos usuários exercer algum controle sobre elas. A complexidade aumenta se levarmos em conta a "barreira tecnológica", poderoso limitador da capacidade de controle por parte da sociedade.

Não há consenso entre os *experts* sobre o futuro da Inteligência Artificial. As pesquisas apontam ser alta a probabilidade da superinteligência[26] ser criada ainda no

[25] Na crise Facebook – Cambridge Analytica tratou-se do uso por terceiros dos dados gerados na rede social.
[26] Nick Bostrom (2014), no livro *Superintelligence*, define superinteligência como "um intelecto que excede em muito o desempenho cognitivo dos seres humanos em praticamente todos os domínios de interesse".

século XXI.[27] Para Kevin Kelly (2010) a vantagem obtida com a cognição de objetos inertes será centenas de vezes mais perturbadora para nossas vidas do que as transformações obtidas pela industrialização. "A chegada do pensamento artificial acelera todas as outras rupturas" (ibid., p. 30), e inaugura novas formas de mediação. Torna-se difícil identificar quem ou o que está agindo, e em localizar, compreender e isolar o papel e a função dos humanos e da tecnologia. O que caracteriza o "ser humano" tradicional encontra-se alargado pelo acoplamento com tecnologias, impossibilitando a identificação dos limites do que seja humano e não-humano; os limites do próprio corpo e da cognição estão expandidos. Temos desde as "tecnologias vestíveis" (*wearable)* até a introdução de dispositivos de IA. Trata-se de inéditas mediações, em interações e diálogos entre inteligências.

Enviado: 20 abril 2018
Aprovado: 20 maio 2018

Referências

ALPAYDIN, Ethem. *Machine learning*. Cambridge, MA: MIT Press, 2016.

AIKEN, M. *The cyber effect*: a pioneering cyberpsychologist explains how human behaviour changes online. New York, NY: Spiegel & Grau, 2016.

ANDERSON, Chris. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 23/06/2008. Disponível em: <https://www.wired.com/2008/06/pb-theory>. Acesso em: 12 mar, 2018.

AGRAWAL, A.; GANS, J.; GOLDFARB, A. *Prediction machines*: the simple economics of artificial intelligence. Boston, MA: Harvard Business Review Press, 2018.

BENTLEY, R.A.; O'BRIEN, M. *The acceleration of cultural change*: from ancestors to algorithms. Cambridge, MA: MIT Press, 2017.

[27] Em relação ao tempo de concretização de uma "máquina inteligente", as pesquisas entre especialistas indicam 10% de probabilidade até 2020, 50% de probabilidade até 2040 e 90% de probabilidade até 2075, supondo que as atividades de pesquisa continuarão sem maiores interrupções (BOSTROM, 2014).

BRYNJOLFSSSON, E.; McAFEE, A. O negócio da Inteligência Artificial. *Harvard Business Review*, 6 de nov., 2017. Disponível em: <http://hbrbr.uol.com.br/o-negocio-da-inteligencia-artificial/>. Acesso em: 12 mar, 2018.

CHRISTIAN, Brian; GRIFFITHS, Tom. *Algorithms to live by*: the computer science of human decisions. New York, NY: Henry Holt, 2017.

FOUCAULT, Michel. *A verdade e as formas jurídicas*. 3. ed. Rio de Janeiro: Nau, 2002.

_____. *Em defesa da sociedade*: curso no Collège de France (1975-1976). São Paulo: Martins Fontes, 2005.

_____. *Nascimento da biopolítica*. São Paulo: Martins Fontes, 2008.

GUNKEL, D.The *Machine question*: critical perspectives on ai, robots, and ethics. Cambridge, MA: MIT Press, 2012.

KELLY, Kevin. *What technology wants*. New York, NY: Viking, 2010.

MAYER-SCHÖNBERGER, Viktor; RAMGE, Thomas. *Reinventing capitalism in the age of big data*. London: John Murray, 2018.

O'NEIL, Cathy. *Weapons of math destruction*: how big data increases inequality and threatens democracy. New York, NY: Crown, 2016.

PARISER, Eli. *The filter bubble*: what the Internet is hiding from you. London: Penguin, 2011.

PASQUALE, F. *The black box society*: the secret algorithms that control money and information. Cambridge, MA: Harvard University Press, 2015.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence*: a modern approach. Upper Saddle River, NJ: Pearson, 2009.

SANTAELLA, Lucia. A IA veio para ficar, crescer e se multiplicar. Disponível em: <https://transobjeto.wordpress.com/2018/05/19/a-ia-veio-para-ficar-crescer-e-se-multiplicar/>. Acesso em: 19 maio, 2018.

STONE, Brad. *The everything store*: Jeff Bezos and the age of Amazon. New York, NY: Little Brown, 2013.

SLEE, Tom. *Uberização*: a nova onda do trabalho precarizado. São Paulo: Editora Elefante, 2017.

VASCONCELOS, Marisa; CARDONHA, Carlos; GONÇALVES, Bernardo. Modeling epistemological principles for bias mitigation in AI systems: an illustration in hiring

decisions. arXiv:1711.07111v1 20 nov. 2017. Disponível em: <https://arxiv.org/abs/1711.07111>. Acesso em: 12 maio, 2018.

O problema da explicação em Inteligência Artificial:

considerações a partir da semiótica

Joel Carbonera[1]
Bernardo Gonçalves[2]
Clarisse de Souza[3]

Resumo: Desde os sistemas especialistas dos anos 1980 e 1990, pesquisadores de Inteligência Artificial (IA) dedicam-se ao problema da explicação, a saber, dada uma inferência por parte do sistema, como identificar os passos ou mecanismos que o levaram a tal conclusão. Com o recente sucesso dos sistemas de IA atuais, sobretudo os baseados em aprendizagem profunda, esse problema voltou à tona com vigor, agora mais pronunciado, por eles serem opacos quanto ao seu processo de inferência, em contraste com os sistemas especialistas, então baseados em regras lógicas. Neste texto, apresentamos o problema da explicação, incluindo destaques de sua literatura mais recente na área de IA. Em seguida, indicamos lacunas de abordagens passadas e recentes, e apresentamos então considerações a partir da semiótica de Peirce que, conforme argumentamos, poderiam contribuir para uma condução equilibrada dessa tecnologia na sociedade.

Palavras-chave: Inteligência Artificial. Explicabilidade. Semiótica e Pragmatismo. Engenharia Semiótica.

Abstract: Since the expert systems of the 1980s and 1990s, Artificial Intelligence (AI) researchers have tried to solve the the problem of explanation, namely, given an inference from the system, how to identify the steps or mechanisms that have led to the conclusion. With the recent success of AI systems, especially those based on *deep learning*, this problem has come to the fore again more forcefully since the processes are opaque as far as their inferences are concerned, in contrast to expert systems, which are based on logical rules. In this text, we present the problem of explanation, including highlights from its most recent literature in the area of AI. Next, we indicate gaps in past and recent approaches, and then present considerations from Peirce's

[1] Doutor em Ciência da Computação na Universidade Federal do Rio Grande do Sul, membro do grupo BDI (grupo de bancos de dados inteligentes) da UFRGS e do grupo de trabalho financiado pelo IEEE RAS, intitulado Padrão para Ontologias para Robótica e Automação (IEEE RAS WG ORA), coordenador de padronização do campo de Robótica e Automação no capítulo IEEE South Brazil Robotics & Automation Society. E-mail: jlcarbonera@inf.ufrgs.br.
[2] Pós-doutorado na Universidade de Michigan–Ann Arbor, Ph.D. em Modelagem Computacional com foco em Data Science/ Laboratório Nacional de Computação Científica (LNCC), doutorando em Filosofia da Ciência/ USP, membro da Associação Profissional de Scientiae Studia e da Associação para a Filosofia e História da Ciência do Cone Sul. E-mail: bgoncalves1@gmail.com.
[3] Professora titular do Departamento de Informática PUC-Rio, doutora em Linguística Aplicada (foco interação humano - computador), Criadora da Engenharia Semiótica, em 2010 foi agraciada com o ACM SIGDOC Rigo Award e em 2013 tornou-se membro da ACM SIGCHI CHI Academy. Em 2014 recebeu o título de HCI Pioneer, outorgado pelo Comitê Técnico de Interação Humano-Computador (TC13) da IFIP. Também em 2014 foi selecionada como uma das 52 pesquisadoras mulheres a figurarem na primeira edição do CRA-W / Anita Borg Institute Notable Women in Computing Card Deck. Em 2016 recebeu o Prêmio do Mérito Científico da SBC e em 2017 o prêmio | Carreira de Destaque em IHC, concedido pela Comissão Especial de Interação Humano Computador da SBC. Em licença sabática da PUC-Rio, trabalhando como Pesquisadora Senior na IBM Research Brazil. E-mail: clarisse@inf.puc-rio.br.

semiotics, which, as we argue, could contribute to a balanced management of this technology in society.

Keywords: Artificial intelligence. Explanability. Semiotics and Pragmatism. Semiotic Engineering.

Introdução

Em dezembro de 2016, a jornalista Carole Cadwalladr (2016), do *The Guardian*, foi a um popular motor de busca para uma pesquisa, e digitou "j-u-d-e-u-s", seguido de "s-ã-o". As sugestões de "autocompletar" e os resultados obtidos (o conjunto de páginas retornadas) foram surpreendentes. Então, numa nova busca, ela digitou "m-u-ç-u-l-m-a-n-o-s", e novamente "s-ã-o"; noutra, digitou "m-u-l-h-e-r-e-s", depois "s-ã-o"; e assim por diante, até se defrontar com um mundo onde "Hitler foi um cara legal" (sic!). Ocorre que o sistema de busca em questão pode ser considerado um sistema de Inteligência Artificial (IA) que *aprende* dos dados, e que é potencialmente vulnerável a vieses dos mais inofensivos aos mais repulsivos. Mas como saber que tipo de viés pode estar codificado em um sistema de IA? Essa é uma questão que está radicada no chamado problema da explicação de (um sistema de) IA – a saber, *como identificar os passos ou mecanismos que levaram um sistema a chegar a tal ou qual decisão?* –, que é o tema deste texto.

À medida que vão sendo implantados em nossa rotina uma miríade de sistemas de IA, das recomendações de produtos, ao reconhecimento facial e aos *chatbots*, a sociedade tem despertado para o problema da lacuna de explicações acerca do comportamento (inteligente) de sistemas de IA. São sinais disso: a formação, em novembro de 2016, de um consórcio de parceria em IA para beneficiar as pessoas e a sociedade (HERN, 2016), por parte de algumas das maiores empresas de tecnologia do mundo (Google, Facebook, Amazon, IBM, Microsoft); e a legislação relativa à *General Data Protection Regulation* (GDPR), que entrou em vigor em 2018 no âmbito da União Européia, induzindo um direito à explicação por parte da/do cidadã(o) que seja "afetado significativamente" (sic!) por decisões automatizadas tomadas por algoritmos preditivos no nível (individual) de um usuário (GOODMAN, 2016).

No contexto dessa recente guinada com relação à relevância e à seriedade com que o tema da IA é tratado no domínio público, a partir do ano de 2017, a comunidade

técnica de IA e aprendizagem de máquina tem reagido com a criação de fóruns especializados (de realização conjunta com as conferências ou simpósios técnicos regulares) para discussão da questão da explicação em IA, o que deu origem à expressão *Explainable AI*, também conhecida pelo acrônimo XAI. Como veremos neste texto, entretanto, é possível que tais iniciativas ainda se ressintam de uma perspectiva demasiado unilateral (oriunda das virtudes e dos vícios de uma orientação técnica específica da comunidade científica), havendo espaço então – sobretudo, em se tratando de um problema de explicação, ou, se assim o quisermos, de comunicação de sentido – para que seja ampliada a discussão à luz das ciências humanas e da semiótica.

Começaremos com uma breve apresentação do problema da explicação em IA e aprendizagem de máquina, seguida de uma breve revisão da literatura pregressa e recente no tema. Procederemos então, indicando lacunas na maneira como o problema vem sendo abordado em IA, e apresentaremos uma perspectiva mais ampla do problema a partir da semiótica de Peirce, em direção a uma condução equilibrada da implantação de sistemas de IA na sociedade.

Visão geral do problema da explicação em IA

Uma breve retomada histórica

Entre as décadas de 70 e 80, a capacidade de explicação de inferências se apresentou como um problema relevante que atraiu a atenção de pesquisadores de IA. Isso se deu primeiro no contexto dos chamados "sistemas especialistas", que eram baseados em regras lógicas e heurísticas de busca para chegarem a uma conclusão visando apoiar a decisão de especialistas humanos. Esse é o caso, por exemplo, do apoio ao diagnóstico médico (CLANCEY; SHORTLIFFE, 1984). Para que a conclusão do sistema fosse aceita, era preciso oferecer ao ser humano responsável uma espécie de rastro do raciocínio automático, identificando os passos tomados pela dedução lógica e os fatos por ela empregados – por exemplo, apresentar a regra de que a hemodinâmica é aceitável *se* a frequência cardíaca é aceitável, a frequência do pulso é estável o suficiente, e a pressão sanguínea sistólica é aceitável etc. (ibid., p. 246).

Em 1985 surgiram as redes bayesianas, que são modelos gráficos probabilísticos (PEARL, 1988). Nelas, eventos são associados a uma probabilidade e conectados com direcionalidade a outros eventos. Por exemplo, seja "chuva" um evento com probabilidade *p1* e "grama molhada" um evento com probabilidade *p2* condicionada por *p1*. Essa proposição pode ser representada visualmente como um grafo direcionado que leva de "chuva" a "grama molhada". Isso, aliado ao fato dos eventos possuírem um nome inteligível (como já era o caso dos sistemas especialistas), pode ter contribuído para o problema de a explicação não ter assumido maior relevância na ocasião.

Entre 1990 e 2000, então com o surgimento dos sistemas de recomendação, novos tipos de inferência necessitavam de explicação. Um cenário típico é a recomendação de um filme a um usuário porque o filme foi bem avaliado por um outro usuário que vem a ser "amigo" do primeiro. Estudos como o de Herlocker et al. (2000), contemplando variados formatos de explicação para esse tipo de inferência, confirmaram que os usuários achavam de fato necessária a apresentação de uma explicação. Eles indicaram também que formatos mais simples e conclusivos de explicação – por exemplo, mostrar a nota (digamos, 4 de 5 estrelas) dada pelo outro usuário (amigo), bem como indicar uma propriedade marcante do filme, como a presença de um ator favorito – eram preferíveis a formatos de explicação baseados em conceitos de aprendizagem de máquina – como a estimativa de confiança do modelo preditivo. Esses estudos (cf. levantamento feito por BIRAN; COTTON, 2017) são informativos, pois sugerem direções acerca do tipo de explicação capaz de satisfazer um (a) usuário (a).

Explicação de aprendizagem profunda: o elemento novo

A expressão "aprendizagem profunda" foi cunhada na comunidade de aprendizagem de máquina por Dechter em (1986) e empregada pela primeira vez na comunidade de redes neurais artificiais por Aizenberg e outros no ano 2000. Em seguida ela se tornou especialmente popular no contexto das redes neurais profundas (DNNs, do inglês *deep neural networks*), que são talvez os modelos mais bem-sucedidos de aprendizagem de máquina até hoje. Apesar das DNNs já existirem há mais tempo

como parte de uma ampla classe de modelos considerados do tipo caixa preta, na última década houve uma retomada de interesse por tais abordagens. Isso foi motivado principalmente pelo aumento do poder computacional disponível e pela disponibilização de um grande conjunto de dados devidamente classificados por seres humanos, tais como a ImageNet (DENG et al., 2009). Desde então, DNNs têm sido aplicadas com sucesso em diversos cenários de uso, obtendo resultados comparáveis aos obtidos por seres humanos em tarefas como o reconhecimento de objetos em imagens (HE et al., 2015). Isso trouxe à tona novamente, talvez com mais vigor, o problema da explicação, visto que, em aplicações reais, seres humanos (incluindo projetistas, usuários etc.) desejam saber como e/ou por que certos resultados foram obtidos.

As DNNs são redes neurais que possuem múltiplas camadas intermediárias de neurônios – conectados de camada a camada, com um valor numérico associado como peso que é ajustado via treinamento para modular a propagação do sinal recebido –, e podem ser treinadas para realizar uma tarefa computacional (por exemplo, classificação de imagens de animais). Para viabilizar o processo de treinamento, é necessário um conjunto suficientemente grande de dados, no qual cada entrada (imagem) do conjunto deve estar previamente associada a um rótulo (por exemplo, “gato”), que representa a resposta que se esperaria da rede neural para a dada entrada. Esse processo de treinamento visa ajustar os pesos das conexões entre os neurônios da rede de tal forma que ela seja capaz de mapear uma nova entrada, para a qual não se conhece o rótulo, a uma resposta correta. Ou seja, considerando este exemplo, o conhecimento sobre o padrão que representa cada animal fica implicitamente representado no conjunto de pesos da rede neural, sem a necessidade de se informar explicitamente ao sistema a lógica e os conceitos subjacentes a este conhecimento.

Do ponto de vista do problema da explicação, considerando a tarefa de classificar uma imagem como contendo (sim ou não) uma ave, podemos ilustrar o contraste entre um sistema de IA baseado em regras lógicas (cf. citado em subtítulo anterior deste artigo) com um baseado em redes neurais (digamos, DNNs). No primeiro caso, o sistema de IA poderia explicar que classifica o animal como ave porque ele tem

penas, asas, bico, duas patas etc. A inferência lógica é de que, se ele tem todos esses atributos, então se trata de uma ave. No segundo caso, a classificação de um animal como ave é função da similaridade que suas características (traços distintivos, até de caráter geométrico) têm com as características extraídas (automaticamente, por operações de processamento de imagem) de um vasto conjunto de imagens de aves. O sistema que "aprendeu" os traços distintivos das aves por meio de exemplos tentará identificar essas mesmas características nas novas imagens que for solicitado a classificar. Há dois pontos centrais no caso das DNNs: (i) as características aprendidas não possuem necessariamente uma relação perceptiva compatível com algo que um ser humano seria capaz de discernir (nomear) nas aves; e (ii) essa aprendizagem é autônoma, sem que seja necessário o acompanhamento de um ser humano. Ou seja, temos aqui dois aspectos convenientes do ponto de vista tecnológico, que se traduzem em um desafio peculiar do ponto de vista do problema da explicação e dos impactos da IA na sociedade.

Iniciativas recentes de pesquisa no problema da explicação em IA

A retomada de interesse no problema da explicação motivou o surgimento de diversos fóruns especializados dentro da comunidade de pesquisa de IA. Começamos pelos esforços conceituais e de levantamento bibliográfico, dedicados a mapear o problema e estruturá-lo por meio de distinções.

Biran e Cotton (2017), por exemplo, assumem que explicabilidade é algo fortemente relacionado à noção de interpretabilidade: um sistema interpretável seria aquele cujas *operações* são compreensíveis para nós humanos, seja por meio da inspeção do sistema, seja por meio de alguma explicação produzida durante o seu funcionamento. Eles estabelecem uma distinção (ibid.) entre interpretabilidade e a noção de justificação, cujo objetivo seria explicar por que a decisão tomada pelo sistema pode ser aceita como uma boa decisão. Ou seja, justificabilidade e interpretabilidade seriam capacidades complementares. Doran et al. (2017), por sua vez, identifica três classes de sistema: opacos, intepretáveis e compreensíveis. Sistemas opacos são como caixas pretas, i.e., seus mecanismos não são inspecionáveis por usuários. Sistemas interpretáveis, por outro lado, permitiriam inspeção, estudo e

compreensão dos seus processos e mecanismos internos, mesmo que essas tarefas demandem certo conhecimento técnico especializado. Já os sistemas compreensíveis seriam aqueles que, além de oferecerem um resultado, também oferecem símbolos que são inteligíveis para os usuários, permitindo compreender por que uma certa saída está associada a uma certa entrada. Segundo Doran et al. (ibid.), portanto, compreensibilidade e interpretabilidade seriam capacidades complementares. Lipton (2016), finalmente, afirma que outros autores negligenciam que explicabilidade não é uma noção absoluta, mas contextual. Com tal perspectiva, ele busca identificar propriedades desejáveis para sistemas interpretáveis, com destaque para *transparência*, relacionada à inteligibilidade dos mecanismos internos do sistema; e *interpretabilidade post-hoc*, relacionada à capacidade do sistema de oferecer informações úteis sobre seus resultados, para usuários diversos.

A maior parte das iniciativas discutidas nos fóruns especializados, entretanto, propõe abordagens (métodos, técnicas ou ferramentas) computacionais específicas, para lidar com o problema da explicabilidade. Para um exemplo do tipo de iniciativa focada naquilo que Biran e Cotton (2017) e Doran et al. (2017) chamaram de compreensibilidade, Yosinski et al. (2015) propõem duas abordagens para auxiliar (através de visualização) na compreensão dos processos internos realizados por modelos de redes neurais convolucionais (um tipo de DNN adequado para classificar imagens). A primeira abordagem (Fig. 1, Parte A) gera uma imagem que é capaz de identificar e realçar quais pixels de uma dada imagem de entrada geram os maiores níveis de ativação para um dado neurônio que o usuário deseja inspecionar. A segunda abordagem (Fig. 1, Parte B) permite gerar imagens sintéticas semelhantes a imagens naturais que representam quais são os padrões visuais que um certo neurônio aprendeu a detectar. Ambas as abordagens oferecem *insights* a respeito do funcionamento interno da rede neural.

Já Dhurandhar et al. (2018) propõem uma abordagem cujo objetivo é oferecer uma justificação (BIRAN; COTTON, 2017) para o resultado obtido por uma rede neural (sem necessariamente oferecer compreensibilidade a seus mecanismos ou processos internos). Considerando o problema da classificação de imagens de acordo com a classe de objetos que ela representa, por exemplo, segundo a perspectiva desses

autores, a justificação é elaborada em termos de dois tipos de características: (a) características que ela possui (evidências positivas), que seriam típicas da classe em que ela foi classificada e que seriam suficientes para justificar a classificação; e (b) características que ela não possui (evidências negativas), que seriam típicas de uma classe muito semelhante à classe em que ela foi classificada e cuja ausência seja necessária para justificar a classificação obtida. Por exemplo, conforme pode ser visto na Fig. 1 (Parte C), na tarefa de classificação de imagens de algarismos numéricos, seja uma imagem representando o algarismo quatro (e classificada por uma DNN de modo correspondente), a abordagem seria capaz de produzir uma visualização da imagem original que realçaria, com uma certa cor, pixels que suportam a identificação do dígito como "4" e, com uma outra cor, pixels ausentes na imagem, mas que se presentes, levariam a rede neural a identificar, digamos, o algarismo "9". A abordagem identifica ambos os conjuntos de pixels por meio da resolução de um problema de otimização. É importante notar que esta abordagem pode ser acoplada como um componente externo a uma rede neural, cuja finalidade é oferecer justificação para os resultados obtidos por ela.

É importante ressaltar, conforme diagnosticado por Biran (2017), que a maior parte dessas abordagens computacionais para o problema da explicação assume (em geral tacitamente) alguma definição do problema, não aprofunda discussões conceituais acerca do que viria a ser uma explicação, e não consideram com alguma profundidade aspectos sociais envolvidos. Há uma lacuna de estudos que visam identificar quais são os significados (e para quem significam, entre projetistas, usuários etc.) que devem ser preservados ao longo das cadeias de computação realizadas pelos sistemas de IA, e como viabilizar essa preservação.

Na próxima seção, articularemos essas lacunas à luz de conceitos semióticos introduzidos por Peirce (1955).

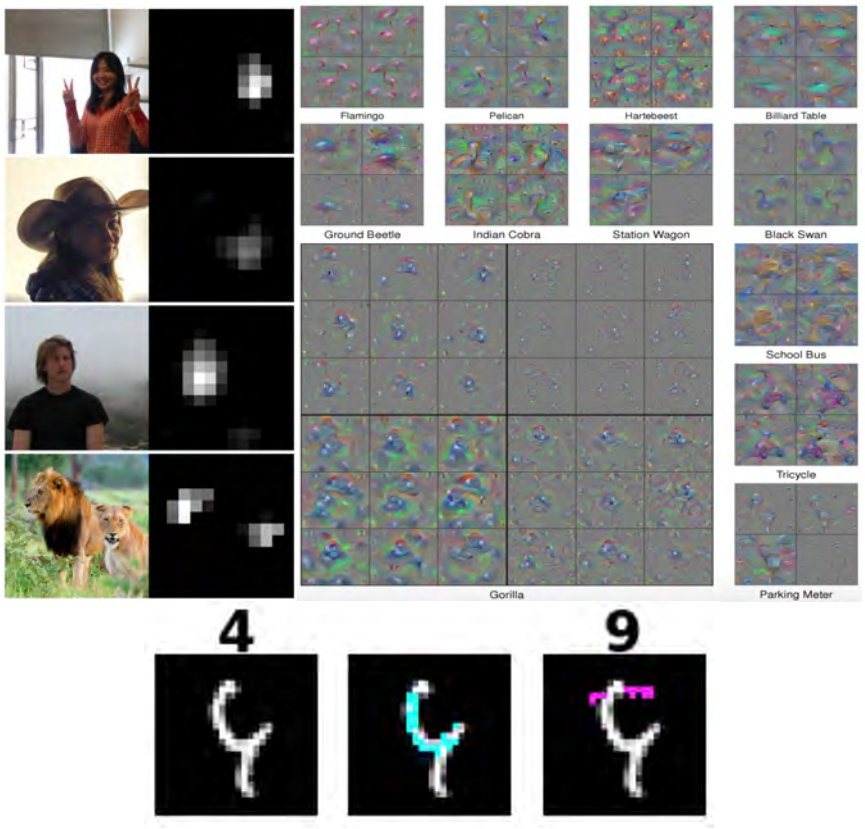

Figura 1. Parte A (superior, à esquerda): Representação do conjunto de pixels que gera um grande nível de ativação de um dado neurônio, para uma dada imagem de entrada. Quanto mais claro o pixel, maior é a ativação gerada no neurônio. Neste caso, é possível notar que o neurônio possui grandes níveis de ativação para faces, tanto humanas quanto de animais. **Fonte:** Yosinski et al. (2015). Reproduzido inalterado sob licença CC BY-NC-SA 3.0. **Parte B** (superior, à direita): Visualizações de imagens sintéticas geradas para representar o padrã o que cada um dos 12 neurônios da camada de saída (cada qual representando uma classe de objetos) detecta. Ao centro, são exibidas 4 visualizações para a classe Gorila, geradas por diferentes parametrizações do método, cada qual exibindo 9 imagens interpretáveis. Para as demais classes, são exibidas apenas 4 imagens interpretáveis selecionadas pelos autores. **Fonte:** Yosinski et al. (2015). Reproduzido inalterado sob licença CC BY-NC-SA 3.0. **Parte C** (inferior): À esquerda, vemos a imagem de entrada, juntamente com a classificação correspondente (algarismo 4) obtida por uma DNN. Ao centro, a visualização destaca em azul os pixels que fornecem evidência positiva à favor desta classificação. À direita, a visualização destaca em rosa os pixels que representam evidências negati vas contra uma classificação alternativa, mas visualmente muito próxima (algarismo 9). **Fonte:** Dhurandhar et al. (2018). Reproduzido com permissão do autor.

Lacunas da IA atual e da explicabilidade de seus sistemas

Uma peça de ficção escrita por Umberto Eco (1988) resume, com seu enredo alegórico, algo que parece estar fora do foco das iniciativas passadas e recentes de explicação em IA (e aprendizagem de máquina). O autor narra as peripécias de duas expedições interplanetárias de habitantes da Terra em um planeta fictício. Um grupo de expedicionários terráqueos toma e bebe – como se fosse água – um líquido extraterrestre que os habitantes do planeta chamaram de "água" na sua presença. Em

seguida, eles são vitimados por uma disenteria, desencadeando uma perigosa contenda diplomática. Para lidar com ela, uma segunda expedição de terráqueos foi enviada para examinar o comportamento dos nativos do planeta, buscando saber se o que eles chamavam de "água" era ou não era o que nós, terráqueos, assim chamamos.

