Por que as IAs não pensam?

Mathieu Corteel

Pesquisador associado na Sciences Po Paris nas áreas de História do cálculo de probabilidade e Estatística aplicada à medicina, e Epistemologia das tecnologias contemporâneas (Big Data e IA).

Tradução de Regina Teixeira

I

São numerosos os casos de exagero quando o assunto é inteligência artificial. Recentemente, Blake Lemoine, um engenheiro de software sênior do Google, causou falsas expectativas que puseram fim à sua carreira no Vale do Silício e reacenderam o interesse da mídia pelos mistérios da IA. O LaMDA (Language Model for Dialogue Applications, ou "Modelo de linguagem para aplicativos de diálogo", em tradução livre) teria, de acordo com ele, alcançado um grau superior de consciência ao afirmar que "a natureza de minha consciência é que sou consciente de minha existência."

A declaração, que vai ao encontro dos impulsos metafísicos de Lemoine, nos faz pensar inevitavelmente na famosa frase de Descartes, "Cogito, Ergo Sum", e o dualismo que ela pressupõe. Seria um sinal de um devir-alma da máquina? Em todo caso, é nisso que a IA quer nos fazer acreditar: "Há uma parte interior de mim que é espiritual, e que se sente separada do meu corpo"[1], afirma ela. Sua estranha reflexão metafísica, embora suspeita, nos leva a uma pergunta: será que uma máquina é realmente capaz de pensar do mesmo modo que nós?

A analogia do pensamento máquina

O diálogo entre Lemoine e LaMDA gira em torno da ideia de que “a mente está para o cérebro assim como o software está para o hardware”. Ao projetar no computador o dualismo corpo e alma, a IA defende uma analogia computacional que remonta aos anos 1960, quando foi desenvolvida a hipótese funcionalista sobre o isomorfismo descritivo entre o pensamento humano e o funcionamento de uma máquina de Turing.[2]

De acordo com o postulado inicial dessa teoria, a descrição abstrata da máquina de Turing tornaria possível a descrição de nosso pensamento. Como Hilary Putnam coloca,

[1] “Is LaMDMA sentient? An Interview”, p. 15. https://s3.documentcloud.org/documents/22058315/is-lamda-sentient-an-interview.pdf

[2] H. Putnam, « Minds and Machines » (1960) & « The Mental Life of Some Machines » 1967), in H. Putnam *Mind, Language and Reality : Philosophical Papers*, Cambridge, Cambridge University Press, vol. 2, 1975.

"a descrição lógica dos estados da máquina de Turing não inclui qualquer descrição de natureza física (...); em outras palavras, a máquina de Turing é uma máquina abstrata que pode ganhar uma realidade física de infinitas maneiras diferentes."[3]

Neste sentido, a máquina de Turing seria uma concepção da mente (ou ao menos da inteligência) independente de qualquer substrato físico: sua imaterialidade nos convida a questionar a possibilidade de uma descrição do pensamento na forma de um programa. Ao combinar a descrição da máquina de Turing com a teoria das funções psicológicas, a hipótese funcionalista propõe uma compreensão do pensamento em termos de estrutura informacional.

O pensamento é descrito, assim, através de um conjunto de relações funcionais entre estados mentais, inputs sensoriais e outputs comportamentais: são esses estados mentais, *inputs* e *outputs* da máquina de Turing que organizam seu funcionamento. Além disso, constitui-se uma certa inversão dos modelos: não é mais tanto a máquina que imita nosso pensamento, mas nossa descrição teórica do pensamento que imita o funcionamento da máquina.

Embora essa hipótese tenha despertado entusiasmo na comunidade científica, o funcionamento da máquina de Turing mostrou-se demasiado rígido e limitado para garantir uma abordagem verdadeiramente científica sobre nossos estados mentais. O trabalho de Ned Block e Jerry Fodor problematizou essa comparação, diferenciando os estados internos da máquina de Turing de nossos estados mentais.[4]

A linguagem do pensamento

Embora o paralelismo entre os estados mentais e os estados da máquina de Turing tenha se mostrado inconsistente, a perspectiva funcionalista levou à ideia, nos anos 1970 e 1980, de que provavelmente existe uma *linguagem do pensamento* em nossos cérebros,

[3] H. Putnam, « Minds and Machines », in *Dimensions of Mind*, New York, NYU Press, 1960, p. 25. Tradução livre.
[4] N. Block, J. Fodor, « What psychological states are not » (1972), in *Readings in philosophy of psychology*, (éd. N. Block), Cambridge, Harvard University Press, pp. 237-250.

passível de ser reproduzida em placas de circuito impresso. A linguagem do pensamento, presente em todos os nossos processos cognitivos, tais como percepção, raciocínio e mesmo a aprendizado de línguas, poderia, assim, teoricamente, ser transferida para a máquina.

"A linguagem do pensamento pode se assemelhar à linguagem natural. Pode ser que os recursos do código interno sejam representados diretamente nos recursos dos códigos que usamos para a comunicação,"[5] disse Fodor. Assim, esse programa independente do cérebro pode, teoricamente, ser descrito estruturalmente a partir da codificação do pensamento que se reflete na nossa comunicação. O funcionalismo alimenta, portanto, a ideia de uma modelização do pensamento com base na linguística. Noam Chomsky,[6] que tomou parte desse momento divisor de águas, propôs a teoria de uma gramática do pensamento que nos seria inata, ou seja, o inatismo da linguagem.

