Artigo Original doi:10.1590/0100-512X2025n16209chb

# A COLONIZAÇÃO DO JUÍZO PELO CÁLCULO*

# *RECKONING'S COLONIZATION OF JUDGMENT*

*Carlos Henrique Barth*
*https://orcid.org/0000-0002-9327-9818*
*carlos@cbarth.me*
*FAJE – Faculdade Jesuíta de Filosofia e Teologia, Belo Horizonte, Minas Gerais, Brasil*

*Elton Vitoriano Ribeiro*
*https://orcid.org/0000-0003-2276-7221*
*elton.ribeiro@faje.asav.org.br*
*FAJE – Faculdade Jesuíta de Filosofia e Teologia, Belo Horizonte, Minas Gerais, Brasil*

**RESUMO** *Há crescente confiança em sistemas algorítmicos, particularmente aqueles baseados em Inteligência Artificial (IA), cujo modo de operação denominamos cálculo. Eles influenciam e substituem tomadores de decisão humanos em domínios cada vez mais sensíveis. Isso ocorre a despeito de profundas divergências no modo como o juízo humano e o cálculo operam, e de persistentes desafios à tentativa de superá-las. Neste artigo, defendemos que essa confiança é desmedida e decorre parcialmente de uma falha em perceber o fenômeno da colonização do juízo pelo cálculo: o que aparece como uma crescente acurácia na simulação do juízo é, em boa medida, fruto de uma crescente influência do cálculo no modo como o juízo opera, delimitando seus*

* Artigo submetido em: 03/04/2025. Aprovado em: 16/10/2025.

*contextos de operação e incitando a adoção de suas trajetórias inferenciais. Isso se mostra em habilidades cognitivas fundamentais, como a de determinar fatores relevantes em quaisquer contextos, inclusive os de deliberação moral.*

**Palavras-chave**: *Inteligência artificial. Juízo. Sensibilidade à relevância. Tipos humanos.*

**ABSTRACT** There is growing reliance in algorithmic systems, particularly those relying on artificial intelligence (AI), whose mode of operation we call reckoning. Despite the profound differences in the way human judgment and reckoning operate, as well as the persistent challenges in trying to overcome them, such systems influence and even replace human decision-makers in increasingly sensitive areas. In this paper, we argue that this reliance is misplaced and stems from the failure to realize how the reckoning is able to colonize judgment. What seems to be a growing accuracy in simulating judgment is, to a large extent, a growing influence of reckoning over judgment, for the reckoning delimits the contexts of operation and primes the adoption of its own inferential trajectories. This reflects on fundamental cognitive skills, such as realizing the relevant factors in any context.

**Keywords**: *Artificial intelligence. Judgment. Relevance realization. Human kinds.*

## Introdução

Este artigo defende duas teses. A primeira é a de que existem hoje pelo menos duas formas fundamentalmente distintas de caracterizar e processar informações sobre o mundo: o *cálculo*, associado à Inteligência Artificial (IA), e o *juízo*, característico da cognição humana.[1] A segunda é a de que o juízo vem sendo *colonizado* pelo cálculo, modulando-o e reestruturando-o de forma profunda.

A primeira tese tem como pano de fundo a IA enquanto empreitada que busca modelar capacidades da inteligência humana em sistemas computacionais. Historicamente, é comum dividir a IA em duas grandes

1 A distinção é inspirada em Smith (2019), que utiliza os termos "reckoning" – usualmente traduzido como "cálculo" – e "judgment" – traduzido aqui como "juízo".

ondas ou gerações: a IA clássica, também conhecida como *GOFAI*[2] (*good old fashioned AI*, ou "boa e velha IA"), que utiliza arquiteturas computacionais tradicionais, e a IA *conexionista*, baseada na conjunção de arquiteturas computacionais neurais e algoritmos de aprendizagem (particularmente os algoritmos de *deep learning*).[3] A presente discussão, contudo, parte de algo comum às duas ondas: o norte contra o qual os avanços da IA costumam ser mensurados. Embora busque simular capacidades associadas à inteligência humana, a IA nunca teve uma grande preocupação em especificar o que a inteligência é. Em geral, ela se guia por parâmetros "comportamentais", isto é, que se concentram no resultado obtido: ao simular uma dada capacidade, se o *output* produzido for suficientemente semelhante ao que teria sido produzido por um ser humano, entende-se que a capacidade está sendo adequadamente simulada.[4] Assim, os avanços da IA são reconhecidos na medida em que se mostram capazes de simular um número cada vez maior de capacidades humanas associadas à inteligência, ou que o façam de modo mais eficiente.

A IA supõe a possibilidade de simular essas capacidades por meios computacionais, mas essa suposição será aqui desafiada no caso de uma capacidade em particular. Para fazê-lo, operaremos com a distinção de Smith (2019) entre cálculo e juízo. As ferramentas disponíveis a sistemas computacionais serão aqui acomodadas sob o termo "cálculo". Por sua vez, as ferramentas disponíveis ao aparato cognitivo humano serão abarcadas pelo termo "juízo". Fornecer uma descrição exaustiva dessas ferramentas é algo claramente fora do alcance de um artigo, e por isso trabalharemos a partir do contraste entre elas. Considere o exemplo de Xu *et al.* (2019). Nele, os autores mostram como sistemas computacionais de reconhecimento facial podem falhar sistematicamente em perceber que estão diante de um rosto, se o alvo estiver usando uma camiseta com determinados padrões de cores. A sensibilidade a diferentes condições adversas indica a presença de operações funcionalmente distintas das realizadas pelo aparato cognitivo humano:

2 A expressão foi introduzida por Haugeland (1985).

3 Há várias formas de traçar a distinção entre as duas ondas da IA, mas todas carregam imprecisões. Por exemplo, é comum afirmar que as arquiteturas tradicionais – tais como os formalismos lógicos (McCarthy, 1968), as estruturas de *frames* (Minsky, 1997), ou os sistemas de produção (Newell, 1994) – são simbólicas, mas esse nem sempre é o caso (Cummins, 1996). Por sua vez, é comum associar algoritmos de aprendizagem (*machine learning*) a arquiteturas neurais, mas diversas técnicas de aprendizagem operam com arquiteturas clássicas (Rudin, 2018). Para os fins deste trabalho, é suficiente notar que a onda contemporânea se baseia na conjunção de um tipo de arquitetura (neural) e um conjunto de técnicas de *machine learning* dedicado a esse tipo de arquitetura (o que se convencionou chamar de *deep learning*).

4 Esta é a intuição que guia o famoso teste de Turing (1950): se o conteúdo linguístico produzido por uma IA for indistinguível do conteúdo linguístico que um ser humano produziria, não teremos razão para rejeitar a atribuição de inteligência àquela IA.

embora não saibamos tudo o que há para saber sobre como identificamos faces, sabemos que nossa *performance* independe da camiseta utilizada pelo alvo.

Apresentaremos aqui uma divergência de tipo semelhante, porém muito mais abrangente, no modo como juízo e cálculo operam. Trata-se de uma distinção envolvendo a capacidade de delimitar o que é relevante – para qualquer dado objetivo – em um número indefinidamente multiplicável de contextos, o que chamamos de *sensibilidade à relevância* (SR). Para um rápido exemplo de como ela se manifesta, considere o problema de simular o senso comum envolvido na compreensão da sentença a seguir:

(1) Deixei a capa de chuva na banheira porque ela estava molhada.

O que estava molhada, a banheira ou a capa de chuva? O léxico não nos permite desambiguar, mas nossa compreensão dos usos típicos desses itens nos leva a apostar na capa de chuva. Essa familiaridade com a gigantesca estrutura de atividades humanas típicas está sempre presente no juízo humano. Navegá-la depende de sermos capazes de detectar o que é contextualmente relevante de modo fluido e adaptativo.

Argumentaremos que o cálculo não é capaz de simular a SR tal como ela se mostra no juízo. Mais precisamente, ainda que o cálculo consiga reproduzir os *outputs* da SR do juízo em um número razoável de contextos, ele não é capaz de simular o modo a partir do qual o juízo opera. As trajetórias percorridas pelo cálculo e pelo juízo ao navegar o ambiente informacional nunca são suficientemente semelhantes.

A despeito dessa profunda diferença, temos colocado cálculo e juízo para interagir em um número crescente de casos. Os exemplos mais usuais envolvem tarefas como determinar o que se mostra em nossas redes sociais, ordenar os resultados de nossas buscas on-line e sugerir conteúdos que nos interessam. Porém, o cálculo já é usado em contextos mais sensíveis. Ele classifica nossas demandas no poder judiciário, avalia o risco de concessão de crédito e ajuda seguradoras e planos de saúde a precificar nossa saúde e nosso comportamento. Nesses casos, eventuais inadequações são mais difíceis de diagnosticar, modular e corrigir, afetando principalmente o réu, o tomador de empréstimo, o paciente.

Não bastasse, o cálculo avança em cenários ainda mais sensíveis: *chatbots* atuam como terapeutas (Haque; Rubya, 2023); carros autônomos em situações adversas escolhem entre manobras que arriscam a vida dos passageiros e manobras que ameaçam pedestres (Hawkings, 2018; Marshall; Davies, 2018);

armas militares atiram caso entendam que o indivíduo à frente é uma ameaça e não um civil inocente (Adam, 2024). Nessas circunstâncias, mesmo pequenos erros têm consequências drásticas e potencialmente irreversíveis.

Este é o cenário da segunda tese: em certos casos, o cálculo consegue delimitar e circunscrever os contextos em que o juízo atua, afetando sua SR. Nesses casos, o que parece uma crescente capacidade de predição do que o juízo julgaria relevante se revela uma crescente capacidade de incitação e delimitação do exercício do juízo. Isso é o que denominamos *colonização do juízo pelo cálculo*. O juízo se limita a operar de modo circunscrito ao que é capturável pelo cálculo em um número crescente de atividades, contribuindo para uma confiança desmedida nas suas capacidades e para a ocultação de suas limitações. No que se segue, defenderemos esse diagnóstico, apresentando os mecanismos por meio dos quais esse efeito se dá.

Procederemos da seguinte forma: na primeira seção, apresentamos mais detalhadamente a SR como uma capacidade fundamental para o exercício tanto do cálculo, quanto do juízo. Nas seções 2 e 3, descrevemos se e como essa capacidade se mostra, no juízo e no cálculo, respectivamente. Na seção 4, detalharemos os mecanismos por meio dos quais a colonização acontece, utilizando para isso alguns elementos do trabalho de Ian Hacking.