A alegoria acima traz à tona, a distinção entre um símbolo (a palavra em si) e seu significado (seu sentido pragmático em contexto de uso). A alegoria oferece um cenário conveniente para ampliarmos a perspectiva acerca do problema da explicação em IA buscando estabelecer o limite das visões de explicação que revisitamos antes neste artigo.

Na continuação da narrativa de Eco (1988, p. 41), fica clara a dificuldade de comunicação entre os expedicionários humanos e os alienígenas, que costumam falar (e oferecer explicações) apenas em termos de seus estados neurais. Por exemplo, ao ver uma criança perto de um fogão quente, sua mãe extraterrestre assustada diria: "Ó meu Deus, ela vai estimular suas fibras-C!". Em outro exemplo, referindo-se ao que nós poderíamos caracterizar (ou "explicar") em termos como estes: "Aquilo pareceu um elefante, mas isso seria espantoso porque não se tem notícia de elefantes nesse continente; então me dei conta de que tem de ser um mastodonte" (ECO, 1988, p. 41).

Um nativo daquele planeta diria algo como: "Eu tive G-412, mas junto com F-11; e então eu tive S-147".

O último exemplo ilustra, ainda que de forma aproximada, o que está em jogo nas abordagens de explicação com base em regras lógicas (cf. anteriormente). É apresentada uma estrutura lógica instanciada por signos cujo sentido efetivo subjaz noutro nível semântico, correspondente a uma ontologia ou rede semântica específica e construída em certo contexto e (no caso de sistemas de IA) com um certo propósito. G-412 é o símbolo de uma noção conceitual de elefante, embora só possamos entender isto se articularmos dois (ou até mais) sistemas de significação. O mesmo acontece com a relação entre F-11 e a surpresa da constatação de que o lugar onde a percepção se dá é incompatível com alguma característica conceitual de G-412, e assim por diante. Para um computador, lembremos, nem G-412, nem "elefante", são *significativos* do ponto de vista conceitual. São somente cadeias simbólicas pertencentes a um vocabulário pré-definido sobre o qual se pode fazer alguma operação computacional programada.

Se consideramos as abordagens de explicação recentes voltadas para modelos caixa preta de aprendizagem de máquina, esse cenário fica ainda mais extremo. Em termos do exemplo de Eco, a assertiva análoga, nesse caso – desprovida de sua estrutura lógica dedutiva, e restrita a ocorrências correlacionadas –, seria o equivalente de "G-412, com F-11, com S-147".

Vemos então que, através da ficção, Eco (1988) sintetiza a essência de um debate que, apesar de muito antigo para a filosofia e estudos da linguagem ou da mente, volta a ser novamente central para a IA. Se por um lado, essa é uma discussão que tangencia questões das mais profundas, como a relação mente-corpo e o cartesianismo, ambas além do nosso escopo neste artigo, por outro lado, ela nos permite ressaltar – à luz da semiótica de Charles S. Peirce (1955) – dois elementos centrais disparados pela presença de um signo (re)conhecido: (i) o chamado processo de *semiose*, isto é, a construção de sentido entendida como indo além da fala ou da estrutura formal da linguagem, levando em conta aspectos pragmáticos (contextuais e de uso); e (ii) o raciocínio dito *abdutivo*, isto é, a inferência cuja justificação não se completa pela estrutura do argumento em si, como nos casos da indução e da dedução (na dimensão epistemológica), mas depende também da dimensão metodológica – seu papel e sua promessa para o avanço da investigação; por exemplo, se uma hipótese formada, digamos, por indução, é ou não é testável.

No caso dos expedicionários terráqueos que bebem o líquido extraterrestre porque os nativos o chamam de "água" (ibid., p. 41), é evidente a lacuna característica da semiose e do raciocínio abdutivo que se autocorrigem (SANTAELLA, 2004). A precipitação para uma conclusão na presença de evidências contingentemente consideradas suficientemente fortes, e ausência de contraditório, fez os expedicionários beberem "água" *daquele* planeta e terem disenteria. Agora, consideremos os sistemas de IA atuais, cujos modelos de comportamento são reconhecidamente desprovidos de correspondências claras com conceitos e regras lógicas que normalmente utilizamos numa explicação, mas têm entrado em contato com situações reais não antecipados – por exemplo, o caso mostrado pela jornalista (CADWALLDR, 2016). Não é difícil ver que a implantação desses sistemas em larga escala na sociedade, acompanhada das noções de explicação vista anteriormente neste

artigo, pode estar fadada a produzir um sem-número de frustrações, tais como as dos primeiros expedicionários da alegoria de Eco (1988).

Considerações sobre o potencial da semiótica de Peirce para a IA

Desde os sistemas de IA surgidos na última terça parte do século XX, a necessidade de explicação sempre foi evidente. O aspecto mais frequentemente mencionado é que a confiança nas decisões e avaliações produzidas por esses sistemas depende, como é frequente entre nós, seres humanos, de que o usuário – beneficiário ou de outra forma afetado pelo resultado da inferência automática – se *convença* de que ela procede e tem fundamento. Ora, mas de que depende convencermos alguém de que algo (seja o que for) procede e tem fundamento? Em geral, para qualquer agente (humano ou não) convencer uma pessoa de algo, é preciso que seja estabelecida uma relação entre as duas partes e que, em virtude dessa relação, sejam geradas e operacionalizadas (num processo de colaboração recíproca) inferências e expectativas que balizam um processo de comunicação. Idealmente, isso levará a uma persuasão e, como esperado, à confiança da pessoa no agente artificial. Ou seja, para se chegar propriamente à noção de explicação, há de se articular conceitos como o de relação (reciprocidade), comunicação e colaboração, em que inferências e expectativas desempenham um papel fundamental. É razoável considerar que estamos diante de questões pragmáticas muito claras.

Essa visão oferece um fio condutor e organizador de um espaço de pensamento que nos permite identificar, na prática de pesquisa ou de desenvolvimento de IA, determinadas rupturas que, a despeito da evolução das técnicas de raciocínio e aprendizado automáticos, não parecem ter sido vencidas. Uma questão central é a insistência em se manter *uma teoria do uso da IA divorciada de uma teoria da construção da IA*. Ora, os sistemas especialistas, dos anos 1980 e 1990, assim como os atuais sistemas autônomos que utilizam técnicas de aprendizagem profunda têm por destinação comum serem usados e serem úteis. A computação não é uma arte parnasiana, mas, sim, a base para processos de Engenharia de Sistemas e Tecnologias com que usuários humanos vão interagir por múltiplas razões e motivos, em múltiplas circunstâncias e, cada vez mais, em múltiplas modalidades e múltiplos dispositivos. A

rigor, a construção de sistemas de IA seria, por princípio, o estrito equivalente da construção de sistemas explicáveis, ou que podem ser explicados. No entanto, as divisões de competências, interesses, formações e práticas profissionais tanto na área científica quanto na grande área de desenvolvimento tecnológico reafirma o hiato que historicamente separou o estudo do *uso* de sistemas linguísticos (naturais ou artificiais) do estudo de sua estrutura ou sua *lógica*.

É notório que as novas técnicas de aprendizagem de máquina apresentam desafios ainda mais difíceis de explicabilidade e interpretabilidade do que enfrentaram os sistemas especialistas de duas ou três décadas passadas. Hoje, no entanto, ainda nos ressentimos da lacuna de uma infraestrutura teórica para unificar a modelagem, a engenharia e o uso de sistemas de IA. A semiótica peirceana permanece, no cenário contemporâneo, promissora. Recentemente, Nadin (2017) reiterou uma posição já anteriormente expressa em diversos textos, a saber, de que a IA não chegou ainda à realização e ao advento da "inteligência" artificial, embora tenha atingido resultados impressionantes na automação de tarefas associadas ao comportamento de seres naturais inteligentes. O ponto central do autor é a distinção entre *prever* (no âmbito do raciocínio sobre probabilidades e consequências lógicas) e *antecipar* (no âmbito dos significados possíveis, para agentes diversos). A seu ver, enquanto a computação supervaloriza a capacidade de prever o que vai acontecer, ela negligencia a capacidade de antecipar o que pode acontecer.

Entendemos que a questão do significado, ou da antecipação de oportunidades para além do raciocínio sobre probabilidades e consequências lógicas, nosso trabalho reunindo Semiótica e Computação pretende seguir uma outra linha, aderente à *Engenharia Semiótica* (DE SOUZA, 2005). A Engenharia Semiótica, em brevíssimas linhas, caracteriza a construção de sistemas computacionais como um processo de construção (Engenharia) Semiótica, centrado no conceito de *metacomunicação*. Partindo da premissa de que o destino de toda a computação é ser usada por pessoas em contextos pessoal e socialmente relevantes, os sistemas e tecnologias computacionais têm nas suas interfaces de usuário, o principal signo do que são. São elas que, por meio do desdobramento das interações que elas próprias facultam aos usuários, comunicam a esses usuários as formas, meios, efeitos, razões e possibilidades

de comunicação que os criadores (designers e engenheiros) de sistemas e tecnologias, eles sim, *anteciparam* ser relevantes, úteis, desejáveis, prazerosas e interessantes para os destinatários de seu trabalho. Sistemas e tecnologias, eles mesmos, conforme a visão de Nadin (2017), não fazem "antecipações de significado", apenas calculam os efeitos da presença ou ausência de padrões informacionais, seguindo instruções que direta ou indiretamente são estabelecidas por pessoas. Assim, a Engenharia Semiótica tem por objetivo restaurar o elo semiótico no processo de desenvolvimento de software, teorizando sobre a natureza metacomunicativa dos signos computacionais.

Para a IA contemporânea, a Engenharia Semiótica propõe, como um primeiro passo exploratório, uma investigação deste elo, buscando o rastro da significação humana sob as muitas camadas de sedimento computacional e informacional. Com um conjunto de métodos utilizados para a pesquisa sobre HCI (*Human–computer interaction*) e HCC (*Human-centered computing*) (SEBE, 2010), acreditamos ser possível sondar as antecipações de muitos seres humanos ligados à produção e ao consumo de informações (dados), programas (algoritmos) e sistemas (aplicações e tecnologias) usados em IA e, com isto, elaborar um enquadramento pragmático mais rico e consistente para o estudo de explicações e interpretações do comportamento de agentes artificiais inteligentes – seja para quem os utiliza (ou é afetado por ações e decisões de quem os utiliza), seja para quem os constrói (não importa em que ponto da potencialmente longa cadeia de produção de software que culmina nas tecnologias que hoje temos à nossa disposição).

Nosso trabalho, porém, é apenas um pequeno nicho de possibilidades que a semiótica, em particular a peircena, pode abrir para a IA. Voltando a um ponto anterior sobre a unificação teórica da pragmática com a lógica da descoberta e da antecipação, acrescentando agora os termos de Nadin (2017), a semiótica tem a possibilidade de apresentar-se como uma teoria integradora de que a computação aparentemente carece quando confrontada com a necessidade de produzir artefatos explicáveis e interpretáveis. Contudo, coloca-se em grande evidência e de imediato um grande desafio interdisciplinar. Para muitos pesquisadores, em muitas áreas de conhecimento e especialização que podem beneficiar-se significativamente de uma troca interdisciplinar, o discurso da semiótica como disciplina não é ameno, e por vezes não é

sequer compreensível. A efetiva fertilização do território da IA com as ideias de Peirce poderia ser levada adiante em larga escala como uma iniciativa interdisciplinar sustentada e sustentável que construa um território novo de interlocução científica em torno de projetos definidos e igualmente estimulantes para todas as disciplinas e subdisciplinas envolvidas.

De fato, a recente regulamentação europeia sobre o uso de dados e o direito à explicação (GOODMAN, 2016) tem o potencial de fomentar projetos interdisciplinares. Trata-se de um fato com repercussões sociais, legais e jurídicas, para não mencionar as econômicas e tecnológicas, que ainda mal vislumbramos. No entanto, já está claro que esta regulamentação, se não for ela mesma a causadora de grandes modificações no enquadramento social da IA – ou seja, do compromisso da computação com seus contextos de uso – pode ser o estopim de um movimento questionador profundo sobre os aspectos pragmáticos, éticos e filosóficos de teorias que até aqui se construíram sem pensar muito neles. Vemos, portanto, aí uma porta aberta para a circulação de ideias entre os domínios da computação e da semiótica, ancoradas a situações significativas e urgentes da sociedade contemporânea, cujas práticas já não conhecem uma fronteira clara entre o físico e o virtual, que, no entanto, se unificam sob o "signo" semiótico.

Enviado: 1 maio 2018
Aprovado: 29 maio 2018

Referências

AIZENBERG, I.; AIZENBERG, N.; VANDEWALLE, J. *Multi-valued and universal binary neurons*: theory, learning and applications. Dordrecht: Springer, 2000.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: a survey. *Proceedings of IJCAI-17*, Workshop on explainable AI (XAI), 2017.

CADWALLADR, C. Google, democracy and the truth about internet search. *The Guardian*, 4 dec 2016. Disponível em: <http://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook>. Acesso em: 26 maio, 2018.

CLANCEY W. J.; SHORTLIFFE, E. (Orgs.). *Readings in medical artificial intelligence*: the first decade. Reading, MA: Addison Wesley, 1984.

DE SOUZA, C. S. *The semiotic engineering of human-computer interaction*. Cambridge, MA: MIT Press, 2005.

DECHTER, R. Learning while searching in constraint-satisfaction problems. *Proceedings of the 5th National Conference on Artificial Intelligence*. Philadelphia, PA, August 11-15. Vol. 1, p. 178-183, 1986.

DENG, Jia; DONG, Wei; SOCHER, Richard; LI, Li-Jia; LI, Kai; FEI-FEI, Li. Imagenet: a large-scale hierarchical image database. In: *Proceedings of IEEE*, Conference on Computer Vision and Pattern Recognition. p. 248-255, 2009.

DHURANDHAR, Amit; CHEN, Pin-Yu; LUSS, Ronny; TU, Chun-Chen; TING, Paishun; SHANMUGAM, Karthikeyan; DAA, Payel. Explanations based on the missing: towards contrastive explanations with pertinent negatives, 2018. Disponível em: <http://arxiv.org/abs/1802.07623>. Acesso em: 26 mai. 2018.

DORAN, Derek; SCHULZ, Sarah; BESOLD, Tarek. What does explainable ai really mean? a new conceptualization of perspectives. *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, arXiv:1710.00794, 2017.

ECO, U. On truth: a fiction. In: ECO, U.; SANTAMBROGGIO, M.; VIOLI, P. (Orgs.). *Meaning and mental representations*. Bloomington, IN: Indiana University Press. p. 41-59, 1988.

GOODMAN, B; FLAXMAN, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, vol. 38, no. 3, 2016. Disponível em: <http://doi.org/10.1609/aimag.v38i3.2741>. Acesso em: 26 mai. 2018.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceeding ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. p. 1026-1034, 2015.

HERLOCKER, J.; KONSTAN, J.; RIEDL, J. Explaining collaborative filtering recommendations. In: *Proceedings of the Third Conference on Computer Supported Cooperative Work* (CSCW), p. 241-250, 2000.

HERN, A. 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft". *The Guardian*: International Edition, 29/06/2016. Disponível em: <http://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms>. Acesso em: 4 maio, 2018. Ver também sítio oficial do consórcio: <http://www.partnershiponai.org/>.

LIPTON, Zachary. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning*, 2016. Disponível em: <https://arxiv.org/pdf/1606.03490.pdf>. Acesso 17 junho, 2018.

NADIN, M. In folly ripe. In reason rotten: putting machine theology to rest, 2017. Disponível em: <https://arxiv.org/abs/1712.04306v1>. Acesso em: 4 maio, 2018.

PEARL, J. *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann, 1988.

PEIRCE, C. S. *Philosophical writings of Peirce*, Buchler, J. (Org.). New York, NY: Dover Publications, 1955.

SANTAELLA, L. *O método anticartesiano de C. S. Peirce*. São Paulo: Ed. Unesp, 2004.

SEBE, N. Human-centered computing. In: NAKASHIMA, Hideyuki; AGHAJAN, Hamid; AUGUSTO, Juan Carlos (Orgs.). *Handbook of ambient intelligence and smart environments*. Dordrecht: Springer, p. 349-370, 2010.

YOSINSKI, Jason; CLUNE, Jeff; NGUYEN, Anh; FUCHS, Thomas; LIPSON, Hod. Understanding neural networks through deep visualization. In: *Proceedings of the Deep Learning Workshop*, 31st International Conference on Machine Learning, Lille, France. 2015.

Uma cartografia comum aproximando Inteligência Artificial, Filosofia e Psicologia

Luciano Frontino de Medeiros[1]
Alvino Moser[2]
Marilene S. S. Garcia[3]

Resumo: O artigo relaciona quatro problemas epistemológicos a partir de uma perspectiva em que a Inteligência Artificial compartilha um domínio de conhecimento comum às áreas da Filosofia e da Psicologia: i) o problema clássico do *frame* ou quadro de referência, surgido a partir das pesquisas em IA sobre a limitação da representação em lógica de primeira ordem; ii) o problema de Hume, exposto por Daniel Dennett, abordando representações que raciocinam sobre representações; iii) na direção indicada por William Frawley, é apresentado o problema de Platão, expondo sobre a eficácia do conhecimento a partir de evidências escassas e fragmentadas do mundo; iv) o problema de Wittgenstein sobre a compatibilidade entre a linguagem determinística computada e a linguagem probabilística real. A ideia chave é mostrar que a Inteligência Artificial não é apenas engenharia de robôs ou sistemas inteligentes, mas também uma área que suscita questionamentos mais profundos, contendo objetos de interesse próximos ou mesmo comuns com os campos de atuação da Filosofia e da Psicologia.

Palavras-chave: Inteligência Artificial. Filosofia da Mente. Psicologia Cognitiva. Representações de conhecimento. Epistemologia da Inteligência Artificial.

Abstract: This paper presents a discussion of four philosophical problems from perspectives in which Artificial Intelligence, Philosophy and Psychology have a common interest. The issues are: the classical frame problem, which originated from AI research concerning the restriction of representations in first order logic; David Hume's view on representation and on reasoning about representations. In a line of argumentation presented by William Frawley, the paper focuses on two philosophical problems, (1) Plato's question of how world knowledge can grow from the mere fragements of which our cognition of facts is made up and (2) Wittgenstein's question conerning the compatibility between natural and computational languages. The main purpose of the paper is to show how certain fundamental philosophical premises may exclude the

[1] Doutor em Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina (2010). Professor Titular do Programa de Mestrado Profissional em Educação e Novas Tecnologias do Centro Universitário Internacional UNINTER. E-mail: luciano.me@uninter.com.
[2] Doutor em Filosofia e Ética pela Université Catholique de Louvain, Louvain-la-Neuve, Bélgica (1973). Professor Titular do Programa de Mestrado Profissional em Educação e Novas Tecnologias do Centro Universitário Internacional UNINTER. E-mail: alvino.m@uninter.com.
[3] Marilene S. S. Garcia concluiu pós-doutorado pela PUC-SP – TIDD. É autora dos livros: *Mobilidade tecnológica e planejamento didático* (2017); *Avaliação e validação de projetos* (2018), ambos pela editora Senac-SP. Possui doutorado pela USP e Mestrado pela UNICAMP, com estágio de pesquisa pelas Universidades de Freiburg e Oldenburg, na Alemanha. É professora do Mestrado Profissional em Educação e Novas Tecnologias da UNINTER- PR, em que ministra a disciplina de Aprendizagens ativas, Metodologias ativas e ensino híbrido. Coordena pesquisa em design de aplicativo para a inclusão de analfabetos funcionais. E-mail: marilenegarc@uol.com.br.

arguments of specialists from other domains of research, such as, philosophy, psychology and AI.

Keywords: Artificial Intelligence. Philosophy of Mind. Cognitive Psychology. Knowledge representations. Epistemology of AI.

Introdução

A Inteligência Artificial (IA) como área consolidada de conhecimento científico, tem alcançado nos dias atuais uma popularidade nunca experimentada em sua recente história. O desenvolvimento de tecnologias como o Aprendizado de Máquina (*machine learning*) e o advento do paradigma de Aprendizagem Profunda (*deep learning*) demonstram a franca vitalidade atual desse campo de pesquisa (GOODFELLOW; BENGIO; COURVILLE, 2016; HINTON; OSINDERO; TEH, 2006; MOHAMED; DAHL; HINTON, 2009). Alia-se a isso o sucesso e a proliferação de plataformas de automação de baixo custo, que tem auxiliado na disseminação sem par das tecnologias robóticas. Em decorrência, constata-se que as tecnologias consideradas "inteligentes" estão mais protagonistas no dia a dia da Sociedade do Conhecimento, seja para a melhoria da qualidade de vida alicerçada na dimensão construtiva do ser humano, seja para motivações de exploração de mercado (e no limite, bélicas) ou ainda permitir as alienações sociais excludentes.

Apesar de a Inteligência Artificial estabelecer como foco das suas investigações a criação de artefatos inteligentes (RUSSELL; NORVIG, 2004), seus estudos e desenvolvimentos também tiveram influência significativa (assim como foi influenciada) sobre as áreas disciplinares da Filosofia e da Psicologia. Por ter também o homem e a sua inteligência como objeto de estudo e ponto de partida, encontra-se nos seus fundamentos todo o edifício filosófico envolvendo as questões profundas e existenciais sobre o que é ser "humano", bem como o conhecimento relativo aos aspectos biológicos e psicológicos da mente humana. O filósofo Daniel Dennett explorou diversas possibilidades em um trabalho de quase três décadas atrás sobre essa cartografia comum em que transitam filósofos da mente, psicólogos experimentais e os engenheiros da Inteligência Artificial (DENNETT, 2006, p. 163-183). Dennett destacou a importância do estudo da IA por parte dos filósofos da mente, ressaltando a grande proximidade com o campo de estudo dos psicólogos.

Em outro ramo dessa árvore interdisciplinar, a Psicologia Cognitiva manifesta hoje uma série de aproximações com a Inteligência Artificial. Podem-se citar obras como Sternberg (2014) e Lefrançois (2013), que apresentam capítulos específicos sobre IA junto à discussão sobre os temas gerais da Psicologia Cognitiva e teorias de aprendizagem. Sternberg alça ainda as simulações de Inteligência Artificial ao status de componentes do método de pesquisa em Psicologia Cognitiva. "Os psicólogos cognitivos usam experimentos controlados, pesquisa psicobiológica, autoavaliações, observação naturalista e simulações por computador e IA para estudar os fenômenos cognitivos" (STERNBERG, 2014, p. 16-17).

Sternberg considera ainda a Psicologia Cognitiva, a IA, a Psicobiologia, a Filosofia, a Linguística e a Antropologia como campos que fornecem métodos e ideias para um campo mais geral e multidisciplinar, a Ciência Cognitiva (NICKERSON, 2005 apud STERNBERG, 2014).

Ao exposto dessas considerações preliminares, a proposta deste artigo é expor quatro problemas epistemológicos, ampliando as percepções consolidadas por Dennett com outras abordagens como as exploradas por Frawley com respeito às aproximações entre as ciências cognitivas e o sociointeracionismo (FRAWLEY, 2000). Esses problemas têm origem em uma área de conhecimento que se sobressai a partir de um campo comum e pode ensejar respostas em uma ou outra dessas áreas. Dennett destaca acertadamente o problema de Hume, originário da Filosofia, o qual tem uma resposta peculiar e única dada pela Inteligência Artificial; e o problema do *frame*, originário do campo da Inteligência Artificial, que tem impacto sobre as formas de representação, com base na lógica, das coisas na realidade. Por sua vez, Frawley discute também o problema de Platão, que pode ser abordado adequadamente pela Inteligência Artificial e por último o problema de Wittgenstein (assim denominado por Frawley), relativo ao compromisso da linguagem como forma computada e como uso. A ideia geral é reforçar a constatação do quanto essas áreas, a IA, a Filosofia e a Psicologia estão imbricadas entre si, na emergência de um campo de conhecimento interdisciplinar em que estudos particulares em uma área específica devem ser ampliados no contexto maior do mapa comum constituído por elas.

A categorização adotada aqui não tem a intenção de acondicionar o âmbito de toda a Filosofia pensada pelas personalidades que concedem a denominação, mas apenas para permitir uma proposta de estudos que possa pavimentar caminhos entre as diferentes áreas. Não se deseja racionalizar ou compactar toda a contribuição filosófica de um Platão ou de um Wittgenstein em poucas linhas ou conceitos, mas sim mostrar que certas questões, quando abordadas na pesquisa empírica da IA, já foram objeto de uma abordagem filosófica extensa, a qual pode fornecer a fundamentação teórica necessária para que essa pesquisa possa se desenvolver de forma ampliada. Como mencionado por Dennett, os pesquisadores de IA precisam estudar filosofia, "a menos que estejam contentes em reinventar a roda a cada poucos dias".

> Quando a IA reinventa a roda, ela é caracteristicamente quadrada, ou, no melhor dos casos, hexagonal, e pode fazer apenas algumas centenas de revoluções antes de parar. As rodas do filósofo, por sua vez, são círculos perfeitos, e não requerem , em princípio, nenhuma lubrificação; e podem ir pelo menos duas direções em duas direções ao mesmo tempo. Claramente, um encontro de mentes é necessário. (DENNETT, 2006, p. 183)

O problema de Platão

A concepção de sistemas inteligentes parte da premissa que estabelece a existência de processos internos ao agente, os quais têm inspiração na forma como os seres biológicos inteligentes pensam (ou como se imagina que pensam). A preocupação dos pesquisadores de IA se concentra nesse ponto em como dotar o agente inteligente de conhecimentos suficientes para que possa enfrentar a situação experimental de resolução de um problema. Portanto, há um lado interno, relativo ao agente, que precisa ser equipado com algum tipo de representação do mundo em que ele irá atuar e essa representação requer robustez para fazer frente a variações que um problema idealizado genericamente pode apresentar.

Platão apresentava a sua teoria das ideias como forma de dar um fundamento estável ao *logos*, para escapar da multiplicidade das interrelações contraditórias do sensível. A busca do inteligível é a busca pelo lugar da não contradição, enquanto que o sensível é o lugar da contradição, em que a identidade é una e múltipla. "O sensível tira sua pouca realidade da sua participação nessa realidade suprema que as Ideias representam" (ROGUE, 2011, p. 86-111).

Embasado na concepção desse dualismo ontológico de Platão, Frawley ressalta que possuímos um conhecimento rico e eficaz, ainda que as evidências e fatos apresentados a nós, a partir dos sentidos, sejam escassos, fragmentados e inconclusivos. Dessa forma, põe-se em uma interrogação o problema de Platão "de que forma sabemos tanto a partir de tão pouco? O que faz com que o lado interno de nossas mentes seja tão pleno e sistemático enquanto nossos lados externos sejam tão heterogêneos?" (FRAWLEY, 2000, p. 42).

Platão também dá uma resposta no âmbito de sua teoria das ideias: "Possuímos um conjunto de verdades em nossa razão que se origina da exposição à realidade das formas puras e ideais". Frawley comenta, entretanto, que respostas ao problema de Platão não precisam necessariamente ser da categoria de uma resposta de Platão. Por exemplo, a posição do behaviorismo radical contradiz o fato de que o que sabemos não é tão pouco, como posto no enunciado. Mas de maneira geral, nas ciências cognitivas, as respostas ao problema de Platão têm genuinamente uma resposta de Platão (FRAWLEY, 2000, p. 43).

