# Os androides sonham com papagaios estocásticos? Pensamento, Saber(es) e Redes*

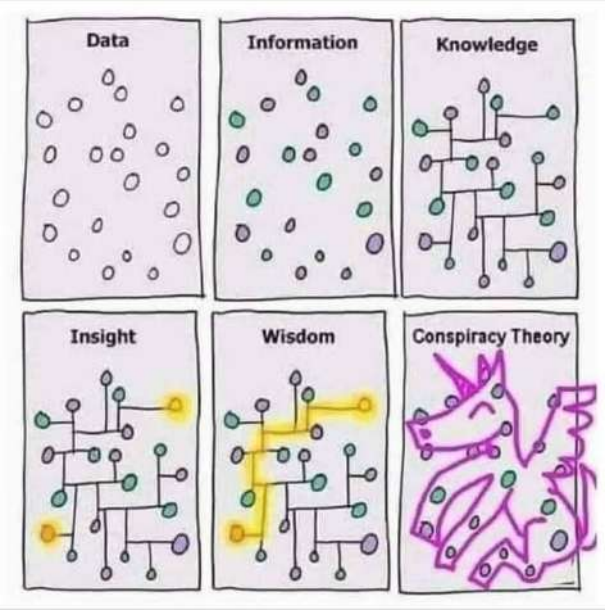
Luca Szaniecki Cocco

*Economista, doutorando da Universidade de Humboldt, Berlim.*

*Isso é uma referência ao livro de P. K. Dick, *Os Androides Sonham com Ovelhas Elétricas?*, que foi a principal inspiração para o filme *Blade Runner*.

Em um documentário recente dedicado ao pensamento de Hannah Arendt (*I am not a monster* por Nelly Ben Hayoun-Stépanian), um empreendedor japonês no ramo da inteligência artificial (Hiroshi Ishiguro) diz que a maior invenção dessa área foi a invenção do conhecimento (*knowledge*), e que antes só tínhamos/haviam *dados* (*datas*). Isso pode parecer inicialmente paradoxal no sentido mais forte do termo, pois vai contra a *doxa*, um consenso pessimista em relação às capacidades dessas máquinas de serem inteligentes. Por exemplo, um artigo recente qualifica os modelos de linguagem (do tipo ChatGPT) de *papagaios estatísticos/estocásticos* (Bender et al. 2021), no sentido de que eles seriam *apenas* capazes de prever o que vem em seguida em uma frase a partir de suas gigantescas bases de dados: no fundo, eles não compreendem nada, apenas copiam probabilisticamente o que já viram. Assim, o objetivo deste pequeno artigo é estudar um pouco mais detalhadamente o vínculo entre construção de redes, o ato de saber (ou conhecer) e o ato de pensar. Mais precisamente, trata-se de mostrar que, apesar de nossas pré-noções, nossa inteligência é mais artificial do que parece (no sentido de artifício, feito de uma série de bricolagens, *patchworks*, tecelagens) e, nesse sentido, a inteligência artificial é mais humana do que parece. Para retomar Nietzsche, a IA seria ao mesmo tempo humana, demasiado humana, e demasiado pouco humana.

Podemos começar com uma provocação graças à ajuda de um meme encontrado na internet, onde o autor (cujo nome não encontrei) tenta explicar uma série de conceitos e raciocínios por meio de pontos e como eles formam redes. Assim, uma série de pontos isolados constituiria apenas informação (ou dados, segundo as palavras do nosso amigo japonês), enquanto o saber seria uma situação onde essa série de pontos está interligada, formando uma rede mais ou menos densa. À primeira vista, isso pode parecer bastante intuitivo e até mesmo remeter ao conceito de inteligência em seu sentido etimológico.