## 1. Sensibilidade à relevância

A cada momento, temos diante de nós um grande volume de recursos cognitivos à disposição: conhecimento teórico e prático, habilidades, experiências acumuladas, percepção dos arredores, etc. Em conjunto, eles permitem um número ilimitado de inferências, mas somente algumas são relevantes para nossos objetivos momentâneos.[5] A SR é essa habilidade de identificá-las em um número ilimitado de situações. Ao teorizarmos sobre, por exemplo, comunicação, racionalidade ou epistemologia, tomamos a SR como um recurso explicativo, ou seja, na medida em que ela participa da teoria em questão, ela o faz enquanto *explanans*. Parte do que torna a comunicação humana possível é a capacidade que falante e ouvinte compartilham de continuamente revisar o que é relevante para a conversa em desenvolvimento. De modo similar, quando se realiza uma abdução – uma inferência para a

5 O termo "inferência" é usado em sentido amplo, isto é, ele não referencia apenas inferências lógicas ou linguísticas, mas também inferências perceptuais ("a iluminação está ruim, mas aquela pessoa parece ser o Igor"), simulações estruturais ("se aquela criança continuar pulando em cima do sofá, o móvel vai quebrar"), e assim por diante.

melhor explicação de um fenômeno –, também supomos a capacidade de delimitar o que conta como uma evidência relevante e o que deve ser relegado a uma cláusula *ceteris paribus*.[6]

Em contraste, ao teorizarmos sobre a cognição, a capacidade de se ater ao que é relevante emerge como um *explanandum*, i.e. como um fenômeno a ser explicado. O cenário em que essa preocupação surge é aquele em que precisamos decidir o que pensar – sobre uma situação particular –, a partir de tudo o que sabemos. Com o advento da IA e das ciências cognitivas, cenários desse tipo se tornaram abundantes. Nessas circunstâncias, a análise não parte de uma dada conclusão, mas de um número ilimitado de axiomas, suposições e hipóteses, levantando a questão de como identificar os que levam a inferências contextualmente relevantes.

A SR subjaz a maior parte da atividade cognitiva humana, fazendo dela um tópico central para a IA e para as ciências cognitivas (Vervaeke; Lillicrap; Richards, 2012). Nesse âmbito, o problema de explicá-la é conhecido como *frame problem*. Sua primeira aparição se deu como uma dificuldade circunscrita à tentativa de expressar cláusulas *ceteris paribus* em certos formalismos lógicos (McCarthy; Hayes, 1969); no entanto, foi rapidamente identificada por Dennett (1981) e outros como um problema muito mais amplo e não circunscrito a esses formalismos. De fato, a amplitude e a profundidade do *frame problem*, i.e. do problema de como modelar a SR, são amplamente discutidas há décadas (Ford; Pylyshyn, 1996; Kiverstein; Wheeler, 2012; Pylyshyn, 1987; Wheeler, 2008). Mas o que há de tão desafiador na SR? É assim que essa capacidade aparece, por exemplo, em Hume:

> Nada é mais admirável que a rapidez com que a imaginação sugere suas ideias, apresentando-as no instante em que elas se tornam necessárias ou úteis. A fantasia percorre o universo de um extremo ao outro, reunindo as ideias que dizem respeito a um determinado assunto. É como se a totalidade do mundo intelectual das ideias fosse a um só tempo exposta à nossa visão, e simplesmente escolhêssemos as mais adequadas a nosso propósito. No entanto, as únicas ideias que podem estar presentes são aquelas que foram reunidas por essa espécie de faculdade mágica da alma, a qual, embora seja sempre a mais perfeita possível nos grandes gênios – constituindo, aliás, precisamente o que denominamos gênio – permanece inexplicável para o entendimento humano, a despeito de todos os seus esforços (Hume, 2000, p. 48).

Em sistemas computacionais, a necessidade de fazer o que Hume descreve como "percorrer o universo de um extremo ao outro" a fim de identificar as ideias mais "adequadas a nosso propósito" leva a explosões combinatoriais.

6 Essas capacidades não são infalíveis, claro, mas nosso desempenho é muito superior ao do acaso.

É preciso encontrar um outro caminho, portanto. As primeiras dificuldades de evitar essa explosão se mostram como uma resistência a explicações por meio da simplificação. Na física, por exemplo, a complexidade do mundo é frequentemente abstraída, como quando, ao calcular a aceleração de um objeto, se ignora a resistência do ar ou as condições de temperatura e pressão atmosférica. Parece uma estratégia promissora: e se ao invés de percorrer o universo de um extremo ao outro, como diz Hume, percorrêssemos apenas modelos simplificados desse universo?

Infelizmente, no caso da SR, toda simplificação parece resultar na perda de informações circunstanciais determinantes. Isso se mostra mesmo em contextos simples, como o de enfeitar a casa para uma época festiva. Diante da necessidade de pregar um enfeite numa porta de madeira, ainda que tenhamos um peso de papel em casa, suas propriedades físicas (densidade, massa, forma, etc.) parecem irrelevantes e, portanto, podem ser desconsideradas, a princípio. No entanto, se um martelo não estiver disponível, fatores circunstanciais, como a urgência, podem fazer com que essas propriedades se tornem relevantes na deliberação ("será que consigo usar o peso de papel no lugar do martelo?"). Isso ilustra como a relevância de uma propriedade ou objeto é determinada por relações de interdependência não antecipáveis *a priori*. Toda (ou quase toda) propriedade da situação concreta é, no mínimo, potencialmente relevante.

Esse caminho nos coloca diante de uma circularidade: para explicar a SR (sem apelar para uma consideração exaustiva das múltiplas relações entre todas as propriedades da situação), é preciso "simplificar" a situação, abstraindo algumas propriedades; mas saber quais podem ser deixadas de lado pressupõe a capacidade de determinar sua relevância.

A dificuldade se intensifica porque toda situação concreta é única e, nesse sentido, inédita. Considere, por exemplo, a simples tarefa de decidir o que beber num restaurante. Certamente há fatores que se repetem em quase toda situação similar: a existência de um cardápio, a disponibilidade da bebida, etc. Mas nossa decisão é também sensível à presença de certas pessoas (estou acompanhado?), à ausência de outras (colegas de trabalho estão presentes?), à temperatura, ao horário, à época do ano, à saúde de um parente distante, e assim por diante. Nenhum desses elementos está diretamente associado ao que é típico ou característico de restaurantes. Um arranjo inédito de um número indefinidamente multiplicável de fatores pode transformar uma decisão em outra.[7]

7 Esse grau de dependência dos contextos concretos impede a modelagem da SR por meio de uma "teoria da relevância" que se pretenda geral. Essa é a principal razão pela qual o trabalho de Sperber e Wilson

A despeito dos desafios envolvidos, parece não haver alternativa a buscar uma explicação adequada. Na ausência da SR, só há dois caminhos possíveis: ou tomar todo e qualquer elemento de toda situação como relevante, ou tomar todo e qualquer elemento de toda situação como irrelevante. Os dois caminhos bloqueiam a deliberação e a ação. O primeiro leva a uma deliberação sem fim, o que Fodor (1987) caracterizou como o problema de Hamlet da perspectiva de um engenheiro: quando parar de pensar? Por sua vez, o segundo leva à inércia, pois nada motiva a ação. Dada nossa capacidade de deliberar e agir, resta evidente que, de algum modo, a faculdade que Hume reputou "mágica" se faz presente, e somos capazes de nos ater ao que é circunstancialmente relevante.

Para nossos propósitos, é importante sublinhar que a SR é indissociável da capacidade de inteligir uma situação, isto é, percebê-la de um certo modo (estou numa aula importante; estou num evento tedioso, mas importante para um ente querido, etc.). Determinar o que é relevante numa situação é parte do que significa compreendê-la em termos das ações que ela possibilita, inibe ou demanda. Uma vez que pesquisadores de IA têm a pretensão de produzir sistemas capazes de interagir autonomamente com seres humanos em contextos não controlados, compartilhar o modo como se compreendem situações concretas parece indispensável. Ambos precisam inteligir o mundo, se não da mesma forma, ao menos com um grau de intersecção suficiente para os fins em questão.

No caso do cálculo, os cenários mais sensíveis são os que envolvem juízos morais. Ainda que se possa embutir um dado princípio ético numa IA, sua aplicação depende da capacidade de determinar os elementos moralmente relevantes em uma dada situação. Princípios com a forma "não mentir a menos que a vida de alguém esteja em risco" não podem ser adequadamente seguidos, a menos que se saiba identificar os elementos circunstancialmente relevantes para determinar se a vida de alguém está em risco.

Nesse cenário, podemos aprender em que medida sistemas artificiais compartilham do mundo humano analisando o modo como determinam o que é relevante em cada situação. Mesmo na ausência de uma explicação científica

(1995), famoso por oferecer uma teoria da relevância aplicável tanto à comunicação quanto à cognição, é insuficiente para resolver o *frame problem* no âmbito da cognição, ou seja, para modelar a SR. Sperber e Wilson caracterizam a relevância em termos de eficiência. Contudo, ter SR envolve ser capaz de perceber quando é o caso de abrir mão da eficiência em virtude de, por exemplo, uma maior abrangência ou uma reconcepção radical da situação – característica central do juízo, tal como se argumentará na próxima seção. Para uma discussão aprofundada da relação entre a teoria da relevância e o *frame problem*, ver Chiappe; Kukla (1996), Sperber; Wilson (1996), Vervaeke; Ferraro (2013) e Barth (2024).

da SR no juízo, é possível contrastar o que já sabemos sobre a cognição humana com o que sabemos sobre o modo de operação das IAs. Com isso em mente, buscaremos analisar como a SR se acomoda no juízo. Isso permitirá um posterior contraste com as operações computacionais do cálculo.

## 2. Juízo

O que denominamos "juízo" tem sua inspiração em Smith (2019):

> Eu uso [a palavra] *juízo* para o ideal normativo com o qual, eu argumento, devemos avaliar a inteligência humana autêntica – uma forma de pensamento deliberativo e desapaixonado, fundamentado em compromisso ético e ação responsável, adequada à situação em que é empregada[8] (Smith, 2019, p. XV).