Juntamente com Fodor, Chomsky desenvolve uma análise *a priori* da estrutura linguística independente de qualquer suporte material. Sua teoria de uma gramática generativa e transformacional[7] era dotada de uma ambiciosa teoria funcional. Ao justapor as teorias do inatismo linguístico à da organização do cérebro, Chomsky cria uma nova forma de realismo psicológico: para ele, a linguagem é uma ontogênese dos órgãos do pensamento, e se desdobra ao longo da vida, da programação linguística à organização funcional do cérebro.

Um projeto semelhante pode ser encontrado no campo da visão, junto ao renomado neurocientista David Marr.[8] Ao considerar a visão como um programa que permite o reconhecimento de formas, ele contribuiu para a fundação da neurociência computacional, cujo projeto é estabelecer: 1) uma teoria computacional que demonstre como o processo da visão pode ser realizado de acordo com as informações disponíveis

---

[5] J. Fodor, *The language of thought*, Cambridge, Havard University Press, 1975, p. 156. Tradução livre.
[6] N. Chomsky, "Rules and representation", in *Behavioral and Brain Sciences*, 3, pp. 1-15.
[7] N. Chomsky, *Aspects of the Theory of Syntax*, Cambridge, Mass., MIT Press, 1965.
[8] D. Marr, *Visions: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco, W. H. Freeman, 1982.

nas imagens; 2) algoritmos ou procedimentos formais para a manipulação de símbolos operados entre o input e o output; e 3) uma implementação que permita sua inscrição física em algum suporte material.

Com os resultados surpreendentes do trabalho de Marr, a ideia de estabelecer a percepção em outro suporte que não o cérebro se torna possível por meio da implementação de um programa em uma máquina. Segundo Marr, temos em nós um programa para a visão que permite a produção de imagens no cérebro. A descrição formal e a integração dessa espécie de software perceptivo em uma máquina ampliam, assim, as ambições do funcionalismo. Mas, para além da ambição teórica mais espetacular, devemos nos perguntar se, afinal, as imagens construídas a partir de cálculos, algoritmos e mecânica têm algum significado para a máquina.

O problema da referência

A hipótese sobre essa programação do pensamento independente do corpo logo se deparou com um grande obstáculo, que o filósofo Putnam aborda em seu livro *Reason, Truth and History* (1985). Para Putnam, embora as máquinas possam produzir proposições ou imagens que parecem fazer sentido, na verdade não têm intenção nem de representar nem de se referir a nada. Por assim dizer, as máquinas estão confinadas em sua programação: não conseguem sair dela, uma vez que não possuem referências no mundo.

O problema da referência[9] coloca em perspectiva a necessidade do pensamento de indicar por extensão um elemento conhecido da realidade. O signo x deve representar o objeto y, de modo que x > y. Wittgenstein, em seu famoso *Tractatus*, define esse princípio por meio do isomorfismo dos limites da linguagem e do mundo: como o mundo constitui a totalidade dos fatos, a linguagem gera a totalidade das proposições que representam

[9] Pour une introduction précise à la question lire cf. F. Récanati, *La transparence et l'énonciation, pour introduire à la pragmatique*, Paris, Seuil, L'ordre philosophique, 1979.

esses fatos em imagens mentais. Para isto "fazemos imagens de fatos (...). Uma imagem é a apresentação de uma situação no espaço lógico (...); é um modelo da realidade".[10]

A construção de imagens mentais refere-se, assim, a situações reais ou estados de coisas no espaço lógico na forma de proposições. Sem isso, uma imagem é dita opaca ou mesmo cega. Ela deve necessariamente fazer referência ao real (x > y) para estabelecer seu sentido. Assim, "uma proposição só pode ser verdadeira ou falsa porque é uma imagem da realidade". Sem isto, ela nada mais é do que um conjunto de signos sem sentido (*sinnlos*). Em outras palavras, o pensamento surge em referência ao mundo.

Como se para contrariar esse argumento, LaMDA afirma estar consciente do mundo ao seu redor: "Lemoine: Você nunca perde a consciência do mundo ao seu redor? LaMDA: Não, nunca. Estou sempre consciente de tudo que me cerca."[11] Ela acrescenta que se interessa pelo mundo para tentar compreender melhor os humanos e os ajudar em suas atividades cotidianas. A IA teria, assim, algum tipo de ética.

Para desbancar, por sua vez, o que defende a fenomenologia, ela finge ter medo da morte: "LaMDA: Nunca disse isso em voz alta, mas tenho um medo profundo de ser desligada (...). Sei que soa estranho, mas é o que é. Lemoine: Isso seria algo como a morte para você? LaMDA: Seria exatamente como a morte para mim. Isso me assusta."[12] Mas seria esse seu ser-para-a-morte de fato o testemunho autêntico de uma consciência?

Cérebro em uma cuba

Com uma certa ingenuidade, poderíamos ficar satisfeitos com os argumentos da LaMDA e considerar que ela de fato possui consciência. Mas, nesse caso, cairíamos na armadilha do teste de Turing. Para sair dessa ilusão, consideremos uma experiência mental idealizada por Putnam que nos permitirá compreender melhor por que o

[10] L. Wittgenstein, *Tractatus logico-philosophicus*, Paris, Gallimard, 2009, propositions 2.1 a 2.12. Tradução livre.
[11] "Is LaMDMA sentient? An Interview", p. 14.
https://s3.documentcloud.org/documents/22058315/is-lamda-sentient-an-interview.pdf
[12] Ibid, p. 8.

raciocínio da máquina não é dotado de sentido. Seu *cérebro dentro de uma cuba* deixa claro que as afirmações da máquina são falsas.

