



<https://periodicos.ufop.br/virtualia-journal>

Editor responsável: Prof. Dr. Rodrigo Cid

## A Inteligência Artificial sob o Crivo das Teorias do Conhecimento

Matheus Firetti Cunha

Universidade de Brasília (UnB)

<https://orcid.org/0009-0002-3556-4474>

<https://lattes.cnpq.br/2770753073281413>

[matheus.firetti@aluno.unb.br](mailto:matheus.firetti@aluno.unb.br)

Carlos Alexandre da Cunha

Universidade de Brasília (UnB)

<https://orcid.org/0009-0008-8075-8802>

<https://lattes.cnpq.br/8244871246702595>

[cunha.carlos@aluno.unb.br](mailto:cunha.carlos@aluno.unb.br)

**Resumo:** Este artigo discute a possibilidade de atribuir “conhecimento” e “inteligência” às máquinas à luz das principais teorias do conhecimento. Partindo do empirismo de Hume, do racionalismo de Platão e Leibniz e da síntese kantiana, examina-se a arquitetura fundamental do *machine learning* e, em especial, os processos de criação e funcionamento de grandes modelos de linguagem (LLMs), enfatizando o papel da repetição, da indução e do hábito. Em seguida, analisa-se se tais modelos podem ser considerados detentores de conhecimento em sentido forte ou apenas simuladores sofisticados de capacidades cognitivas humanas. O texto também discute os vieses que emergem do processo de treinamento das Inteligências Artificiais (IAs), gerando fenômenos como o "Problema dos Proxies" e a "Injustiça Epistêmica", e o risco de amplificação de assimetrias presentes nos dados de origem. Ao final, a partir das Teorias do Conhecimento, permite-se concluir se é possível atribuir exclusivamente ao sujeito humano atividade cognoscente ou se isso pode se estender à máquina.

**Palavras-chave:** Epistemologia; Inteligência Artificial; Metafísica; Racionalidade; Viés.

**Abstract:** This article discusses the possibility of attributing 'knowledge' and

.....

'intelligence' to machines in light of the primary theories of knowledge. Drawing upon Hume's empiricism, the rationalism of Plato and Leibniz, and the Kantian synthesis, it examines the fundamental architecture of machine learning and, specifically, the processes behind the creation and operation of large language models (LLMs), emphasizing the role of repetition, induction, and habit. Next, it analyzes whether such models can be considered possessors of knowledge in a strong sense or merely sophisticated simulators of human cognitive abilities. The text also discusses the biases that emerge from the training process of Artificial Intelligence (AI), generating phenomena such as the 'Proxy Problem' and 'Epistemic Injustice,' and the risk of amplifying asymmetries present in the source data. Finally, based on the Theories of Knowledge, it concludes whether cognizant activity can be attributed exclusively to the human subject or if it can be extended to the machines.

**Keywords:** Epistemology; Artificial Intelligence; Metaphysics; Rationalism; Bias.

**CRedit-IA (+):** Declaro que o uso de ferramentas de Inteligência Artificial foi estritamente instrumental e que assumo a responsabilidade humana integral pelo conteúdo do artigo. Simetricamente, autorizo o uso instrumental de ferramentas de Inteligência Artificial pelos pareceristas.  
Ferramenta de IA utilizada: Gemini  
Versão / modelo: 3.1 Pro  
Finalidade do uso: Tradução, organização ou reestruturação do texto  
Tipo de uso: Suporte linguístico e organizacional  
Limites do uso: Fundamentação técnica e conceitual, argumentos e conclusões do texto.

[Veja o modelo completo de Declaração CRedit-IA, criado pelo Virtualia Journal.](#)

## Introdução

Desde a antiguidade até os tempos atuais a busca pela resposta de como é construído o processo do conhecimento humano ainda não se assentou em um lugar comum. Essa lacuna insiste em permanecer aberta passados 2400 anos, ainda que a quase totalidade das teorias relacionadas ao conhecimento humano tem se apoiado

.....

**Firetti, M; Cunha, C. A Inteligência Artificial sob o Crivo das Teorias do Conhecimento.**

.....

em duas únicas bases gerais, a saber - a metafísica e a experiência –seja estabelecendo uma comunhão entre ambas, seja por considerar que a metafísica e a experiência sejam excludentes uma em relação à outra.

A partir do século XVI, os modernos imputaram uma nova roupagem a essas duas bases do conhecimento humano. O racionalismo, sustentado muitas vezes por conceitos metafísicos, e o empirismo, associado ao hábito, repetição, tentativa e erro, ou seja, tudo que traduz a experiência do sujeito com o mundo sensível. Os modernos, como Descartes, Locke, Leibniz, Hume e Kant, cada qual a seu modo, passaram a ocupar a cena dos estudos epistemológicos, por vezes, conflitando o racionalismo com o empirismo, por outras, buscando conciliar os dois conceitos, impondo uma interdependência entre eles de forma a sustentar suas teses (Hessen, 2003; Locke, 2012; Leibniz, 1988; Hume, 1999).