Como se nota, o que Smith tem em mente não é uma propriedade funcional do aparato cognitivo. Ele não descreve mecanismos mentais, como anseiam as ciências cognitivas, mas um norte contra o qual avaliamos o comportamento humano, na medida em que este é considerado fruto da inteligência. Trata-se de algo culturalmente sedimentado, desenvolvido ao longo de nossa história enquanto espécie e enquanto sociedade. Partilhamos dos seus objetivos na seguinte medida: Smith quer caracterizar o que precisa ser simulado pelo cálculo a fim de simular a inteligência humana. Tomaremos aqui o mesmo ponto de partida, mas nos limitaremos a apontar elementos que consideramos indispensáveis para simular a SR.

De saída, convém esclarecer como o termo "desapaixonado" deve ser interpretado. Smith é explícito ao dizer que não tem em mente uma deliberação dissociada da dimensão afetiva.[9] O ser humano se importa consigo e com vários elementos do mundo, e o juízo não pode descuidar disso. Assim como inferências irrelevantes, inferências que ignorem a dimensão afetiva frequentemente constituem juízos inadequados. Quando deixamos uma criança aos cuidados de alguém, por exemplo, nossas instruções pressupõem que esse alguém se importa com a criança e com as consequências de um trabalho inadequado. Não seria aceitável ouvir algo como "faz horas que não sei onde seu filho está, mas fique tranquilo, pois estou seguindo todas as instruções que

8 I use judgment for the normative ideal to which I argue we should hold full-blooded human intelligence – a form of dispassionate deliberative thought, grounded in ethical commitment and responsible action, appropriate to the situation in which it is deployed.

9 I neither mean nor intend to suggest that judgment should (or can) lack in care or commitment. On the contrary [...], judgment must be simultaneously passionate, dispassionate and compassionate. (Smith, 2019, Nr. 2, p. XV).

você deu". O ideal de pensamento desapaixonado não prescreve a negação do afeto, portanto, mas que seu papel no raciocínio seja circunstancialmente adequado.

Buscaremos agora mostrar que a SR, tal como se apresenta no juízo, envolve 1) a capacidade de inteligir o mundo de múltiplas formas; 2) a capacidade de alterar a forma como o mundo é compreendido em função de como os fenômenos se apresentam (i.e., do mundo como ele é); e 3) o fato de que a diferença entre o modo como o mundo se mostra e o modo como o inteligimos importa para nós.

### *2.1. Um mundo, múltiplos modelos*

É comum pensar a inteligência como uma capacidade de *resolver* problemas. Mas, como veremos, as discussões da IA sugerem algo diverso. As arquiteturas computacionais típicas da IA clássica, por exemplo, dependiam da formalização de tarefas. Isso é simples para problemas como "decidir qual a próxima jogada" no xadrez, mas muito difícil no caso de "decidir se um dado comportamento num aeroporto é suspeito". Esse é um dos fatores que levou a IA a adotar arquiteturas neurais (i.e., conexionistas), que capturam regularidades sem depender da explicitação de regras formais, exibindo assim maior tolerância à ambiguidade e à variação. Contudo, arquiteturas neurais têm dificuldade em lidar com casos que demandam raciocínios lógicos longos e mais articulados (Smith, 2019; Zhang *et al.*, 2022).

Isso sugere que a inteligência humana acomoda, de alguma forma, o melhor dos dois mundos: ela tem a flexibilidade característica das arquiteturas neurais, mas também exibe um elevado grau de articulação envolvendo relações lógicas, características das arquiteturas tradicionais (implicações, quantificações, raciocínios hipotéticos, etc.). Tudo indica que, por alguma via ainda não compreendida, somos capazes de articular diferentes estratégias, e a abordagem mais adequada depende de como um dado problema é concebido. Dito de outro modo, em vez de técnicas avançadas para solucionar problemas, a inteligência se caracteriza antes pela capacidade de *formular* problemas de modo a simplificar sua solução.

Considere o exemplo do *problema do tabuleiro mutilado*, encontrado em Kaplan e Simon (1990). Começamos com um tabuleiro de xadrez comum, com 8 linhas e 8 colunas, e a pergunta: é possível cobrir todas as 64 casas do tabuleiro com 32 peças de dominó, sem que uma se sobreponha à outra, e sem que nenhuma peça fique com parte alguma para fora do tabuleiro? Considerando que cada peça ocupa duas casas consecutivas, e que o tabuleiro

tem um número par de linhas e colunas, parece evidente que sim. Basta, por exemplo, posicionar quatro peças na horizontal, lado a lado, em cada linha.

Considere agora uma nova versão do tabuleiro em que duas casas de diagonais opostas foram removidas:

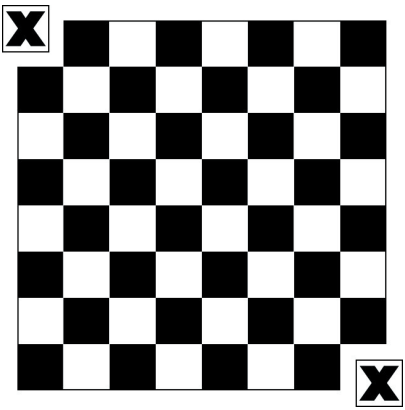

*Figura 1 – Tabuleiro mutilado*

Será possível cobrir os 62 espaços remanescentes com 31 peças de dominó? A resposta parece exigir o teste de todas as possíveis configurações das peças até que se esgotem ou que uma solução seja encontrada. Mas o número de configurações possíveis é gigantesco, e não é viável exauri-las. É preciso achar um outro caminho, e isso envolve reformular o problema. Tabuleiros físicos têm uma propriedade até aqui ignorada: a distribuição alternada de casas pretas e brancas. Essa propriedade, que Kaplan denomina *paridade*, era irrelevante para resolver o primeiro problema, mas no caso do tabuleiro mutilado ela torna saliente o fato de que as duas casas removidas têm a mesma cor. Como cada peça de dominó cobre, necessariamente, casas de cores distintas, fica claro que não há combinação possível.

Como explicar esse *insight* que leva da formulação do problema sem paridade para uma formulação que admite essa propriedade? Não é plausível deliberar sobre todas as propriedades do tabuleiro, pois isso também leva a uma explosão combinatorial. Afinal, se é verdade que o tabuleiro tem paridade, também é verdade que ele é quadrado, simétrico, possui quatro diagonais principais que o atravessam de um extremo ao outro, pode ser representado em 3D, pode ser matematicamente expresso como uma matriz 8x8, pode ser dividido em quadrantes menores e assim por diante. Nenhuma

dessas propriedades parece relevante para a solução do problema posto, mas como saber disso antes de elencá-las e verificá-las, uma a uma?

No caso do juízo, parte da resposta está no modo como se compreende a situação no interior da qual o problema é formulado. Diferentes compreensões envolvem a aplicação de diferentes *modelos do mundo*. Modelos podem ser tanto internos ao aparato cognitivo (e.g. padrões de ativação neuronal estruturalmente isomórficos ao caminho que se faz ao trabalho) quanto públicos (e.g. mapas impressos do caminho até o trabalho). Ambos fazem as vezes de alguma estrutura do mundo, permitindo raciocinar sobre ela mesmo na sua ausência, isto é, ambos nos permitem pensar sobre o caminho ao trabalho, mesmo quando não estamos caminhando até o trabalho. Há vários tipos de modelo e há várias formas de categorizá-los, mas alguns exemplos familiares são mapas, paletas de cores, notações musicais, diagramas elétricos, etc. Podemos também pensar em modelos que simulam determinados domínios causais, como os modelos de sistemas meteorológicos.

Diferentes modelos salientam e ocultam diferentes propriedades, restringindo assim as possíveis formulações de tarefas ou problemas.[10] Quando deixamos de lado um modelo geométrico do tabuleiro e adotamos um modelo que acomoda propriedades físicas como a cor, algumas possibilidades de investigação se abrem e outras se fecham. Evidentemente, o exemplo do tabuleiro é artificialmente simplificado para ilustrar esse efeito. Já situações concretas geralmente envolvem a aplicação de diversos modelos de modo concomitante e articulado. O simples ato de atravessar a rua pode envolver modelos estruturais e causais do fluxo de veículos, de comportamentos típicos de pedestres, cores, sinais, etc. Com efeito, o que caracteriza um *insight* no juízo é conceber uma situação (i.e., aplicar um modelo) de um modo que torne saliente as propriedades relevantes aos nossos objetivos. Podemos com isso perceber se, e quando, propriedades não geométricas, como a cor, podem ajudar a resolver problemas originalmente concebidos como puramente geométricos.

Mudar a forma como se compreende uma situação é profundamente distinto de buscar formas eficientes de realizar uma tarefa sem reformulá-la. Podemos elaborar algoritmos para automatizar a análise das possíveis configurações de peças de dominó, mas ainda que traga resultado positivo, esse caminho envolve tempo, memória e velocidade de processamento bastante distintos dos envolvidos no juízo. Além disso, o resultado seria

10 Cukier, Mayer-Schoenberger e Vericourt (2021) apresentam vários exemplos de como isso se mostra, mesmo em tarefas cotidianas.

profundamente dependente do modo como o problema é formulado. Pequenas modificações na disposição das casas ou nas peças de dominó (que podem ter formato em L, por exemplo) já trariam à tona a necessidade de formular um novo algoritmo. Com efeito, sistemas incapazes de reformular problemas ao modo do juízo têm aplicação mais restrita.[11]

Em síntese, em vez de uma grande capacidade de solucionar problemas, o que caracteriza o juízo é a capacidade de articular diferentes modelos de mundo. Isso permite reformular problemas complexos em versões simplificadas, mais afeitas ao que é circunstancialmente relevante e ao poder de raciocínio disponibilizado pelo nosso aparato cognitivo.

### *2.2. O mundo como ele é*

Passemos agora ao item (2): o papel do mundo como ele é.

Vimos como o problema do tabuleiro mutilado pode ser solucionado pelo *insight* que leva à reformulação da tarefa e à consequente aplicação de um modelo distinto. Parte do que torna essa reformulação possível é a capacidade de entender que mundo e modelo de mundo podem divergir. Ou seja, a capacidade de conceber que o modo como as coisas nos parecem não é necessariamente o modo como as coisas são. Somos capazes de pensar o que aconteceria se empurrássemos o copo para fora da mesa, mas somos também capazes de empurrá-lo e contrastar o resultado obtido com o resultado pensado.[12] Isso é parte do que nos leva a perceber se e quando um problema precisa ser reformulado, e isso supõe a capacidade de se relacionar diretamente com o mundo, isto é, de experienciá-lo. Ou seja, somos capazes de rever nossos próprios modelos mentais a partir de experiências do mundo. Disso emerge a distinção entre *de re* e *de dicto*: nosso pensamento pode ser orientado tanto às coisas elas mesmas quanto à forma como as modelamos.