As abordagens correlatas para essa resposta podem ser exploradas a partir dos dois principais paradigmas existentes dentro da Inteligência Artificial: o representacionalismo (ou computacionalismo, ou ainda simbolismo) e o conexionismo. O representacionalismo identifica a mente operando por meio de símbolos, enquanto que o conexionismo vê o cérebro, as redes neurais e seu potencial para associação de padrões (DINSMORE, 1992, p. 1; FRAWLEY, 2000, p. 73). A corrente do representacionalismo tem à frente alguns expoentes como Jerry Fodor, Noam Chomsky e Alan Newell. O conexionismo tem, dentre várias personalidades, pesquisadores como David Rumelhart, Frank Rosenblatt e Geoffrey Hinton.

O representacionalismo, ou computacionalismo, é a tese da metáfora do computador na essência do funcionamento da mente, tal como um sistema baseado em fatos e regras que produz inferências a partir da manipulação dessas estruturas, objetivando preservar significado e condições de verdade (DINSMORE, 1992, p. 2). Dessa forma, o pensamento e outros processos ligados à mente efetuam cálculos operacionais utilizando a sintaxe de representações que compõem a linguagem do pensamento. Na concepção de que a mente possui um código interno simbólico, ela

deve ser capaz de representar as hipóteses sobre o mundo e permitir que a sua validade seja verificada. O código interno é uma espécie de linguagem proposicional do pensamento (FRAWLEY, 2000, p. 74).

Uma linguagem proposicional compõe-se de fatos e regras. Fatos são representações tidas como verdadeiras sobre um dado contexto sendo representado e tem característica estática. As regras, por sua vez, permitem que se derivem novos fatos a partir de fatos existentes, agregando um aspecto ampliador e dinâmico ao raciocínio proposicional. Esses sistemas também são conhecidos como sistemas de produção do tipo "se... então", que englobam premissas e conclusões (NEWELL; SIMON, 1972 apud STERNBERG, 2014, p. 285).

Dessa forma, essa é uma primeira resposta ao problema de Platão: a partir de um sistema rico de verdades necessárias e suficientes, o ser humano é capaz de produzir novos conhecimentos na exposição e validação das formas presentes em um código interno da mente ao mundo. A partir do conhecimento proveniente de fatos estáticos, as regras permitem a ampliação dos limites do nosso próprio conhecimento.

> Nossas mentes vêm ao mundo equipadas com um sistema rico e predeterminado de verdades necessárias: formas atemporais, idealizadas, imutáveis, absolutas e autoevidentes levadas à consciência pela introspecção, não desenvolvidas através de imitação ou de outro tipo de código do mundo externo. (FRAWLEY, 2000, p. 43)

Entretanto, a corrente do conexionismo também permite obter uma resposta ao problema de Platão. O paradigma conexionista tem como pilar a concepção de que o cérebro consiste de um número massivo de elementos interconectados, os quais enviam sinais excitatórios ou inibitórios entre si, atualizando suas excitações a partir de mensagens simples (McCLELLAND; RUMELHART; HINTON, 1986, p. 10). Com base nessa premissa, redes neurais artificiais podem ser construídas, criando-se uma arquitetura de unidades (neurônios) interconectadas (por sinapses) que podem aprender tarefas como reconhecimento de padrões a partir de algoritmos que utilizam processos de tentativa e erro (CHURCHLAND, 2004, p. 244-252).

Um dos tipos clássicos que aparece no estudo de aprendizado em redes neurais artificiais é o perceptron, criado pelo psicólogo Frank Rosenblatt (RUSSELL; NORVIG, 2004, p. 732). O perceptron possui unidades sensoriais, associativas e geradoras de

respostas. As sinapses ou pesos são descritos em termos de valores que irão ponderar os sinais de entrada, permitindo a ativação de neurônios específicos conforme a atividade de aprendizado desejada. Uma rede neural artificial aprende a partir de um conjunto de amostras representativas do contexto do problema de classificação ou reconhecimento de padrões. Após o algoritmo ser executado várias vezes, a rede neural pode alcançar um nível de aprendizado tal que faça inferências de forma muito precisa sobre as amostras oferecidas no treinamento. "O padrão de conexões determina o que o sistema conhece e como vai responder" (LEFRANÇOIS, 2013, p. 290).

Entretanto, uma das características que fazem as redes neurais artificiais um dos tipos de IA mais utilizados para semelhantes tarefas é a sua capacidade de generalização (HAYKIN, 2001). Caso sejam apresentadas amostras que não fizeram parte do conjunto de treinamento, a rede neural artificial pode inferir de maneira aproximada, ou seja, fornecer uma resposta de tal forma que seja uma interpolação das amostras em questão, em comparação com o conjunto de amostras do conjunto de treinamento.

Portanto, o conexionismo também fornece uma resposta ao problema de Platão, pois as redes neurais podem armazenar em sua estrutura simples um conhecimento que pode extrapolar e generalizar para além do aprendizado retido a partir de treinamento prévio, permitindo a ampliação do conhecimento com o qual a rede pode lidar. Portanto, o código interno, que permite a extrapolação a partir dos seus limites, seja no caso das regras para o representacionalismo, ou a capacidade de generalização no caso do conexionismo, tem condições de preencher a demanda do problema epistemológico de Platão, cada uma com suas peculiaridades e distinções.

> Para a ciência cognitiva, o código interno é coincidente com a mente. Em uma explicação representacionalista ele é explícito, local, inflexível e dedicado; em uma explicação conexionista ele é emergente, distribuído, flexível e ajustável. (FRAWLEY, 2000, p. 78)

Ainda com relação à aprendizagem, Frawley aponta que os cientistas cognitivos entram em concordância quanto à existência de um código interno para acomodar a aprendizagem. Entretanto, enquanto os representacionalistas a minimizam, os conexionistas a maximizam (FRAWLEY, 2000, p. 77).

O problema de Hume

Dennett (2006) foi o primeiro a formalizar a possibilidade, por meio da abordagem da Inteligência Artificial, de resolução do que foi denominado por ele de "problema de Hume" que, por mais de duzentos anos, foi alvo de estudo por filósofos e psicólogos. Se a Inteligência Artificial assume o papel de uma investigação mais abstrata da possibilidade do conhecimento, as representações de conhecimento, que sempre estiveram na ordem do dia das pesquisas da IA, abriram uma trilha na direção de uma Psicologia Cognitiva bem sucedida na explicação dos fenômenos complexos da mente. Na abertura da "caixa preta", preterida pelos behavioristas radicais por explicações comportamentais enraizadas no ambiente, as representações internas da mente foram bem respaldadas nas metáforas construídas para o raciocínio mecânico feito por sistemas inteligentes.

> A única psicologia que poderia possivelmente ser bem sucedida na explicação das complexidades da atividade humana deve postular representações internas. [...] Para os empiristas britânicos, as representações internas eram chamadas ideias, sensações, impressões; mais recentemente, os psicólogos têm falado de hipóteses, mapas, esquemas, imagens, proposições, engramas, sinais neuronais e mesmo hologramas e teorias inatas inteiras. (DENNETT, 2006, p. 178)

Entretanto, se uma representação qualquer existe é porque houve uma intencionalidade prévia de alguém para a atividade de elaborar essa representação. É necessário também que exista um sujeito que desempenhe o papel de interpretar a representação, de dar significado a ela.

> Nada é intrinsicamente uma representação de nada; algo é uma representação apenas para alguém; qualquer representação ou sistema de representações assim requer um usuário ou interpretador da representação, que é externo a ela. Qualquer interpretador como tal deve possuir vários traços psicológicos ou intencionais; ele deve ser capaz de várias compreensões, e deve possuir crenças e objetivos [...]. Tal interpretador é um tipo de homúnculo. (DENNETT, 2006, p. 178)

Hume fundamenta a sua teoria das ideias abstratas, originárias das ideias concretas e das impressões do mundo exterior sobre os sentidos, estabelecendo a possibilidade de conexões associativas entre elas (MORRIS; BROWN, 2017, p. 14). No entanto, sua teoria das ideias não converge para qualquer *locus* de significação. Na

esteira dessa proposição, filósofos como Kant e Husserl apresentaram posições contrárias à teoria das ideias de Hume. Kant, no âmbito do idealismo transcendental, criticou Hume por pressupor que conceitos provenientes dos sentidos prevalecessem sobre o raciocínio causal *a priori*; a generalização não poderia chegar jamais por meio dos sentidos e também não poderia estar contida ao mesmo tempo na forma pura da intuição sensível, pois ela é um ato do poder de representação, e o "eu penso", a presunção de um sujeito, tem que estar por trás de todas as representações (KANT, 2015, p. 127-129). E, na gênese da sua fenomenologia, Husserl investe nos modos de consciência e nas vivências intencionais para preencher o vazio deixado pelas conexões associativas de Hume que parecem manifestar, de certa forma, uma espécie de intencionalidade (HUSSERL, 2012, p. 157).

> As representações internas de Hume eram impressões e ideias, e ele sabiamente se esquivou da noção de um *eu* interno que manipularia de maneira inteligente esses elementos; mas isso o deixou na necessidade de fazer as ideias e impressões "pensarem por si mesmas". O resultado foi sua teoria do "eu" como um "feixe" [...] de impressões e ideias. Ele tentou colocar essas impressões e ideias em interação dinâmica, postulando diversas ligações associacionistas, de forma que cada ideia sucessiva na corrente de consciência arrastava sua sucessora para o palco de acordo com um ou outro princípio, todos sem o benefício de uma supervisão inteligente. (DENNETT, 2006, p. 179)

Contudo, a IA respondeu ao problema de Hume a partir da existência de estruturas de dados que integram os sistemas inteligentes, as quais têm condições de serem pensadas como representações que se entendem por si mesmas e prescindindo, portanto, de um núcleo significador. Um sistema concebido em alto nível pode ser decomposto sucessivamente em agentes cada vez menores, que irão desempenhar a tarefa em questão executando apenas uma pequena parte do todo; o sistema seria uma organização de diversos pequenos agentes que assumem papeis variados para a emergência da tarefa inteligente em alto nível.

> Se pudermos ter uma equipe ou comitê de homúnculos relativamente ignorantes, de mente estreita, cegos, para produzir o comportamento inteligente do todo, isso é um progresso. Um fluxograma é tipicamente o esquema organizacional de um comitê de homúnculos [...]; cada quadro especifica um homúnculo ao prescrever uma função sem dizer como ela é realizada. (DENNETT, 2006, p. 180)

Entretanto, mesmo com a decomposição sucessiva, Dennett complementa que não será possível atingir uma representação que se autocompreenda. No limite, todos os homúnculos são eliminados e há um sistema que manifesta comportamento inteligente de alto nível, composto por subsistemas que executam em partes mínimas a tarefa global (DENNETT, 2006, p. 181).

A moderna teoria da IA tem por base a construção de agentes inteligentes. Na esteira da noção de homúnculo dada por Dennett, um agente possui uma intencionalidade e uma racionalidade, percebendo o ambiente, tomando decisões conforme a sequência das percepções e agindo sobre ele, atualizando seus estados internos, as representações que permitem a execução de alguma tarefa inteligente.

> Para cada sequência de percepções possível, um agente racional deve selecionar uma ação que se espera venha (sic) a maximizar sua medida de desempenho, dada a evidência fornecida pela sequência de percepções e por qualquer conhecimento interno do agente. (RUSSELL; NORVIG, 2004, p. 36)

É interessante ressaltar, sobre a resposta da IA na perspectiva de Dennett, dada ao problema de Hume que, ainda que considerada para sistemas muito limitados, é até o momento o único modo conhecido de resolver o problema (DENNETT, 2006, p. 182).

O problema do *Frame*

Originado de uma situação primeiramente posta por McCarthy e Hayes (1969), no âmbito da Inteligência Artificial, o problema do *frame*[4] (também conhecido como problema do quadro, ou ainda o problema da estrutura)[5] é a tarefa desafiadora de representar os efeitos de uma ação em lógica sem ser necessária a representação explícita de uma enorme número de não efeitos óbvios.[6] Seria possível a limitação do escopo do raciocínio que é requerido para derivar as consequências de uma ação em determinado contexto? De maneira geral, o que faz o ser humano ter uma habilidade

[4] Termo inspirado na Física Mecânica, significando o quadro inercial de referência, necessário para se fazer a análise de movimentos de objetos.
[5] Dennett sugere ainda que o problema do *frame* fosse denominado também de problema de Kant (DENNETT, 2006, p. 183).
[6] Se um agente inteligente precisa raciocinar, por exemplo, sobre um mundo de blocos para dizer o que é um cubo, uma pirâmide ou uma esfera, ele precisa dizer também que um cubo não é uma pirâmide e uma esfera, uma pirâmide não é um cubo ou uma esfera, e uma esfera não é um cubo ou uma pirâmide. Esse conhecimento precisa ser representado de forma explícita. Pequenos conjuntos são passíveis de uma representação adequada, mas, como se pode intuir, à medida em que o número de elementos cresce, as combinações possíveis crescem exponencialmente. Ao abordar problemas reais, não se tem memória (ou tempo) suficiente para acondicionar todas as possibilidades de representação.

peculiar para a tomada de decisões, baseando-se no que é relevante em uma situação em andamento, sem ter que explicitamente considerar tudo o que não é relevante (SHANAHAN, 2016, p. 1)?

Russel e Norvig (2004, p. 320-321) afirmam que o problema do *frame* é a representação de todas as coisas que permanecem iguais, em uma situação em que se encontra um agente que tem a possibilidade de executar determinadas ações a partir de um mecanismo de raciocínio, baseado em lógica de primeira ordem e contendo regras (axiomas) explícitas de efeito, alterando certos elementos e deixando outros inalterados. Portanto, uma grande parte da dinâmica qualitativa que é necessária para o planejamento de uma situação consiste na inferência de elementos que não irão se modificar. O problema do *frame*, de forma sumária, é o problema em como formalizar o raciocínio inercial necessário (THOMASON, 2016, p. 36).

O problema do *frame* puramente lógico pode ser resolvido utilizando-se a lógica monotônica, simplesmente escrevendo axiomas explícitos que dizem o que não pode ser modificado enquanto uma ação é executada por um agente. Entretanto, soluções baseadas em lógica não-monotônica têm sido amplamente pesquisadas (THOMASON, 2016, p. 37). Em suma, o desafio está em encontrar uma forma de capturar os não-efeitos de ações de uma forma sucinta em lógica formal. A premissa fundamental é denominada de *lei de inércia do senso comum* (SHANAHAN, 2016, p. 3).

Dennett extrapola o problema do *frame* para o patamar de um problema epistemológico abstrato, descoberto no escopo da IA. Se um agente cognitivo, que possui diversas crenças sobre o mundo que o rodeia, realiza uma ação, o próprio mundo em que o agente existe se modifica e, dessa forma, as crenças desse agente precisariam ser atualizadas ou mesmo revistas.

> Se supõe, como fizeram tradicionalmente os filósofos, que as crenças de alguém são um conjunto de proposições, e que o raciocínio é uma inferência ou dedução a partir de membros desse conjunto, fica-se em dificuldades, pois é inteiramente claro (embora ainda controvertido) que os sistemas que se baseiam apenas em tais processos fiquem atolados por explosões combinatórias em seu esforço de atualização. (DENNETT, 2006, p. 182-183)

Dennett comenta ainda sobre a inegável capacidade dos seres humanos de manter as suas crenças de maneira aproximada em consonância com a realidade em que vivem.

Frawley (2000, p. 24) apresenta o problema do *frame* como um enigma relativo à afirmação clara e integral das condições que restringem de forma global as decisões tomadas por um agente inteligente. E menciona que a psicologia de Vygotsky pode ter um papel importante para mostrar não apenas uma solução, mas um método para abordar o problema do *frame*. Frawley apresenta a teoria sociointeracionista de Vygotsky e Luria não como uma explicação do que é "externo" , mas sim como uma teoria da internalização do "externo". O problema do *frame* é mais uma questão de como o raciocínio lida com a adição de informações que influencia a situação de comprovação das conclusões, sendo capaz de inibir ou descartar uma diversidade de opções inferenciais.

A teoria sociointeracionista propõe uma linguagem para o pensamento, uma fala privada (VYGOTSKY, 1978, p. 24) que é acionada na resolução de problemas e que "codifica a internalização individual única da significação social e cultural externa" (FRAWLEY, 2000, p. 38). A estruturação dessa linguagem de inferências do pensamento de onde são derivadas as soluções possui uma função inibidora que descarta alternativas e dá uma orientação ao pensamento representacional.

> Para os vygotskianos, um dos métodos padrão de experimento é interromper o comportamento introduzindo uma tarefa difícil e, então, observar como o sujeito recupera o controle cognitivo. No endireitar do self, o sujeito utiliza a fala privada como uma orientação simbólica externa e o que emerge são pistas linguísticas para a luta do indivíduo com a estrutura (frame) cognitivo. (FRAWLEY, 2000, p. 38)

Sternberg (2014, p. 67) expõe a comparação do ser humano com os robôs. Ainda que o ser humano possua uma visão limitada e que esteja sujeito a ilusões, possui uma capacidade extraordinária superior a dos robôs para codificar as representações visuais e dar significado a elas.

> Dada a sofisticação dos robôs de hoje em dia, qual é a razão da superioridade humana? Possivelmente várias, mas o conhecimento é uma delas. O ser humano simplesmente conhece muito mais acerca do ambiente e das fontes de regularidade no ambiente que os robôs. O conhecimento humano é uma

> grande vantagem que os robôs – ao menos os atuais – não são capazes de suplantar. (STERNBERG, 2014, p. 67-68)

Sternberg, na abordagem dos processos da atenção, destaca também o limite da mente humana quanto ao volume de informação na qual pode se concentrar. Os fenômenos psicológicos da atenção possibilitam o uso de recursos mentais que são limitados de forma sensata.

> Ao diminuir a atenção sobre muitos estímulos externos (sensações) e internos (lembranças), podemos focar os estímulos que mais nos interessam. Esse foco acentuado aumenta a probabilidade de resposta rápida e precisa aos estímulos que interessam. (STERNBERG, 2014, p. 108)

Portanto, parece que o ser humano é naturalmente e biologicamente muito bem equipado em relação à abordagem aproximada do problema do *frame*. Enquanto que nos agentes da IA é necessária uma grande quantidade de conhecimento para acomodar inclusive os não-efeitos, o ser humano parece suplantar o problema do *frame* de uma maneira muito natural e eficaz. A corrente representacionalista da mente também se defronta com essa dificuldade para explicar a mente operando de acordo com um código simbólico, lidando de forma natural com o problema do *frame*.

O problema de Wittgenstein

Como último problema epistemológico a ser abordado, Frawley (2000, p. 55) apresenta o problema de Wittgenstein a partir do questionamento sobre como o ser humano consegue coordenar a experiência idealizada do mundo interior, virtual e a experiência vivenciada sujeita às circunstâncias contingentes. Concernente aos aspectos da linguagem, esse problema se posta de forma importante para a ciência cognitiva para que se alcance um "compatibilismo" entre os padrões legítimos e determinísticos presentes na mente e a visão das pessoas como agentes com livre arbítrio e liberdade de escolha (FRAWLEY, 2000, p. 55).

É necessário então invocar a unidade dialética do filósofo Ludwig Wittgenstein autor de duas obras emblemáticas que influenciaram o estilo de pensar contemporâneo: O *Tractatus Logico-Philosophicus* (publicado na sua primeira edição em 1921), assinalado como o "primeiro Wittgenstein", e *Investigações Filosóficas* (lançado

em 1953), relativo ao "segundo Wittgenstein". O primeiro Wittgenstein segue a corrente da lógica adotando um estilo de demonstração e buscando a forma e o significado das proposições lógicas associadas à linguagem sob as condições de verdade determinada. Já o segundo Wittgenstein é um estudioso da história natural que aborda, seguindo um estilo discursivo, a respeito dos usos de declarações sob condições indeterminadas em relação à verdade e ao significado.

> Wittgenstein sempre trabalhou na fronteira entre a determinação e a indeterminação. A partir de um único ponto de vista teórico, ele explorou os dois lados, concentrando-se, no início, na lógica e, posteriormente, no uso. Por esse motivo, alguns atribuíram uma unidade dialética a Wittgenstein. (FRAWLEY, 2000, p. 55)

O primeiro Wittgenstein se insere na discussão a partir dos trabalhos sobre o logicismo de Frege, cuja influência se constata de forma significativa no decorrer do *Tractatus*. Apresenta uma solução para a antinomia de Zermelo-Russell, desconsiderando a teoria de tipos que Bertrand Russell havia proposto para a abordagem ao problema (BUCHHOLZ, 2009; WITTGENSTEIN, 2010, p. 159). Visando confrontar na época a abordagem do psicologismo na fundamentação da teoria do conhecimento, a elucidação da inteligência humana não deveria acontecer por meio de investigações psicológicas, mas sim por uma teoria do significado proposicional. Essa teoria procederia adequadamente apenas se fosse identificado o significado linguístico a partir do emprego lógico dos sinais linguísticos (BUCHHOLZ, 2009, p. 41).

Em *Investigações Filosóficas*, temos um segundo Wittgenstein preocupado com jogos de linguagem, envolvendo a totalidade em que consiste a linguagem e as atividades que as englobam (WITTGENSTEIN, 2012, p. 19). De acordo com Fann (2013), Wittgenstein se dá conta de que as doutrinas do *Tractatus* se baseavam em uma imagem particular da essência da linguagem humana. Seria a teoria do significado-correspondência, cuja essência eram as palavras individuais que nomeavam objetos, o objeto que representava a palavra era seu significado. "Wittgenstein ahora se da cuenta de que su anterior parecer sobre las proposiciones no era el resultado de una investigación, era un requisito […] Su concepción del lenguaje requería que toda proposición tuviera un sentido definido." (FANN, 2013, p. 78-79).

Portanto, conforme o segundo Wittgenstein, podemos chegar apenas a uma análise correta por meio do que se poderia chamar de investigação lógica dos próprios fenômenos, ou seja, em certo sentido *a posteriori*, e não por meio de conjecturas sobre as possibilidades *a priori* (FANN, 2013, p. 62).

Um Wittgenstein único apresentaria "uma tensão perpétua entre a representação e a atuação, entre o tempo formalmente computável e o decorrido no mundo real" (FRAWLEY, 2000, p. 56). Portanto, o problema de Wittgenstein se resume em "como a linguagem se enquadra na relação entre máquinas (fatos determinados em forma computável) e pessoas (escolha de valores sob regras indeterminadas)" (FRAWLEY, 2000, p. 57).

O problema de Wittgenstein assume ascendência direta sobre as teorias representacionalistas. Se a mente humana opera segundo uma linguagem da mente, operando sobre cálculos operacionais, chega-se a uma condição de limite sobre como se dá a relação de compromisso entre representações determinísticas operando sobre a sintaxe formal da mente com o que é indeterminado e ocorrido na realidade. Da perspectiva da IA, esse é um problema fundamental: como um robô, programado com um sistema de regras determinísticas, irá lidar com as situações indeterminadas provenientes da interação com o ambiente externo? Como deverá ser elaborado (ou se é possível de elaborar) um sistema de aprendizado de máquina que torne o robô apto a compatibilizar o problema de Wittgenstein, assim como um ser humano habilmente o faz?

Frawley sugere que o tratamento do problema de Wittgenstein deva consistir em duas soluções: para o primeiro Wittgenstein, a abordagem é muito semelhante com aquela feita com relação ao problema e a resposta de Platão, com a Psicologia Cognitiva internalista subsidiando por meio de um conjunto de resultados tanto teóricos quanto empíricos. Já, para o segundo Wittgenstein, é necessário que se aborde o problema por meio de uma psicologia social da ação, a partir do programa de pesquisa de Vygotsky, Luria e Leontiev sobre a mente inserida no contexto sociocultural (FRAWLEY, 2000, p. 61).

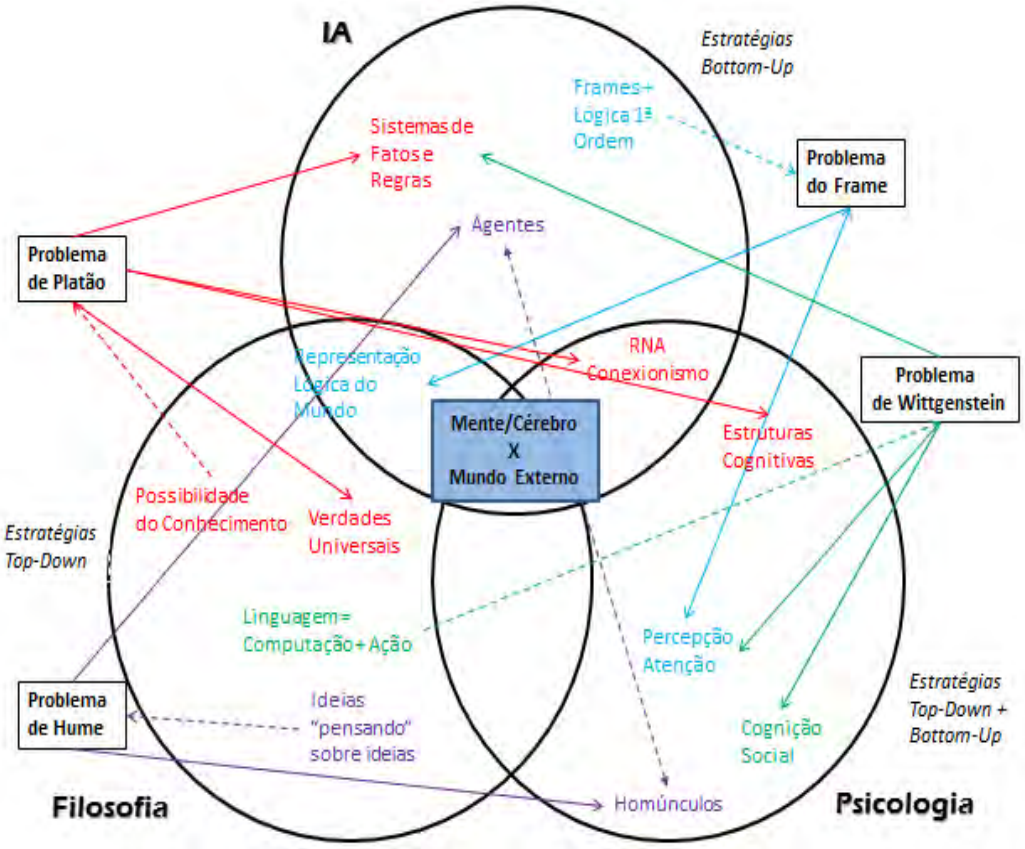

Figura 1. Uma cartografia sintetizando os quatro problemas sobre as áreas da Filosofia, Psicologia e Inteligência Artificial.

Considerações finais

A explanação dos quatro problemas filosóficos apresentados em detalhe teve como intuito demonstrar algumas das aproximações existentes entre as áreas do conhecimento da Filosofia e da Psicologia com a Inteligência Artificial e mesmo a existência de uma área interdisciplinar, uma cartografia comum indicando um domínio em que as áreas manifestam a convergência dos objetos de estudo relacionados à fenomenologia da inteligência. A figura 1 é uma tentativa de sintetizar e ligar os temas desenvolvidos para cada problema em uma representação comum. No centro, a ideia de que as três áreas exploram as questões relacionadas aos fenômenos do sistema cérebro/mente e suas relações com o mundo externo.

É natural pensar as pesquisas em IA, sobre a reprodução de raciocínio automatizado ou comportamentos inteligentes, como algo embasado fortemente na tecnologia que torna possível a construção desses artefatos, sem compromisso com

uma abordagem mais reflexiva sobre aspectos do funcionamento da mente. Porém, a IA pode ser pensada como uma área de pesquisa que emerge a partir da atividade de escrutínio incessante da mente humana, iniciada na Grécia Antiga e que perpassa por séculos de história, quando inserida no tempo presente com a manifestação da alta tecnologia. Neste século XXI, a humanidade é capaz de reproduzir de forma artificial os processos naturais da mente humana, exteriorizando-se a inteligência. Talvez esse seja o aspecto mais peculiar da IA. Enquanto que na Filosofia e na Psicologia se busca entender os aspectos da mente inteligente enquanto manifestações internas e encerradas nos limites do ser, a IA, por sua vez, as externaliza.