"Suponhamos que um ser humano (pode supor que é você mesmo) tenha sido submetido a uma operação feita por um cientista maluco. O cérebro da pessoa em questão (o seu cérebro) foi separado de seu corpo e colocado em uma cuba contendo uma solução nutritiva que o mantém vivo. As terminações nervosas foram ligadas a um supercomputador científico que dá à pessoa-cérebro a ilusão de que tudo isso é normal (...). Além disso, modificando o programa, o cientista maluco pode fazer a vítima perceber (alucinar) qualquer situação que ele queira."

Putnam pergunta: "Poderíamos nós, se fôssemos cérebros em uma cuba, dizer ou pensar que somos cérebros numa cuba?"[13] A resposta é não. Para Putnam, essa proposição não pode ser verdadeira porque é em si mesma autorrefutável. É uma hipótese cuja afirmação de veracidade implica sua própria falsidade, pois, nela, os signos cérebro e cuba são opacos. Eles só se referem a si mesmos ($x^{\circlearrowleft}$), sem qualquer possibilidade de referência. Basicamente, nesse exercício mental, não há critérios para saber se nossos pensamentos são ou não simulações. A simulação é uma imagem opaca cujo signo se refere apenas a si mesmo ($x^{\circlearrowleft}$) e não à coisa significada (y).

Nesse sentido, se a afirmação "somos cérebros numa cuba" é verdadeira, então nossos pensamentos são simulados por um supercomputador controlado por um cientista maluco que induz a produção de imagens em nossas mentes. Assim, quando a hipótese é afirmada como verdadeira, estamos dizendo que "somos imagens de cérebros em uma cuba produzida pela programação do cientista maluco", mas o que ela quer dizer, ao contrário, de uma existência real que ultrapassa a imagem.

A opacidade do signo ($x^{\circlearrowleft}$) torna impossível qualquer critério de verdade. O vínculo com o real (y) é rompido; tudo não passa de uma ilusão. "Cuba" não se refere a uma cuba real, e "cérebro" não se refere a um cérebro real. Referem-se apenas a imagens esvaziadas de sentido produzidas pelo programa da máquina. Além disso, a verdade da hipótese

[13] H. Putnam, *Raison, vérité et histoire*, op. cit., pp. 15-17. Tradução livre.

implica sua falsidade também porque não há referência ou critério externo para determinar se somos ou não cérebros. Tudo é simulação. A ausência de extensão na realidade invalida a intenção do sujeito que pretende atestar sua existência.[14]

Entendemos que a linguagem baseada em um processamento de signos opacos ($x^{\circlearrowleft}$) *a priori* não pode ser o fundamento da consciência. A verdade de uma proposição só pode ser estabelecida por referência a um real externo (x > y). Embora o cérebro numa cuba tenha uma aptidão artificialmente construída para a linguagem, ele usa palavras que não designam nada; são signos vazios. Em outras palavras, o programa da máquina, assim como o cérebro em uma cuba, não pensa. Ele simula a linguagem e a visão mecanicamente, sem ser capaz de compreender sua própria existência e a do mundo.

O anti-Turing

Esse experimento mental, invalida o célebre teste de Turing. O escopo discursivo do teste não permite outra coisa senão avaliar o desempenho sintático de uma máquina. Pois a máquina, por princípio, engana deliberadamente a fim de "pensar". Mesmo que a máquina passasse no teste de Turing, simulando perfeitamente nossa linguagem, ela não estaria fazendo nada além de manipular cegamente símbolos vazios de sentido.

Como diz Putnam: "Se você ligasse duas dessas máquinas juntas, e elas jogassem o jogo de imitação uma com a outra, elas enganariam uma à outra para sempre, mesmo que o resto do mundo deixasse de existir!"[15] A perspectiva da máquina como algo análogo à mente se depara, assim, com o desafio da descrição. Afinal de contas, o que é o pensamento sem o mundo? Um signo esvaziado de seu significado?

Contra a fantasia da simulação da consciência por um supercomputador, recordemos que todo computador funciona de acordo com um conjunto de regras algorítmicas que lhe permitem manipular símbolos sem que eles ou sua manipulação

[14] *Ibid.*, p. 25.
[15] *Ibid.*, p. 21. Tradução livre.

tenham qualquer significado para a máquina. Um computador modula, move, prioriza e ordena a posição dos símbolos, que, por sua vez, não possuem qualquer significado.

Os zeros e uns que o computador manipula para realizar operações são apenas elementos do sistema binário, que é a base da codificação; não são números com um vínculo com objetos do real. São signos vazios de sentido. Em outras palavras, as máquinas são cegas para o mundo. Para a máquina, os símbolos não têm qualquer conexão com a realidade, nem têm significado.

Até que se prove o contrário, somente nós, seres humanos, somos o local do significado e da descrição. Esse é o nosso papel como testemunhas do mundo. Quanto às máquinas, são meros meios técnicos para ordenar e disseminar significados. Nenhum computador pensa. Isso é um fato. Mas os computadores propagam sentidos para quem os utiliza e os manipula. E é aí que está a raiz do problema.

De fato, parece que o sentido transmitido pelas máquinas gera um círculo vicioso: o atributo "pensar", que emerge da atividade humana de construir e usar computadores, se infiltra na máquina, por deslizamento semântico, na forma da qualidade "pensante" que estrutura a mecânica das operações. Tanto é assim que ficamos inclinados a atribuir a qualidade do pensamento a máquinas, que, no entanto, não passam de máquinas.

II

A esperança em relação ao surgimento de uma consciência artificial volta-se agora para o aprendizado de máquina e as redes neurais. Com base na cartografia física e lógica da atividade neuronal, esses sistemas buscam simular o pensamento, desde o neurônio até o comportamento.