No caso do tabuleiro, percebe-se que uma propriedade de tabuleiros físicos, até então ignorada, pode levar a uma revisão no modelo aplicado. Essa capacidade de adaptar modelos ao mundo está por trás também de vários *insights* ao longo da história. Galileu, por exemplo, demonstrou que formas geométricas podem ser usadas para calcular velocidades, e não apenas

11 Como veremos, essa abordagem subjaz tanto o atual sucesso da IA contemporânea quanto suas atuais limitações.

12 Esta descrição é neutra em relação ao debate entre representacionistas e antirrepresentacionistas no âmbito das ciências cognitivas. Não se trata de afirmar que o aparato cognitivo precisa ser capaz de operar na ausência de representações mentais. Contudo, ela não é neutra quanto à natureza incorporada e integrada da cognição humana.

espaços.[13] Na mesma linha, após Descartes, chegamos ao extremo de tomar a matemática como preocupada tão somente com relações abstratas entre objetos, sem assumir uma conexão intrínseca a qualquer domínio que seja (formas geométricas, movimentos, forças, etc.). Esses são exemplos de novas formas de inteligir diferentes domínios, jogando luz sobre diferentes aspectos do mundo, assim como se jogou luz sobre um aspecto diferente do tabuleiro.[14]

Um experimento mental inspirado em Cummins (2010) pode ajudar a deixar esse ponto mais claro. Imagine um pequeno cubo que pode assumir qualquer matiz de cor visível. Milhares desses pequenos cubos são então organizados como um gigantesco cubo de *Rubik*: em vez de 3 x 3 x 3, ele tem, digamos, 10000 x 10000 x 10000. Seria possível usar esse cubo como um sistema para representar objetos, bastando colorir os cubos na proporção e na ordem adequadas. Suponha-se agora que esse esquema seja usado como um modelo de como a informação visual é registrada no aparato cognitivo. Isso significa que esse aparato operará de forma constrangida por uma série de suposições:

(1) Todo objeto é colorido.
(2) Todo objeto tem um tamanho e uma forma.
(3) Dois objetos não podem ocupar o mesmo lugar ao mesmo tempo.
(4) Todo objeto tem um local determinado relativo aos demais objetos.
(5) Todo objeto está a uma distância determinada de outros objetos.

Em outras palavras, um sistema visual que se guie por esse esquema será incapaz de representar objetos transparentes, será incapaz de representar objetos sem tamanho e forma definidos, e assim por diante. Um aparato visual como esse permite a uma criatura (biológica ou artificial) operar no mundo. Ela reconhecerá objetos, suas cores, sua forma, sua posição, e poderá fazer inferências adequadas em diversos contextos ("vou nessa direção porque isso me afasta do objeto indesejado").

Não obstante, nosso conhecimento contemporâneo de física mostra que, se tomadas como descrições do mundo tal como ele é, todas as suposições (1)-(5) são falsas. Diante de fenômenos que escapem a, ou que contradigam essas suposições, uma criatura dotada de juízo tem diante de si a possibilidade de repensar o modo como o mundo deve ser modelado. Ela pode ser incapaz de

13 Trata-se do Teorema 1, encontrado em Galilei (1954).

14 Não se quer sugerir que esses *insights* emergiram num vácuo cultural, nem que foram fruto de pura genialidade individual dessas figuras históricas. Certamente ambos beberam das ferramentas conceituais que herdaram, bem como do modo de pensar vigente no seu tempo. Busca-se apenas enfatizar a necessidade (sem discutir a suficiência) de um contato direto com o mundo para que *insights* dessa natureza ocorram.

alterar seu aparato visual, então o mundo continuará parecendo a ela como se fosse restrito por (1)-(5), mas sua compreensão pode ser expandida pelo uso de modelos cognitivos que permitam acomodar os fenômenos de forma mais próxima do real. Movimentos como esse estão por trás do que nos levou a acomodar a física quântica como uma descrição mais acurada dos fenômenos em certos contextos físicos.

O mundo, e não só o modo como o inteligimos, pode se mostrar relevante para nossos propósitos. Se o juízo fosse incapaz de perceber que pode haver diferença entre aparência e realidade, rever os modelos que aplica ao mundo não seria uma opção. Os avanços aqui expostos seriam impossíveis, pois fenômenos que não pudessem ser acomodados sob os modelos já disponíveis constituiriam, para sempre, um espaço incompreensível e incognoscível.

### *2.3. O mundo que importa*

Chegamos assim ao item (3): as capacidades (1) e (2) são exercitadas sempre de modo permeado pela nossa dimensão afetiva. Isso se mostra tanto numa deferência epistêmica ao mundo quanto num compromisso existencial com nossa compreensão de mundo. Nós não nos importamos com o mundo apenas tal como ele aparece para nós. Desejamos que nossa compreensão se articule com o mundo tal como ele é. Ou seja, a realidade *importa* para nós.[15]

Considere a situação de perda de um ente querido. Diante do sofrimento, caso fosse oferecida uma opção terapêutica que nos fizesse crer piamente numa explicação alternativa para seu desaparecimento ("ele virou um monge"), ela seria insuficiente. Não queremos apenas que o mundo nos *pareça* ser de certo modo. Queremos que o mundo *seja* daquele modo, por mais que a diferença nos faça sofrer. Isso caracteriza uma profunda deferência epistêmica ao mundo. Com a exceção de casos patológicos, sabemos que o mundo sempre vence.[16]

15 Essa deferência ao mundo não deve ser confundida com uma deferência às pesquisas científicas. Essa última envolve uma confiança nas instituições que produzem conhecimento, e essa relação de confiança pode ser posta em xeque até mesmo em nome da deferência ao mundo. O terraplanista não busca rejeitar o mundo como é. Ele rejeita a intermediação da NASA e dos departamentos de Física. O problema com sua abordagem não é o descaso pela verdade, mas ignorar o conhecimento humano sedimentado. Para uma discussão desse ponto, ver Perini-Santos (2023).

16 Fenômenos como o da pós-verdade podem fazer com que essa afirmação seja recebida com algum ceticismo. Porém, a tensão é apenas aparente. O debate em torno desse fenômeno não diz respeito à nossa deferência epistêmica ao mundo como aqui tratada, mas sim ao que (ou quem) pode ser aceito como intermediário na interpretação do mundo. Não por acaso, é comum encontrar negacionistas afirmando que informações da imprensa são insuportavelmente enviesadas e que devemos, por isso, buscar a verdade nos "dados brutos". Para uma discussão desse ponto, ver Andrejevic (2017).

Por sua vez, quando uma experiência ou um fenômeno desafia nossa compreensão de mundo, isso também nos impacta afetivamente, e tendemos a defendê-la. Se alguém acredita que sistemas de IA não podem simular a capacidade humana de escrever poemas, uma demonstração em contrário soa como um desafio (Porter; Machery, 2024). O mesmo acontece diante de um fenômeno que consideramos impossível. Se o copo diante de nós começar a flutuar, nossa deferência ao mundo nos leva a buscar formas de acomodar a experiência: alucinação? Ilusão de ótica? Pegadinha? Na mesma linha, se algo considerado impossível nos é relatado, tendemos a duvidar da testemunha ou de sua capacidade avaliativa. Isso sugere um forte compromisso afetivo com nossa compreensão do mundo. Mudanças radicais são sabidamente dolorosas, e os fenômenos precisam resistir duramente a nossas tentativas de conciliação, antes de sequer cogitarmos rever nossos modelos.

Parte da razão para isso é que nossa compreensão de mundo adentra a esfera existencial: ela é indissociável da nossa autocompreensão. Entender a si mesmo como prudente, frágil ou medroso se associa a uma compreensão do mundo como um lugar perigoso. Da mesma forma, se nos sentimos confiantes, o mundo nos aparece como um desafio manejável. Se somos gentis, estruturamos o mundo em termos de oportunidades para o exercício de gentilezas. Essa articulação se mostra em qualquer categoria aplicada à autocompreensão: “pai dedicado”, “pessoa engraçada”, “especialista em Hegel”, etc. Essas formas de autocompreensão são atuantes, ainda que não consigamos sintetizá-las linguisticamente. Quando uma situação coloca nossa autoimagem em xeque, experienciamos tensão ou vergonha, mesmo que não saibamos explicitar o motivo.

Por fim, uma vez que o mundo humano é profundamente social, compreendê-lo é também compreender o outro. Entendemos quando alguém passa vergonha, ou quando está em uma situação emocionalmente demandante, porque podemos nos imaginar naquelas circunstâncias. Isso se aplica também à compreensão de personagens fictícios. Por isso, Haugeland (1998a) defende que a dimensão existencial é necessária, até mesmo para a compreensão de textos literários. Entendemos as motivações, os anseios e as dificuldades dos personagens, na medida em que partilhamos de sua compreensão de mundo.

Antes de prosseguirmos para a discussão do cálculo, cabe sintetizar rapidamente o que já foi analisado: o modo como o juízo se mostra sensível ao que é circunstancialmente relevante se caracteriza pela navegação habilidosa de um grande número de formas de compreensão do mundo (i.e., articulação de distintos modelos de mundo). Essa navegação sempre envolve tanto uma dimensão afetiva (o que importa para nós) quanto a capacidade de distinguir

aparência e realidade. Diante disso, discutiremos agora se e como a SR pode ser simulada pelo cálculo. Isso nos permitirá compreender se, e de que forma, cálculo e juízo divergem em relação a essa capacidade.

## 3. Cálculo

Como visto na introdução, o termo "cálculo" acomoda os diversos modos pelos quais a IA computa informações a fim de simular comportamento inteligente. Ele é usado em um sentido neutro em relação às diferentes ondas da IA (clássica e conexionista), e acomodará – seguindo o uso feito por Smith (2019) – as capacidades algorítmicas exibidas por quaisquer sistemas de IA, tais como computação regida por regras, procedimentos formais, estatística, otimização, busca, aproximação de funções, etc. O cálculo acomoda, portanto, formas de processar informação presentes tanto em arquiteturas clássicas quanto em arquiteturas neurais, independentemente dos fins específicos de cada sistema, tais como gerar conteúdos, classificar informações ou tomar decisões.