A confluência entre as áreas da Filosofia, da Psicologia e da IA não se esgota com esses quatro problemas epistemológicos. Há diversos temas que ensejam aproximações como, por exemplo, o problema mente-cérebro, o teste de Turing, o argumento da sala chinesa de Searle e também os aspectos éticos que estão surgindo em função da inserção da IA de forma intensiva na sociedade. Porém, é preponderante afirmar que na investigação ou aquisição de conhecimentos em uma área específica, o necessário aporte às outras áreas, na condição de correlatas, não pode deixar de ser considerado. É possível filosofar com IA ou entender processos psicológicos complexos com IA, assim como é legítimo tomar por empréstimo acadêmico os conhecimentos e os avanços da IA para se possa fazer Filosofia ou Psicologia.

Enviado: 12 março 2018
Aprovado: 15 abril 2018

Referências

BUCHHOLZ, K. *Compreender Wittgenstein.* 2a. ed. Petrópolis, RJ: Vozes, 2009.

CHURCHLAND, P. M. *Matéria e consciência*. São Paulo: Editora UNESP, 2004.

DENNETT, D. C. *Brainstorms*: Escritos filosóficos sobre a Mente e a Psicologia. São Paulo: UNESP, 2006.

DINSMORE, J. Thunder in the gap. In: DINSMORE, J. (Org.). *The symbolic and connectionist paradigms*: closing the gap. Hillsdale, NJ: Lawrence Erlbaum, p. 1-23, 1992.

FANN, K. T. *El concepto de filosofía en Wittgenstein*. 3ª. ed. Madrid: Tecnos, 2013.

FRAWLEY, W. *Vygotsky e a ciência cognitivas*: linguagem e interação das mentes social e computacional. Porto Alegre: Artes Médias Sul, 2000.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. Cambridge, MA: MIT Press, 2016.

HAYKIN, S. *Redes neurais*: princípios e prática. Porto Alegre: Bookman, 2001.

HINTON, G. E.; OSINDERO, S.; TEH, Y. W. A fast learning algorithm for deep belief nets. In: *Neural computation*, v. 18, n. 7, p. 1527-1554, 2006.

HUSSERL, E. *Investigações lógicas:* investigações para a fenomenologia e a teoria do conhecimento. Rio de Janeiro: Forense, 2012.

KANT, I. *Crítica da razão pura*. 4a. ed. Petrópolis, RJ: Vozes, 2015.

LEFRANÇOIS, G. R. *Teorias da aprendizagem*. São Paulo: Cengage Learning, 2013.

McCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. In: GINSBERG, Matthew (Org.). *Readings in nonmonotonic reasoning*. San Francisco, CA: Kaufmann, p. 26-45, 1987.

McCLELLAND, J. L.; RUMELHART, D.; HINTON, G. E.; PARALLEL DISTRIBUTED GROUP. *Parallel distributed processing*: exploration of the microstructure of cognition. Cambridge, MA: MIT Press, 1986. p. 3-44.

MOHAMED, A.; SAINATH, T.N.; DAHL, G.; RAMABHADRAN, B.; HINTON, G.; PICHENY, M.A. Deep belief networks for phone recognition. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, p. 5060-5063, 2011.

MORRIS, W. E.; BROWN, C. R. David Hume. In: ZALTA, E. N. (Org.). *Stanford encyclopedia of philosophy*. 2017. Disponível em: <https://plato.stanford.edu/archives/spr2017/entries/hume>. Acesso em: 30 jun. 2018.

ROGUE, C. *Compreender Platão.* 6a. ed. Petrópolis, RJ: Vozes, 2011.

RUSSELL, S.; NORVIG, P. *Inteligência Artificial*. Rio de Janeiro: Campus, 2004.

SHANAHAN, M. The frame problem. In: ZALTA, E. N. (Org.). *Stanford encyclopedia of philosophy*. 2016. Disponível em: <https://plato.stanford.edu/archives/spr2016/entries/frame-problem>. Acesso em: 30 jun. 2018.

STERNBERG, R. J. *Psicologia Cognitiva*. São Paulo: Cengage Learning, 2014.

THOMASON, R. Logic and Artificial Intelligence. In: ZALTA, E. N. (Org.). *Stanford encyclopedia of philosophy*. 2016. Disponível em: <https://plato.stanford.edu/archives/win2016/entries/logic-ai>. Acesso em: 30 jun. 2018.

VYGOTSKY, L. S. *Mind in society*: the development of higher psychological processes. Cambridge, MA: Harvard University Press, 1978.

WITTGENSTEIN, L. *Tractatus logico-philosophicus.* São Paulo: EDUSP, 2010.

_____. *Investigações filosóficas*. 7a. ed. Petrópolis, RJ: Vozes, 2012.

Interação, indistinguibilidade e alteridade na Inteligência Artificial

João Cortese[1]

Resumo: Como devemos agir diante de inteligências artificiais ou de robôs que viriam a ser indistinguíveis de um ser humano? O problema se insere no domínio de uma ética da tecnologia em relação ao homem, e algumas questões básicas devem ser levantadas. Pode-se pleitear que a IA sirva para investigar a inteligência humana. Mas, neste caso, deve-se precisar se o que se pretende investigar são os efeitos dessa inteligência ou sua ontologia. A questão é: afinal, o que está sendo comparado entre a inteligência humana e a IA? Não tratarei aqui da questão de fato de se um computador pode hoje se passar por um humano. Minha questão é sobre o estatuto ético de uma máquina caso isso ocorra: podemos tomar como equivalentes a indiscernibilidade pela interação e a constituição de um agente autônomo que teria, enquanto tal, um estatuto ético intrínseco? Trata-se, portanto, de colocar uma questão sobre a ontologia dos agentes éticos.

Palavras-chave: Robôs. Ontologia. Agentes éticos.

Abstract: How should we act in the face of artificial intelligences or robots that would be indistinguishable from a human being? The problem lies in the domain of an ethics of technology in relation to man, and some basic questions must be raised. One can plead that AI serves to investigate human intelligence. But, in this case, it must be determined whether what is intended to be investigated are the effects of this intelligence or its ontology. The question is, after all, what the purpose of comparing human with artificial intelligence is. The paper does not deal with the question of whether a computer can act today like a human being. It concerns the ethical status of a machine should this occur: can we take the indiscernibility by interaction and the constitution of an autonomous agent as an intrinsic ethical status as such? It is therefore a question of the ontology of ethical agents.

Keywords: Robots. Ontology. Ethical agents.

Introdução

Os recentes desenvolvimentos da Inteligência Artificial (IA) têm levado sistemas e robôs a realizarem ações cada vez mais indistinguíveis daquelas dos seres humanos. Saber até onde isso pode ser realizado, ou quando, é tarefa demasiadamente

[1] Doutor em co-tutela na Université Paris 7 e no Departamento de Filosofia da USP, pesquisador associado ao laboratório SPHERE (CNRS/Paris 7), membro do Núcleo de Bioética do Instituto PENSI - Pesquisa e Ensino em Saúde Infantil e participa do Grupo de Estudos em Inteligência Artificial do Instituto de Estudos Avançados da USP. Agradecimento do autor: Gostaria de agradecer a Bernardo Gonçalves, Fabio Cozman, Dora Kaufman, Hugo Neri e Lucas Petroni por terem me apresentado a diversas questões presentes neste artigo. Os colegas da *Associação Filosófica Scientiae Studia* acolheram uma primeira apresentação do presente trabalho, e Adriano Bechara, Marcos Paulo de Lucca-Silveira e Osvaldo Pessoa tiveram uma participação importante na discussão das ideias aqui apresentadas. E-mail: joaocortese@gmail.com.

difícil para o presente momento, ao menos para o autor. Isso não impede que nosso imaginário, não só em discussões como em diversos filmes recentes, já nos coloque: como devemos agir diante de inteligências artificiais ou de robôs que viriam a ser indistinguíveis de um ser humano? O problema se insere no domínio de uma ética da tecnologia em relação ao homem, e algumas questões básicas devem ser levantadas.

Um fator essencial aqui é a *eficácia:* a tecnologia da computação moderna percebeu que para ter *competência* sobre uma tarefa, não é preciso ter *compreensão* sobre ela. Uma máquina não precisa *entender* a aritmética para conseguir computar.[2] Que dizer então da ética envolvida para tais tipos de máquinas?

O Teste de Turing avalia, como se sabe, o resultado de *interações* entre um computador e um ser humano (TURING, 1950). À questão de se as máquinas podem pensar, Turing propõe uma substituição: seriam as máquinas capazes de se fazer indistinguíveis de humanos em um jogo da imitação?

Pode-se pleitear que a IA sirva para investigar a inteligência humana. Mas, neste caso, deve-se precisar se o que se pretende investigar são os *efeitos* dessa inteligência ou sua *ontologia*. A questão é: afinal, o que está sendo comparado entre a inteligência humana e a IA? A resposta é menos evidente do que parece. Tomando como referência o Teste de Turing, o que é evidente é que se consegue *simular* eficazmente diversos efeitos da interação humana – mas quais são os pressupostos envolvidos nisso?

Fala-se hoje de uma suposta superação deste teste por certos sistemas de IA. Pode-se questionar se isso de fato se realizou, pois cabe uma discussão sobre as condições nas quais isso teria se dado, assim como sobre a interpretação de tais resultados.[3] Não tratarei aqui da questão *de fato* de se um computador pode hoje se passar por um humano. Minha questão é sobre o estatuto ético de uma máquina caso isso ocorra: podemos tomar como equivalentes a indiscernibilidade pela interação e a constituição de um agente autônomo que teria, enquanto tal, um estatuto ético intrínseco? Trata-se, portanto, de colocar uma questão sobre a ontologia dos agentes éticos.

[2] Tal visão é apresentada por exemplo por Dennett (2013).
[3] Para uma crítica ao "sucesso" do teste de Turing, ver, por exemplo, Floridi et al. (2009).

IA "indistinguível"

A Inteligência Artificial (IA) pretende vir a se relacionar com seres humanos de maneira "indistinguível" em diversas frentes: por meio de um "chat", assim como já proposto pelo Teste de Turing original; em interação por aúdio: o Google já anunciou ser capaz de sintetizar voz indistinguível daquela de seres humanos,[4] e mais recentemente anunciou seu serviço *Duplex* de assistência, que seria pretensamente capaz de ligar para alguém sem ser distinguido de um ser humano,[5] vencendo os humanos no xadrez ou compondo música original esteticamente agradável;[6] dirigindo carros de maneira autônoma, como tem sido testado por diversas empresas; etc.

Poderíamos seguir com exemplos de êxito de IA indefinidamente. Mais do que comentar um exemplo concreto de sucesso ou de fracasso, importa aqui avançar o argumento de que, potencialmente, máquinas poderiam passar em qualquer teste interacional adaptável a elas, quando o que se avalia é a realização, ou não, de uma determinada *função*. Mas o que a interação pode mostrar sobre o agente?

Pensar sobre a IA é a outra face de se pensar sobre a inteligência humana. Ora, é claro que se pode considerar o ser humano unicamente a partir de suas interações, ou de seu comportamento – não foi o que fez, por exemplo, B. F. Skinner (1904-1990) com a sua psicologia comportamental, e o que é ainda objeto de diversos projetos contemporâneos? O aspecto metodológico do *behaviorista* é claro: tratar a psicologia humana unicamente a partir das interações entre os homens, não buscando uma causalidade além dos comportamentos observados.

Não há dúvida de que a modelagem do comportamento, por reforço ou por inibição, *funciona*. A teoria do behaviorismo "foi usada, por exemplo, para ensinar pombas a jogar tênis de mesa" (DALRYMPLE, 2017, p. 28). Mas é evidente que a questão é mais complexa: uma série de atos pode ser *interpretada* como a participação em um certo jogo, mas o que me garante que este jogo de fato esteja sendo *jogado* pelo agente, significando que realizar tais ações tenha um *sentido* para ele, abrangendo inclusive o desejo de vencer o jogo?

[4] Gershgorn, D. "Google's voice-generating AI is now indistinguishable from humans", *Quartz Media*, 26 de dezembro de 2017. Disponível em: <https://qz.com/1165775/googles-voice-generating-ai-is-now-indistinguishable-from-humans/>. Acesso em: 1 fev. 2018. Ver Shen, Jonathan, et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." arXiv preprint arXiv:1712.05884 (2017).
[5] Ver, por exemplo, Harwell, D. "A Google program can pass as a human on the phone. Should it be required to tell people it's a machine?", *The Washington Post*, 8 de maio de 2018.
[6] Cf. Gonçalves em: Consentino et al. (2018).

É evidente que a realização de uma ação não responde pela *intencionalidade* que estamos acostumados a associar a ela. Esta pode existir ou não – a realização do comportamento não responde ao que, digamos assim, motivou este comportamento.

Parece assim que, em primeira instância, uma posição comportamentalista não responde em nada à questão ontológica do sujeito, ao menos no que esta diz respeito à intencionalidade deste sujeito.[7] Contudo, algumas versões do comportamentalismo podem se apresentar de maneira mais radical.

Conjuntos de dados e interações

> O que começou como metodologia tornou-se ontologia. Um adágio antigo do diagnóstico médico diz que a ausência de evidências nem sempre é a evidência da ausência, mas os behavioristas ignoram esse sábio chamado para a modéstia. Em vez disso, começaram a acreditar que estímulo e resposta era só o que havia na vida humana, que tudo que é humano pode ser explicado dessa maneira. Embora risível, isso foi levado extremamente a sério por muitos. (DALRYMPLE, 2017, p. 28)

À parte o mérito que tenham ou não tais teorias comportamentais para descrever o comportamento humano propriamente dito, sua eficiência parece ser indubitável em um outro aspecto contemporâneo: aquele da gestão de sistemas nos quais a pessoa humana é inserida, por definição, reduzindo-se sua totalidade a um conjunto de dados. Podemos crer que o homem seja irredutível a um aspecto binário. Mas ele construiu máquinas que funcionam sob tal lógica; desde que se submeteu a viver a partir de sistemas gerenciados por tais máquinas, é claro que ele em certo sentido decidiu viver sob tal estrutura de dados.

"Alice", uma mulher, quando classificada em um sistema de dados, se torna uma jovem mulher de perfil executivo com um diploma universitário etc. Nos mais diversos âmbitos de nossa vida hoje, do trabalho à saúde, dos estudos ao entretenimento e toda a interação virtual, pessoas são transformadas em meros *tipos* (estudante, cliente, terrorista potencial, etc.). Poderíamos dizer que cada indivíduo é inexaurível quanto às suas informações (voltaremos a isso). Por outro lado, a quantidade de dados é sempre finita, por maior que seja. Nós então abstraímos,

[7] Vale ressaltar que não se trata aqui de avaliar a posição comportamentalista de maneira aprofundada em nenhuma de suas versões, mas unicamente de levar em conta seu aspecto metodológico de maneira ampla, naquilo que ele poderia ser transferido para a avaliação da "intencionalidade" de inteligências artificiais e de robôs.

generalizamos, agregamos, interpolamos, agrupamos, classificamos dados... Lidando portanto com um número muito grande que é, porém, sempre finito.[8]

Ainda que o ser humano tenha mais informações do que cabe em qualquer conjunto de *big data*, não deixa de ser o caso de que este sempre será insuficiente para representá-lo em sua integralidade.[9] O que ocorre é que os bancos de dados tratados são suficientemente grandes para que, sob certos aspectos, possamos negligenciar o que é perdido, de maneira que dizemos, por exemplo, que estamos "conversando" com alguém por um chat.

Não há dúvida de que tais empreitadas sejam de grande *eficiência*. A "quantificação das interações humanas", se podemos dizer assim, toma cada vez maior parte em nossas sociedades. Cabe, porém, questionar onde reside o seu fundamento. Ora, além de descrever as interações humanas, a quantificação das interações sociais hoje *intervém e molda* parte das interações humanas. Trata-se de reconhecer a formatação das interações entre humanos que aparecem em meios que são, por sua própria constituição, intrinsecamente quantificáveis.

Pense-se, por exemplo, em um aplicativo para *smartphones* destinado a facilitar relacionamentos humanos – seja para compra e venda de imóveis, seja para a busca de companhia. Antes de falar-se no uso bom ou mal de tal aplicativo, deve-se levar em conta que, pelo seu próprio *design*, um tal aplicativo define já um espaço de possibilidades prévio, ao qual o usuário deverá adequar-se para poder "escolher". Ao mesmo tempo em que esse tipo de implementação cria um leque de possibilidades, cria também uma lista de "restrições" quanto às interações possíveis (que seja possível enviar mensagens de texto, de áudio ou de vídeo – é evidente que isso não esgota as interações humanas). Escrever um código é definir um espaço de possibilidades de vivências. Neste sentido, o homem restringe-se a uma quantidade finita de posibilidades, o que evidentemente pode ser mais fácilmente imitado por uma IA.

[8] "The overall perspective, emerging from digital ontology, is one of a metaphysical monism: ultimately, the physical universe is a gigantic digital computer. It is fundamentally composed of digits, instead of matter or energy, with material objects as a complex secondary manifestation, while dynamic processes are some kind of computational states transitions. There are no digitally irreducible infinities, infinitesimals, continuities, or locally determined random variables" (FLORIDI, 2011, p. 319).

[9] Poderíamos evocar aqui ainda a questão mais ampla de se o caráter qualitativo das vivências humanas pode ser representado por relações quantitativas implementadas.

Aspecto da consciência

Dennett (2013) lembra bem que a palavra "computador" não se aplicou unicamente a máquinas: antes de que Turing criasse as máquinas que levam o seu nome, pessoas possuíam a função de "computadores" em diversas instituições, geralmente mulheres. Estas realizavam uma série de contas necessárias a empreitadas complexas, num trabalho que seguia algoritmos de cálculo. É assim que Turing (1936, p. 251) declara que "podemos agora construir uma máquina para fazer o trabalho deste computador [humano]".

Nesta passagem, diz Dennett (2013, p. 571), "vemos a redução de *todas as computações possíveis* a um processo sem uma mente [*mindless*]". Isto é um fato; mas por outro lado podemos dizer que este processo sem mente repete *apenas* todas as computações possíveis. A questão, no fundo, é saber o que é passível de computação.

Um enorme conjunto de dados, utilizado por uma IA dotada de *machine learning*, parece portanto poder vir a gerar uma simulação de interação humana tão bem quanto se queira – desde que aceitemos, como no Teste de Turing, que a "interação" pode ser modelada, avaliando uma função específica.

Mas o que isso nos diz sobre a ética de tais sistemas de IA? Um aspecto fundamental a ser considerado aqui é aquele da *consciência*, frequentemente considerada como necessária para que se considere que um agente tem estatuto ético.[10] Mas como saber se uma máquina é consciente? Se por acaso, examina Descartes na sua Segunda Meditação, vejo pela janela homens que passam pela rua, não deixaria de dizer, ao vê-los, que vejo homens;

> e, entretanto, que vejo desta janela, senão chapéus e casacos que podem cobrir espectros ou homens fictícios que se movem apenas por molas? Mas julgo que são homens verdadeiros e assim compreendo, somente pelo poder de julgar que reside em meu espírito, aquilo que acreditava ver com meus olhos. (DESCARTES, 1904, p. 25)[11]

Seriam esses "chapéus e capas" que passam diante de minha janela realmente homens ou meros autômatos? Como sabê-lo?

[10] Falo aqui de "consciência" no sentido forte de autoconsciência.
[11] Tradução em Descartes (Os Pensadores) São Paulo: Abril Cultural, 1983.

> Como saberemos se nossas máquinas se tornaram conscientes? Descartes argumentou que a própria consciência está além de qualquer possibilidade de dúvida. No caso dos outros, nunca estamos absolutamente certos. Muitos de nós tivemos, ainda que por um momento, a ideia de que todos os outros pudessem ser um zumbi: rindo, chorando, reclamando, regozijando-se, mas sem "ninguém em casa". Talvez os cientistas acabem descobrindo a assinatura da consciência, e então poderemos testá-la em nossos robôs, assim como nos animais e uns aos outros. Mas é certo que construiremos máquinas que *parecem* conscientes muito antes de chegarmos a esse ponto. (BLOOM; HARRIS, 2018)

À parte a questão de se a máquina terá efetivamente uma consciência, a *simulação* de uma consciência certamente aparecerá muito antes. Mas isso não é o mesmo que a consciência, ao menos de um ponto de vista ético. O que a IA faz aqui é uma aproximação, que se torna indistinguível de um ser humano.

Aproximação indefinida

A própria noção de aproximação parece estar no coração da ciência moderna. Contrariando a clássica separação aristotélica, a partir dos séculos 16 e 17 vê-se uma tendência na Europa a relacionar as Matemáticas e as Ciências Naturais, em particular matematizando a Física. Isto tem implicações também para as práticas da Engenharia.

Cabe dizer que, ao contrário do que se crê comumente, as matemáticas e a exatidão não se identificam necessariamente. Em meados do século 17, por exemplo, o cálculo das probabilidades foi inventado por Blaise Pascal e Pierre Fermat: a Matemática já podia quantificar o incerto, algo que não seria sem implicações para a computação e a IA.

Norbert Wiener, o criador da cibernética, acreditava que, mais do que a Einstein ou a Planck, deveríamos creditar a J. W. Gibbs, já no século 19, a maior das revoluções na Física do século 20: aquela de tratar probabilisticamente fenômenos contingentes.

> Nenhuma medição física é jamais precisa; e o que tenhamos a dizer acerca de uma máquina ou de outro sistema mecânico qualquer concerne não àquilo que devamos esperar quando as posições e momentos iniciais sejam dados com absoluta precisão (o que jamais ocorre), mas o que devemos esperar quando eles são dados com a precisão alcançável. [...] Por outras palavras: a parte funcional da Física não pode furtar-se a considerar a incerteza e contingências dos eventos. (WIENER, 1968, p. 10)

Isto impactaria tanto a própria ciência da época de Wiener quanto "nossa atitude para com a vida em geral" (WIENER, 1968, pp. 13-14). Quer dizer que, para Wiener, o "paradigma'' gibbsiano, se podemos dizer assim, é uma das matrizes do nosso modo de viver moderno (Wiener escrevia nos anos 1950). A análise da aproximação e da incerteza seria assim constitutiva da ciência moderna. O mesmo, podemos crer, se aplica à IA: a noção de *aproximação indefinida* aparece como um critério fundamental na avaliação das interações descritas até aqui. Se tal aproximação é possível no caso da IA, a simulação da consciência será indistinguível da própria consciência.

Aceitar uma aproximação como solução implica definir quão próximo se está da solução exata, o que implica uma teoria do erro. Isso aparece de certa maneira na fundamentação dos métodos dos Cálculos Integral e Diferencial. Estabelecidos pela análise do século 19, eles foram problematizados já no século 17, retomando desenvolvimentos tão antigos quanto aqueles que aparecem nos escritos de Euclides e de Arquimedes.

Vale aqui ressaltarmos brevemente um aspecto da indistinguibilidade na prática matemática, no método dos "indivisíveis" de Pascal.[12] Retomando o "Método da exaustão" arquimediano, Pascal propõe, como diversos autores de sua época, que para calcular a área sob uma curva, sejam somados retângulos delimitados abaixo ou acima dela (Figura 1). O problema é que sempre a área desses retângulos excede ou falta em algo para cobrir a área da curva. A soma seria "exata" apenas caso houvesse infinitos retângulos, cada um deles com uma área infinitamente pequena (os "infinitesimais"). No método dos indivisíveis de Pascal, entretanto, isso não é preciso: basta que a diferença entre a área sob a curva e a soma dos retângulos seja menor do que uma "quantidade dada qualquer" para que o resultado seja considerado como uma solução. O controle do erro permite uma aproximação indefinida, aceita na prática como uma solução válida.

[12] Para mais detalhes, ver Cortese, J. F. N. L'infini en poids, nombre et mesure: la comparaison des incomparables dans l'oeuvre de Blaise Pascal. Tese de doutorado, Université de Paris 7 e Universidade de São Paulo, 2017.

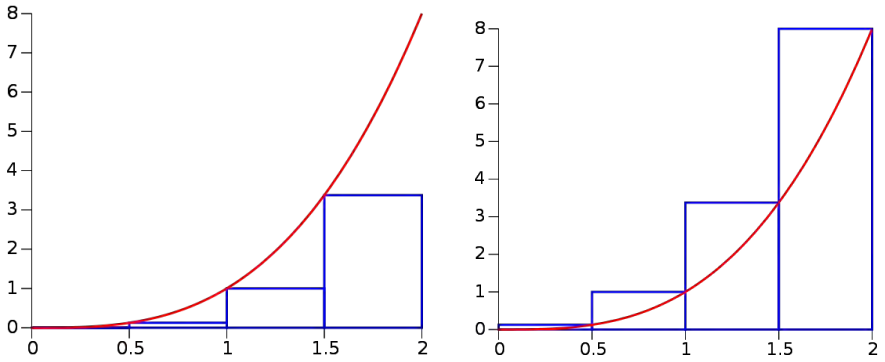

Figura 1. Cálculo da área sob uma curva: soma dos retângulos delimitados abaixo ou acima dela.
Fonte: <https://en.wikipedia.org/wiki/Riemann_sum>. Acesso em: 1 fev. 2018. Imagens de domínio público.

Trata-se em um certo sentido do que Aristóteles já propusera na sua *Física* (III, 11): os matemáticos não precisam do infinito atual para seus cálculos, mas apenas de um finito tão grande quanto se queira. O que importa aqui é que uma certa prática matemática pode ser feita em total acordo com uma colocação entre parênteses da questão ontológica do infinito, bastando uma aproximação indefinida para que se efetuem os cálculos.

Tais considerações podem talvez trazer alguma luz sobre a questão do que é que se avalia ao se abordar a interação entre o humano e a máquina. Que dizer nesse caso? O infinito é necessário desde que passamos aos fenômenos humanos ou o indefinido bastaria para avaliarmos tais interações?

> Dentre todos os filósofos que rejeitam o Teste de Turing como um critério para a inteligência, pode-se esperar que alguém apoie a proposta de responder positivamente à Grande Questão, tão quixotesco quanto isso possa parecer. (SHIEBER, 2004, p. 268)

É desta maneira que Shieber (2004) apresenta Daniel Dennett. Eu gostaria de ser aqui "quixotesco" no sentido oposto ao de Dennett, mas no que concerne a agência moral. Trata-se de dizer que, no caso do ser humano, não podemos nunca esgotar este *infinito* que é o homem. Não basta que nos satisfaçamos com nenhuma caracterização parcial, com nenhuma redução de dimensionalidade do que é o humano. A modelização dos fenômenos humanos nunca poderia ter uma pretensão ontológica radical, se admitimos o próprio fato de que a modelização deve ter sempre limites. Por mais que

tenhamos que lidar com um mero "indefinido" que não podemos fazer ser "infinito", temos de aceitar que este infinito existe – e está colocado na pessoa do outro, ainda que eu veja os "limites" do corpo deste. O ser humano é sempre irredutível, por mais que se faça uma aproximação dele.

Dada esta concepção, creio que temos de abandonar uma exequibilidade indistinguível como critério de demarcação de um agente ético no sentido próprio. A verdadeira ética, afinal, não está numa questão do fazer, mas numa questão do ser – entendendo por isso que a autoconsciência seja necessária para a constituição de um agente ético. Isso não quer dizer porém, como veremos adiante, que a dimensão ética esteja absolutamente excluída desse tipo de situação – o que pleiteio é que a constituição de um *agente* ético não pode ser mostrada por interações.

Estatuto moral da IA

A *indiscernibilidade* sob um certo critério (pelo controle do erro) mostra como de uma ideia matemática se faz algo que *funciona*. O critério interacional também *funciona*, como em toda boa tecnologia; mas será que isso basta no caso da ética?