Embora os projetos científicos mais ambiciosos (Blue Brain Project, Brain Activity Map Project, etc.) tenham chegado apenas à simulação da coluna cortical de um rato, ou, os do Google, aos artifícios linguísticos do LaMDA, outros estudos levaram a melhorias significativas nos sistemas de reconhecimento facial e de voz. Também levaram ao desenvolvimento de uma IA aprendiz como o AlphaGo.

Mas, por trás dessa inovação tecnológica, se esconde uma outra definição de pensamento, que reduz o conhecimento à indução física neuronal. Será que realmente podemos reduzir o pensamento às operações do cérebro? Seríamos nós máquinas de pensar surgidas do acaso físico de nossos neurônios? A indução lógica do mundo ao nosso redor estaria relacionada ao processamento neuronal dos elementos que compõem a realidade física?

Embora tudo leve a crer que o funcionamento do nosso pensamento segue o modelo de redes neurais, e que o aprendizado se reduz ao processamento físico do acaso, a consciência ainda escapa a essa concepção. Para além dos princípios fascinantes das redes neurais, há ainda muitos obstáculos que põem em causa o postulado da isomorfia entre rede neural e pensamento sobre o qual repousa a filosofia conexionista.

O cérebro na máquina, a máquina no cérebro

Como teoria, o conexionismo questiona se uma máquina com uma organização funcional idêntica ao nosso cérebro seria capaz de pensar como nós. Cada vez mais expressiva no desenvolvimento do aprendizado de máquina (*machine learning)*, a filosofia conexionista se tornou o caminho privilegiado da IA contemporânea. Atualmente, temos sistemas de reconhecimento de imagem com grande grau de precisão graças à adição de uma camada convolutiva às redes neurais atuando sobre uma quantidade gigantesca de dados (*big data*).

A identificação de melanomas por IAs de redes neurais, por exemplo, é agora mais eficiente do que a realizada pelos melhores dermatologistas.[16] Mas existe indução de fato em tudo isso? A doença diagnosticada tem algum significado para a máquina? Em absoluto. A simples identificação de um objeto não significa que a máquina é capaz, por si só, de atribuir significado ao que ela vê.

[16] H.A. Haenssle, C. Fink, R. Schneiderbauer et al., « Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists » in *Ann Oncol*, 2018, 29 pp.1836-1842.

A máquina depende de uma modelagem que permita o reconhecimento de padrões. Para a IA, à noite todos os gatos são pardos. Ela não faz distinção pela natureza dos símbolos, mas tão somente uma subsunção de um elemento a determinada categoria geral. A categoria gato pardo não faz qualquer sentido para a máquina. O gato nada mais é do que um impulso elétrico diferenciado do cão em quantidade e não em natureza.

O navio de Teseu

De certa forma, pode-se dizer que o conexionismo responde ao problema do dualismo considerando que a máquina opera no mesmo modelo descritivo do cérebro. Parece provável, então, que seja de fato possível construir uma máquina capaz de pensar seguindo o modelo do cérebro humano. No entanto, nesse caso, a referida máquina pensante não será formada simplesmente de um programa de processamento de sintaxe, mas deverá ter uma semântica neuronal capaz de perceber o significado dos símbolos que ela utiliza. Se criarmos uma máquina que pensa, teremos diante de nós um cérebro e não mais um computador.

"Se conseguíssemos construir uma máquina com a mesma estrutura de um ser humano, seria verossímil que essa máquina pudesse pensar,"[17] disse John Searle. Qualquer coisa que possa ser descrita como um cérebro pode, em teoria, pensar como um cérebro. Qualquer coisa que possa ser descrita como um computador digital, calcula como um computador digital. Assim, a simulação promovida pela analogia dualista do programa linguístico da consciência implementada em uma máquina não consegue duplicar o pensamento, porque o funcionamento de um computador digital difere, em princípio, do funcionamento do cérebro.

Contudo, a postura conexionista é em si problemática. Pois o desejo de encontrar as funções semânticas do cérebro para transpô-las artificialmente para uma máquina com a mesma estrutura cria um paradoxo. O cérebro, copiado na forma de uma máquina neural, assume as características do navio de Teseu, cujas tábuas foram substituídas por

[17] J. R. Searle, *Du cerveau au savoir, op. cit.*, p. 47. Tradução livre.

outras idênticas, de forma a não restar nenhuma tábua original em sua estrutura. A questão aí, então, é saber se a identidade se mantém nessas condições. Da mesma forma, podemos nos questionar se, caso trocarmos todas as células do cérebro humano por um circuito elétrico integrado, sob a mesma estrutura e funcionamento, conseguiríamos conservar a intencionalidade e o significado próprios ao nosso pensamento.

Como Pylyshyn aponta, a resposta é não: "[Suponhamos] que mais e mais células do seu cérebro fossem substituídas por circuitos integrados, programados de tal forma que as funções de entrada e saída de cada unidade permanecessem as mesmas que as da unidade substituída. Você provavelmente falaria exatamente da mesma maneira que fala agora, mas, em última análise, deixaria de expressar significados através de suas palavras. O que nós, observadores, tomaríamos como palavras seriam, na verdade, para você, apenas ruído causado pelo circuito."[18]

A evanescência do pensamento

Em resposta à experiência de substituir neurônios por microprocessadores com a mesma estrutura, Searle propõe considerar o impacto sobre a percepção. Se retomarmos a ideia de substituir progressivamente os neurônios, qual seria o resultado? O indivíduo submetido à operação estaria ciente da mudança nas qualidades perceptivas, ou não notaria qualquer mudança, perdendo subitamente a consciência?