Pode o cálculo, munido somente dessas capacidades, simular alguma forma de SR? Vimos como, em larga medida, determinar o que é relevante se traduz na capacidade de inteligir (modelar) e navegar o mundo de modo semelhante ao do juízo. Algumas das pretensões mais ambiciosas da IA certamente assumem que este é um feito possível. Esse é o caso da tentativa de construir o que se convencionou denominar AGI (para *Artificial General Intelligence*, ou inteligência artificial geral). Esse objetivo é rastreável até os primórdios da IA, que almejava criar sistemas capazes de operar inteligentemente em qualquer domínio, resolvendo problemas formulados de modos tão diversos quanto aqueles que encontramos no juízo (McCarthy, 1968).

A SR é necessária para a AGI. Abrir mão dela implica considerar inteligente um sistema que se paralisa diante de qualquer tarefa, seja por deliberar indefinidamente sobre tudo (afinal, tudo é relevante para tudo), seja por nunca deliberar sobre nada (afinal, nada é relevante para nada). Contudo, a despeito de todos os avanços, os sistemas existentes só conseguem operar no interior de domínios bem delimitados. Tais sistemas "driblam" o requisito da SR, por serem projetados em contextos em que os fatores relevantes são dados de antemão: uma IA dedicada ao xadrez, por exemplo, pode ignorar, por princípio (sem deliberação), qualquer fator não explicitado nas regras do jogo. Em contraste, uma AGI precisaria ser capaz de delimitar o que é relevante, caso a caso, por si mesma.

A afirmação de que todo sistema de IA opera no interior de domínios específicos pode suscitar dúvida. Considere, por exemplo, os grandes modelos linguísticos (LLMs, para *Large Language Models*), como o utilizado pelo ChatGPT, que são a mais popular e avançada manifestação contemporânea do cálculo.[17] Eles parecem capazes de gerar conteúdo textual sobre um conjunto ilimitado de tópicos e domínios. Além disso, parecem parcialmente capazes de inferir o que importa para nós, como quando processam nossas instruções em linguagem natural. Há quem afirme que, dada sua capacidade de gerar conteúdo sobre qualquer tópico, LLMs já seriam uma forma de AGI ou, no mínimo, um sinal de que o objetivo está prestes a ser alcançado.[18] Deveríamos admitir então que um LLM não opera no interior de um domínio bem delimitado e, consequentemente, que exibe alguma forma de SR?

No que se segue, argumentaremos que nenhuma característica dos LLMs sinaliza avanço significativo rumo a uma AGI, tampouco implica avanço rumo a uma simulação adequada da SR do juízo. Para isso, analisaremos as implicações de três diferenças já familiares. LLMs são incapazes de: (1) aplicar múltiplos modelos de mundo a uma situação; (2) distinguir entre aparência e realidade; (3) operar segundo uma dimensão afetiva. Como veremos, essas características ensejam desafios distintos. Enquanto alguns podem ser superados, outros parecem demandar uma ruptura radical com o paradigma que subjaz as pesquisas contemporâneas.

### *3.1. Uma pergunta, múltiplos contextos*

Em suas críticas dos anos 1960, Dreyfus (1972) apostou que uma IA jamais jogaria xadrez de alto nível. Esse ceticismo se mostrou insustentável quando, em 1997, o sistema Deep Blue venceu Kasparov. Apesar de reconhecer que não antecipara aquela possibilidade, Dreyfus insistia que isso não afetava as razões centrais do seu pessimismo (Dreyfus; Dennett, 2005).

As críticas de Dreyfus se ancoravam na sua concepção fenomenológica da *expertise* humana (Dreyfus; Dreyfus, 1986). Em sua visão, jogar xadrez demanda a capacidade de conceber a situação que se coloca no tabuleiro de múltiplas formas (rainha em situação frágil, domínio das casas centrais ameaçado, etc.) e reformular continuamente a tarefa a cumprir (e.g. de "decidir

17 LLMs são sistemas que usam arquiteturas conexionistas geradas a partir de algoritmos de aprendizagem de *deep learning*, e se dedicam primariamente à classificação e à geração de conteúdos textuais.

18 Ver, por exemplo, Norvig e Arcas (2023). Os autores afirmam que os sistemas atuais já constituem uma forma de AGI, mas sua justificativa envolve substancial revisão dos critérios tipicamente adotados pela comunidade da IA. Por isso a ampla maioria dos pesquisadores continuam tratando a AGI como um objetivo a ser alcançado.

qual a melhor jogada" para "decidir qual a melhor forma de proteger a rainha"), de maneira conforme, fluida e adaptativa. Ou seja, jogar xadrez demanda uma AGI, e não há AGI sem SR. Em contraste, computadores se limitam a buscar soluções apenas para os problemas tal como previamente formulados pelos desenvolvedores. Como o xadrez leva a uma explosão combinatorial de formulações possíveis, segue-se que caberia aos desenvolvedores um trabalho potencialmente interminável de programar soluções para um número ilimitado de situações. Mas se é o caso, como é possível que um computador jogue xadrez de forma excelente?

Sistemas de xadrez usam heurísticas, i.e., regras práticas que orientam as decisões sem exaurir todas as possibilidades. Embora acertem na maioria das vezes, podem falhar em casos incomuns. Por exemplo, um sistema pode supor que as primeiras jogadas devem priorizar o controle sobre o centro do tabuleiro, mas, ao fazê-lo, pode deixar passar uma chance clara de capturar uma peça adversária prematuramente exposta. Heurísticas funcionam no xadrez porque, a despeito de sua complexidade, o xadrez constitui um domínio autônomo. Tais domínios caracterizam atividades que podem ser realizadas de forma isolada do resto do mundo. As decisões tomadas no interior desses domínios levam em conta apenas conhecimento da sua dinâmica interna (i.e., as regras do jogo). Com efeito, fatores internos ao domínio são sempre relevantes, e fatores externos ao domínio são sempre irrelevantes, tornando a SR desnecessária.[19]

Dreyfus não considerou as possibilidades advindas da expansão do poder computacional, levando-o a subestimar o alcance da abordagem heurística. Ela é, de fato, insuficiente para lidar com domínios abertos, que não são navegáveis de forma autônoma. Situações concretas como a de escolher o que beber num restaurante não constituem domínios autônomos porque, como vimos, essas decisões envolvem um número potencialmente *ilimitado* de elementos. Não há delimitação clara do que está dentro ou fora do domínio, portanto. Mas a abordagem heurística se mostrou capaz de lidar adequadamente com tarefas específicas em domínios autônomos, ainda que grandes e complexos. Quando uma IA é treinada para navegar um dado domínio, aprende a identificar um grande número de heurísticas e de condições para sua aplicação. O uso dessas heurísticas em escala e velocidade inumanas acaba compensando, em alguma medida, a incapacidade do sistema de reformular os problemas nos moldes do juízo.

19 Um modo mais preciso de afirmar o mesmo ponto é que o recorte desses domínios repousa sobre a SR dos desenvolvedores, que decidem pelo sistema tudo o que será levado em conta.

A história do erro de Dreyfus fornece uma chave para entender o atual sucesso dos LLMs, ao mesmo tempo em que joga luz sobre seus limites: nosso comportamento verbal corrente (i.e., os padrões de uso da linguagem natural) é regular e estável o suficiente para permitir que seja tratado pelo cálculo como um domínio autônomo.[20]

Dreyfus acreditava que jogar bom xadrez só era possível em conjunto com a capacidade de navegar o mundo humano, ou seja, essa não seria o tipo de habilidade que poderia ser alcançada de forma isolada das demais capacidades humanas. Ele estava errado. A linguagem costuma ser pensada de forma semelhante: produzir conteúdo linguístico de forma efetiva só é possível em conjunto com a capacidade de navegar o mundo humano. LLMs sugerem fortemente que esse não é o caso.

A principal evidência empírica está no fato de que LLMs não se limitam a replicar regularidades sedimentadas. Elas conseguem extrapolar (generalizar estatisticamente), a partir dessas regularidades, de modo efetivo, produzindo respostas linguísticas adequadas a *inputs* com os quais não teve contato prévio (i.e., nem o *output* nem o *input* participavam do *corpus* de dados utilizados no seu treinamento).[21] Dito de outro modo, LLMs levaram ao achado empírico de que é possível simular a produtividade[22] linguística humana sem simular a compreensão linguística.[23] Eles conseguem navegar uma estrutura que codifica parte da forma humana de navegar o mundo, ainda que indiretamente, isto é, por meio das expressões linguísticas produzidas por nós no decorrer dessa navegação.

Isso leva a uma nova versão do que Haugeland (1985) denominou *mote do formalista*: cuide da sintaxe e a semântica cuidará de si mesma. A "sintaxe" em questão não é mais a estrutura da linguagem,[24] mas sim a estrutura probabilística que descreve o modo como costumamos fazer uso das palavras. Ao se deixarem nortear por essa estrutura, LLMs são capazes de processar e produzir conteúdo que, quando interpretado por nós, faz sentido, ainda que LLMs não compreendam texto algum. Parece, portanto, que a IA

20 Não se trata de afirmar que o domínio do xadrez e o domínio que caracteriza as regularidades do uso da linguagem são estabilizados da mesma forma. O domínio no interior do qual jogos de xadrez ocorrem se estabiliza a partir do compromisso contínuo de seus participantes com um certo conjunto de regras. Por sua vez, o domínio do uso da linguagem emerge de regularidades nas nossas práticas linguísticas.

21 Disso não se segue que a capacidade de extrapolação linguística presente nas LLMs seja equivalente à humana, mas seus limites ainda não são claros.

22 Por "produtividade", referimo-nos à capacidade de compreender frases inéditas, desde que bem estruturadas.

23 Essa é uma diferença fundamental entre *chatbots* contemporâneos e sistemas antigos como o ELIZA (Weizenbaum, 1966).

24 Originalmente, a frase de Haugeland faz referência à estratégia de processamento linguístico adotada pela IA clássica dos anos 1960-1980.

contemporânea pode navegar algumas dimensões da linguagem natural, ainda que usando um barco distinto do juízo.

LLMs respondem a uma única pergunta, formulada sempre do mesmo modo: dado o que foi dito até aqui, o que dizer em seguida?[25] Para responder a essa pergunta, LLMs fazem uso massivo de heurísticas. Elas mapeiam probabilisticamente os discursos que produzimos anteriormente, identificando padrões estatísticos em nossas decisões, comportamentos e pensamentos expressos linguisticamente. Em outras palavras, captam regularidades na forma como entendemos o mundo no passado, extrapolando a partir desses padrões. A aparente SR do cálculo é, portanto, fruto da estruturação probabilística do que o juízo considerou relevante no passado.