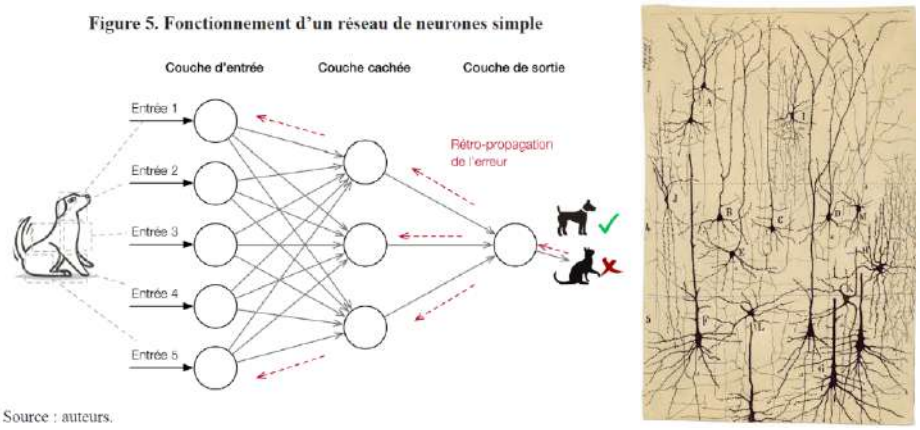



No entanto, essa conclusão não é tão evidente assim, e para isso precisamos introduzir algumas noções a mais, notadamente a de pensamento. Mais precisamente, defenderemos também a ideia de que, mesmo que essa tecelagem seja fundamental no conhecimento e no pensamento, trata-se de uma tecelagem muito particular.

Em primeiro lugar, para ilustrar essa dificuldade, podemos remeter a um debate próprio das ciências cognitivas e do próprio campo da inteligência artificial. Durante muito tempo, a doxa oriunda da cibernética estava longe de admitir essa noção de conhecimento como oriunda de uma rede de informações. Pelo contrário, a corrente simbólica insistia mais nas capacidades lógicas que determinam nosso pensamento. Segundo essa concepção, *pensar é calcular símbolos que têm ao mesmo tempo uma realidade material e um valor semântico de representação* (Cardon et al. 2018). Essa visão do pensamento tem uma estranha familiaridade com o pensamento cartesiano, no sentido em que o homem pode se conhecer racionalmente por introspecção e onde conhecer é criar um modelo (no sentido matemático) dos fenômenos para realizar manipulações lógicas (Dupuy 2013). Aliás, não é por acaso que o próprio Descartes também era fascinado pela imagem da máquina como modelo mecânico do corpo humano[1].

[1] Dito isso, admitimos ter uma visão bastante caricatural do pensamento de Descartes, o que parece ser algo bastante recorrente nas ciências cognitivas (cf. *O Erro de Descartes* por Damasio). Para uma visão mais nuançada do pensamento de Descartes e sua natureza paradoxalmente "dialogal", seria necessário ler Jean-

No entanto, as máquinas que conhecemos hoje não são as mesmas que Descartes concebia. A corrente conexionista, outrora marginalizada e hoje dominante, defende uma concepção mais *emergente* onde o pensamento poderia emergir de uma aglomeração de pequenos elementos que não são necessariamente pensantes por si mesmos. Esse pensamento inspirará novos métodos estatísticos, como as redes neurais (as de Yann LeCun, por exemplo), inspirados pelos trabalhos pioneiros de McCullough e Pitts (entre outros). Segundo essa concepção, *pensar assemelha-se a um cálculo massivamente paralelo de funções elementares (...) cujos comportamentos significativos aparecem ao nível coletivo apenas como um efeito emergente das interações produzidas por essas operações elementares* (Andler 1992). Inspirado pela teoria da informação, a informação em si não precisa estar associada a um sentido preciso, como supunha a corrente simbólica. Em termos de raciocínio, trata-se de máquinas que privilegiam o cálculo indutivo em vez do raciocínio dedutivo das máquinas lógico-simbólicas.




Fonte: À esquerda, um exemplo de rede de neurônios extraído de Cardon et al. (2018). À direita, um desenho de neurônios de Ramon y Cajal do início do século XX.

Luc Marion (ou ouvi-lo na France Culture). Para uma crítica mais ácida e completa de Descartes, seria mais apropriado ler Daniel Dennett.