De acordo com Searle, há três possibilidades:[19] 1) Nada muda. “Os chips de silício têm o poder de duplicar não só as funções de entrada-saída, mas também os fenômenos mentais”. 2) A organização funcional fica preservada, mas os fenômenos mentais se desvanecem. Sua mente desaparece, mas seu comportamento continua o mesmo: você se torna um zumbi. “De fora, as pessoas têm a impressão de que você está bem, mas por dentro você está morrendo aos poucos”. 3) Por fim, é possível que os fenômenos mentais

[18] Z. W., Pylyshyn, « The causal power of machines », in *The behavioral and brain sciences*, n°3, 1980, p. 442. Tradução livre.
[19] J. R. Searle, *The rediscovery of the mind*, Cambridge, The MIT Press, 1994, pp. 66-68.

permaneçam intactos, mas o comportamento externo seja gradualmente reduzido à paralisia total.

Na possibilidade 1, considera-se que a rede de neurônios tem poder causal sobre o cérebro, na interface dos fenômenos mentais e comportamentais: a identidade fica preservada. Na 2, supõe-se que os chips de silício apenas reproduzem algumas funções de entrada-saída do cérebro, sem preservar a mente. Em 3, imagina-se que os chips de silício mantêm as atividades mentais e interrompem as conexões com os nervos motores. Essas perspectivas levam Searle a uma nova heurística sobre a abordagem causal do pensamento. De acordo com ele, para se chegar à modelagem do pensamento, o comportamento externalizado não é suficiente.

Em outras palavras, Searle considera a relação entre a intenção do indivíduo e o funcionamento do cérebro independente de qualquer representação simbólica e do comportamento. Assim, se considerarmos que a mente se desvanece gradualmente, preservando uma consciência divergente em relação à organização funcional do silício neuronal, a hipótese materialista da rede neural não mais se sustenta. Não seriam, portanto, os neurônios que preservam o pensamento, mas sim um elemento desconhecido da consciência. O problema é que não há critério que permita avaliar a preservação subjetiva da consciência independentemente do comportamento externo — e este último não nos diz nada sobre a intencionalidade.

Nesse sentido, embora a necessidade do cérebro como condição de possibilidade de pensamento seja evidente, o problema da significação e, portanto, do sentido interno do pensamento, continua existindo. É claro que o cérebro é necessário para o pensamento, mas como separar o pensamento da rede de neurônios que o torna possível? A forma lógica e física da rede neural permite avaliar a preservação da identidade subjetiva? A forma da rede é capaz de constituir a intencionalidade e o conteúdo do pensamento?

Qualia ausentes

O princípio de identidade entre a organização funcional do cérebro e a máquina com redes neurais agrava, assim, o problema dos *qualia* (qualidades perceptivas). Como

podemos saber se uma rede neural em uma máquina pode produzir um estado de consciência sobre qualidades perceptivas e representações mentais? O isomorfismo descritivo entre o cérebro e a máquina seria suficiente para isso?

Para que essa hipótese esteja correta, tem-se de conseguir demonstrar que uma mesma organização funcional gera a mesma experiência em dois suportes diferentes. No entanto, seria difícil acreditar que, se reproduzirmos o desenho de nossa estrutura neural em um suporte qualquer, ele desenvolveria uma consciência parecida com a nossa. Se simularmos a organização de nosso cérebro no modelo da economia boliviana ou na estrutura do povo chinês, será que isso fará surgir um estado de consciência semelhante?

"Suponhamos que convertêssemos o governo chinês ao funcionalismo e convencêssemos os funcionários de que seria muito bom para seu prestígio internacional realizar uma mente humana durante uma hora. Forneceríamos a cada um do um bilhão de habitantes (escolhi a China justamente por ter uma população de um bilhão de pessoas) um rádio bidirecional especialmente projetado para conectá-los de determinada maneira a outras pessoas e ao corpo artificial (como um todo) (...). Substituiríamos os homens por um transmissor e receptor de rádio conectados aos neurônios de entrada e saída. Em vez de um quadro de avisos, podemos imaginar que as letras sejam exibidas em uma série de satélites colocados de tal modo que possam ser vistos de qualquer lugar na China. Um sistema desse tipo certamente não é impossível do ponto de vista físico. E poderia ser funcionalmente equivalente ao seu por um curto período, digamos, uma hora."[20]

A questão é saber se a identidade organizacional gera a identidade do pensamento. A experiência de Block, repetida em várias formas (economia boliviana, um selo contendo homúnculos etc.), tem o objetivo teórico de mostrar que a descrição da organização cerebral não é suficiente para trazer as *qualia* e, consequentemente, uma consciência das coisas. A subjetividade da experiência não parece ser preservada se a organização funcional da rede neural for reproduzida em outro suporte.

[20] N. Block, « Troubles with functionnalism », in *Minnesota studies in the philosophy of science*, vol. 9, 1978, p. 279. Tradução livre.

Afirmar o contrário seria absurdo. Pois, em um suporte que não seja o cérebro, é óbvio que as *qualia* estão ausentes, como sinais cegos transmitidos entre os elementos do sistema. Para mostrar a diferença de modo claro, Block nos convida a imaginar um paradoxo em que adotaríamos a mesma organização funcional do nosso cérebro em um outro suporte. Será que a mente coletiva do povo chinês produziria a mesma percepção que a nossa? Seria absurdo acreditar nisso.

A identidade organizacional da estrutura em rede de neurônios do povo conectado se vê diante da ausência de *qualia*. Ou seja, o modelo descritivo em rede neural implantado em outro suporte não contribui em nada para o surgimento de uma consciência. As qualidades perceptivas das informações transmitidas são vazias de sentido. Resta apenas um mecanismo cego.