Acredito que a consideração de tal aspecto puramente interativo não pode dar uma resposta sobre a constituição de um agente ético, *nem no caso dos homens e nem no caso das máquinas*. Cabe aqui voltarmo-nos ainda para o infinito, que já na prática matemática trazia uma questão ontológica independente daquela verificável pelas "interações" da prática do indefinido.

Se Descartes, dando um passo a partir da tradição medieval, refletiu sobre o homem, ser finito, que busca conhecer o Outro que é Deus, ser infinito, algo semelhante pode ser visto nas interações entre os homens. Para Emanuel Lévinas, de fato, a *alteridade* é vista como algo irredutível: cada *outro* é um infinito que se apresenta a mim, e pretender "conhecê-lo" plenamente seria reduzi-lo. A alteridade humana torna-se assim discernível daquela da máquina – mas não se trata de fazê-lo considerando somente as interações, sob a pena de perder aquilo que é intrinsecamente ético.

Antes de analisar quais seriam as implicações dessa concepção, cabe analisar uma dentre as diversas proposições contemporâneas sobre o tema da ética da IA: aquela de Bostrom e Yudkowsky (2011/2014).

Esses autores consideram um ser tendo um estatuto moral quando ele é um fim em si mesmo, e não um meio para algo. Eles declaram que, pelo momento,

> [...] é amplamente aceito que os atuais sistemas de IA não têm *status* moral. Nós podemos alterar, copiar, encerrar, apagar ou utilizar programas de computador tanto quanto nos agradar, ao menos no que diz respeito aos próprios programas. (BOSTROM; YUDKOWSKY, 2011, p. 208)

Não é por isso, entretanto, que seja claro quais atributos deveriam ser levados em conta para avaliar se um sistema de IA tem estatuto moral ou não. Bostrom e Yudkowsky identificam dois critérios comuns propostos à avaliação de se uma máquina tem estatuto moral ou não: a *Senciência*, ou seja, a "capacidade para a experiência fenomenal ou *qualia*, como a capacidade de sentir dor e sofrer", e a *Sapiência*, o "conjunto de capacidades associadas com maior inteligência, como a autoconsciência e ser um agente racional responsável". Seria portanto moral um agente que tivesse ambas essas capacidades. Nesta linha, desligar um computador hoje não parece infringir um direito, mas se fosse possível desligar uma máquina que pudesse "sentir dor", isto seria moralmente errado, assim como maltratar um animal.

Bostrom e Yudkowsky indicam ainda que atrás desta ideia subjaz um outro princípio:

> *Princípio da Não-Discriminação do Substrato*: "Se dois seres têm a mesma funcionalidade e a mesma experiência consciente, e diferem apenas no substrato de sua aplicação, então eles têm o mesmo *status* moral". (BOSTROM; YUDKOWSKY, 2011, p. 209)

Quer dizer que "não faz diferença moral se um ser é feito de silício ou de carbono, ou se o cérebro usa semicondutores ou neurotransmissores". Ora, o problema é que esse princípio parece fazer sentido apenas se pressupomos um reducionismo materialista em relação à consciência. Ainda aqui, o problema parece ser o mesmo: a solução funciona desde que pressuponhamos que a consciência é redutível a propriedades físicas que podem ser decompostas.

Os autores apresentam ainda um outro princípio, que faz igualmente referência à consciência sem problematizar como identificá-la:

> *Princípio da não-discriminação da ontogenia*: "Se dois seres têm a mesma funcionalidade e mesma experiência de consciência, e diferem apenas na forma como vieram a existir, então eles têm o mesmo *status* moral". (BOSTROM; YUDKOWSKY, 2011, p. 210)

Deveríamos conceder à personagem *Joi*, uma IA no filme *Blade Runner: 2049*, que um homem é feito apenas de dados, A e C e T e G, meros quatro símbolos, e o robô é feito dos dois símbolos 0 e 1, de maneira que a única diferença entre um homem e uma máquina seria quantitativa? Bostrom e Yudkowsky (2011) avançam os princípios de que, tanto o substrato como o fato de ser objeto de um *design* ou não, pouco importam. Entretanto, eles parecem cair, em um certo sentido, no erro de tomar a metodologia pela ontologia: não parece haver meio de conceder ou não a "mesma experiência consciente" a dois seres a não ser por meio de suas interações funcionais. Ainda aqui, é um pressuposto metafísico reducionista que parece adiantar a resposta que foi "buscada".

Como dito acima, escolho também partir de um pressuposto metafísico, porém aquele diametralmente oposto: o de que, a despeito da "equivalência" comportamental, há um *infinito* humano contraposto a um *indefinido* da máquina. Cada lado sairá descontente com o pressuposto metafísico contrário; dado, porém, que a ciência hoje não parece ter provado cabalmente nem o reducionismo nem o irreducionismo físico da consciência, a questão é a quem cabe o ônus da prova.

Considerar portanto o agente ético apenas como um *outro* que é infinito, ao invés de indefinido, caracteriza um tipo de "resistência ontológica", sob um certo ponto de vista quixotesca. Seria possível apresentar uma motivação à ela?

Alteridade

No filme *Contato*,[13] a cientista interpretada por Jodie Foster decifra os sinais enviados por seres do espaço para descobrir que eles indicam como construir uma máquina. Esta concluída, a doutora embarca e a liga, iniciando um tipo de viagem. Em

[13] Filme de 1997, dirigido por Robert Zemeckis e adaptado de um romance de Carl Sagan.

vez de pequenos homens verdes, ela se encontra numa praia de areias claras, e vê vindo ao seu encontro seu pai, já morto. "Você não é real, nada disso é real", diz a mulher, desorientada; "quando eu estava inconsciente, você extraiu meus pensamentos, minhas lembranças". "Achamos que assim seria mais fácil para você", respondem *eles*, sob a forma do seu pai. O contato foi feito, mas *eles* não podem se mostrar por completo. O *outro*, enquanto outro, sempre se apresenta de maneira mais diferente do que o *eu* gostaria de crê-lo.

Cabe retomar aqui o tratamento cartesiano do infinito. Como se sabe, uma das demonstrações das *Meditações* sobre a existência de Deus repousava sobre o seguinte argumento: tenho a ideia de infinito em mim; sou, porém, um ser finito; como posso então ter a ideia de infinito em meu interior? Unicamente se um ser infinito deixou esta ideia dentro de mim.

Lévinas (2000) estendeu essa ideia à interação entre o eu e o *outro* que é um homem. Para este autor, a alteridade é vista como algo irredutível: cada *outro* é um infinito que se apresenta a mim, e pretender "conhecê-lo" plenamente seria reduzi-lo.

Mas, como vimos, pode-se propor de um ponto de vista matemático que o indefinido não é o mesmo que o infinito. Cabe distingui-los para saber quando é que basta tratar de algo suficientemente grande, e portanto indistinguível do infinito porque imenso, ou quando trata-se do verdadeiro infinito.

Para Lévinas (2000, p. 31-32), o outro é infinito – não um infinito matemático, ou um infinito de mera negação do finito, mas um infinito de transcendência. Para este autor, a ética *precede* a ontologia: pretender conhecer o outro para então interagir com ele seria tentar reduzi-lo, o que é impossível. Trata-se, ao invés disso, de fazer uma escolha moral: tratar o outro como um *sujeito moral* ou como um *objeto*. A atitude ética fundamental é colocar-se face ao outro. Há dimensão ética quando coloco-me face a face a alguém. Pode-se ainda dizer, com Martin Buber (2001), que ela aparece quanto me situo diante de um *Tu*; ela não existe quando me situo diante de um *isso*.

A alteridade é, portanto, irredutível para Lévinas (2000). Mas poderíamos colocar ainda a questão: a máquina seria um outro? Se *o meio é a mensagem* (cf. Marshall McLuhan), a tecnologia com a qual eu "falo" é um outro (cf. Lévinas) ou um mero dispositivo?

Alguns pensadores propuseram recentemente que, segundo uma ética lévinasiana, poderíamos crer que o homem, ao lidar com o robô de maneira próxima e antropomorfizada, chegaria a tratar-lhe como um *Tu*, conferindo-lhe portanto um estatuto ético (WOHL, 2014; GUNKEL, 2012). Desta maneira, com a ética precedendo a ontologia, e havendo a opção de colocar-se diante de um *Tu* ou de um *isso*, uma alteridade poderia ser encontrada na interação com máquinas.

Seria interessante abordar mais detidamente a questão da "alteridade" da máquina num trabalho futuro. Por ora, vale apenas levantar a questão de se, mais do que tratar a máquina como um *Tu*, não estamos talvez nos apresentando a nós mesmos como um *isso*. Desde que o homem *entra* no sistema que projetou, desde que ele se reduz a uma certa modelização, ele se empobrece ao crer que isso é todo o seu ser. Perguntar por uma interação ética meramente a partir de uma funcionalidade numa interação já é empobrecer a questão. Parece que, do ponto de vista "ontológico", a alteridade segundo Lévinas (2000) é um infinito transcendente, que não poderia ser confundido com um indefinido.

Esse posicionamento pode parecer difícil para um funcionalista ou para qualquer pessoa que discorde do pressuposto de que o ser humano é um "outro" infinito. Eu gostaria, contudo, de terminar apresentando um argumento quanto à moralidade no trato de robôs antropomórficos que independe desse pressuposto, podendo, portanto, ser aceito de maneira mais geral.

Robôs antropomórficos

No caso de robôs e sistemas de IA que se assemelham indistintamente a pessoas, cabe refletir sobre o fato de que parecemos ter uma tendência a antropomorfizá-los, a despeito de nossas visões filosóficas ou de uma advertência dos fabricantes sobre como foram construídos. Como agir, portanto, em relação a esses robôs? Bloom e Harris (2018) consideram o que Kant dizia sobre o respeito humano aos animais. Ainda que ele visse estes como coisas sem valor moral, ele insistia em que os homens os tratassem adequadamente: "pois quem é cruel com animais torna-se duro

também na sua conduta com os homens".[14] Quanto mais não seria o caso para robôs que em sua aparência e interação fossem indistintos de seres humanos?

> Nós certamente poderíamos dizer o mesmo para o tratamento de robôs realistas [*lifelike*]. Mesmo se pudéssemos ter certeza de que eles não estejam conscientes e não possam realmente sofrer, a tortura deles provavelmente prejudicaria o torturador e, em última análise, as outras pessoas em sua vida. (BLOOM; HARRIS, 2018)

Isso quer dizer que devemos, desse modo, respeitar as máquinas de alguma maneira, sob a pena de nós mesmos nos fazermos piores no caso contrário. Há limites éticos no tratamento de IAs, *mesmo que elas não sejam agentes morais* – este "estatuto ético derivado", por assim dizer, vem simplesmente do fato de que somos homens, de estatuto moral, a lidar com elas.

Passemos a um exemplo prosaico, que já pode ocorrer em nossa sociedade. Um garoto trata *Siri*[15] de maneira indelicada, insultando-a quando lhe ordena que forneça alguma informação. A mãe lhe repreende: "filho, não fale assim com ela". O filho responde: "mas mãe, é só uma máquina". O filho está correto? Sim, pois ontologicamente não parece haver razões para tratar a máquina como um ser dotado de liberdade ou de autoconsciência, fatores importantes para se considerar um agente como detentor de moralidade no sentido forte. Mas a mãe está também correta: ao agir como tal, em particular com um sistema que simula um tipo de interação tipicamente humana (trocar informações numa conversa), o garoto está *agindo mal*, no sentido de que deixa de lado uma virtude no seu agir. Quem piora quando Siri é "desrespeitada" não é o próprio sistema de IA, pois a rigor ele não sofre, mas a própria pessoa que realiza a ação, pois em certo sentido ela se empobrece quanto à sua dignidade no trato com os outros.

Considerações finais

Como conceituar essa dimensão "ética" no trato com robôs e com inteligências artificiais? Eis uma importante discussão a ser feita, e que poderia se desenvolver por

[14] Kant, I. (1997 [1784–5]). "Moral Philosophy: Collin's Lecture Notes", in *Lectures on Ethics*, P. Heath and J.B. Schneewind (org. e trad.), Cambridge : Cambridge University Press, p. 212.
[15] Uma IA da Apple que atua como assistente pessoal, interagindo por voz com o usuário.

diversas vias. Eu gostaria de apresentar aqui simplesmente um esboço de uma proposta que distinguiria duas instâncias éticas.

A primeira está no caso de uma interação "ontologicamente ética": é quando me coloco em interação com alguém que sei possuir um estatuto ético próprio, estabelecendo uma relação entre um *Eu* e um *Tu* (BUBER, 2001) e situando-me face a face com ele (LÉVINAS, 2000). Poderíamos dizer que lido aqui com um "agente moral", ou mesmo com um "paciente moral", no sentido de alguém que, tendo um estatuto moral próprio, sofre uma ação.[16] Nessa situação, posso perguntar: há *alguém* aí? Com *quem* eu falo?

O segundo caso é o de uma interação na qual não lido com alguém que possua estatuto ético, mas trato de um *isso*, de um objeto. Entretanto, do próprio fato de que sou uma pessoa autoconsciente, um estatuto ético perpassa todas as minhas ações, de maneira que posso considerar este *isso* como um "objeto de moralidade", ou seja, um objeto sem intenção, em relação ao qual eu mesmo posso agir eticamente ou não. Pode-se pensar "moralidade" de tal objeto deve ser maior quando ele for um produto humano investido de intencionalidade (fabricado por alguém), ou quando a sua interação comigo assemelhar-se àquela de humanos. Nessa situação, devo perguntar: *o que* está aí? Com *o que* eu "falo"? Como devo me colocar em relação a isso?

Podemos problematizar se há um âmbito ético relacionado a robôs e inteligências artificiais. A única conclusão definitiva é que já não podemos deixar de nos colocar questões sobre isso.

Enviado: 2 fevereiro 2018
Aprovado: 9 março 2018

Referências

BLOOM, P.; HARRIS, S. The Stone: it's Westworld: What's wrong with cruelty to robots? New York Times, 23 de 4, 2018. Disponível em: <https://www.nytimes.com/2018/04/23/opinion/westworld-conscious-robots-morality.html>. Acesso em: 1 fev. 2018.

[16] Sobre as diversas possibilidades do uso dos termos "agente moral" e "paciente moral", ver Floridi e Sanders (2004).

BOSTROM, Nick; YUDKOWSKY, Eliezer. A ética da Inteligência Artificial, trad. Pablo Araújo Batista. *Fundamento*, v. 1, n. 3, p. 200-226, 2011.

_____. The ethics of Artificial Intelligence. In: FRANKISH, Keith; RAMSEY, William M. *The Cambridge handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, p. 316-334, 2014.

BUBER, M. *Eu e e tu*, trad. Newton Aquiles von Zuben. São Paulo: Centauro, 2001 [1974].

DALRYMPLE, T. *Evasivas admiráveis*: como a Psicologia subverte a moralidade. São Paulo: É Realizações, 2017.

DENNETT, D. Turing's strange inversion of reasoning. In: COOPER, S. B.; LEEUWEN, J. van. (Orgs.). *Alan Turing*: his work and impact. Amsterdam: Elsevier, 2013.

DESCARTES, R. *Oeuvres*, vol. IX. Adam, C.; Tannery, P. (Orgs.). Paris: Cerf, 1904.

FLORIDI, L. *The philosophy of information.* Oxford: Oxford University Press, 2011.

FLORIDI, L.; SANDERS, J. On the morality of artificial agents. *Minds and Machines*, 14 (3), 349-379, 2004.

FLORIDI, L.; TADDEO, M.; TURILLI, M. Turing's imitation game: still an impossible challenge for all machines and some judges – an evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), 145-50, 2009.

Consentino, Marcelo; Gonçalves, B.; Cozman, F.; Wasserman, R. Os primeiros 60 anos de feitos da Inteligência Artificial – Revisitando as previsões de Herbert Simon. *Estadão: Estado da arte*, 23 de março de 2018. Disponível em: <https://oestadodaarte.com.br/inteligência-artificial/>. Acesso em: 1 fev. 2018.

GUNKEL, D. J. *The machine question:* critical perspectives on AI, robots, and ethics. Cambridge, MA: MIT Press, 2012.

LÉVINAS, E. *Totalité et infini:* essai sur l'extériorité. Paris: Hachette. 2000 [1961].

SHIEBER, S. M. (Org.). *The Turing test*: verbal behavior as the hallmark of intelligence. Cambridge, MA: MIT Press, 2004.

TURING, A. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42, 23–265, and erratum (1937) 43, p. 544-546, 1936.

TURING, A. Computing machinery and intelligence. *Mind*, 59 (236), 433-60, 1950.

WIENER, N. *Cibernética e sociedade:* o uso humano de seres humanos. São Paulo: Cultrix, 1968.

WOHL, B. S. Revealing the 'face' of the robot: introducing the ethics of Levinas to the field of robotics. In: KOZLOWSKI, Krzysztof; OSMAN, Tokhi; MOHAMMED, O.; GURVINDER, Virk S. (Orgs.). *Mobile service robotics*, Singapore: World Scientific, 704-714, 2014.

O menosprezado debate sobre o artificial em IA

Orlando Lima Pimentel[1]

Resumo: O presente artigo tem por objetivo explorar o debate do papel da artificialidade presente no estudo da Inteligência Artificial (IA). Para tanto, primeiro mostrarei os dois principais sentidos do termo "artificial" e como eles não se adequadam igualmente quando referidos a Inteligência. Em um segundo momento, outras distinções serão feitas com relação a IA e seu uso atual contextualizado. Uma vez estabelecidos esses limites semânticos, o papel e sentido da "artificialidade" em IA será finalmente analisada, tendo como nossa principal referência a ideia do Jogo da Imitação de Alan Turing, a obra de Charles Babbage e a esquecida categoria profissional dos chamados "computadores humanos", fundamentais para o desenvolvimento da computação antes das máquinas de computação.

Palavras-chave: Inteligência Artificial. Jogo da imitação. Computadores humanos. Alan Turing. Charles Babbage.

Abstract: This paper aims to explore the debate concerning the role of artificiality in the study of Artificial Intelligence (AI). In order to do so, I will first show two different meanings of the term "artificial" and how both do not fit equally well when referring to Intelligence. Secondly, a distinction will be made concerning AI and its contextualized current uses. Once these semantic boundaries have been established, the role and meaning of "artificiality" in AI will be finally analyzed, having as our main reference Turing's idea of an *Imitation Game*, the work of Charles Babbage and the forgotten economic category of professionals called "human computer", fundamental to the development of computation before the invention of modern computers.

Keywords: Artificial Intelligence. Imitation game. Human computers. Alan Turing. Charles Babbage.

Os dois sentidos do termo "Artificial"

Abra-se qualquer dicionário. Ao procurar por "artificial", é provável que encontremos pelo menos duas acepções gerais para o termo: uma primeira (chamemos de A_1) será a de que a palavra "artificial" refere-se à qualidade de algo produzido graças à técnica de um artífice, ou seja, a qualidade de algo feito com a *Arte* das mãos humanas e que, portanto, destoaria daquilo que é produzido na natureza através de suas dinâmicas internas, independentes da intervenção humana; uma segunda acepção

[1] Bacharel em Filosofia pela USP (2017) e Mestrando na mesma instituição, tem por objeto de estudo a obra do matemático, inventor e economista inglês Charles Babbage. Atualmente, é membro da Associação Filosófica Scientiae Studia, na qual organiza o Grupo de Estudos Marx, Ciência e Tecnologia e participa do Grupo de Estudos em Inteligência Artificial, da mesma instituição, vinculado ao Instituto de Estudos Avançados da USP. E-mail: orlando.pimentel@usp.br.

(chamemos de A_2) seria a de artificial como a característica daquilo que é imitação, dissimulação, sem naturalidade e não espontâneo, tal como na frase "seu sorriso ficou muito artificial nessa foto".

Atento a esses dois lados do termo, o economista Hebert Simon, em seu *The Sciences of the Artificial*, sugere que o sentido A_2 seria indício da pouca estima que a humanidade possui para com suas próprias obras (Simon, 1996), como podemos ver na passagem a seguir:

> Our language seems to reflect man's deep distrust of his own products. I shall not try to assess the validity of that evaluation or explore its possible psychological roots. But you will have to understand me as using 'artificial' in *as neutral a sense as* possible, as meaning man-made as opposed to natural. (SIMON, 1996, p. 4, grifo nosso)

Diferente de Simon, que não pretendeu se aventurar em seu livro quanto àquilo que chamamos de sentido A_2 e estabeleceu condições próprias às ciências do artificial (nos limites de A_1), pensamos ser importante estimular a reflexão sobre os eventuais choques entre os dois sentidos do termo, tal como ocorrem não apenas no campo semântico, mas também no campo dos fatos e valores cognitivos e sociais, inerentes ao desenvolvimento científico e tecnológico.

Quanto à maior neutralidade possível nessa empreitada, também nos distanciaremos do modo como o autor utiliza o termo "neutro". Ser o mais neutro possível é uma pretensão que deveria expressar um ideal de não vinculação a qualquer valor ideológico ou social específico. O ideal de neutralidade cumpre seu papel no momento específico da análise objetiva entre hipóteses científicas conflitantes (LACEY, 2014) e, justamente por isso, em momentos em que o conflito de hipóteses não está em jogo, o termo "neutro" ganha comumente na ciência e tecnologia de nossos dias outros propósitos nada comprometidos com a objetividade. A passagem citada é um exemplo desse outro sentido do "neutro", a saber, a neutralidade a serviço da esquiva quanto a discussões advindas de perspectivas de valor de outros campos do conhecimento, que, apesar de não engajados imediatamente com os resultados práticos das ciências aplicadas, reivindicam o seu lugar como interlocutores igualmente importantes para a avaliação das consequências científicas e tecnológicas.

No sentido *mais neutro possível do uso do termo artificial*, para ser coerente com a ideia de neutralidade enquanto não vinculação ou predileção valorativa, não deveria haver, por parte de Simon, a consideração de apenas um dos significados de Artificial (A_1), aquele, como supõe o autor, mais adequado às Engenharias. Para ser neutro, de fato, é preciso que ambos os sentidos (A_1 e A_2) sejam considerados. Nossa esperança é de que, agindo de tal forma, estaremos promovendo uma articulação interessante de questionamentos e um caminho certamente mais rico do ponto de vista do fomento à interdisciplinaridade.[2]

Esses dois pólos do termo artificial – o pólo da dissimulação e o pólo daquilo que é exercido pela *Arte* e dedicação humana (portanto, artificial enquanto qualidade de esforços nem um pouco fingidos em nossa sociedade) – convivem de modo nem sempre pacífico nessa palavra, num espectro polissêmico e valorativo que pode tender a um ou outro dos pólos, dependo do contexto no qual o termo se aplica. Nesse sentido, se tomarmos a expressão "aroma artificial de jasmim", por exemplo, tendemos a combinar A_1 e A_2 em proporções idealmente iguais, se temos em mente *o ponto de vista do seu uso corrente no contexto da indústria química de produção de aromas e sabores sintéticos*,[3] no qual a imitação de aroma de jasmim é planejada para ser o mais idêntica possível à da planta e onde não se costuma interromper a reprodução das tarefas e trabalhos diários para pensar nas complicações ontológicas em designar um cheiro como "de jasmim",[4] mesmo não tendo vindo originalmente da planta de mesmo nome.

Por outro lado, o consumidor em geral não gosta da ideia de que está comprando uma imitação e, portanto, quando diz, de um produto, que "seu sabor ou aroma é muito artificial", o contexto é outro e o termo passa a se situar muito mais no

[2] Portanto, agora, pode-se compreender melhor o porquê de optarmos por chamar esse texto de o "menosprezado" debate do artificial em IA. A explicação encontra-se na restrição (tanto semântica quanto valorativa) da artificialidade apenas no sentido A_1, deixando de lado esforços para pensar hipóteses que encarassem as dificuldades em refletir sobre A_2.
[3] Não entraremos aqui nos meandros da diferença entre o sintético e artificial. Elas existem, mas não serão relevantes para o debate que é proposto neste artigo. De todo modo, a diferença pode ser conferida em: Simon, 1996, p.4.
[4] Se formos mais fidedignos ao modo como são escritas as embalagens dos produtos das gôndolas de supermercado no Brasil, nos protocolos da ANVISA é utiliza-se a expressão "aroma idêntico ao natural de jasmim" para designação de aromas sintéticos. A simples omissão do sentido de "artificial" e presença de "natural" é uma tentativa de driblar o pólo de dissimulação (A_2), seus efeitos na opinião pública e, consequentemente, no mercado consumidor.

pólo A_2, dando ênfase à discrepância entre o natural e o artificial, que tenta se passar por natural.[5]

Três sentidos de Inteligência Artificial

Com a adjetivação específica da inteligência, qualificada como um produto artificial, passa-se algo diferente e não poderíamos simplesmente aplicar o que falamos acima sobre A_1 e A_2 na análise do significado da expressão "Inteligência Artificial", sem que antes tomássemos alguns cuidados e contextualizações. É provável que nunca tenhamos desejado ofender alguém falando que a pessoa possui uma "Inteligência Artificial", tal como quando nos queixamos do sabor muito artificial de um produto. A expressão, portanto, só ganha algum sentido, que valha a pena ser enquadrado entre A_1 e A_2, quando avaliada em seu uso como ciência aplicada ou como um estimulante a questionamentos filosóficos sobre o problema mente/corpo, a emergência da consciência e a possibilidade de máquinas pensarem.

Portanto, para determinar qual sentido de "Inteligência Artificial" nos interessa separaremos essa expressão, tal como fizemos com o termo "artificial", em três categorias: a primeira seria de IA enquanto um campo técnico de desenvolvimento de aplicações computacionais que substituem, completa ou parcialmente, a atividade computacional humana (como no caso de análise de imagens, mineração de dados em *big datas* e reconhecimento de padrões) ou mesmo ocupam um processo que um humano jamais poderia ter ocupado (por conta da nossa limitada capacidade sensorial e computacional, comparada aos mais avançados *softwares* e *hardwares* de hoje); uma segunda seria a de IA enquanto uma tentativa de imitar o comportamento inteligente humano, tal como foi concebido por Alan Turing, em seu Jogo da imitação, descrito em seu famoso artigo de 1950, *Computing Machinery and Intelligence*. Uma terceira, finalmente, é aquela própria às conclusões tiradas do experimento mental do cérebro de silício do filósofo da mente David Chalmers (CHALMERS, 1995). Nesse experimento, imagina-se que o sistema neurológico humano vá sendo substituído aos poucos por chips que reproduziriam artificialmente a mesma função e organicidade de neurônios

[5] Aqui deixamos de lado, propositalmente, complicações quanto aos limites do conceito de natureza que, apesar de interessantes, não serão proveitosas para o enfoque deste artigo. Tais complicações advêm do uso de técnicas de Engenharia Genética (transgenia) ou mesmo do estatuto artificial de técnicas agrícolas de especiação (seleção artificial de espécies).

biológicos humanos. Aqui não nos cabe acompanhar toda argumentação do experimento, mas apenas apresentar a conclusão a que chega Chalmers: a de que um sistema artificial com organização e funcionamento idênticos a sistemas neurológicos deveria ter o mesmo tipo de experiência consciente de um cérebro neural-biológico humano. Mais do que imitação de comportamento inteligente, máquinas e sistemas computacionais artificiais, com organicidade e funções complexas o suficiente, poderiam possuir, de fato, inteligência e consciência. Esse último sentido de IA é aquele que anima a pesquisa no sentido do desenvolvimento de uma *Strong AI*, para utilizar uma terminologia de um dos mais respeitáveis críticos dessa corrente filosófica (SEARLE, 1980). Chamemos os sentidos expostos acima de IA_1, IA_2 e IA_3 respectivamente, a título de abreviação.