Nesse cenário, duas perguntas importantes emergem. Primeiro, em que medida o discurso produzido pelo juízo no passado é um bom indicador daquilo que ele considerou relevante? Segundo, o que acontece se as situações encontradas pelo cálculo divergirem o suficiente das situações do passado?

A primeira pergunta é pertinente porque a linguagem tem um caráter complementar à situação concreta.[26] Verbalizamos apenas o que não está dado no contexto. Por isso é possível dizer a) "eu já almocei" para responder a b) "você está com fome?". O modo como (a) serve de resposta para (b) não pode ser inferido apenas do léxico. O juízo é capaz de compreender esse uso porque envolve senso comum e familiaridade com o mundo humano.[27] A caracterização dos contextos em que as atividades humanas se desenvolvem acomoda os três elementos discutidos na seção 2 (juízo): o modelo aplicado (i.e., como a situação foi inteligida), a distinção entre o real e o aparente, e a dimensão afetiva. Por sua vez, LLMs precisam decidir o que deve ser dito, mesmo na ausência desses fatores. Sendo assim, sempre há risco de que o LLM se guie por elementos contextuais irrelevantes, ignore elementos relevantes e, mesmo quando houver coincidência dos fatores considerados, atribua pesos inadequados a cada um. Isso pode ocorrer mesmo em casos marcadamente presentes no *corpus* de dados utilizado para treinar o LLM. Dito de outro modo, o LLM pode alucinar[28] elementos contextuais inadequados, levando a um caminho distinto do trilhado pelo juízo.

25 De modo mais preciso: considerando uma dada sequência de *tokens*, qual *token* é melhor candidato a figurar como próximo item? No caso de LLMs, o termo "token" pode coincidir com uma palavra em linguagem natural, mas pode também referenciar uma sílaba ou um morfema.

26 Essa afirmação remete à posição contextualista no debate entre contextualistas e minimalistas em filosofia da linguagem. Ver Perini-Santos (2014).

27 Como visto no exemplo da introdução.

28 "Alucinação" é um termo técnico que designa casos em que LLMs geram conteúdo não factual. Ver Huang *et al.* (2023).

A segunda pergunta leva a um problema ainda maior. Como discutido, toda situação é, em alguma medida, inédita. Situações novas podem exigir atenção a elementos que não estão contemplados na distribuição probabilística de usos pretéritos de construções linguísticas utilizada para treinar o LLM. Nesses casos, o LLM não tem qualquer outro norte que possa guiar suas inferências. Ele precisa de um fundamento independente da distribuição probabilística, mas qual? Na IA, as buscas por fundamentos inferenciais independentes é conhecida como *out-of-distribution extrapolation* (Liu *et al.*, 2021).

Para uma ilustração mais intuitiva do desafio, considere um exemplo de processamento visual. Se uma IA for treinada para reconhecer vacas apenas em fotos com grama, ela pode falhar diante de vacas num cenário sem grama, assumindo erroneamente que a grama é essencial para identificá-las. Isso ocorre quando sua única fonte de pistas para delimitar o que é parte de uma vaca é a distribuição probabilística das propriedades das fotos usadas no treinamento. Se onde houver vacas houver grama, aquilo que o juízo recorta como duas categorias pode ser tomado pelo cálculo como uma única categoria. Esse problema não se limita ao reconhecimento visual, afetando tarefas como a de categorizar petições jurídicas, identificar humanos diante de um carro autônomo em situação de risco ou reconhecer condições inesperadas em cirurgias. Assim, se uma IA enfrentar situações não previstas nos dados de treinamento, ela pode adotar caminhos inadequados.

Em síntese, a aparente SR de um LLM (que contemporaneamente representa o estágio mais desenvolvido do cálculo) decorre das distribuições probabilísticas dos discursos anteriores. Em vez de adaptar contextualmente modelos de mundo e formular problemas em termos circunstancialmente adequados, ele aplica heurísticas para resolver sempre o mesmo problema: prever o conteúdo linguístico mais provável, com base no discurso passado. Embora aplicável a qualquer conteúdo linguístico, essa abordagem não simula o juízo, pois não envolve a compreensão do discurso. Além disso, não é claro em que medida seria possível reproduzir a performance de um LLM por meio da aplicação de modelos de mundo, uma vez que isso envolve abrir mão do achado empírico que explica o desempenho atual (i.e. que a estrutura dos comportamentos verbais oriundos do juízo pode ser tratada como um domínio autônomo).[29]

29 Há quem defenda a hipótese de que LLMs não usam apenas heurísticas, mas geram modelos de mundo de forma espontânea (Li *et al.*, 2022; Spies *et al.*, 2024). Contudo, as evidências disponíveis são mais favoráveis à interpretação heurística (Nikankin *et al.*, 2024; Vafa *et al.*, 2024).

*3.2. Que mundo?*

Passemos agora aos itens (2) e (3). Para emular (2) a distinção entre aparência e realidade, é preciso que o cálculo seja incorporado (*embodied*) e integrado ao ambiente (*embedded*). Isso vai além de dar um corpo robótico a sistemas existentes. Corpo e ambiente devem ter papel constitutivo no modo como o cálculo processa informação sobre o mundo.

Para atender esses requisitos, o cálculo precisa ser capaz de interagir *ativamente* com o seu ambiente e aprender de modo *contínuo*, a partir dos efeitos dessa interação. Ao explorar o ambiente, ele o instiga a produzir novas informações. Assim, as reações do ambiente viram normas contra as quais a adequação dos modelos pode ser mensurada. Resultados inesperados ou ineficazes podem levá-lo a reavaliar seus modelos de mundo, resultando em adaptação contínua.

O papel fundamental do corpo e do ambiente na estruturação da inteligência não é novidade (Brooks, 1991; Dreyfus, 1972; Haugeland, 1998b). Apesar disso, os principais programas de pesquisa contemporâneos o ignoram, priorizando aprendizagem não contínua e aplicações desincorporadas (*chatbots*, ordenação de conteúdos, processamento de textos, etc.). Com efeito, embora existam pesquisas que busquem desenvolver IAs incorporadas, elas estão distantes dos interesses centrais contemporâneos, e não é claro se, ou quando, haverá uma guinada nessa direção.

Por sua vez, (3) a ausência da dimensão afetiva parece ainda mais difícil de tratar. Haugeland entende que essa falta justifica um forte ceticismo quanto à possibilidade de o cálculo simular fielmente o juízo. Em sua famosa síntese: *"o problema da inteligência artificial é que computadores não estão nem aí"*[30] (Haugeland, 1998a, p. 47). De fato, é difícil imaginar como seria possível agregar uma dimensão afetiva a um conjunto de transistores.

Uma das poucas propostas disponíveis é a de Minsky (2006). Ele argumenta que emoções podem ser analisadas em termos de ativação de recursos cognitivos funcionalmente especificáveis. Raiva, por exemplo, seria equivalente a ativar recursos que resultam em hostilidade e agressividade e suprimir recursos que resultam em simpatia e prudência. Porém, essa ideia vai na contramão das pesquisas contemporâneas sobre o papel cognitivo e a natureza das emoções humanas, que têm caráter fortemente incorporado e integrado (Carvalho, 2019; Mesquita; Boiger, 2014; Pessoa, 2022). Nessa perspectiva, a raiva não pode ser analisada de forma isolada do contexto,

30 The trouble with artificial intelligence is that computers don't give a damn.

como quer Minsky: estar com raiva do filho é uma emoção diferente de estar com raiva do chefe.

É possível que um programa de pesquisa em IA incorporada abra caminho para simular o papel da dimensão afetiva. No atual estado da arte, porém, isso é pouco mais que uma especulação. Muito depende de futuros estudos que aprofundem o conhecimento sobre a relação entre a dimensão afetiva e a cognição, além de uma considerável reviravolta nas prioridades das pesquisas em IA. Ao menos por ora, a maior parte da atenção está voltada a abordagens desincorporadas.

## 4. Efeitos do Cálculo sobre o Juízo

Temos um cenário em que há profundas diferenças no modo como juízo e cálculo delimitam a que atentarão no decorrer de suas operações. O juízo faz uso de diversos modelos de mundo, distingue entre o real e o aparente ao avaliar o resultado das aplicações desses modelos, e opera sempre de modo permeado por uma dimensão afetiva. Por sua vez, o cálculo conta apenas com heurísticas que realizam extrapolações estatísticas sobre *outputs* previamente realizados pelo juízo (i.e., sobre conteúdo linguístico produzido no passado), não consegue distinguir entre real e aparente e não possui dimensão afetiva. Ele sequer tenta simular o modo como a SR se dá no juízo, portanto. Essa diferença leva a diversas preocupações, mas no que se segue, vamos nos concentrar na seguinte questão: diante de distinções tão profundas, como podem as capacidades do cálculo gozar de tamanha confiança em um número crescente de atividades decisórias cada vez mais sensíveis?

### *4.1. A reflexividade nas atividades humanas*

Partimos da diferença entre tipos naturais, como polietileno, plutônio ou água, e tipos humanos, como adolescente, imigrante ou autista. Essa distinção costuma ensejar disputas entre perspectivas realistas e construtivistas, nas quais os tipos naturais são tipicamente tomados como recortes do real e os tipos humanos como construções sociais. Contudo, o argumento aqui avançado parte da análise de Hacking (1996), que é tangencial a essa discussão. Hacking argumenta que tipos humanos se distinguem dos tipos naturais por ensejarem um ciclo de *reflexividade* nas atividades humanas.

Tipos humanos são categorias usadas para especificar comportamentos, ações e temperamentos. Em função disso, eles podem levar à projeção de tipos

de pessoas, como quando dever dinheiro repetidamente enseja o enquadramento de alguém como mau pagador. Com efeito, tipos humanos têm conotação moral. Ser tipificado, ou entender a si mesmo como um exemplar de um tipo determinado, pode ser desejável ou indesejável. Isso nos motiva a conhecer os princípios e as condições dos comportamentos associados, permitindo que comportamentos indesejáveis sejam prevenidos e comportamentos desejáveis sejam incentivados.