Uma rede de neurônios pode ser concebida da seguinte forma, graças a dois elementos principais: unidades (que podemos chamar por analogia de neurônios) e conexões. Em termos de estrutura, as unidades são organizadas em camadas interligadas. Podemos distinguir uma camada de entrada (onde cada entrada pode corresponder a uma *unidade perceptual*, por exemplo), camadas *ocultas* e uma camada de saída (um julgamento, por exemplo). Cada neurônio na camada de entrada seria responsável por uma componente particular do input de uma imagem, por exemplo, e cada neurônio está conectado à próxima camada com um determinado peso. Por *peso*, queremos dizer um coeficiente que determina a importância dessa conexão na ativação da próxima camada e assim por diante até a camada de saída. A distribuição dos pesos é inicialmente relativamente arbitrária (ou até aleatória em alguns casos), mas se adapta de acordo com a resposta dada pela camada de saída (por um sistema de *feedback*). Por exemplo, a máquina pode dizer que a imagem corresponde a um gato, quando na verdade é um cachorro; nesse caso, há uma mensagem de erro que será propagada na direção oposta (da saída para a entrada, ou *retropropagação*) para alterar o valor dos pesos conforme uma função matemática dada (o que não nos interessa necessariamente aqui). Por exemplo, imaginemos que o neurônio responsável pela detecção de uma cauda tem um peso bastante importante, mas cães e gatos têm caudas; assim, se houver um erro, o peso desse neurônio pode diminuir para dar mais peso a neurônios que seriam responsáveis por propriedades mais distintivas de um cachorro (o focinho, por exemplo). Assim, pouco a pouco, por meio de treinamento, os neurônios *aprendem* e a máquina abrangente será cada vez mais eficiente no reconhecimento de um objeto (ou uma série de objetos)[2]. No entanto, isso ainda não é suficiente para determinar se uma máquina pensa como nós.

Para determinar isso, achamos necessário introduzir uma distinção entre pensamento e conhecimento, que também é encontrada na filosofia política, particularmente em Hannah Arendt. No mesmo documentário citado anteriormente, um

[2] Devo dizer que trata-se de uma caricatura do que se passa realmente em modelos de neurônios artificiais, sem contar que modelos de linguagem como ChatGPT são mais complexos que essas camadas de neurônios, porém explicá-los necessitaria maior tecnicalidade.

dos alunos de Arendt (Richard J. Bernstein) distingue esses dois conceitos explicando que o conhecimento é exclusivamente voltado para a verdade, enquanto o pensamento se interessa pela busca de sentido. Mesmo que essa distinção faça sentido e seja interessante, talvez subestimemos como podemos chegar ao pensamento a partir do conhecimento puro de maneira emergente, à maneira dos conexionistas.

Vamos considerar o famoso teste de Turing (Turing 1950). Trata-se de um jogo a três (uma máquina, um ser humano e um interrogador) separados por uma tela e sem poder se ver ou ouvir. O jogo é simples: o jogador principal (o ser humano, um homem) deve fingir ser uma mulher para o interrogador. Assim, se substituirmos o homem pela máquina, a máquina deve não apenas fingir ser uma mulher, mas também simular o ato de simular (da parte do homem). Se o interrogador for incapaz de diferenciar o homem da máquina, a máquina passou no teste de Turing e é considerada inteligente segundo essa concepção.

Mesmo que esse teste tenha se difundido rapidamente, especialmente na cultura *pop*, ele foi criticado rapidamente, seguindo o argumento de que uma máquina poderia passar no teste sem entender o que diz; isso não seria prova de uma capacidade de pensamento. Uma imagem conhecida é a da *sala chinesa* de Searle (Searle 1999). Ele descreve uma sala isolada na qual há um homem com acesso a uma quantidade indefinida e supostamente (quase) infinita de conhecimento. De fora, alguém desliza uma mensagem em chinês; o homem não sabe nada de chinês, mas dispõe de um conjunto de regras, procedimentos a seguir (ou algoritmos) que lhe permitem responder em caracteres chineses, seguindo-os à risca, mesmo sem entender o que diz. Assim, Searle defende que uma máquina dessas poderia passar no teste de Turing mesmo que só realizasse uma manipulação vazia de símbolos. Mas, primeiro, o que seria compreender, afinal? Primeiramente, um conhecimento objetivo das relações entre objetos, mas também uma

dimensão subjetiva ligada ao fato de conhecer em primeira pessoa; a máquina de Searle é capaz do primeiro, mas não do segundo[3].

Mesmo que o argumento de Searle seja interessante, achamos que ele superestima a importância da dimensão subjetiva e que não critica Turing da maneira correta. Alguns diriam até mesmo que a consciência-de-si (*self-consciousness*) foi enganosamente associada à noção de consciência na tradição da fenomenologia (cf. Millière, R. "*Constitutive Self-Consciousness*"). De qualquer forma, a crítica de Searle é encontrada indiretamente em outros autores, como D. Chalmers, que descreve um problema *fácil* e *difícil* da consciência (Chalmers 1995). O problema fácil trata de estudar como a consciência se manifesta por meio das relações entre nossos cinco sentidos de forma relativamente objetiva (que pode ser estudada pelo método científico positivista). Ao contrário, o problema difícil trata de estudar o que faz da nossa consciência uma experiência puramente subjetiva (um *qualia*, segundo o jargão dessa literatura): para retomar o título de um famoso artigo de Nagel, podemos compreender objetivamente como funciona o sentido de orientação de um morcego, mas nunca teremos acesso à sensação de *ser* um morcego (Nagel 1974). No entanto, segundo D. Dennett, essa distinção é mal formulada, pois essas duas propriedades da consciência não podem ser separadas. Nossa tendência a querer defender a particularidade intrinsecamente subjetiva e inacessível da nossa consciência remete em parte ao que Dennett chama de *postura intencional*, o ato de querer incessantemente atribuir a outrem (ou a um coletivo) estados mentais (de tipo intencional). Frente aos novos desenvolvimentos da Inteligência Artificial, essa estratégia serve principalmente como autodefesa, combinada a um forte