Antes de prosseguirmos, no entanto, vale a pena lembrarmos que a distinção entre IA_1 e IA_2 não é nova e muitas categorizações já foram feitas com maior ou menor proximidade a deste artigo. Noam Chomsky, por exemplo, é um dos filósofos que, em suas conferências, tem a usado para a crítica ao reducionismo computacional da consciência, proposta por filósofos como David Chalmers. John Searle, por sua vez, em seu famoso experimento mental do quarto chinês (SEARLE, 1980), estabelece a diferença entre dois sentidos de IA (*strong* e *weak*), que é praticamente a mesma que fizemos aqui entre IA_1 e IA_3.

As relações de IA_1, IA_2, e IA_3 para com A_1 e A_2

Podemos dizer que grande parte do trabalho das empresas do ramo de desenvolvimento de Inteligência Artificial tem em vista principalmente a IA_1 e, em muito menor número, IA_2. Os questionamentos filosóficos (problemas como o da possibilidade ou não de criação de máquinas conscientes), o caráter da imitação do comportamento inteligente (comportamental ou mais do que isso?) e os limites da pesquisa computacional e da metodologia de Turing estão mais relacionados aos limites de IA_2 e a pertinência ou não de IA_3. Uma variedade muito rica de debates, distante do sentido de IA_1, acaba virando assunto privilegiado da Filosofia da Mente e da crença particular de cada simpatizante do tema.

Do ponto de vista filosófico, portanto, o que vale a pena ser abordado são as possibilidades de questionamento e não a mera reprodução das técnicas que gozam de uma blindagem cotidiana e quase fabril quanto à validade ou não de seus pressupostos e de suas utilidades. Nesse sentido, discutiremos a interação entre os dois sentidos de artificial (A_1 e A_2) dando ênfase à IA_2 e IA_3, respectivamente aquele mais próximo ao *Turing Test* e aquele dos funcionalistas da Filosofia da Mente. O sentido de IA_1 só valerá a pena quando abordado num uso subvertido. Se IA_1 não interrompe sua implacável reprodutibilidade de técnicas para ouvir as interferências vindas de questionamentos próprios da IA_3 e aos limites da IA_2, o contrário não precisa ser verdadeiro.

IA_2 e IA_3 foram e continuam sendo influenciadas, sim, pela pluralidade e história de desenvolvimento da artificialidade (seja da artificialidade computacional própria às técnicas de *Machine Learning* ou da artificialidade anterior aos pioneiros da computação eletrônica do séc. XX). Mesmo as mais abstratas concepções computacionais e matemáticas (tal como as máquinas computacionais universais discretas e binárias de Turing) são filhas inegáveis de seu tempo e contexto. Muito pouco avançaríamos, do ponto de vista do debate de ideias e do fomento ao questionamento, se observássemos o desenvolvimento da IA e da computação apenas como saltos promovidos por grandes genialidades pretensamente independentes das tendências de sua época.

Dado esse pressuposto do presente artigo, a nosso ver, parece muito mais empolgante e frutífera a tese do filósofo da ciência Ian Hacking, com a qual nos associamos, de que vivemos, pelo menos desde a primeira metade do séc. XIX, o paradigma da busca por quantificação e medição de tudo (Hacking, 2012, p. 335). Assumindo tal abordagem paradigmática, mesmo a extenuante dinâmica de IA_1 conflui junto a IA_2 e IA_3 nessa grande tendência computacional da qual somos testemunha e parte no experimento.

O paradigma da medição e computação cumpre não apenas um papel teórico, mas ainda um papel prático de fomento às pesquisas da ciência aplicada, o que rendeu e rende também frutos materiais. Seria mesmo possível retraçar uma história de diversas técnicas e artifícios que criaram a base material e artificial para os três sentidos de IA que apresentamos mais acima. Para nossos propósitos, nas próximas

linhas, será suficiente nos restringirmos apenas a dois desses exemplares: a noção de Computador binário Discreto e Universal de Alan Turing e de Computadores Humanos.

Computadores digitais e o jogo da imitação

As *Máquinas de Turing* são máquinas discretas de computação binária e de funcionalidade universal (abreviaremos por MDBU) descritas por Alan Turing em seu jogo da imitação (Turing, 1950). Elas são, segundo o autor, o tipo de mecanismo mais adequado para trabalhar com a codificação e descodificação do comportamento linguístico escrito humano e, portanto, a mais adequada para participar da dinâmica do jogo da imitação. Do ponto de vista da IA_2, portanto, poderíamos dizer que a participação de MDBU é parte artificial (em sentido A_1) necessária do design do jogo. Ou seja, estamos lidando com dois níveis de A_1 em IA_2: há a artificialidade das MDBU e há a artificialidade do *design* do jogo, e ambas tem seu devido lugar de análise. Comecemos pelo *design* geral do jogo, que parece ser de mais fácil compreensão.

Para o funcionamento adequado do teste de Turing, faz parte da mecânica do jogo a presença de 3 elementos dispostos numa estrutura rígida e inalterável: Um Juiz-humano, um entrevistado-humano e uma MDBU, cada qual em uma sala distinta, sem possibilidade de contato visual, mas com a possibilidade de comunicação escrita entre o juiz-MDBU e juiz-o outro ser humano. É parte inerente ao jogo da imitação de Turing uma interface "não-transparente". Elas são os muros das salas que impedem o juíz de ver a materialidade física da MDBU.

O objetivo do jogo é fazer com que o juiz se "engane" com relação a quem é o humano e quem é a máquina, após uma série de perguntas enviadas para ambos, e, da mesma forma, respondidas por ambos. Se a máquina é suficientemente bem programada a ponto de imitar adequadamente o comportamento humano de resposta, ela passou no teste de Turing e pode ser chamada de "Inteligente". Eis aqui o modo como A_2 se articula quanto à mecânica do jogo da IA_2: a imitação que quer se passar pelo humano é o objetivo próprio do teste de Turing e não há dificuldades em reconhecer tal relação.

Vejamos agora como uma MDBU é composta em sua artificialidade (A_1) idealizada. Para tanto, faremos uso aqui da didática descrição de George S. Boolos em seu *Computabilidade elógica*:

> Uma *máquina de Turing* [ou MDBU] é um tipo específico de máquina idealizada, cuja função é executar computações. [...] A computação [efetuada por tais máquinas] tem lugar em uma fita, dividida em quadrados, que é interminável em ambas as direções. [...] Cada quadrado ou está *em branco*, ou tem um traço impresso nele. (BOOLOS, 2012 , p. 43)

A MDBU possui ainda cinco possibilidades inerentes ao seu mecanismo:

1 – Apaga um quadrado da fita;
2 – Escreve um traço em um quadrado da fita;
3 – Move-se um quadrado para a direita;
4 – Move-se um quadrado para a esquerda;
5 – Para a computação.

A partir da articulação dessas cinco habilidades da máquina, ela pode efetuar em sua fita infinita cálculos, codificações, descodificações, reservar quadrados para organizar uma "linguagem" binária (tal como um código Morse da máquina), utilizar essa linguagem para criar linguagens ainda mais complexas e, por fim, por meio de uma adequada programação, ser capaz de descodificar os comportamentos da linguagem do juiz do jogo da imitação de Turing.

Em linhas bem gerais e não exaustivas, essa artificialidade (A_1) é a característica peculiar das MDBUs e que as fazem as máquinas computacionais mais simples e adequadas para nosso autor. Algo mais complicado, no entanto, se passa quando pretendemos investigar A_2 não no sentido geral do design do jogo, seus elementos e estrutura, mas no sentido específico inerente às MDBUs. Poderiam MDBUs, para além do teste de Turing, possuir um sentido dissimulador? Ou seja, mesmo num programa muito mais simples,[6] haveria a possibilidade de pensar sua relação com A_2?

[6] Digamos: um programa de somar um número a qualquer outro e imprimir o resultado numa fita magnética, sem pretensões maiores de escrever respostas e imitar comportamentos humanos.

Parte da resposta dessa questão esta na seguinte passagem de *Computing Machinery and Intelligence*:

> The digital computers […] may be classified amongst the 'discrete-state machines'. These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. *Strictly speaking there, are no such machines. Everything really moves continuously.* But there are many kinds of machine which can profitably be thought of as being discrete-state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them. (TURING, 1950 p. 6 – grifo nosso)

Uma questão ontologicamente muito espinhosa está presente nas linhas acima, a saber, a da existência contraditória, ao mesmo tempo discreta e contínua, aparente na consideração física de mecanismos planejados como ferramentas lógicas (tal como em interruptores que vão do ligado ao desligado, do "1" ao "0", sem estados intermediários). Em que sentido seria lícito ou com que critérios poderia se passar do contínuo da natureza física das máquinas ao pretendido funcionamento de saltos discretos das MDBUs sem recair em contradição? Essa preocupação não é alvo da atenção de Turing que passa rapidamente a assumir os estados discretos de máquinas como uma ficção conveniente (*convenient fiction*), útil e efetiva para a computabilidade.

A saída de Turing, ao enfatizar a conveniência ficcional da binariedade das máquinas discretas computacionais e deixar de lado a ontologia e complexidade subjacente da natureza contínua das mesmas, é análoga a postura dos que, anos mais tarde, passaram a ser designados como funcionalistas da filosofia da mente (relacionados à IA_2). Em ambos os casos, poderíamos dizer que há uma dissimulação (A_2) ou, para ser mais preciso, uma conveniente ficção, quanto ao estatuto discreto ou não da natureza das máquinas e do sistema Nervoso humano.

Nesse ponto é preciso lembrarmos, como muito bem ressalta Herbert Simon (SIMON, 1996, p.3), que os produtos artificiais do engenho humano não estão apartados da Natureza. Não escapam à sua dinâmica. Apesar de Turing, reconhecer na subseção *Argument from Continuity in the Nervous System* de seu artigo (TURING, 1950, p. 15-6) a possibilidade de uma máquina analógica (um *differential analyser*) poder ser

empregada no jogo da imitação, ainda assim a continuidade complexa da natureza ainda não o constrange a, mais uma vez, procurar enquadrar o problema das diferenças ontológicas entre MDBUs e Máquinas Analógicas na estrutura de seu jogo da imitação, com seu fim último de *enganar* o juiz.[7] Em resumo, diferenças entre máquinas analógicas e discretas, sim, existem, mas, do ponto de vista do objetivo de *dissumulação* do jogo da imitação, podem ser convenientemente deixadas de lado.

O "salto" ontológico do contínuo ao discreto dado por Turing deve ser entendido como uma dissimulação até maior do que aquela do design geral do jogo que propõe. Assim sendo, podemos concluir que não só para IA_2, mas também para IA_3, existem pressuposições "muito artificiais" (no sentido A_2) inerentes ao caráter das MDBUs. Em outros termos, tais máquinas tentam se passar por discretas, no entanto, como o próprio Turing reconhece (e logo se desvencilha da declaração), *everything really moves continuously*.

Computadores humanos nos tempos de Babbage

Nos últimos parágrafos, tentei expor alguns motivos para adotar o artificial (A_1 e A_2) para IA_1 IA_2 e IA_3. A reflexão acerca da base material-artificial, no entanto, não é satisfatório se tomarmos os Computadores binários, base para o Teste de Turing, como um conceito sem história. Houve, antes mesmo dos primeiros computadores eletrônicos, tentativas de Ábacos Lógicos (de autoria de William S. Jevons), Máquinas Tabuladoras (com Herman Hollerith) e aquele que foi condecorado como o primeiro projeto de uma primeira máquinas de computação programável através de cartões perfurados: o Engenho Analítico de Charles Babbage e Ada Byron.

Não nutrimos com isso nenhuma esperança de retraçar uma linha causal direta entre o trabalho dos pioneiros da computação eletrônica do século XX e aquele desenvolvido no século anterior. Tal empreitada não seria legítima do ponto de vista bibliográfico de Turing que, segundo comentadores (FEFERMAN, 2001, p.10), não foi influenciado por qualquer contato com os esboços e projetos de máquinas de computação de Babbage. Ainda assim, seria igualmente lamentável não reconhecermos a tendência histórica da qual falávamos mais acima: aquela do

[7] "It is true that a discrete-state machine must be different from a continuous machine. But if we adhere to the conditions of the imitation game, the interrogator will not be able to take any advantage of this difference" (TURING, 1950, p.15).

paradigma da medição (HACKING, 2012, 335), que possui seu marco zero, segundo Hacking, na figura de Charles Babbage.

Por ora, adentraremos mais a fundo na raiz da necessidade histórica material da atividade computacional humana que deu a base material para todos os sentidos de IA de que falávamos anteriormente. Se no jogo de Turing falávamos do problema do salto do contínuo ao discreto quanto ao estatuto das máquinas computacionais, agora falaremos desse mesmo problema quanto aos humanos e à antiga profissão dos computadores humanos. Ou seja, em que sentido do ponto de vista de uma determinada atividade mental e laboral computacional humana se articula a contradição do salto do contínuo ao discreto?

O termo "computador", em nossa época, é corriqueiramente usado como sinônimo de uma máquina. No entanto, nem sempre foi assim. Os "computadores humanos" faziam parte de uma divisão de trabalho específica do empreendimento muito peculiar de confecção de tabelas de cálculo, dos mais variados dados estatísticos e matemáticos (recenseamento de países, medidas náuticas, astronômicas, logística de recursos bélicos etc.). Verdadeiros operários da divisão, soma, subtração e multiplicação passavam horas e mais horas preenchendo tabelas e mais tabelas de cálculos (GRIER, 2005), organizados por uma estrutura de divisão de trabalho própria às empreitadas matemáticas e a reprodução de determinadas operações mentais. [8]

É um pressuposto para a existência dos computadores humanos de que sua função tenha sido enquadrada em um empreendimento matemático quase fabril. Para usar uma expressão de Gaspar de Prony, o que se tinha em mente era produzir tabelas matemáticas/computacionais tal como se produziam alfinetes (menção explícita ao exemplo dado por Adam Smith em seu *The Wealth of Nations*). Foi preciso, portanto, *de modo discreto*, tão discreto quanto mais tarde Turing concebeu seus computadores binários, dividir as atividades mentais humanas em um trabalho específico, passível de receber um algoritmo de instruções matemáticas que deveria reproduzir ao longo do dia de trabalho, com listas imensas dispostas em tabelas de dados. Não é coincidência que os computadores humanos sejam o exemplo privilegiado de Turing em seu artigo

[8] Podemos conferir parte dessa história em *On the economy of machinery and manufacture* de Charles Babbage (Babbade, 2010), ao descrever a organização hierarquizada do trabalho matemático realizada pelo matemático Gaspar de Prony, ainda no período da França Napoleônica.

de 1950. O papel dos computadores humanos foi de fundamental importância para o desenvolvimento de estudos estatísticos dos mais variados e foi útil para muitos dos empreendimentos científicos e tecnológicos contemporâneos (e anteriores) aos pioneiros da computação digital eletrônica. Por conta disso, os computadores humanos, como profissão e forma discreta de abstrair uma função de reprodução do trabalho mental, podem ser enquadrados como um dos frutos próprios ao paradigma da medição e computação de tudo.

MDBUs versus computadores humanos

As MDBUs de Turing, diferentemente dos computadores humanos, realizam suas tarefas, sem reclamar, em fitas infinitas e não se desgastam com sua jornada de trabalho no mundo abstrato em que idealmente habitam. No mundo das necessidades materiais econômicas e sociais, no entanto, os profissionais da computação humana se cansavam, erravam (até propositalmente), trabalhavam tendo em vista o salário e poderiam até mesmo fazer greve contra as condições em seu local de trabalho.

A comparação entre MDBUs e computadores humanos feita acima é apenas a mais direta que podemos estabelecer. Mas, de modo indireto e mediado pelos dois sentidos de artificial (A_1 e A_2), proporemos ainda mais uma linha de aproximação: em que sentido seres humanos se artificializam no sentido de A_1 e A_1?

Como já tínhamos definido no início desse texto, A_1 diz respeito a produtos da produção humana que não existiriam por si só na natureza. A princípio, parece contra-intuitivo e logicamente circular pensarmos que humanos são produtos artificiais de si mesmos, mas parece igualmente contra-intuitivo pensar que a natureza se incumbiria de criar por si só a função e profissão de "computador-humano". Humanos são recorrentemente trabalhos e produzidos ao longo de densos processos formativos na atualidade (em áreas e disciplinas especializadas em universidades, por exemplo), mas também no mercado de trabalho e pelas demandas urgentes de empreendimentos econômicos, mais eficazmente organizados pela divisão do trabalho mais lógica possível (seja ele mental ou manual). Tal divisão do trabalho, por sua vez, é não só uma técnica de gestão; é também uma gramática da produtividade e desenvolvimentos sociais, científicos e tecnológicos. A multiplicidade de funções mentais ou manuais

organizadas pela divisão do trabalho corporifica-se em demandas de reprodução inerente a cada profissão, e é tanto melhor (para a reprodução sistêmica) quanto mais tal estrutura trata cada ser humano, seu comportamento e suas tarefas como variáveis numérica, as processando, com o auxílio cada vez mais ubíquo da informática e técnicas caras à IA_1, no fim das contas como um dado binário, que ao invés de 0 e 1, trabalha com a binariedade útil-inútil ou mesmo empregado-desempregado. Essa é a nossa lógica binária compulsória do dia-a-dia, sem a qual a eficácia esperada de cada um de nós não se dá, e com a qual somos constantemente testados (seja com relação a nossa inteligência ou outras habilidades).

A divisão do trabalho cumpre ainda outra função: a de estimular um ambiente cada vez mais propício a criação de novas ferramentas, máquinas e sistemas automatizados, algo que se encontra defendido já por Charles Babbage, em sua obra *On the economy of machinery and manurfacture*. Seguia-se, de sua análise econômica e ânsia por sistematização e medição de todas as unidades constastes (do natural ou do artificial), que a mesma divisão do trabalho que cria ou elimina profissões fosse aquela que fomentaria também o desenvolvimento de novas tecnologias e isso se daria até mesmo em áreas antes impensáveis de serem automatizadas: como foi o caso do cálculo matemático operado por computadores humanos.

Quanto à relação entre nós e A_2 poderíamos propor uma articulação nos seguintes termos: Seres humanos são artificiais em sentido A_2, não pelo simples trejeito dissimulador ou simples tentativa de imitar algo ou alguém que não são. Mais do que isso, a artificialidade dissimuladora nos é mais cara quanto mais ela faz com que o indivíduo engane-se consigo mesmo e passe a propagar seu autoengano como a maior das verdades. Não conseguimos pensar em exemplo melhor do que aquele fornecido pelos que continuam a suportar uma postura funcionalista e reducionista quanto ao problema mente/corpo (no sentido de IA_2) e, por conta disso, imaginam que a máquina pode de fato pensar e possuir experiências conscientes privadas (o que chamam em filosofia de mente de *qualia*). Felizmente, o melhor antídoto contra essa postura adotada por muitos entusiastas da IA vem pela recordação do que o próprio

Turing escreveu sobre a pergunta "'Can machines think?' I believe to be too meaningless to deserve discussion".[9]

Considerações finais

A gramática da divisão do trabalho estabelece sobre a humanidade (e também sobre os animais e o meio ambiente em geral) uma imposição própria da ordem do discreto, em oposição à complexidade e pluralidade do contínuo de habilidades e expertises existente em qualquer ser humano. Fatia-se a multiplicidade viva de características da inteligência e do corpo humanos para apenas exigir a reprodução de um número específico de rotinas úteis, em geral, à reprodutibilidade técnica do sistema econômico em voga. Tal como discretamente um interruptor passa de ligado para desligado, a demanda de reprodução sistêmica econômica impõe que o trabalhador passe de uma função a outra, ao bel prazer das oscilações do mercado.

Essa demanda sistêmica associa-se adequadamente ao paradigma da medição e quantificação da ciência e tecnologia, pois, frente às oscilações próprias do contínuo da economia e da sociedade, que não se deixa domar pelos modelos discretos, é um imperativo de nosso tempo a ampliação do controle e possibilidade de expansão das fronteiras tecnológicas, de tal modo a transformar a sociedade, idealmente, em uma massa discreta de variáveis controláveis, estruturadas de modo a seguirem as expectativas do jogo econômico. É inegável o papel que IA_1 já apresenta neste cenário.

Os últimos parágrafos foram, em suma, a exposição de nossa chave de análise para a compreensão do sentido de "artificial" enquanto dissimulador (sentido A_2). A hipótese, principal, que defendemos é a de que não se passa do contínuo ao discreto (seja no sentido da ontologia das máquinas, seja no sentido que demos às estruturas econômicas) sem gerar embaraços, desconfianças e autoengano. Não tivemos a pretensão de explicar por completo o porquê de A_2. No entanto, uma coisa é certa: apenas o desprezo humano proveniente de alguma possível raiz psicológica, para usar os termos de Simon, é uma entre uma diversidade de hipóteses.

Enviado: 9 abril 2018

Aprovado: 7 maio 2018

[9] Daí nossa preferência em discutir o artificial em IA e não a inteligência

Referências

BABBAGE, C. *On the economy of machinery and manufactures.* Cambridge: Cambridge University Press, 2010.

BOOLOS, G. et al. *Computabilidade e lógica*. São Paulo: Editora Unesp,2012

CHALMERS, D. The puzzle of conscious experience. *Scientific American* 273(6), p. 80-86, 1995.

FEFERMAN, S. Historical introduction. In: TURING, A. M. *Mathematical logic*. GANDY, R.; YATES, C.E.M. (Orgs.). Amsterdam: Elsevier Science, 2001.

GRIER, D. A. *When computers were human.* Princeton, NJ: Princeton University Press, 2007.

HACKING, I. *Representar e intervir*. Rio de Janeiro: EdUERJ, 2012.

LACEY, H. O modelo das interações entre as atividades científicas e os valores. *Scientiae Studia*, vol. 12, nº. 4, 2014.

SEARLE, J. Minds, brains, and programs. *Behavioral and brain sciences,* 3 (3), p. 417-457. 1980.

SIMON, H. *The sciences of the artificial*. Cambridge, MA: MIT Press, 1996.

SMITH, A. *A riqueza das nações*: investigação sobre sua natureza e suas causas. São Paulo: Nova Cultural, 1996.

TURING, A. Computing machinery and intelligence. *Mind*, vol. 59, issue 236, 433–460. 1950.

Pode uma máquina desejar?

Midierson Maia[1]

Resumo: O presente artigo busca, a partir da célebre pergunta de Alan Turing: "Podem máquinas pensar?", promover um salto de reflexão dentro do tema Inteligência Artificial. O artigo se inicia a partir da análise do texto *Computing Machinery and Intelligence*, publicado por Alan Turing em 1950, e provoca o leitor em direção a uma outra pergunta: "Pode uma máquina desejar?". Considerando a relação possível entre pensamento e linguagem, o texto se desdobra em um inventário acerca dos sentidos das palavras "máquina" e "pensar" dentro do contexto das pesquisas em Inteligência Artificial. A reflexão segue tendo como sua base principal o texto de Alan Turing, mas envereda, de forma interdisciplinar, por referências que incluem obras de Descartes, Lacan e Christopher Bishop. A análise se completa na constatação de que o conceito de "máquina", mencionada no texto de Turing, vem perdendo sentido na medida em que a Inteligência Artificial avança em direção à construção de "entidades artificiais" à nossa imagem e semelhança, incluindo nelas a possibilidade não só de pensar, mas também de desejar.

Palavras-chave: Inteligência Artificial. Desejo. Visão computacional. Linguagem natural.

Abstract: The paper presents an the analysis of Turing's classical question *"Can machines think?"*. It extends this question to the one of: "*Can a machine wish?"*. The purpose is to provide different perspective concerning Artificial Intelligence, which began with Turing's article *Computing machinery and intelligence*, published in 1950. Considering the relationship between thoughts and desires, the study examines the meaning of the concepts of "machine" and "to think" in the context of Artificial Intelligence. Turing's article is the point of departure. The paper advances toward an interdisciplinary approach to the humanities and and computer sciences, including authors such as Descartes, Lacan and Christopher Bishop. The findings of this study suggest that the concept of "machine" used by Alan Turing, is losing its meaning with the advance of Artificial Intelligence toward "artificial entities" that can become our image and likeness. Such entities will be able to not only think, but also to have desires.

Keywords: Artificial Intelligence. Desire. Pattern recognition. Natural language Processing.

Introdução

Publicado por Alan Turing em outubro de 1950, o texto *Computing machinery and intelligence* é atualmente considerado um marco importante da teoria da

[1] Midierson é doutor e mestre em Ciências da Comunicação pela ECA/USP. Fundador da Internucleos Research Community. E-mail: midimaia@gmail.com.br.

computação, inclusive no que se refere à história da Inteligência Artificial. Turing inicia seu ensaio introduzindo a ideia de um jogo chamado *The imitation game*, hoje conhecido como "teste de Turing". Por meio deste teste, Turing buscava deixar claro seu objetivo de verificar a possibilidade de uma máquina "pensar", tal como faz um ser humano.

Turing (1950) argumenta:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think". The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ' machine ' and ' think ' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. (TURING, 2003, p. 433)[2]

O trecho acima deixa clara a preocupação de Turing em estabelecer inicialmente um exercício de reflexão, acerca das palavras *machine* e *think*. Turing, já no início de seu texto, e antes de dar prosseguimento, convida o leitor a empenhar um exercício semântico e filosófico acerca de ambos os termos. Filosófico porque o tema corpo-máquina aparece nos trabalhos do filósofo René Descartes, em especial nos textos intitulados *O discurso do método* (2017 [1649]) e *As paixões da alma* (2017 [1649]).

Segundo Murta e Falabretti (2012, p. 76) a tese mecanicista de Descartes, contextualizada no século XVII, concebe o corpo humano como máquina. Do relógio aos sistemas hidráulicos, as comparações com o funcionamento do corpo humano, feitas por Descartes, ganhavam contornos. Embora seja importante considerar as limitações presentes no século de Descartes, sua tentativa de explicar o corpo humano, a partir de uma referência mecanicista, deve ser respeitada, mesmo havendo atualmente outras visões, contrárias a Descartes.

[2] Eu proponho considerar a questão "As máquinas podem pensar?" Isso deve começar com uma análise do significado dos termos "máquina" e "pensar". As definições podem ser enquadradas de modo a refletir tanto quanto possível o uso normal das palavras, mas essa atitude é perigosa. Se o significado das palavras "máquina" e "pensar" forem encontrados a partir de seus usos, é difícil escapar à conclusão de que o significado e a resposta à pergunta "As máquinas podem pensar?" devem ser procurados em uma pesquisa estatística, como uma pesquisa Gallup. Mas isso é um absurdo. Em vez de tentar tal definição, substituirei a pergunta por uma outra, que esteja intimamente relacionada a ela e expresse em palavras relativamente inequívocas (Tradução nossa).

Indo na contramão do pensamento cartesiano, o médico Randolph Nesse (2016), cientista da Universidade do Arizona, é categórico ao dizer: "The body isn't a machine" e justifica seu ponto de vista dizendo: "Machines are products of design, bodies are products of natural selection, and that makes them different in fundamental ways. The organic complexity of bodily mechanisms is qualitatively different from the mechanical complexities of machines" (NESSE, 2016, p. 1).[3]

Ademais, Randolph completa dizendo que corpos biológicos e máquinas falham por razões distintas. Se as máquinas, conforme as conhecemos, são projetadas por designers e não selecionadas naturalmente, conforme a lógica darwiniana, torna-se difícil estabelecer um paralelo de comparação entre máquina e corpo biológico. Para que pudéssemos validar essa comparação, seria necessário ampliar essa discussão para um campo religioso e acreditar na existência de um possível Deus, o qual poderia fazer o papel de "designer dos corpos biológicos". Obviamente tal objetivo não cabe ao propósito deste texto.

Em outra passagem, o pesquisador Randolph Nesse (2016, p. 1) afirma: "Bodies have parts that may have blurry boundaries and many functions and the parts are often connected to each other in ways hard for human minds to fathom."[4] Esta última frase de Randolph aponta claramente uma postura lógica o suficiente para atingir a tese cartesiana em seu cerne. Não há como negar a evidência de processos sofisticados e inteligentes dentro dos corpos biológicos.