Hacking exemplifica como essas propriedades geram um ciclo reflexivo. As ciências humanas buscam estudar tipos humanos, em particular aqueles que possuem forte conotação moral, como os que se caracterizam por desvios do "normal". Estudá-los gera conhecimento e muda nossa compreensão deles. Isso afeta o modo como compreendemos a nós mesmos (quando nos entendemos como exemplares de um dado tipo, ou quando passamos a ser tratados de forma diferente por quem nos tipifica daquele modo) e os demais (quando tipificamos outras pessoas e alteramos o modo como as tratamos). Essas mudanças podem ser profundas como quando, parafraseando Hacking, experienciamos novos passados (Hacking, 1996, p. 368), isto é, reinterpretamos eventos pretéritos à luz de novos conhecimentos e produzimos um novo juízo. Porém, o ciclo continua: como indivíduos tipificados alteram seu comportamento, segue-se que agora há novos comportamentos, valores e expectativas associados àquele tipo, ou seja, há novos elementos a estudar, há mais a aprender. Uma vez estudados, geram-se novos conhecimentos, e isso afeta novamente os indivíduos tipificados, num ciclo de reflexividade aparentemente sem termo.

De modo importante para nosso argumento, a reflexividade se mostra mesmo nos casos em que não estamos cientes da tipificação. Ou seja, os critérios de tipificação e os próprios tipos utilizados podem nos ser inacessíveis, desconhecidos e mesmo incompreensíveis. Nesses casos, o efeito sobre o agente classificado é indireto, por meio de alterações no seu ambiente. Ele não sabe, mas seu horizonte de ações é afetado pela tipificação. Esse é o caso de crianças tipificadas como autistas, por exemplo. A reflexividade opera pelo modo como ela é tratada por familiares, profissionais de saúde e instituições em seu entorno. Ela então adapta seu comportamento em função do modo como as possibilidades, os constrangimentos e as demandas se apresentam.

Se Hacking tem razão, a reflexividade influencia o juízo. Seu exercício depende de como nossas compreensões de mundo e de nós mesmos se articulam e se influenciam mutuamente, e a reflexividade afeta os tipos usados nessa dinâmica. Nesse cenário, se as classificações trazidas à tona pelo cálculo ensejarem efeitos de reflexividade, elas também podem influenciar o juízo.

Mas em que sentido podemos relacioná-las aos tipos humanos de que trata Hacking?

Os usos do cálculo que envolvem geração de perfis nos permitem ver claramente como as classificações aplicadas podem partilhar características dos tipos humanos. Essa prática ocorre em quase todas as plataformas informacionais (redes sociais, serviços de *streaming*, e-mail…), bem como em análises para contratação, concessão de crédito, cobertura de saúde, etc. Primeiro, temos interesse em saber que classificações são essas e o que capturam. Esse interesse é ilustrado na demanda por explicabilidade em sistemas de IA, isto é, a vontade de jogar luz sobre as heurísticas aplicadas e suas respectivas classificações, gerando conhecimento sobre elas (Linardatos; Papastefanopoulos; Kotsiantis, 2020). Isso importa não só na medida em que os resultados são úteis, mas também na medida em que têm conotação moral, pois o uso de uma dada classificação para tomar decisões a nosso respeito pode ser desejável ou indesejável. Em segundo lugar, essas classificações podem ser usadas para projetar tipos de pessoas: uma reiterada recusa do cálculo em conceder crédito a alguém pode levar à sua categorização como mau pagador. Dito de outro modo, as classificações do cálculo podem participar da forma como prevenimos, incentivamos e modulamos comportamentos humanos. Elas partilham, portanto, de características que Hacking considera essenciais aos tipos humanos.

Resta mostrar como as classificações do cálculo podem gerar efeitos de reflexividade. Há pelo menos dois caminhos: primeiro, podemos adotar conscientemente uma classificação "descoberta" pelo cálculo. Ele pode notar, por exemplo, que consumidores de um produto X costumam consumir Y. Ao percebermos esse padrão, ele pode rapidamente se tornar um tipo humano: o "consumidor-XY". Essa adoção ocorre sem necessidade de explicar a relação entre os produtos, ou seja, não sabemos a razão que leva consumidores de X a buscar Y. Pode ser um efeito local, restrito a certo período, ou mesmo uma correlação espúria. Uma vez adotado como tipo humano, porém, ele se generaliza e modifica o que o juízo considera relevante. O fato de alguém consumir X torna-se relevante para predizer seu interesse por Y.

Em segundo lugar, o cálculo pode modular nosso ambiente informacional a partir de classificações inacessíveis, isto é, de tipificações que não necessariamente compreendemos. Ele estrutura o espaço em que nossos comportamentos se desenrolam, rearranjando nosso horizonte informacional e epistêmico, i.e., delimitando *por nós* o que é relevante. Ele não vai fornecer explicações dos motivos pelos quais considera que temos interesse em dado conteúdo, por exemplo. O conteúdo simplesmente aparece em destaque – ao

custo da supressão de outros conteúdos –, suscitando uma resposta irrefletida de nossa parte. É possível, por exemplo, que o cálculo alucine correlações (e.g. "consumidores de balas de menta tendem a gostar de música clássica") e induza comportamentos a partir disso, incitando o consumo de um produto a quem se mostra interessado no outro e fazendo com que a regularidade se torne real. Esse efeito transcende o espaço das plataformas informacionais, porque afeta nossa autocompreensão: passamos a nos tomar como pessoas que gostam de balas de menta e/ou de musica clássica, por exemplo, e isso tem efeito sobre o que consideramos relevante ao formularmos objetivos e deliberarmos sobre formas de alcançá-los. Com isso, o cálculo delimita o contexto das nossas ações, sugerindo e influenciando determinados cursos.

Os dois caminhos mostram como nossa SR é permeável às classificações do cálculo. Na medida em que estrutura nosso espaço de ação, ele modula nossos comportamentos, que são então modificados, gerando novos rastros e novos dados sobre nós. Esses dados levam o cálculo a revisar nosso perfil. A importância de certas classificações pode ser mitigada, enquanto a de outras é reforçada, levando a uma nova modulação do nosso comportamento e à continuidade do ciclo de reflexividade.

O cenário descrito é potencializado pela quantidade gigantesca de informação gerada diariamente (notícias, artigos científicos, vídeos, etc.). A estruturação do cálculo tornou-se um recurso indispensável para determinar o que merece atenção. Com efeito, escapar do ciclo de reflexividade parece cada vez mais difícil. As consequências para o exercício do juízo são profundas. Ela delimita o ferramental que o juízo tem para compreender suas próprias experiências. Além disso, dilui a distinção que o juízo faz entre aparência e realidade, assim como dilui a distinção entre autocompreensão e perfil.

Uma ilustração importante desse fenômeno pode ser encontrada em Schuster e Lazar (2024). Eles destacam que o uso judicioso da atenção constitui uma habilidade moral, e a dependência do cálculo para navegar o mar de demandas por atenção nos priva de oportunidades para o exercício e o aprimoramento dessa habilidade. Com efeito, na medida em que nossa atenção é crescentemente modulada pelo cálculo, também o são nossos juízos morais. Dadas as profundas distinções entre o modo como o cálculo e o juízo fazem inferências, o campo no qual se desdobram nossas deliberações morais é inevitavelmente afetado.

Obtemos, enfim, os fundamentos para uma resposta à pergunta que nos guia. O que fomenta a confiança desmedida nas capacidades do cálculo, a despeito de suas limitações e do seu modo de operar profundamente distinto do juízo? Parte substancial da resposta é que, como nosso juízo opera em um

espaço cada vez mais estruturado e delimitado segundo as classificações do cálculo, os efeitos de reflexividade mudam a forma como pensamos. Temos, assim, a ilusão de que o cálculo é capaz de simular a SR do juízo de modo cada vez mais acurado. Esse aparente sucesso, no entanto, é antes fruto de uma capacidade cada vez maior de incitar o juízo a operar de modo conforme ao cálculo, fazendo com que o juízo tenda a tomar como pouco ou nada relevante tudo o que está fora do alcance do cálculo. Ao não enxergarmos a SR como um desafio a ser superado pelo cálculo, sua capacidade de incitação é reforçada e facilitada, pois não percebemos como nossas ações se dão, de modo crescente, num espaço delimitado e estruturado pelo que o cálculo oferece. Isso leva a uma crescente aceitação dos seus efeitos e a uma profunda dificuldade de enxergar suas limitações.

## Conclusão

O cálculo consegue modular a forma como a SR opera. Até certo ponto, isso pode ser compreendido como mais um episódio em que o ser humano se reestrutura a partir de suas próprias criações. Estamos familiarizados com a necessidade de reestruturar nossos contextos de atividade de modo a acomodar novas tecnologias, como quando usamos um GPS para transitar por caminhos que antes precisávamos memorizar. Estamos igualmente familiarizados com casos em que o próprio juízo humano tenta modular essas estruturas de modo a promover interesses particulares (como quando uma instituição ou indivíduo usa o poder econômico para promover ou inibir comportamentos). Contudo, se o diagnóstico aqui apresentado estiver correto, há uma diferença importante: o cálculo já consegue executar essa reestruturação por nós, e geralmente o faz trabalhando com parâmetros que permanecem obscuros. Essa opacidade torna a tarefa de criticar o processo de reestruturação ainda mais difícil do que costuma ser nos casos mais familiares.

Chamam a atenção também a escala e a profundidade em que esse fenômeno pode ocorrer. Ele não se limita a mudanças pontuais circunscritas a um tipo de atividade. Mudanças em nossa SR tendem a afetar nosso comportamento em um número aberto de contextos. Com efeito, podemos estar diante de uma reestruturação profunda e involuntária do juízo.

Como lidar com esse fenômeno? Há amplo espectro de possibilidades entre um conservadorismo tecnofóbico que rejeita completamente o uso do cálculo e a dócil aceitação de uma realidade supostamente inevitável. O primeiro polo se traduz numa jornada árdua, solitária e insustentável. O

cálculo opera hoje em quase todas as dimensões da vida humana, e rejeitá-lo de modo radical se aproxima da rejeição do modo de vida contemporâneo.

Por sua vez, tendências que se aproximam do segundo polo podem ser adotadas em larga escala com mais facilidade. Pode-se, por exemplo, aceitar a influência do cálculo, desde que ele simule a adoção de princípios éticos que nos são caros (Kerns; Roth, 2019). Nessa perspectiva, a colonização só é um problema se cálculo e juízo operarem de forma disjunta. Porém, se a análise aqui apresentada vai na direção certa, o problema com essa sugestão deve estar claro: ela repousa sobre a esperança de que o cálculo pode capturar o mundo tal como estruturado pelo juízo, subestimando os desafios envolvidos na superação do profundo vão entre os dois. O fenômeno da colonização é fruto desse vão. Superá-lo envolve encontrar formas de distinguir uma crescente capacidade de simulação de uma crescente capacidade de incitação.