[3] Eu retomo aqui em parte os argumentos de Monsieur Phi: https://www.youtube.com/watch?v=_XJsAQsT0Bo

sentimento de ansiedade social ligado à possibilidade de um *grande substituição*[4] do trabalho humano pelas máquinas[5].

Isso dito, Searle pode ter razão ao criticar o teste de Turing, que claramente já não é mais suficiente para o estado atual de nossas máquinas (um bom resumo do que pesquisadores tentam realizar hoje em dia pode ser lido em Srivastav et al. (2022)). De fato, mesmo que o argumento de Searle não seja o mais adequado, a crítica permanece a mesma segundo Dennett: Turing também superestimou a capacidade de nossas novas máquinas (redes neurais, etc.) de utilizar dados em grande escala de maneira explorável sem a verdadeira capacidade de compreensão (Dennett). Mas podemos ir mais longe, defendendo a ideia de que essa exploração de dados é, na prática, muito próxima do que chamamos de compreensão, uma vez que não é concebida de forma monolítica. Segundo a tipologia de Marconi (1997), podemos entender isso pelo menos de duas maneiras: de forma *referencial* (o sentido das palavras a partir de que elas fazem referência a) ou *inferencial*, onde as relações entre palavras são fundamentais, inferindo o sentido da rede que elas formam com outras palavras. Mesmo que essa concepção pareça bastante recente, pois mais probabilística do que simbólica (e para alguns ela é até característica dos modelos de linguagem como o ChatGPT), ela não é tão estranha às antigas concepções na filosofia da linguagem (por exemplo, os eixos sintagmáticos e paradigmáticos de Ferdinand de Saussure).

Assim, a conclusão seria que entender a estrutura das relações entre palavras reflete em grande parte a estrutura da nossa compreensão do mundo e, nesse sentido, o ChatGPT tem capacidades, sim, semânticas[6]. Certamente, essas capacidades são limitadas e devem ser melhor compreendidas também para melhor regulamentá-las, em vez de

[4] https://www.lemonde.fr/economie/article/2023/12/13/prix-du-livre-d-economie-avec-l-intelligence-artificielle-pres-d-un-francais-sur-deux-craint-pour-son-emploi_6205546_3234.html

[5] Não queremos dizer que esses sentimentos sejam ilegítimos, mas sim que não são fundamentados na realidade da situação e, portanto, não são necessariamente muito eficazes.

[6] Nós retomamos aqui em grande parte os argumentos de Raphaël Millière: https://www.youtube.com/watch?v=aUJOcVPdDvg

negar sua existência para proteger nossa frágil noção de subjetividade. Retomando as críticas clássicas em relação ao ChatGPT: é verdade que o modelo tem dificuldade para compreender o que é *simples* para nós (Mitchell 2009), coisas advindas do *senso comum*, mas quando o número de parâmetros utilizados é aumentado (*scaling*), esses modelos explicam cada vez melhor coisas "complexas", como piadas. No quesito da imitação, embora o jogo de Turing seja criticado por sua simplicidade, imitar alguém ou alguma coisa nos parece uma etapa fundamental no processo de compreensão: por exemplo, quando tentamos imitar o sotaque das pessoas ao aprender uma nova língua. Além do mais, tais perguntas não são estrangeiras a outros dominios da cultura humana, inclusive na propria criação teatral, como com a emergência do metodo Stanilavski (mais "naturalista) em contraste com a atuação dramática clássica. Outra crítica, a dos *papagaios estocásticos*, também subestima o poder da previsão: certamente, é apenas a previsão, mas prever o próximo movimento do número 1 mundial no xadrez é bastante notável e supera amplamente as capacidades de muitos de nós (eu incluído). Enfim, estes são apenas alguns exemplos de como subestimamos a IA, e para uma visão mais ampla do fenômeno recomendo ler Millière & Buckner (2024).