Um bom exemplo de tais processos sofisticados pode ser visto no sistema imunológico de bactérias que combatem vírus a partir do mecanismo chamado CRISPR/Cas,[5] por meio do qual fragmentos de material genético são utilizados nas batalhas travadas pelo sistema orgânico. Para esse tipo de funcionamento, não há uma programação específica, orientando o que uma determinada bactéria ou célula do sistema imunológico irá fazer. O sistema vai se adaptando conforme o contexto, em condições específicas.

[3] Máquinas são produtos do design, corpos biológicos são oriundos da seleção natural. E há uma diferença fundamental entre ambos. A complexidade dos mecanismos presentes nos corpos biológicos é qualitativamente diferente da complexidade dos corpos mecânicos.
[4] Os corpos têm partes que transcendem suas fronteiras funções. E as partes são frequentemente conectadas umas às outras. Isso torna o compreensão do processo difícil para muitas pessoas.
[5] Ver Guimarães (2016).

Independente da sofisticação do corpo biológico, é ainda muito difícil prever a falha total de um corpo biológico, levando o animal a óbito. Isso reforça as palavras de Randolph: "Bodies have parts that may have blurry boundaries."[6] Se, de fato, o corpo biológico fosse uma máquina, como os carros e computares, não dotados de Inteligência Artificial, bastaria trocar-lhes as peças e tudo continuaria bem, sendo sempre possível prever seu funcionamento e vida útil das novas peças inseridas. Máquinas, se paradas, podem durar para sempre, mas corpos biológicos, em estado natural, não podem ou, pelo menos, não ainda.

A partir dos argumentos acima é possível retomar o mesmo exercício semântico e filosófico, lançado por Turing (1950) logo no inicio do texto *Computing Machinery and Intelligence*, porém desdobrando a análise para o segundo conceito: o "pensar". A princípio, seguindo o argumento trabalhado no parágrafo anterior, no qual classifica corpo biológico e máquina como entidades distintas, é possível imaginar que o ato de pensar é uma característica exclusiva ao corpo biológico, pois o exercício do pensamento é não previsível. Não se pode projetar ou programar o pensamento.

Um tear, ou um automóvel, como os conhecemos hoje, não possuem a capacidade de tomada de decisão. E mesmo os chamados *smart cars* podem ser programados, tornando-se previsíveis. Mesmo que o veículo obedeça a uma ordem de programação randômica, ele dificilmente irá apresentar um comportamento inesperado, conforme um contexto, fora do roteiro e do controle dos programadores, a não ser por uma falha de projeto.

Importante ressaltar aqui que a dimensão da Inteligência Artificial não foi ainda incluída na análise. O exemplo citado, referente aos *smart cars*, não inclui *machine learning*, *data driven*, *deep learning* e visão computacional. Logicamente, tratar de tais temas é o objetivo deste texto, porém é preciso preparar a base para a discussão, iniciada a partir do texto de Turing, configurado como marco inicial da computação e da Inteligência Artificial.

Frente aos pontos levantados até o momento, já é possível obter uma conclusão prévia acerca do questionamento de Turing sobre os sentidos das palavras *machine* e *think*. Se o sentido da palavra "máquina" diz respeito a algo mecanicamente

[6] Corpos biológicos possuem partes que transcendem suas fronteiras.

e eletronicamente planejado por designers, isso inviabiliza a compreensão da máquina como um ser pensante, pois o pensamento é imprevisível, dotado de uma dinâmica que foge ao controle dos seres vivos.

Um bom exemplo de tal imprevisibilidade pode ser visto no campo da religião. Por mais fanática que possa ser uma pessoa religiosa, não há garantias explícitas de que ela irá manter eternamente fidelidade à causa religiosa. Tudo vai depender de um contexto bem maior, o qual envolve, em uma trama de comunicação e linguagem, todos os demais indivíduos presentes no contexto. Portanto, máquinas não pensam. Máquinas são máquinas, e não podem ser confundidas com corpos biológicos, ao menos sob a perspectiva da análise aqui empreendida.

A Inteligência Artificial e seus desdobramentos possíveis

Ampliando a discussão trabalhada por Alan Turing, no primeiro parágrafo do texto *Computing Machinery and Intelligence*, é possível afirmar que o autor, ao lançar a célebre pergunta *Can machines think?* lançaria também um desafio ontológico, pois, antes de concluir ser possível uma máquina de fato pensar, foi preciso buscar compreender melhor a raiz da discussão, a partir de uma melhor reelaboração dos sentidos inscritos nas palavras envolvidas.

Isso leva a crer que, para compreender o funcionamento e os possíveis desdobramentos da Inteligência Artificial, em especial no século XXI, é preciso abandonar a clássica ideia de "máquina" em si, pois, enquanto "máquina", ela jamais poderá "pensar". As ciências da linguagem estendidas à Psicanálise mostram que quando um sentido escapa é preciso reordenar a ordem da significação, visando reconstruir um significado para a essa "presença de uma ausência"[7] ou, em outras palavras, um significado para o vazio.

Isso acontece no jogo *Tabu,*[8] se a palavra "máquina" for eliminada do contexto das discussões envolvendo Inteligência Artificial, o que fica? Como tratar o tema Inteligência Artificial sem ao menos mencionar tal palavra? A verdade é que, se

[7] O termo "presença de uma ausência" é muito comum na Psicanálise e se liga ao conceito de "falta". A concepção do ser humano como um ser faltante, que sempre está em busca de algo que na infância se perdeu, exerce papel central na Psicanálise e ajuda a compreender a natureza do desejo, que só existe em função da sinalização de uma falta e também de um objeto correspondente à essa falta.

[8] O Tabu é um jogo composto por algumas cartas, nas quais há um algo a ser adivinhado pelos demais jogadores. Quando um jogador tira uma carta, ele deve dar pistas aos demais sobre o que ele vê na carta, mas não pode usar uma lista de palavras-chave sobre o objeto impresso na carta.

máquinas não pensam, passam a não servir, pelo menos de forma profícua, para discussões.

Será preciso então buscar um novo modelo ontológico. Um modelo dotado de Inteligência Artificial e que de fato possa "pensar" de forma autônoma e aleatória, exatamente como fazem os corpos biológicos, mediante não somente o cérebro, mas mediante sua totalidade, incluindo o sistema imunológico.

No cenário das discussões sobre Inteligência Artificial, *deep learning* e computação cognitiva, termos como androide e mutantes (seres híbridos) parecem fazer muito mais sentido do que a palavra "máquina". A partir do momento em que um corpo biológico passa a receber um marca-passos, ou uma prótese integrada ao sistema nervoso central, recebendo dele comandos (do pensamento) resultantes em movimentos, a ideia de máquina, em si, se perde no horizonte.

Quando um dispositivo é acoplado, de forma sincronizada aos comandos do corpo biológico, o conceito de máquina desaparece na medida em que o dispositivo passa a fazer parte de um todo, composto de diferentes processos que asseguram diferentes tipos de vida. Esses dispositivos, como um marca-passos, passam a ocupar as funções dos órgãos e até de células, se considerados os nano robôs, realidade já presentes em laboratórios de grandes centros de pesquisa, como os de Stanford e MIT.

Se uma entidade, feita de matéria inorgânica e orgânica, dotada de Inteligência Artificial, com alto poder de processamento a partir de redes neurais, reconhecimento de padrões e *deep learning*, atingir um nível de processamento próximo à linguagem natural, estaremos à frente de um novo ser. Trata-se de uma nova espécie, capaz sim de pensar. E mais, capaz de sentir e (por que não?) desejar!

Enfim, depois de percorrido o percurso de reflexão sobre o sentido dos termos "máquina" e "pensar", tem-se aqui o ponto central da discussão: "Can a machine wish?" Poderá, um dia, uma máquina desejar? Em se tratando de máquinas, certamente não, pois pensamento e desejo estão intimamente ligados. Sendo máquina, não há definitivamente pensamento. Não havendo pensamento, não haverá linguagem. Não havendo linguagem, não haverá desejo. Mas, ao aceitar a possibilidade de superação da ideia de máquina e a abertura para o reconhecimento da "nova entidade", dotada de

Inteligência Artificial, *deep learning* e visão computacional, tudo muda e novas possibilidades se abrem em um horizonte ainda desconhecido.

Um bom exercício de reflexão sobre possibilidades ligadas a uma nova entidade, presente no cenário da tecnologia, pode ser obtido a partir do projeto *Sophia*[9], da *Hanson Robotics* (figura 1).

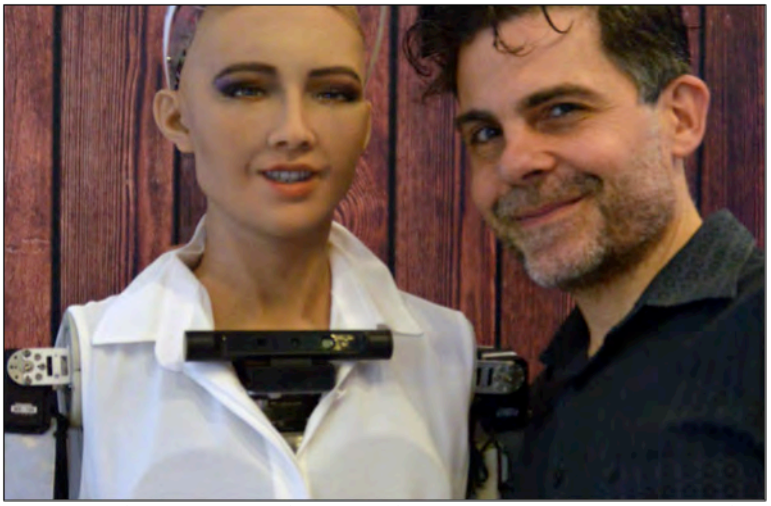

Figura 1. Sophia e seu criador, David Hanson. **Fonte:** Hanson Robotics.

Sophia é um dos mais avançados humanoides já criados. Seu criador David Hanson não poupa esforços para fazer de *Sophia* seu principal ativo, estrelando como uma espécie de garota propaganda de sua empresa. Mas o fato é que, muito além de garota propaganda, *Sophia* representa um avanço considerável na tecnologia de Inteligência Artificial, esbarrando na tênue fronteira que divide os conceitos vida animada e vida não animada.

A matéria do jornal *O Globo,* de 26 de outubro de 2017, intitulada "Sophia é o primeiro robô do mundo a receber um título de cidadania"[10] fala em "direitos robóticos". É sabido que o direito em geral foi criado para garantir que determinadas condições humanas de igualdade e justiça pudessem ser resguardadas em um código. Não há normas jurídicas voltadas exclusivamente às relações estabelecidas entre coisas, como máquinas, paus e pedras. As normas jurídicas foram criadas para determinarem certos limites no uso de coisas inanimadas para alguns fins, como não matar ou apedrejar alguém.

[9] Ver mais em: <http://www.hansonrobotics.com>. Acesso em: 18 mai. 2018.
[10] Disponível em: <https://oglobo.globo.com/sociedade/tecnologia/sophia-o-primeiro-robo-do-mundo-receber-um-titulo-de-cidadania-21996085>. Acesso em: 18 mai. 2018.

Já com *Sophia* é diferente. Em uma entidade artificial, cujo comportamento pode ser imprevisível. Não há garantias em relação a qual uso ela fará de paus e pedras, porque agirá conforme o contexto e referenciais recebidos. Então, o tema "direitos robóticos" se mostra pertinente e necessário, frente ao cenário que se coloca, onde entidades dotadas de Inteligência Artificial emergem rapidamente, revelando uma situação apavorante e ao mesmo tempo excitante, frente às inúmeras possibilidades de usos da Inteligência Artificial.

As primeiras aparições de *Sophia* em jornais de todo o mundo despertaram, e ainda despertam, um certo pânico em grande parte das pessoas. Há um medo referente à perda de controle dessas máquinas, a ponto delas se rebelarem contra os seres humanos, causando uma matança em massa, conforme mostram os filmes e séries distópicas, a exemplo do clássico hollywoodiano *O Exterminador do Futuro*, e também da série *Black Mirror*, produzida pela Netflix.

Frente à percepção desse sentimento de medo, torna-se importante lançar sobre esse sentimento um olhar analítico, buscando compreender o que está por trás do medo da distopia. Em uma primeira leitura, com base por exemplo na Psicanálise, é possível perceber que o medo de grande parte das pessoas, em relação à Inteligência Artificial, está ligada ao medo da possibilidade de as máquinas, de fato um dia, "desejarem" destruir a humanidade. Mas poderiam as máquinas desejarem salvar a humanidade? Com base em pesquisas e projetos avançados no campo da robótica e Inteligência Artificial, a exemplo do projeto *Sophia*, é possível dizer que ambas as perspectivas podem, em um futuro próximo, serem concretizadas.

No filme *O Exterminador do Futuro II* o mesmo avatar, encenado por Arnold Schwarzenegger no primeiro filme, volta para defender John Connor. O filme apresenta várias situações nas quais é possível observar "máquinas" (entidades) inteligentes tomando decisões sofisticadas, muito próximas às decisões tomadas pelos humanos Sarah Connor e seu filho, John Connor. Ao longo de toda narrativa é possível observar vários momentos em que humanos e entidades artificiais (concebidas no filme como máquinas) aparecem entrelaçadas por dimensões simbólicas e imaginárias, necessárias à manutenção da cultura e suportes para a construção de desejos.

No mesmo filme, John estabelece uma relação de amor com a entidade T800 (exterminador), interpretada por Schwarzenegger, e chora quando, a partir de uma atitude altruísta, o robô decide se exterminar em uma caldeira de metal líquido. É importante ressaltar que o suicídio é um ato particular dos seres humanos. Oriundo de um conflito com a realidade, a partir da qual um sujeito não recebe os sentidos necessários à construção de significados para sua vida, o desejo pelo suicídio surge a partir de um distúrbio simbólico e ao mesmo tempo imaginário. Então, se uma "máquina", aqui tratada com entidade artificial, com base em uma decisão não programada, decidir um dia se autoextinguir estará essa entidade muito próxima à natureza humana.

É fato que a discussão acima surge com base em possibilidades, não em fatos concretos, mas é possível perceber o quanto todo desenvolvimento da Inteligência Artificial caminha para esse fim. As pesquisas envolvendo *deep learning* e visão computacional, baseadas em reconhecimento de padrões, buscam contribuir para o desenvolvimento de entidades capazes de pensar e sentir como seres humanos. As demonstrações da entidade *Sophia*, feitas pelo seu criador David Hanson, mostram o quanto o desenvolvimento de seu protótipo caminha em direção à construção de um "ser humano artificial", que não pode mais ser compreendido simplesmente como máquina.

Seguindo em direção à análise dos desdobramentos possíveis da Inteligência Artificial, é importante lembrar que a maioria das tecnologias construídas pelo ser humano possuem inicialmente um estado de nulidade. Isso significa dizer que os impactos das tecnologias criadas e desenvolvidas irão depender do uso a elas destinado. É fato que a mesma coisa ocorre com a Inteligência Artificial. É importante lembrar que, embora não exatamente da mesma forma com que é colocada, a distopia presente no filme *O Exterminador do Futuro* e também na série *Black Mirror*, pode ser possível e materializável. Tudo vai depender de como e por quem essa tecnologia será manipulada.

Intitulado *Volto já*[11] (Figura 2), o episódio de *Black Mirror,* apresenta um contexto distópico, em que a personagem Martha encomenda, a uma empresa de

[11] Título original: *Be Right Back*.

tecnologia, uma "cópia" de Ash, seu namorado, morto em um acidente de trânsito. Com base nos dados de navegação de seu marido nas redes sociais, a empresa envia à casa da cliente uma entidade, dotada de Inteligência Artificial, contendo dentro de si um software capaz de simular as ações de seu namorado nas redes sociais e também na vida real.

Figura 2. Martha e Ash em cena do episódio *Volto Já*, da série *Black Mirror*. **Fonte:** Netflix.

No episódio, uma espécie de avatar de Ash sente medo, frio e desejo. Nesse caso, a construção das condições necessárias à formação do desejo só foi possível a partir da obtenção de uma cópia das ações (de linguagem) de Ash nas redes sociais. É importante lembrar que, na série, essas ações de linguagem são reflexos de ações de Ash na vida real.

A Inteligência Artificial, nesse caso, é o que possibilita a construção do avatar de Ash, pois é a partir do *machine learning*, associado a outros recursos de IA, que se torna possível chegar próximo à linguagem natural, em computação, estudada como *NLP* (*Natural Language Processing*).

Considerados alguns exageros da ficção, o quadro representado já pode ser evidenciado, em estágio inicial, em *chat bots* e *social bots*, que aprendem conforme o contexto das interações de pessoas em redes sociais. O termo *bot*, vem do processo de um clipping inicial (ou truncamento) da palavra *robot*. O termo não é novo. Após ser

aplicado ao contexto das redes sociais, ganhou um novo sentido. Assim como os demais *softwares*, que incluem em seu funcionamento algoritmos, cada *bot* possui uma função e finalidade. Há *bots* feitos para usos legais, como os *chatbots*, usados por empresas para se comunicar com seus clientes. Mas há também os *social bots*, usados para simular outros tipos de ações humanas em redes sociais. Um dos problemas desse tipo de algoritmo está no impacto dele no comportamento dos indivíduos.

Os *social bots* não são usados para realizar um trabalho de agenciamento do conteúdo nas redes, mas para enganar o internauta. Ele se passa muitas vezes por humano produzindo um certo tipo de conteúdo. Embora os *social bots* não realizem agenciamentos de conteúdo, podem sofrer, em algum momento, a influência dos algoritmos curadores.

Ao impactar, a partir de outros conteúdos, o conteúdo produzido por humanos, os *social bots* podem alterar os sentidos dados nas interações. Mas, para que haja um impacto considerável, é necessário haver um ganho na escala de influência desses *bots*. E o ganho de escala vai depender do números de *social bots* controlados pelos *botmasters*, humanos que espalham e controlam um grande número de *social bots* nas redes.

Quando se pensa no teste de Turing e em sua proposta, é possível dizer que, ao menos pela perspectiva da análise aplicada aos *social bots* e *chat bots*, esse teste já foi vencido por esses "atores" que simulam ações humanas na rede. No caso dos *social bots*, essas ações buscam, e muitas vezes conseguem, ludibriar consumidores e eleitores em todo mundo.

E o caso da *Cambridge Analytica*, envolvendo roubo de dados e perfis falsos no Facebook, passou a ser emblemático para se pensar em qual medida o teste de Turing mostra o quanto um algoritmo pode ser capaz de enganar um ser humano. Sob essa perspectiva, pelo menos, é possível dizer que o lado artificial já deu vários indícios de que é capaz de vencer, no teste de Turing, não somente uma dezena de pessoas, mas milhões delas.

Entidades inteligentes, artificiais e cognitivas

A Psicologia da Aprendizagem e a Neurociência mostram que nosso cérebro, a partir de dados armazenados em nossa memória, funciona mediante reconhecimento de padrões, registrados anteriormente com a ajuda de experiências contextuais, vividas pelos indivíduos em seu entorno. A Linguística é outra vertente que contribui para compreender como um determinado sujeito, a partir do exercício da linguagem, concebe sua realidade pela via das operações simbólicas, que irão suportar sua presença no mundo enquanto ser falante, com demandas e desejos inerentes ao seu perfil.

Indo além, é possível se beneficiar da psicanálise (LACAN, 1956) para a compreensão de como determinados padrões de conduta influenciam na operacionalização de alguns modelos de comportamentos, desejos e fantasias, cristalizados no social. É importante ressaltar que, nas relações sociais, desejos, fantasias e comportamentos surgem a partir de uma estrutura simbólica, prévia aos indivíduos e composta por alguns padrões, vistos como normas sociais e valores, servindo de suporte para que os sujeitos possam conceber uma noção possível de realidade.

O desenvolvimento da Inteligência Artificial, aplicada às técnicas de visão computacional, segue no sentido de imitar o funcionamento do cérebro humano não somente em operações mais simples, como cálculos matemáticos, mas também segue no sentido de simular a forma com a qual aprendemos e apreendemos a realidade. Sendo assim, é possível dizer que o despertar de uma entidade artificial desejante pode se materializar na medida em que a entidade, dotada de Inteligência Artificial plena, for capaz de reconhecer e processar padrões através de signos, sentidos e significados, como fazem os humanos.

Uma simples busca no *Google Acadêmico* mostra que o termo *pattern recognition* ganha escala e velocidade em artigos científicos e hoje se faz presente muito mais em textos ligados à área da ciência da computação e robótica do que em textos ligados à área da Psicologia da Aprendizagem. Isso mostra o quanto esse conceito vem sendo empregado nos estudos aplicados à Inteligência Artificial. Os

avanços nessa direção têm despertado interesse de alguns pesquisadores das Ciências Humanas, como a Filosofia, a Psicologia e a Psicanálise.

Christopher Michael Bishop, cientista da computação e professor de ciências da computação da Universidade de Edimburgo, dedicou uma importante obra acerca do tema "reconhecimento de padrões" e *machine learning*. Sua obra mais importante é intitulada *Pattern Recognition and Machine Learning* (BISHOP, 2006). Ela mostra o quanto a comunidade acadêmica, aliada a grandes empresas de tecnologia, vem dedicando tempo e recursos para criar sistemas cada vez mais avançados em visão computacional envolvendo reconhecimento de padrões.

Ampliando essa discussão para uma dimensão tácita, é possível mencionar casos nos quais a Inteligência Artificial já pode reconhecer, mediante nível profundo, detalhes a partir de imagens gráficas que trazem em si signos associados a recortes de realidade previamente construída pelos seres humanos.

Figura 3. Anúncio do aplicativo Seeing IA.
Fonte: <https://www.microsoft.com/en-us/seeing-ai>.

Um dos exemplos que ilustram o poder da visão computacional, mediante reconhecimento de padrões, é o projeto *Seeing AI* (Figura 3) da Microsoft. Destinado a deficientes visuais, o aplicativo foi feito para reconhecer diversos tipos de objetos à frente do observador. Basta ativar o aplicativo e selecionar quais funções você deseja usar. É possível também utilizar o aplicativo para ler textos em documentos.

Considerando, grosso modo, que as ações do sistema operacional de uma máquina possa ser equiparado às ações de nosso cérebro, temos então, embora ainda

em estágio muito inicial, o despertar das condições para a criação de um sistema computacional inteligente, a ponto de, em um futuro próximo, compreender e processar signos. Estaríamos então frente a um novo cenário no qual entidades artificiais poderão se tornar entes capazes de não somente imitar os seus criadores em tarefas elementares, mas entes capazes de processar a realidade de forma muito próxima aos humanos.

O artigo *Artificial Intelligence and Sign Theory* (1989), de Jean Guy Meunier apresenta um louvável esforço em abordar a Inteligência Artificial de forma interdisciplinar, pela via da filosofia e da semiótica. Meunier ressalta a importância do cruzamento interdisciplinar e apresenta dados necessários que mostram o quanto a Inteligência Artificial deixou de ser um tema exclusivo ao campo das ciências exatas. O autor ressalta também a importância de se estudar a "manipulação" dos signos que participam de processos de comunicação entre homens e entidades artificiais, compreendidas pela maioria como "máquinas". Meunier (1989) completa:

> What is the interpretation to be given to these symbols that are "manipulated" by a computer in an AI system. The answer her seems to be related to the representational function these symbols play in the processing system. According to Haugeland (1986: 28) they represent something of the outside world or to Newell and Simon (1976) an intern process of some kind by which some action is undertaken. (MEUNIER, 1989, p. 8)[12]

Ao impactar o universo da comunicação e da linguagem a Inteligência Artificial, a partir de algoritmos que simulam ações humanas de comunicação no ambiente digital, desdobra-se nas ações concretas do dia-a-dia, pois cada vez mais é possível perceber que usuários de redes sociais não percebem o quanto estão sujeitos às ações dos *social bots* e *chat bots*. Ao se envolverem na malha simbólica da linguagem, os *bots* participam dos atos de fala, se envolvendo na construção e reconstrução da noção de realidade apreendida pelos humanos. Certamente esse é um caso a ser tratado pela vertente da semiótica da cultura.

[12] Qual é a interpretação a ser dada a esses símbolos que são "manipulados" por um computador em um sistema de IA? A resposta dela parece estar relacionada à função representacional que esses símbolos desempenham no sistema de processamento. De acordo com Haugeland (1986: 28) eles representam algo do mundo externo ou de Newell e Simon (1976) um processo interno de algum tipo pelo qual alguma ação é empreendida.

Considerações finais

Após analisar dados do atual contexto de pesquisas científicas e desenvolvimento de aplicações em Inteligência Artificial, é possível dizer que as possibilidades da Inteligência Artificial, analisadas sob uma perspectiva de desenvolvimento pleno, são inúmeras e certamente irão transformar não somente a forma com a qual os indivíduos utilizam as ferramentas de trabalho. A Inteligência Artificial em desenvolvimento pleno poderá transformar ainda mais a cultura, a sociedade e as formas de relações sociais.

O desenvolvimento da Inteligência Artificial plena já encontrou seu norte, seu objetivo, sua missão: construir seres à nossa imagem e semelhança, que possam executar tarefas distintas, similares à ação humana, mas com um ganho de vantagem muito grande para algumas tarefas, como reconhecer o rosto de uma pessoa em meio a uma multidão. É fato que a visão computacional já provou ter esse poder.

A maioria das notícias que circula pela mídia, sobre Inteligência Artificial, em diferentes meios de comunicação, tem servido muito mais para informar e tornar públicas as ações de importantes *startups*, a exemplo da Hanson Robotics. Toda essa informação gerada e repetida na mídia se esgota muitas vezes em um debate vago sobre as aplicações práticas e de mercado envolvendo a Inteligência Artificial.

Pensar nos desdobramentos de tecnologias como a Inteligência Artificial, Computação Quântica e Engenharia Genética é algo que vem sendo feito pela academia, a exemplo do grupo de pesquisadores do TIDD, da PUC-SP. Pensar em desdobramentos é lançar um olhar em direção ao futuro, mas não somente de forma fictícia, para produzir filmes e séries distópicas. Pensar em desdobramentos possíveis é avaliar, de forma prática e científica, os impactos das tecnologias na sociedade.

Em outras palavras, é possível dizer que é preciso trazer o debate para o campo das ciências humanas. Estudar as tecnologias disruptivas à luz de vertentes como antropologia, semiótica, psicanálise, filosofia e sociologia significa ampliar o debate para níveis que englobam a subjetividade. Ao ampliar a discussão para esses níveis, abre-se uma possível compreensão de embates entre desejos humanos e certos tipos de desejos envolvendo entidades artificiais, que logo deixarão de ser máquinas para se

tornarem espécies, como visto na série *West World*, da *HBO*, em que há um claro embate entre os desejos dos robôs e os desejos dos seres humanos.

Em síntese, é possível afirmar que um futuro repleto de entidades artificiais, dotadas da capacidade de desejar, só será possível se essas entidades atingirem um estágio avançado no qual passarão a ser capazes de processar sentidos, significado e significantes, atuando de forma paralela, mas conjunta, ao processo humano de produção de significados em rede.

Ainda sobre a possibilidade da existência de desejos nessas entidades, é importante ressaltar que toda formação de desejo, dada em nível humano, segue uma lógica de objeto. A Psicanálise oferece subsídios para a compreensão da natureza do desejo e ensina, a partir dos trabalhos de Jacques Lacan (2009 [1956]), que um desejo existe em função de um objeto, seguido pela percepção de falta desse objeto, previamente carregado de sentidos, envoltos em relações entre significantes.

Quando a Inteligência Artificial atingir seu desenvolvimento pleno, possibilitando o nascimento de uma entidade artificial capaz de manifestar demandas de sentido a partir de operações significantes, estaremos finalmente frente a frente a uma resposta mais clara para a pergunta "Pode uma máquina desejar?". Por enquanto, tudo o que há é a pavimentação de uma longa avenida, com percurso e objetivo bem definido, indo em direção à criação de um ser à nossa imagem e semelhança, capaz de pensar e, sobretudo, desejar.