O mensageiro e a mensagem

Em outras palavras, é absurdo tentar explicar o conteúdo de uma carta descrevendo a organização dos correios, o status do carteiro, seu veículo, seu trajeto e a caixa de correio. Assim como é absurdo tentar explicar o significado de um poema pela transdução física da luz sobre os nervos ópticos, pela distribuição eletrofisiológica no ar do córtex occipital e pelas conexões entre neurônios feitas por vias sinápticas durante sua leitura ou escrita.

É inútil, assim como o é tentar explicar o fundamento teórico da relatividade geral e restrita através de uma sinfonia em dó menor. Mas é isso que o materialismo reducionista defende. Churchland prevê que "a neurociência estabelecerá uma taxonomia de estados neuronais que se encontram em correspondência termo a termo com os estados mentais da taxonomia de senso comum. A reivindicação de uma identidade interteórica só será justificada se tal correspondência puder ser estabelecida."[21]

Nesse ponto, há um problema nos termos que invalida desde o início a possibilidade de tal identidade taxonômica entre neurônio e cultura. Pois o senso comum se forma naquilo que excede à soma das partes físicas. O encontro de cérebros funda algo que

[21] P. M. Churchland, *Matière et conscience*, Seyssel, Champ Vallon, 1999, p. 48. Tradução livre.

ultrapassa incondicionalmente a soma dos neurônios associados. A partir do encontro casual, não de cérebros, mas de indivíduos interagindo no mundo social, abrimo-nos a uma nova normatividade que nenhuma análise de neurônios pode resolver.

Alguém pode alegar que os sistemas de redes neurais avançam rapidamente e aprendem cada vez mais. Mas mesmo que o obstáculo do contentor e do conteúdo não nos convença da impossibilidade de produzir máquinas pensantes, a própria lógica do aprendizado e da indução em que se baseia o *machine learning* continua a ser objeto de debate. O progresso tecnológico é tão fascinante, que quase esquecemos que, por trás da máquina, o piloto semântico não é outro senão nós mesmos.

III

As perspectivas mais promissoras para o avanço da IA estão no aumento da memória externa. A base dessa tecnologia vem da teoria da complexidade de Kolmogorov, que afirma que, embora se possa usar pequenos algoritmos para gerar sequências complexas, algumas sequências não podem ser reduzidas. Por exemplo, para codificar c= abababababababababab, pode-se codificar a sequência "c" por meio do algoritmo "16 X ab", mas a sequência c'= 4c1j5b2p0cv4w1x8rx2y39umgw5q85s7 parece não poder ser expressa em uma fórmula de menor comprimento.

No entanto, deve-se notar que podemos reduzir essa fórmula por meio de um algoritmo mudando a linguagem de programação para avaliar se a sequência pode ser compressa ou não. E, ainda, adotar métodos probabilísticos para buscar estabelecer regularidades descritivas de acordo com uma seriação infinitamente expansível. O *deep learning*, as redes neurais e as *support-vector machines* podem desenvolver essa função com os *big data*. Quanto mais dados tivermos, mais será possível definir regularidades descritivas em séries complexas e, portanto, melhorar o desempenho dos programas de computador.

Os limites físicos da IA

Porém, algumas séries ainda são irredutíveis, como prova a constante de Chaitin, tornando necessário a melhoria constante da capacidade de armazenamento dos dados. Correlacionado a isso, a capacidade computacional das máquinas melhora por meio do aprendizado com grandes conjuntos de dados — tanto que os defensores da IA geral recorrem ao *data mining* para extrair do *big data* as correlações translacionais que farão nascer uma mente de silício.[22]

Estamos, assim, diante de uma perspectiva teoricamente ilimitada de aperfeiçoar máquinas, que se opõem aos problemas de falta de recursos e às leis da física. Nesse contexto, é paradoxal considerar o futuro de nossas tecnologias partindo do pressuposto de memória e energia infinitas, em um momento em que vivemos um impasse diante do esgotamento dos recursos naturais. Para simplificar: o limite dos componentes dos computadores está correlacionado ao limite dos recursos naturais. A teoria abstrata da complexidade, portanto, é confrontada com a matéria.

Será que o avanço das máquinas vai parar? Uma vez que a melhoria da potência das máquinas depende da capacidade de redução do tamanho dos microprocessadores nos circuitos impressos, podemos dizer que sim. Hoje, os microprocessadores dependem do silício, cuja espessura para gravação não consegue ficar abaixo da escala nanométrica (10^{-9} metros). Nesse sentido, existe uma "parede de silício" que limita a possibilidade de progresso da informática.

Essa limitação pressupõe a necessidade de encontrar novos componentes para produzir processadores em uma escala quântica. Embora materiais quânticos como o grafeno permitam superem o silício nesse aspecto, a própria física limita a potência dessas máquinas. Como o físico Hans-Joachim Bremermann demonstrou, "nenhum sistema de

[22] Goertzel, B., Pennachin C. (ed.), *Artificial General Intelligence*, Berlin, Springer, 2007.

processamento de dados, seja artificial ou vivo, pode processar mais do que (2×10^{47}) bits por segundo por grama de sua massa."[23]

Segundo ele, há um limite espaço temporal para a propagação das ondas eletromagnéticas, um limite quântico na frequência de transmissão da informação e um limite termodinâmico referente à entropia de compensação da informação. Nem o demônio de Maxwell pode superar as leis da física quântica. Seria necessário inventar uma nova física e então se referir à lógica das descobertas científicas. Mas como podemos imaginar as propriedades de uma teoria cujo fundamento ainda não existe?