Podemos concluir este artigo ressaltando o que não queremos dizer. Em primeiro lugar, não queremos defender que nossas máquinas sejam superiores e mais inteligentes do que nós e que estamos próximos de uma singularidade (Kurzweil 2005). Além disso, não queremos defender a onipotência das máquinas conexionistas como a nova fronteira em relação à corrente simbólica. No entanto, acreditamos que as críticas e defesas em relação aos novos desenvolvimentos da inteligência artificial são bastante caricaturais e falham em abordar os verdadeiros problemas. Tentamos explicar como essas críticas se inscrevem em uma tendência *natural* que serve como autodefesa de nossa própria concepção de subjetividade, mesmo que errada. De certa forma, isso se inscreve na continuidade das *feridas narcísicas* de Freud: após a revolução copernicana (a Terra não está no centro do universo), a de Darwin (o homem é um animal como os outros) e a *descoberta* do inconsciente (o homem *não é mais senhor de sua própria casa*), vem a

inteligência artificial que nos lembra que a tecnologia, assim como a biologia, não existe na ausência de evolução, como a própria vida. Conhecer melhor o que a IA é capaz ou não é fundamental para podermos regulá-la melhor. A filosofia deve, portanto, desempenhar um papel mais importante; ela pode servir como uma ferramenta de compreensão desses novos desafios e de seus problemas. Mais precisamente, pode ajudar a entender como o indivíduo não tem mais o monopólio de certos atributos da subjetividade, que agora constituem efeitos emergentes de uma aglomeração de redes (Dupuy 2013). No final, segundo as palavras de J. Monod, isso remete a um debate plurissecular (pelo menos na filosofia ocidental) entre os partidários de uma concepção em que a realidade do universo reside em suas formas imutáveis e invariáveis por essência, e aqueles que defendem que a verdade repousa nos fluxos e na evolução (Monod 1974). O debate atual em torno da IA é apenas um capítulo adicional dessa longa história que é a história da inteligência e de como o homem se vê a si mesmo: se por muito tempo ele se apoiou nessa concepção estável do logos, talvez seja hora de explorar o domínio da *mètis*, uma inteligência que prioriza o devir e a plasticidade (Malabou 2017).

Ou seja, antes de atacar a inteligência artificial, devemos primeiramente nos atacar ao questionamento do que constitui a inteligência humana[7], por muito tempo prisioneira de diversos eugenismos. ChatGPT pode ser um papagaio estocástico, mas nos também o somos de certa forma. Sobretudo, escrever milhares de artigos sobre o quanto esses modelos são estúpidos, por mais corretos que eles sejam (às vezes), é direcionar nossas energias em perguntas erradas. Como dizia Henri Bergson, uma grande parte da filosofia é saber fazer as perguntas corretas.

Bibliografia

Andler, D. (1992). From paleo-to neo-connectionism. In *New Perspectives on Cybernetics: Self-Organization, Autonomy and Connectionism* (pp. 125-146). Dordrecht: Springer Netherlands.

[7] For uma tentative recente, por um ator no ramo da IA: Chollet (2019). Um resumo pode ser achado no seguinte podcast: https://www.youtube.com/watch?v=rTh3UcPj_7o.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Cardon, D., Cointet, J. P., & Mazières, A. (2018). La revanche des neurones. *Réseaux*, *211*(5), 173-220.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200-219.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Dupuy, J. P. (2013). *Aux origines des sciences cognitives*. La découverte.

Kurzweil, R. (2005). The singularity is near. In *Ethics and emerging technologies* (pp. 393-406). London: Palgrave Macmillan UK.

Malabou, C. (2017). *Métamorphoses de l'intelligence: que faire de leur cerveau bleu?*. puf.
Marconi, D. (1997). *Lexical competence*. MIT press.

Millière, R. Constitutive Self-Consciousness.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models-Part II: The Way Forward. *arXiv preprint arXiv:2405.03207*.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models-Part II: The Way Forward. *arXiv preprint arXiv:2405.03207*.

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.

Monod, J. (1974). *On chance and necessity* (pp. 357-375). Macmillan Education UK.

Nagel, T. (1974). What is it to be a bat. *Philosophical Review*, *83*(4), 435-450.

Searle, J. (1999). The Chinese Room.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.