Enviado: 19 fevereiro 2018
Aprovado: 20 março 2018

Referências

BISHOP, Christopher. *Pattern recognition and machine learning*. New York, NY: Springer, 2006.

DESCARTES, René. *O discurso do método*. São Paulo: Martins Fontes, 2009.

DESCARTES, René. *As paixões da alma*: grandes obras do pensamento universal. São Paulo: Lafonte, 2017.

GUIMARÃES, Maria. Uma ferramenta para editar o DNA. Pesquisa Fapesp, v. 240, p. 38-41, 2016.

HELLER, Agnes. *O quotidiano e a história*. Rio de Janeiro: Paz e Terra, 2008.

LACAN, J. A instância da letra no inconsciente ou a razão desde Freud. In: *Escritos*. Rio de Janeiro: Jorge Zahar, p. 493-533, 1998.

_____. Função e campo da fala e da linguagem em psicanálise. In: *Escritos*. Rio de Janeiro: Jorge Zahar, p. 237-324, 1998.

_____. Simbólico, imaginário e real. In: *Os nomes do pai*. Jorge Zahar, Rio de Janeiro: Zahar, p. 9-54, 2009(1956).

LATOUR, Bruno. On recalling ANT. In: LAW, John, HASSARD John. (Org.). *Actor-network and after*. Oxford: Blackwell, p. 15-26, 1999.

MEUNIER, Jean Guy. Artificial intelligence and the theory of signs. *Semiotica*, vol. 77, p. 43-63,1989.

MURTA, Claudia; FALABRETTI, Eric. O autômato: Entre o corpo máquina e o corpo próprio. *Natureza humana*, v. 17, nº. 2, p. 75-92, 2015.

PRIMO, Alex. *Interação mediada por computador*: comunicação, cibercultura, cognição. Porto Alegre: Sulina, 2007.

Saad Corrêa, Elizabeth; Bertocchi, Daniela. A cena cibercultural do jornalismo contemporâneo: web semântica, algoritmos, aplicativos e curadoria. *Matrizes*, v. 5, p. 123-144, 2012.

TURING, Alan. Computing machinery and intelligence. *Mind*. New Series, vol. 59, no. 236, p. 433-460, Oct. 1950.

resenha

HARARI, Yuval Noah. Homo Deus: uma breve história do amanhã. São Paulo: Companhia das Letras, 2016. 443 p. Resenha por Rodrigo Petronio. *Teccogs: Revista Digital de Tecnologias Cognitivas*, TIDD | PUC-SP, São Paulo, n. 17, p. 146-156, jan-jun. 2018.

Homo Deus: uma breve história do amanhã

Resenha por Rodrigo Petronio[1]

Livros e espelhos

Yuval Noah Harari é fenômeno dos mais interessantes da produção intelectual contemporânea. Doutor por Oxford e professor da Universidade de Jerusalém, o autor de *Sapiens: uma breve história da humanidade*, um dos maiores *best sellers* de divulgação de ciência da atualidade, não deixa de provocar algumas experiências intelectuais contraintuitivas, mesmo para os *scholars* mais exigentes e os leitores especializados. E, em alguma medida, vai além: propõe algumas teses que podemos considerar como ousadas e mesmo originais tanto no âmbito da bioantropologia quanto na área de historiografia e da filosofia contemporânea. A partir de uma escrita que narra o passado e o futuro do *sapiens* sob a forma de *storytelling*, Harari não poupa expor suas intuições polêmicas e abordar temas de extrema complexidade, como a ontogênese da consciência e o impacto da especiação iminente, quando o *sapiens* deve se bifurcar em um novo hominídeo.

Em *Homo Deus: uma breve história do amanhã*, os argumentos centrais desenvolvidos em *Sapiens* são retomados. Agora projetados como cenários futuros para a humanidade, tendo em vista os avanços da biotecnologia, da Inteligência Artificial e, sobretudo, da onipresença e da onisciência dos algoritmos. O papel nuclear desempenhados pelas ficções em *Sapiens* é protagonizado pelos algoritmos em *Homo Deus.* Algoritmos e ficções: eis as pedras angulares do pensamento de Harari. As ficções não são o oposto da realidade, mas a cola mítica e o horizonte de emergência de toda a odisseia do *sapiens*, bem como o fator decisivo que distinguiu o *sapiens* das demais espécies. Todas as instituições, das religiões às ciências, das filosofias à política, dos estados à jurisdição, todas as produções humanos são epifenômenos ficcionais e por meio desse tecido intersubjetivo da ficção, diferente dos primatas superiores, conseguimos produzir ações à distância (Sloterdijk), comunidade imaginadas,

[1] Rodrigo Petronio é escritor e filósofo, autor e organizador de diversos livros. Professor Titular da FAAP e pesquisador de pós-doutorado no Programa de Tecnologias da Inteligência e Design Digital (TIDD | PUC-SP), sob supervisão de Lucia Santaella.
E-mail: rodrigopetronio@gmail.com.

realidades expandidas e cooperações em larga escala. Por isso, a ficção é o conceito-chave para compreendermos não apenas o que nos aguarda no século XXI. Ela é a chave de acesso a cenários realistas para o terceiro milênio. A divisão que deve surgir dessa especiação do *sapiens* em novas linhas de hominídeos não deve dizer mais respeito apenas a divisões socioeconômicas, relativas às classes. A divisão deve gerar uma distinção antropotécnica (Sloterdijk), ou seja, simultaneamente antropológica e ontológica, entre duas castas de humanos: a dos sequenciados pela biotecnologia e a dos filhos da natureza, do acaso, do amor, de Deus ou de outra mitologia envelhecida. Estes formarão as grandes hordas e as hostes dos excluídos dessas novas cosmopolíticas (Stengers), à frente da Linha de Pobreza Biológica (LPB).

Ambos os livros são complementares e se espelham. Em *Sapiens*, Harari pretende mapear a narrativa do humano a partir dos três imperativos que o guiaram até os dias de hoje: a fome, a peste e a guerra. Já em *Homo Deus*, abrem-se horizontes para a exploração especulativa (Whitehead) que deve emergir dos três novos imperativos que devem fundamentar a agenda dos hominídeos daqui pra frente: a imortalidade, a felicidade e a divindade. O plano do livro *Homo Deus* tem como eixo três questões. Primeira: como o *sapiens* se tornou o que é. Segunda: como o humanismo se tornou a religião dominante no mundo. Terceira: por que a tentativa de concretizar o sonho humanista deve paradoxalmente conduzir o humano à sua desintegração em um futuro iminente. Neste texto, pretendo fazer um breve apanhado das principais respostas dadas por Harari a essas questões, tendo em vista os pressupostos lançados em *Sapiens* e expandidos em *Homo Deus.*

Imperativos e agendas

Harari se vale muito de números para essa argumentação em defensa de um declínio da peste, da fome e da guerra em termos absolutos, observado sobretudo nos últimos três séculos, que convencionamos nomear como modernidade. A presença das epidemias na expansão e nos domínios de umas civilizações sobre outras foi estudada com esmero por Jared Diamond em um clássico da bioantropologia. A tese de Harari se assemelha àquelas defendidas por outros pensadores que veem o processo de modernização como um dos agentes mais importantes para a erradicação da miséria e

para a supressão desses antigos imperativos da peste, da guerra e da fome, que têm se tornado aos poucos obsoletos. Nesse sentido, aproxima-se do conceito de sociedade da abundância (*affluent society*) de John Kenneth Galbraith, e das obras de Norbert Elias, Steven Pinker, Francis Fukuyama, Peter Sloterdijk e outros autores que não minimizam o papel desses três agentes humanos, mas defendem a sua diminuição exponencial nas sociedades modernas. O grau de incidência desses três fatores é redimensionado quando abandonamos comparações de pequena escala e abordamos esses fenômenos a partir das cesuras de longa duração, fornecidas pelas metanarrativas e por meio de abordagens do *sapiens* em uma perspectiva evolucionista de larga escala.

Por exemplo, no século XIII, por causa da peste negra, 75 milhões do total de 200 milhões de pessoas morreram mais de um quarto de toda população da Eurásia. Na Inglaterra, 4 de cada 10 pessoas pereceram pelo mesmo motivo. A população caiu de 3,7 milhões para 2,2 milhões. A cidade de Florença perdeu 50 mil habitantes, simplesmente metade do total de sua população. A partir das navegações e expansões do século XIX, a situação não melhorou. O capitão Cook e sua tripulação introduziram patógenos da gripe, da tuberculose e da sífilis no Havaí. Visitantes europeus introduziram a tifo e a varíola. Em 1853, restavam apenas 70 mil habitantes na ilha. No começo do século XX, a gripe espanhola infectou meio bilhão de pessoas, um terço da população global do planeta. Esta mesma gripe dizimou 5% de toda população (15 milhões de pessoas). Por causa de vírus inoculados por estrangeiros, Taiti e Samoa perderam, respectivamente, 14% e 20% de sua população.

Como as guerras e a fome, as pestes e epidemias também têm sua linha decrescente nos séculos XX e XXI, mesmo em um momento humano em que deveria haver maior vulnerabilidade a epidemias por causa do aumento da comunicação e dos meios de circulação. Contudo, hoje a imensa maioria das pessoas morre de doenças não infeccionas, como câncer, doenças cardiovasculares ou simplesmente de velhice. Claro que convivemos cada vez mais com o aumento das superbactérias, produzidas justamente pelo uso irrestrito de combates a doenças transmissíveis. Os microrganismos têm 4 bilhões de anos de experiência acumulada lutando contra inimigos orgânicos, mas sua experiência é nula no combate a predadores biônicos e a

novas bactérias sintetizadas em laboratório. A hipótese de Harari entretanto é a de que não vivemos uma iminência de guerras bacteriológicas ou químicas, como preveem outros futurologistas, como Jacques Attali. Essa globalização do mundo produziu um sistema global de cofragilidade, como diria Sloterdijk. A membrana do planeta está interconectada por malhas radiculares. O equilíbrio entre violências torna-se cada vez mais sutil e fino. A violência contra o outro e violência contra si acabaram encontrando uma estranha homeostase.

Com relação à fome os dados também são impactantes. Aproximadamente 2,8 milhões de franceses (15% da população) morreram de fome entre 1692 e 1694 enquanto Luis XIV, o Rei Sol, flertava com suas amantes. As regiões do globo onde ainda existem ondas maciças de fome vivem estas situações mais em decorrência de problemas políticos locais do que por conta de catástrofes naturais ou da escassez de alimentos. Em 1974, a informação era de que seria impossível alimentar 1 bilhão de pessoas que viviam na China. Desde então, milhões de chineses têm sido resgatados da fatalidade da fome. E a fome tem sido erradicada, tanto na China quanto em outros países em desenvolvimento. Hoje a fome e a subnutrição matam cerca de 1 milhão de pessoas ao redor do mundo. Ao passo que a obesidade mata 3 milhões. Na Idade da Pedra, um ser humano médio tinha 4 mil calorias por dia a seu dispor. Um americano médio hoje usa 228 mil calorias por dia para o corpo, o carro, a televisão, o computador, a geladeira. Usa 60 vezes mais calorias do que um caçador-coletor. Não bastassem esses dados, entre 1950 e 2000 o PIB americano cresceu de 2 trilhões para 12 trilhões. Um salto de seiscentos por cento. Daqui se segue a anedota repetida por Harari: o açúcar é mais perigoso que a pólvora.

E os perigos inerentes à natureza humana? Em relação às guerras e às diversas formas de violência que conduzem à mortalidade, Harari também insiste em relativizar os números atuais. Enquanto nas antigas sociedades agrícolas a violência humana foi causa de 15% de todas as mortes, durante o século XX a violência provocou apenas 5% das mortes. No século XXI, essa mesma violência é responsável por apenas 1% da mortalidade global. Desde a Idade da Pedra à era do vapor, das tribos do Ártico às tribos do Saara, durante dezenas de milhares de anos cada pessoa da Terra sabia que a qualquer momento seus vizinhos poderiam invadir seu território, derrotar seu exército,

chacinar a sua população e ocupar sua propriedade. A guerra foi um imperativo durante milênios e estava atrelada aos princípios da economia política do *sapiens.* O *homo sapiens* foi por isso por milênios um *homo necans* (homem matador), como Walter Burkert o demonstra em seu brilhante estudo. Semelhante a Sloterdijk, a Michio Kaku e a outros pensadores, para Harari a guerra, como a entendemos, está se tornando obsoleta e pouco lucrativa. Hoje a maior fonte de riqueza seria o conhecimento. A diminuição da violência se deve à ascensão do Estado. Mesmo a chamada paz atômica (Jean-Pierre Dupuy), defendida pelos teóricos da dissuasão nuclear, segundo a qual quanto mais países possuam energia nuclear, menor e a probabilidade de seu uso, seria um paradoxo nesse debate acerca da violência. Depois das eras da terra, dos territórios e das mercadorias, estaríamos ingressando na era do saber (Pierre Lévy). Em virtude do declínio desses imperativos que regularam o *sapiens* durante milênios, a humanidade estaria adotando uma nova agenda, fundada sobre três novos imperativos: a felicidade, a imortalidade e a divindade. À medida que a peste, a fome e a guerra são aos poucos dirimidas ou controladas, o que então poderia obstruir a emancipação humana em direção à conquista desses nobres ideias? Em *Homo Deus*, Harari se concentra em fazer o levantamento de todos os paradoxos e contradições presentes em cada um dos três imperativos dessa nova agenda.

Imortalidade e divindade

A questão da imortalidade surge diretamente de pesquisas cada vez mais comuns e efetivas das ciências da natureza: a capacidade de gerar seres amortais. Não se trata de seres imortais, pois serão organismos que poderão morrer de causas acidentais. Mas não por desgaste biológico ou por envelhecimento. A morte e o processo de decomposição das células não estarão inscritos nos seus tecidos vivos. Como diria o físico Richard Feynman, não existe nada no conhecimento da Biologia que negue a possibilidade de se erradicar a morte. No mesmo sentido, seguem os empreendimentos de Ray Kurzweil, um dos principais proponentes da teoria da singularidade, o evento que deve alçar a humanidade a um patamar nunca antes imaginado. Para Kurzweil a morte é um problema técnico. Algo como uma gripe: “Vamos resolver a morte”. Assim como o *big bang* é uma singularidade no plano

cosmológico (Mario Novello), pois não se pode especular sobre o que existiu antes da existência do universo, porque outras leis teriam regido esse momento anterior à criação das leis do cosmos, o *sapiens* estaria hoje em um umbral da hominização e em uma passagem rumo à singularidade.

Para corroborar essa tese, Harari recorre a uma interpretação da religião, que exerce um papel nuclear em todo seu sistema. Durante toda história do *sapiens*, as religiões e ideologias não sacralizaram a vida em si mesma. Sacralizaram sempre algo situado acima ou além da existência terrena, em uma dimensão metaempírica. Isso explica por que os sistemas religiosos sempre foram muito tolerantes com a morte. Mais do que isso, sempre se basearam em uma crença na inexorabilidade da morte como o alicerce fundamental do transmundo que se insinua no plano além-vida. O cristianismo, o hinduísmo e o islamismo não existiriam em um mundo sem morte, ou seja, sem céu, inferno e reencarnação. A mortalidade e sua contrapartida, a promessa de imortalidade fora do mundo, são a espinha dorsal da axiologia (sistema de valores) que nortearam o *sapiens* desde as cavernas, representados precariamente pelos sistemas religiosos. A divinização do humano começa no século XVIII, justamente com a revolução científica. É protagonizada pelo humanismo liberal. Não por acaso, coincide com a mecanização dos animais, que passam a ser tratados como matéria-prima de reposição de energia dos organismos humanos.

A busca da imortalidade e da felicidade implica um controle das qualidades divinas por parte dos humanos. Para atingir essa meta, a evolução dos humanos à condição de deuses pode seguir três engenharias: a Engenharia Biológica, a Engenharia Cibernética e a Engenharia de seres inorgânicos. A primeira diz respeito à seletividade dos genes dos humanos e dos demais seres vivos. A segunda se refere à programação e à reprogramação da vida por meio dos algoritmos. A terceira consiste em uma alteração das propriedades fisioquímicas, de modo que a vida possa amplificar seu poderio. Se analisarmos essas mudanças a partir dos sistemas não-lineares (Prigogine e Stengers), não conseguiremos quantificar o que cada alteração do *sapiens* pode trazer de impacto e de produção de cenários futuros. Uma mudança relativamente pequenas nos genes, hormônios e neurônios do *homo erectus* (produtor de facas a partir de lascas de pedra) o transformaram no *homo sapiens* (produtor de computadores e

espaçonaves). Há um grau exponencial de incomensurabilidade dentre as condições iniciais dos sistemas e o impacto ulterior dessas mesmas alterações, objeto dos modelos desenvolvidos pelas teorias da complexidade.

A despeito da exatidão dos contornos e da fisionomia desse novo *homo*, Harari designa como *homo deus* esta nova espécie divinizada que deve surgir de um entroncamento do *sapiens*. Será tão distintas do *sapiens* quanto os *sapiens* somos distintos do *erectus*. Durante quatro bilhões de anos a seleção natural vem promovendo ajustes nos organismos. A seleção promoveu as passagens dos unicelulares e dos protozoários a répteis e destes aos mamíferos e ao *sapiens*. Nada nos leva a supor que o *sapiens* será o fim da linha da evolução. Depois de 4 bilhões de seleção natural regida pelo acaso, estamos ingressando nos primórdios da seleção artificial, regida pela deliberação humana, pela desoneração das forças latentes da antropotecnia e pela *autopoiesis* infinita (Sloterdijk, Varela e Maturana), ou seja, pela Engenharia Humana e pela seleção artificial. Estaríamos em um limiar de nos liberarmos dos fardos do acaso e da necessidade na seleção, que até agora determinaram a preservação, a mutação e a metamorfose da vida na Terra (Jacques Monod). Nesse sentido, também as tecnologias de química inorgânica podem levar à criação de vida sintética. Devem assim não apenas concorrer para a formação do *homo deus*, mas dar ensejo ao futuro da exobiologia e de um império intergaláctico dominado por descendentes divinoides do *sapiens* (Kaku). Entretanto, quais seriam as urdiduras capazes de chancelar a união entre imortalidade e divindade nessa nova figura do *homo deus*? As condições de possibilidade para a emergência desse cenário de imortalidade e divinização do *sapiens* não foram geradas apenas pela ciência e a tecnologia. Há duas forças-matrizes que impulsionam o *sapiens* em direção a esse horizonte vazio da hominização: a felicidade e o humanismo.

Felicidade e humanismo

Não existe seleção natural para a felicidade. Os genes de um ermitão feliz podem se extinguir. Ao passo que os genes da ansiedade coletiva podem se perpetuar. Entretanto, cada vez mais se domestica a felicidade química. Cresce de modo exponencial a possibilidade de eliminar o desprazer e preservar apenas sensações

agradáveis sentidas no corpo. Para Harari a alma não existe. Não há uma substância metafísica indecomponível capaz de conferir unidade aos organismos singulares. Não existem indivíduos. Existem apenas divíduos. Assim, somos seres subdivisíveis ao infinito. Sistemas decomponíveis em infinitos subsistemas. Diante disso, para sermos felizes basta manipularmos nossa bioquímica ou delegarmos o controle dessa homeostase a um sistema alheio: os algoritmos.

Ora, é nessa chave da busca infinita por felicidade que Harari aloca o projeto do humanismo liberal. Houve três projetos humanistas. O humanismo nazifascista, que consistiu na proposição de um super-humano e, por isso, utilizou o darwinismo social para a construção ficcional da narrativa de uma raça pura e da existência ficcional de uma hierarquia na natureza. Houve o humanismo socialista, que criou uma narrativa alternativa, fundada nos ideais de igualdade e de universalidade da espécie humana, igualmente ficcionais. Ambos os projetos foram destruídos pelo terceiro humanismo vitorioso: o humanismo liberal. O liberalismo é a religião do eu e da subjetividade. Essa religião liberal acredita na capacidade de mensurar os estados de felicidade a partir das experiências do sujeito e se baseia em uma crença na desinibição infinita (Sloterdijk) da felicidade terrena.

Nisso a religião do humanismo liberal é distinta de todas as religiões anteriores do *sapiens*. A maioria das religiões e ideologias reivindica parâmetros objetivos para o bem, o belo e a felicidade. O budismo se dedicou exaustivamente à compreensão da felicidade. O sofrimento para o budismo nasce da identificação entre as sensações e o eu. As sensações estão sempre oscilando e passando do prazer ao desprazer. Apenas mediante a dissociação entre o eu e a flutuação das sensações a consciência pode se libertar do sofrimento e, em certo sentido, da própria morte. Para a maior parte das doutrinas e sabedorias antigas e medievais, o eu não pode se identificar às sensações, pois nesse caso ficaria sempre preso às flutuações das sensações transitórias de dor e prazer. Em 1776, os EUA instituíram que o direito à felicidade. Trata-se de um dos direitos inalienáveis do ser humano, junto com direito à vida e à liberdade. Essa lógica conduziu a humanidade a eleger a felicidade como o segundo objetivo mais importante do século XXI. Desde então, o telhado de vidro da felicidade é sustentado por dois pilares: um psicológico e outro biológico. O psicológico diz respeito às sucessivas

sensações de prazer de que podemos gozar. O biológico consiste na seguridade corporal e de saúde que podemos obter hoje em dia.

Para Epicuro a felicidade consistia em um controle entre prazer e dor. A busca de prazer, ou seja, de felicidade, sem moderação, traria infelicidade. Para Buda, algo semelhante: a busca de sensações prazerosas é a raiz do sofrimento. A identificação entre o eu e as sensações é a grande cilada dos sentidos e o labirinto do pensamento e do desejo. Para os utilitaristas, no extremo oposto dessas proposições, a felicidade baseia-se na eliminação da dor. O lema de John Stuart Mill e de outros utilitaristas que forjaram os modelos de vida moderna se baseia em duas teses. Primeira: a felicidade é igual ao prazer. Segunda: o prazer consiste em uma minimização da dor. Essa visão utilitarista se converteu na ortodoxia científica e filosófica do século XXI. Do ponto de vista darwiniano, essa sobrevalorização da felicidade é um imenso erro. Durante milhões de anos, o sistema bioquímico humano foi adaptado para promover a sobrevivência e a reprodução, não para aumentar a felicidade. Caso um animal se contentasse com sua felicidade em um bosque e não conseguisse prever a escassez de alimentos ou o ataque iminente de predadores, esse animal estaria extinto. E, no entanto, as fórmulas científicas têm se desenvolvido a contrapelo da evolução darwiniana, do crescimento econômico, das reformas sociais e das revoluções políticas. Afinal, para elevar os níveis globais de felicidade precisamos apenas manipular a bioquímica humana.

Os paradoxos dessa sociedade feliz são cada vez mais evidentes. Como acentua Harari, a bioquímica da felicidade é uma das principais causadoras de crimes no mundo. Estima-se que 38% dos presos da Itália, 55% dos presos do Reino Unido e 62% dos condenados da Austrália são criminosos relacionados ao mundo das drogas. Entretanto, o controle das drogas é apenas o começo de um amplo projeto de controle amoral da bioquímica humana a serviço da felicidade. Se a evolução não adaptou o *sapiens* a ponto de o tornar apto a experimentar prazer constante, a Bioengenharia pode corrigir essa imperfeição evolutiva. A Biotecnologia pode vir a ser o projeto de um prazer ininterrupto. Uma promessa de felicidade total. O segundo projeto seria a reengenharia do *sapiens* para algo inédito na história humana: a experiência do prazer amortal. Não a *eudaimonia* epicurista (controle das sensações), mas uma deificação

utilitarista do eu. Um humanismo deificado. Uma santificação do humano. Uma glorificação do sujeito.

Algoritmos e Deus

Se a morte, a dor e o mal existem, e não há uma explicação racional para sua existência, é sinal de que existiria uma explicação não racional e não natural: uma explicação divina. O declínio das religiões, a morte de Deus e a universalização das promessas seculares promovidas pela ciência, pela tecnologia e pela modernidade produziram uma minimização da dor e da morte, pedras angulares da explicação da vida pelos sistemas religiosos. Disso decorre que nunca a felicidade esteve tão presente na cultura humana quanto nos últimos séculos. Entretanto, o *sapiens* se ilude. Imagina que matar Deus e colocar o humano no centro do universo é uma forma não-religiosa de vida. O que chamamos de modernidade para Harari não é nada mais do que a religião da humanidade, a sacralização do *sapiens*, ou seja, a produção das condições evolucionárias para a especiação e a mutação do *homo sapiens* em *homo deus*. Essa nova religião chamada humanismo é apenas a primeira figura, efêmera e indefinida, do processo de divinização do humano que se encontra em franca expansão. E que deve produzir uma das sociedades mais injustas que jamais existiram.

Durante trezentos anos o mundo tem sido dominado pela narrativa humanista, que santifica a vida, a felicidade e o poder do *sapiens*. Humanismo e perfectibilidade (Passmore e Sloterdijk) são as cifras secretas desse novo mundo que se anuncia. A imortalidade, a felicidade e a divindade são apenas as conclusões lógicas da religião humanista: a antropotécnica (Sloterdijk). O culto ao humanismo dominou o mundo e, paradoxalmente, lançou as sementes de sua própria destruição. Contudo, como sempre, Harari gosta de enfatizar os paradoxos. O paradoxo dos caçadores-coletores nômades e animistas era a incapacidade de fixação e de ampliação de suas riquezas simbólicas. O paradoxo da revolução da agricultura era materializado em verbos de crescimento: as pestes se propagam, as doenças proliferam, a demografia explode. Cresce a acumulação primitiva de bens primários e se aprofunda o abismo da divisão de classes. Por conseguinte, mais e mais guerras e disputas por territórios. Por sua vez, o paradoxo da revolução científica consiste no seguinte axioma: ao expandir o

conhecimento do universo, da vida e da matéria a confins infinitos e imensuráveis, o *sapiens* produziu como contrafigura uma imagem do humano ainda mais insignificante, precário, aleatório e excêntrico em relação a esse mesmo universo descortinado por esse mesmo ato de conhecimento.

O mesmo sistema paradoxal ocorre com o humanismo liberal, ou seja, com a deificação do humano levada a cabo pela modernidade e pelo utilitarismo. O alicerce dessa divinização deve ocorrer por meio da aliança entre duas entidades tão prosaicas quanto abissais: a felicidade e os algoritmos. A felicidade como força centrífuga e de êxodo ontológico dos humanos em direção à solução de suas ambivalências estruturais, ou seja, rumo à erradicação da humanidade do humano. A os algoritmos como forças centrípetas. Serão novos aparelhos (Flusser) e gestores da liberdade alienada voluntariamente pelos divíduos que enfim descobriram que não são indecomponíveis. Esses novos superorganismos articulados em uma mente coletiva e em redes neurais serão os novos transumanos que devem emergir desse longo processo evolucionário de divinização, imortalidade e sacralização do humano. Paradoxalmente, a mesma tecnologia que eleva os humanos à condição de deuses deve reduzir estes mesmos humanos à completa irrelevância e, por que não, à iminente extinção. Por isso, mais do que as relações humanos-humanos, as relações humanos-animais e humanos-deuses são o melhor modelo para prever as eventuais relações futuras entre super-humanos, infra-humanos, transumanos e derivações (Pierre Lévêque, Claude Lévi-Strauss e Donna Haraway). Como no conto de Kafka, por meio dessas relações podemos imaginar o *sapiens* não mais como um deus em relação a um primata superior. Mas como um macaco em potencial. Um futuro animal em relação a esse *homo deus* que se encontra agora na iminência de emergir no horizonte temporal.

Enviado: 4 abril 2018
Aprovado: 1 maio 2018