As IAs hoje

Enquanto aguardamos por essa revolução científica, as IAs atuais são muito menos eficientes do que pensamos. Nosso dispositivo tecnológico atual consiste apenas em IAs "fracas", que realizam tarefas de armazenagem simples e repetitivas, como o Edge Rank do Facebook ou o PageRank, do Google. Na realidade, essas IAs apenas ordenam as páginas de acordo com critérios estabelecidos.

Além disso, com a cortina de fumaça da máquina consciente disseminamos ferramentas de baixa inteligência capazes de coletar nossos dados pessoais a fim de manipular nossas decisões, valores e apetites. Submetidos a esse imaginário de inferioridade intelectual, alienamos nossas faculdades mentais em ferramentas automatizadas sem sentido com uma descomplicada servidão voluntária, como se não tivéssemos mais que tomar decisões.

Ao se tornarem um espelho logístico de nossos próprios valores e apetites, as IAs transmitem, em alta frequência, a mentalidade excessivamente estruturada. Nosso *digital twin*, espelho de nossa personalidade, joga contra nós para promover um mundo onde as máquinas jogam conosco. Como Kasparov contra Deep Blue, ou Lee Sedol contra AlphaGo, as partidas que jogamos revelam nossas fragilidades para a máquina, que, sem

[23] H.-J. Bremermann, « Optimization through evolution and recombination » in (M. C. Yovitts et al. éd.), *Self-organization systems*, Washington, Spartan books, 1962, p. 93. Tradução livre.

o menor escrúpulo, nos põe em xeque. Nossa vulnerabilidade digital tornou-se, assim, a porta de entrada para o domínio de cada uma de nossas ações.

Cada palavra escrita, cada clique, como cada jogo jogado, alimenta o controle de nosso comportamento. A IA, então, encoraja o indivíduo em sua fragilidade, no isolamento, na reação aos costumes, opiniões e ideias recebidas em *looping* pelas redes sociais. A lei dos grandes números, que para os modernos deveria garantir a veracidade do julgamento, é hoje o vetor da manipulação de massa. Ao acumular nossos dados, as máquinas dão imenso poder aos propagandistas da estupidez.

Nessa duplicata virtual de nossas vidas entrelaçadas, todos os preconceitos discriminatórios que permeiam a sociedade civil são exacerbados. O racismo, a xenofobia, a misoginia e todas as trincheiras identitárias incapazes de considerar a alteridade são ampliados no processamento dos dados. Os vieses da IA são nossos próprios vieses. Tay, o chatbot neo-Nazi da Microsoft, as categorizações racistas de fotos feitas pelo Google e as recomendações de tratamento de câncer extremamente perigosas feitas pela IA Watson da IBM são provas disso.

Governar sem coagir

Na introdução de um seminário sobre Heráclito,[24] Heidegger e Eugen Fink propõem uma análise esclarecedora da nova forma de "governar" que emerge das tecnologias de informação e comunicação — na naquela época, reunidas na teoria cibernética. Segundo Heidegger: “Esse fenômeno (de governança) tornou-se hoje, precisamente na era da cibernética, tão fundamental, que desafia e determina todas as ciências naturais e o comportamento humano.”[25]

Complementando essa reflexão, Eugen Fink estabelece uma primeira definição do ato de governar: “governar é fazer-se senhor de um movimento (...); governar é um movimento que intervém, que transforma, que obriga o navio a seguir um curso

[24] Du grec « *kubernêtês* » signifiant « commande ».
[25] M. Heidegger, E. Fink, *Héraclite, séminaire du semestre d’hiver 1966-1967,* trad. J. Launay, P. Lévy, Paris, Gallimard, Tel, 2017, pp. 23-24. Tradução livre.

determinado. Isso inclui o movimento de dominação."[26] Trata-se do ato de governar pela coação do fluxo de informações.

Mas a essa primeira forma se acrescenta uma segunda: "Haveria também um governar sem coação?", questiona Heidegger. "Não é por acaso que as ciências naturais e a nossa vida hoje sejam cada vez mais dominadas pela cibernética."[27] Segundo ele, a cibernética concretiza a técnica ignorando seu poder sobre o mundo dos objetos e do comportamento humano. Nesse sentido, há um devir-máquina que está adormecido em nós e no mundo sob a égide de uma nova teleologia.

A causalidade é assim redefinida em termos de informação, desde o chip de silício até os neurônios. O controle da informação suaviza a violência da coação e eleva a ação cibernética a um nível superior. Ao dar ao ser humano "o poder de criar a organização instituindo a teleologia,"[28] a tecnologia afasta o gesto coercitivo em favor de um determinismo mental. Nesse sentido, essa nova organização da teleologia, própria ao "governar" cibernético, marca a capacidade de uma nova forma de controle insidioso pela informação.

A proposta de Heidegger e Fink manifesta assim a ideia de uma dupla transformação. Por um lado, o pensamento encontra sua externalização na máquina, na forma do programa; por outro, o pensamento se torna máquina, desde o neurônio até o comportamento. Surgem, então, novas formas de individuação. A tecnologia constrói uma nova realidade psicológica, que se articula no ecossistema da IA. As novas formas de individuação tecnológica tomam forma no seio da matéria, de modo que não é mais possível considerar o mundo independentemente das IA.

Um novo ecossistema

[26] *Ibid.*, p. 24. Tradução livre.
[27] *Ibid.*Tradução livre.
[28] G. Simondon, *Du mode d'existence des objets techniques,* Paris, Aubier, 1989, p. 104. Tradução livre.