Por fim, mesmo que esse desafio seja superado em algum momento, ainda deveríamos nos perguntar se os comportamentos guiados pelo cálculo teriam o mesmo valor dos comportamentos produzidos pelo juízo. Visitar um parente no hospital por orientação de LLMs é qualitativamente distinto de visitá-lo por se guiar pelo juízo, que leva em conta afeto, amizade, parceria, etc. O juízo está sempre permeado de coisas que importam para nós, e talvez não queiramos abrir mão disso. O melhor caminho, portanto, é compreender o fenômeno da colonização do modo mais minucioso que pudermos, para que possamos modular seus efeitos em conformidade com o que realmente importa para nós.

**Disponibilidade de dados:**
Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

**Ausência de conflito de interesses:**
O autor declara que não há conflito de interesses.

**Editores responsáveis:**
Mauro Luiz Engelmann

## Referências

ADAM, D. “Lethal AI weapons are here: how can we control them?” *Nature*, Vol. 629, Nr. 8012, pp. 521–523, 2024.
ANDREJEVIC, M. B. “Framelessness, or the cultural logic of big data”. In: DAUBS, M., MANZEROLLE, V. (eds.). *Mobile and Ubiquitous Media*. Digital Formations. Germany: Peter Lang Publishing, 2017. pp. 251–267.
BARTH, C. “Representational cognitive pluralism: towards a cognitive-science of relevance-sensitivity”. PhD thesis — Belo Horizonte: Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, 2024.
BROOKS, R. A. “Intelligence without representation”. *Artificial Intelligence*, Nr. 47, pp. 139–159, 1991.
CARVALHO, F. N. De. “O Papel do contexto na percepção das emoções”. *Perspectiva Filosófica*, Vol. 46, Nr. 2, pp. 116–142, 2019.
CHIAPPE, D. L., KUKLA, A. “Context selection and the frame problem”. *Behavioral and brain sciences*, Vol. 19, Nr. 3, pp. 529–530, 1996.
CUKIER, K., MAYER-SCHOENBERGER, V., VERICOURT, F. De. “Framers: Human Advantage in an Age of Technology and Turmoil”. WH Allen, 2021.
CUMMINS, R. “Representations, targets and atitudes”. MIT Press, 1996.
CUMMINS, R. “Representational specialization: the synthetic a priori revisited”. In: CUMMINS, R. (ed.). *The world in the head*. Oxford, 2010. pp. 194–209.
DENNETT, D. “Artificial Intelligence as Philosophy and as Psychology”. In: *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, 1981.
DREYFUS, H. “What computers can’t do: a critique of artificial reason”. New York: Harper & Row, 1972.
DREYFUS, H., DENNETT, D. “Did Deep Blue’s win over Kasparov prove that Artificial Intelligence has succeeded? A debate”. In: FRANCHI, S., GÜZELDERE, G. (eds.). *Mechanical bodies, computational minds: artificial intelligence from automata to cyborgs*. [s.l: s.n.]. pp. 266–279.
DREYFUS, H.; DREYFUS, S. “Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer”. Free Pass, 1986.
FODOR, J. A. “Modules, frames, fridgeons, sleeping dogs and the music of the spheres”. In: PYLYSHYN, Z. W. (ed.). *The robot’s dilemma: The frame problem in artificial intelligence*. Ablex, 1987. pp. 139–149.
FORD, K. M., PYLYSHYN, Z. W. (eds.). “The Robot’s Dilemma Revisited: The Frame Problem in Artificial Intelligence”. Norwood, NJ, USA: Ablex Publishing Corp., 1996.
GALILEI, G. “Dialogues concerning two new sciences”. Tradução de Henry Crewand; Alfonso De Salvio. Dover, 1954.
HACKING, I. “The looping effects of human kinds”. In: *Causal Cognition*. Oxford University Press, 1996. p. 351–383.
HAQUE, M. D. R., RUBYA, S. “An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews”. *JMIR mHealth and uHealth*, Vol. 11, maio 2023.

HAUGELAND, J. “Artificial Intelligence: The Very Idea”. Bradford, 1985.
_____. “Understanding Natural Language”. In: HAUGELAND, J. (ed.). *Having thought*. Cambridge: Harvard University Press, 1998a. pp. 47–60.
_____. “Mind embodied and embedded”. In: HAUGELAND, J. (ed.). *Having thought*. Cambridge: Harvard University Press, 1998b. pp. 207-237.
HAWKINGS, A. J. “Uber self-driving car saw pedestrian but didn’t brake before fatal crash, feds say”. *The Verge*, 2018. Disponível em https://www.theverge.com/2018/5/24/17388696/uber-self-driving-crash-ntsb-report. Acessado em 11 de março de 2025
HUANG, L. *et al.* “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. *arXiv:2311.05232*, 2023.
HUME, D. “Tratado da natureza humana”. Tradução de Déborah Danowski. São Paulo: Unesp, 2000.
KAPLAN, C. A., SIMON, H. A. “In search of insight. *Cognitive Psychology*”. Vol. 22, Nr. 3, pp. 374–419, jul. 1990.
KERNS, M., ROTH, A. “The Ethical Algorithm: The Science of Socially Aware Algorithm Design”. Oxford University Press, USA, 2019.
KIVERSTEIN, J., WHEELER, M. “Heidegger and cognitive Science”. New York: Palgrave Macmillan, 2012.
LI, K. *et al.* “Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task”. *arXiv:2210.13382*, 2022.
LINARDATOS, P., PAPASTEFANOPOULOS, V., KOTSIANTIS, S. Explainable “AI: A Review of Machine Learning Interpretability Methods”. *Entropy*, Vol. 23, Nr. 1, p. 18, 2020.
LIU, J. *et al.* “Towards Out-Of-Distribution Generalization: A Survey”. *arXiv:2108.13624*, 2021.
MARSHALL, A., DAVIES, A. “Uber’s Self-Driving Car Saw the Woman It Killed, Report Says”. *Wired*, 2018. Disponível em https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/. Acessado em 11 de março 2025
MCCARTHY, J. “Programs with common-sense”. In: MINSKY, Marvin (ed.). *Semantic information processing*. Cambridge: MIT Press, 1968. pp. 403–418.
MCCARTHY, J., HAYES, P. J. “Some philosophical problems from the standpoint of artificial intelligence”. *Machine Intelligence*, Vol. 4, pp. 463–502, 1969.
MESQUITA, B., BOIGER, M. “Emotions in Context: A Sociodynamic Model of Emotions”. *Emotion Review*, Vol. 6, Nr. 4, pp. 298–302, 2014.
MINSKY, M. “A framework for representing knowledge”. In: HAUGELAND, J. (ed.). *Mind design II: phylosophy, psychology, artificial intelligence*. MIT Press, 1997. pp. 111–142.
MINSKY, M. “The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind”. New York: Simon; Schuster, 2006.
NEWELL, A. “Unified Theories of Cognition”. Harvard University Press, 1994.

NIKANKIN, Y. *et al.* “Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics”. *arXiv:2410.21272*, 2024.
NORVIG, P., ARCAS, B. A. Y. “Artificial General Intelligence is already here”. *Noema*, 2023. Disponível em https://www.noemamag.com/artificial-general-intelligence-is-already-here/. Acessado em 24 de março de 2025.
PERINI-SANTOS, E. “Contextualismo”. In: BRANQUINHO, J., SANTOS, R. (eds.). *Compêndio em linha de problemas de filosofia analítica*. Centro de Filosofia da Universidade de Lisboa, 2014.
PERINI-SANTOS, E. “Espaço social da dúvida: negacionismo, ceticismo e a construção do conhecimento”. *Estudos de Sociologia*, Vol. 28, Nr. esp. 1, ago. 2023.
PESSOA, L. “The entangled brain: how perception, cognition, and emotion are woven together”. The MIT Press, 2022.
PORTER, B., MACHERY, E. “AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably”. *Scientific Reports*, Vol. 14, Nr. 1, 2024.
PYLYSHYN, Z. W. (ed.). “The Robots Dilemma: The Frame Problem in Artificial Intelligence”. Ablex, 1987.
RUDIN, C. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. *arXiv:1811.10154*, 2018.
SCHUSTER, N., LAZAR, S. “Attention, moral skill, and algorithmic recommendation”. *Philosophical Studies*, 2024.
SMITH, B. C. “The Promise of Artificial Intelligence: Reckoning and Judgment”. MIT Press, 2019.
SPERBER, D.; WILSON, D. “Relevance: communication and cognition”. John Wiley; Sons, 1995.
SPERBER, D., WILSON, D. “Fodor’s frame problem and relevance theory”. *Behavioral and brain sciences*, Vol. 19, Nr. 3, pp. 530-532, 1996.
SPIES, A. F. *et al.* “Transformers Use Causal World Models in Maze-Solving Tasks”. *arXiv:2412.11867*, 2024.
TURING, A. “Computing Machinery and Intelligence”. *Mind*, Vol. LIX, Nr. 236, pp. 433-460, 1950.
VAFA, K. *et al.* “Evaluating the World Model Implicit in a Generative Model”. *arXiv:2406.03689*, 2024.
VERVAEKE, J., FERRARO, L. “Relevance Realization and the Neurodynamics and Neuroconnectivity of General Intelligence”. In: *SmartData*. Springer New York, 2013. pp. 57-68.
VERVAEKE, J., LILLICRAP, T. P., RICHARDS, B. A. “Relevance Realization and the Emerging Framework in Cognitive Science”. *Journal of Logic and Computation*, Vol. 22, Nr. 1, pp. 79-99, 2012.
WEIZENBAUM, J. “ELIZA — a computer program for the study of natural language communication between man and machine”. *Communications of the ACM*, Vol. 9, Nr. 1, pp. 36-45, 1966.

WHEELER, M. "Cognition in context: phenomenology, situated robotics and the frame problem". *International journal of philosophical studies*, Vol. 16, Nr. 3, pp. 323-349, 2008.
XU, K. *et al.* "Adversarial T-shirt! Evading Person Detectors in A Physical World". *arXiv:1910.11099*, 2019.
ZHANG, H. *et al.* "On the Paradox of Learning to Reason from Data". *arXiv:2205.11502*, 2022.