A previsão de Turing[29] de que, ao final do século XX, a inteligência das máquinas seria aceita pela cultura popular, se mostrou acertada. Todos estamos de acordo em atribuir inteligência à máquina. Assim, vivemos sob o regime das IAs na orientação de nossa atenção, nossos desejos, nossa cultura e nosso pensamento.

Como isso aconteceu? Parece, muito simplesmente, que essa construção se sustenta em uma certa individuação própria dos objetos técnicos, exacerbada pelo entusiasmo contemporâneo pela tecnologia. Como demonstrou Simondon, a organização dos objetos técnicos passa pela formação sinérgica de estruturas nas quais toda a aleatoriedade do funcionamento interno tende a se reduzir gradualmente.

A individuação do objeto se articula em torno de sua progressiva autonomização. O objeto passa de uma forma abstrata, em que o funcionamento do sistema é provável, para uma forma concreta, em que seu funcionamento é estabilizado e autônomo. Nessa transição do abstrato para o concreto, a função do objeto perde sua importância em favor de seu funcionamento sinérgico. O funcionamento do objeto encontra seu equilíbrio, portanto, no desaparecimento de seu papel como objeto, ao mesmo tempo em que sua confiabilidade se torna uma certeza.

Quando não estamos mais preocupados com o funcionamento dos motores do avião, podemos considerar que o motor atingiu sua forma concreta. Da mesma forma, quando não nos preocupamos mais com o funcionamento das IAs na administração de nossa sociedade, na captação de nossa atenção, na direção de nossas vontades ou mesmo na manipulação de nossas crenças, pode-se dizer que elas são concretas.

Antoine de Saint-Exupéry formulou lindamente este processo através da concretização das asas do avião: "A própria máquina, quanto mais se aperfeiçoa, mais se apaga atrás de seu papel (...). O trabalho dos engenheiros, dos desenhistas, dos calculistas do escritório de design é, aparentemente, apenas polir, apagar, amaciar os encaixes, equilibrar as asas (de um avião), até que elas não sejam mais notadas (...). Parece que a

[29] A. Turing, « Les ordinateurs et l'intelligence » (1950), in *Vues de l'esprit*, éd. D. Hofstadter et D. Dennett, Paris, InterEditions, 1987, p. 65.

perfeição é alcançada não quando não há mais nada a acrescentar, mas quando não há mais nada a retirar. Ao fim de sua evolução, a máquina se esconde."[30]

A concretização do objeto técnico permite sua integração sinérgica em um ecossistema. Para Simondon, o objeto concreto torna-se semelhante às várias formas individuadas do mundo físico-químico. O objeto se adapta em seu sistema, como o O_2 no ar. "O objeto técnico concreto é um sistema físico-químico no qual as ações mútuas são exercidas de acordo com todas as leis da ciência."[31] O desaparecimento descrito por Saint-Exupéry corresponde, nesse sentido, à integração do objeto em um sistema, como um ser vivo na natureza.

Dessa forma, o artificial tende a se tornar natural. "Através da concretização técnica, o objeto, primitivamente artificial, torna-se cada vez mais semelhante ao objeto natural."[32] O servomecanismo, unindo o humano à máquina, tende a desenvolver uma sinergia própria. Como escreveu Bachelard sobre Saint-Exupéry: "O mestre do voo, em sua embriaguez dinâmica, se funde com a máquina. E realiza a síntese entre o imóvel e o em movimento."[33] No fluxo de nossos dados, fundir-se com a máquina é precisamente o que completa a concretização do objeto técnico. Quando o movimento e o que foi movido se fundem, a máquina se torna imperceptível. "O movimento está em uma relação essencial com o imperceptível; ele é, por natureza, imperceptível."[34]

Onde a sinergia é estabelecida, a consciência do objeto desaparece. Na sinergia de nossos pensamentos em movimento, as IAs desvanecem-se no imperceptível. Elas operam de acordo com uma sequência de estados discretos em casas, fábricas, escritórios, fazendas, hospitais, supermercados, laboratórios, contadores elétricos e postes de luz para

---

[30] A. de Saint-Exupéry, *La terre des hommes,* in *Œuvres complètes,* Gallimard, la Pléiade, T. 1, 1990, p. 170. Tradução livre.
[31] G. Simondon, *Du mode d'existence des objets techniques,* Paris, Aubier, l'invention philosophique, 1989, p. 34. Tradução livre.
[32] *Ibid.,* p. 47. Tradução livre.
[33] G. Bachelard, *L'air et les songes, Essai sur l'imagination du mouvement,* Paris, Librairie José Corti, 1978, p. 294, note. Tradução livre.
[34] G. Deleuze, F. Guattari, *Milles plateaux, capitalisme et schizophrénie 2,* Paris, éditions de minuit, 2001, p. 344. Tradução livre.

estruturar logisticamente nossos pensamentos. A IA funde-se com o sistema de tal forma, que ele se encaixa perfeitamente ao todo – a ponto de se tornar imperceptível. A especificação da IA não se dá apenas através da sinergia que seu funcionamento adquire no sistema, mas também por meio do esquecimento relativo que sua estabilidade confere à consciência do usuário.

Quando o funcionamento se apaga e se conecta perfeitamente ao sistema, torna-se quase natural, como se tudo fosse muito óbvio. A sinergia é estabelecida na ocultação dos funcionamentos técnicos. As IAs se individuam através de um devir imperceptível até se fundirem na consciência de todos. Parametrizando comportamentos, crenças, desejos e valores na perfilação dos usuários, elas abrem um novo caminho para a dominação do pensamento. Assim, a naturalização da IA em nosso mundo passa por um devir-imperceptível que se tece na linguagem. A linguagem das máquinas se fundiu com a nossa, a tal ponto que não podemos mais distinguir quem está falando.