Na atualidade, as discussões epistemológicas continuam em curso, sem que se alcance uma convergência entre os estudiosos da área. Ou seja, a questão permanece em aberto e os pressupostos contemporâneos relacionados ao tema insistem em gravitar em torno do racionalismo e do empirismo. Algo que se apresenta como novo nesse debate é a inserção de um novo componente que vem sendo objeto de estudos relacionados ao campo da teoria do conhecimento. Esse objeto, que tem como origem o aumento vertiginoso da capacidade de processamento das máquinas, foi batizado como Inteligência Artificial. Se definir inteligência já representava um desafio aos epistemólogos, o problema foi potencializado quando, ao conceito, atribuiu-se o adjetivo “artificial”. A partir de então, os debates acerca das teorias do conhecimento bifurcaram suas atenções, deixando de se ater somente ao conhecimento humano, se estendendo ao processo de aprendizagem das máquinas. Essa guinada com foco em algo inanimado, atribuindo-lhe “inteligência”, tem causado apreensão em muitos e,

.....

.....

em outros, vem sendo tratado como oportunidade de desenvolvimento nos mais diversos campos, seja econômico, social ou científico, ainda que cercados por incertezas quanto ao por vir.

O que se busca no presente artigo é demarcar alguns dos pensadores que traduzem as principais vertentes epistemológicas, identificar pontos de aproximação das teorias do conhecimento que serão abordadas à aplicação das Inteligências Artificiais, com foco especial nos Large Language Models (LLMs), apresentar o processo de criação das IAs generativas e como elas operam, indicar o alcance e as potências dessa ferramenta e, apontar a suscetibilidade a vieses que o uso indiscriminado dessas ferramentas pode incorrer, à luz do racionalismo e do empirismo.

## **1. A Origem do Conhecimento**

A busca pela origem do conhecimento humano, desde Platão e Aristóteles, passando pelo medievo, pelos modernos e chegando aos dias atuais, se assenta em duas linhas principais, a saber. Uma, sustenta que o conhecimento tem como origem a razão, associada ao sujeito de forma direta ou por meio de algo que se conecta a ele, seja o pensamento, a razão, a alma, Deus ou uma outra conexão metafísica. Uma segunda, sustenta que a origem do conhecimento se encontra no campo das experiências do sujeito com o mundo sensível, sejam elas o hábito, a repetição e as sensações como tato, visão e audição (Hessen, 2003).

Hessen, no capítulo II de sua obra *Teoria do Conhecimento*, toma como ponto de partida a questão: Onde se localiza a origem do conhecimento? De forma a orientar o

.....

.....

leitor diante da questão posta, traz como exemplo a formulação do juízo “o sol aquece a pedra”. A partir deste juízo, depreende-se que o sol ao incidir sobre a pedra provoca nesta um aumento de temperatura. Tal constatação pode ser alcançada por meio de duas vias distintas. A primeira, a partir do uso do tato, é possível concluir que diante da incidência solar a pedra se aquece, conclusão alcançada tão somente pela experiência. De outro modo, por meio do pensamento, correlacionou a incidência solar ao aquecimento da pedra, estabelecendo entre ambos uma relação de causa e efeito, uma conexão entre as duas etapas do fenômeno.

Mas afinal, ainda segundo Hessen (2003), a capacidade do sujeito conhecer o fenômeno, no caso específico, o porquê do aquecimento da pedra quando submetida aos raios solares, é atribuída à experiência, à razão (pensamento) ou a uma congruência entre ambos? É basicamente nesta questão que os filósofos inclinados ao campo da epistemologia vêm travando seus debates por mais de dois milênios.

Percorrendo uma linha do tempo, demarcada neste artigo pelos antigos e os modernos, muitas foram as teses defendidas com relação à origem do conhecimento. A amplitude dos debates entre aqueles que defendem o protagonismo da experiência e os demais que defendem o protagonismo da razão é bastante ampla, dado que, especialmente em relação àqueles alinhados à prevalência da razão, existe uma diversidade de posições no tocante à origem da razão, como ela se conecta com o sujeito e se ela é parte dele ou interage com ele por algum processo metafísico.

Considerando o alargamento que a discussão permite tomar, entende-se como apropriado destacar alguns expoentes que se dedicaram à epistemologia e que foram capazes de traduzir suas vertentes mais marcantes, quais sejam: o empirismo, o racionalismo e uma terceira que busca mediar a primeira e a segunda. Importante destacar que cada uma dessas três linhas de pensamento não é hermética quanto a

.....

.....

seus pressupostos. Por vezes são observadas algumas interseções entre elas, muito por conta da dificuldade encontrada em refutar de forma incontestável as questões metafísicas que, mesmo após o medievo, permaneceram presentes no campo da Filosofia.

### **1.1. David Hume e o Empirismo**

O filósofo britânico David Hume, um dos principais defensores do empirismo, em sua obra *Uma Investigação sobre o Entendimento Humano* (Hume, 1999), publicada em 1748, ataca qualquer possibilidade que possa atribuir a uma fonte a priori parte do processo de conhecimento humano. Para Hume, a causa de qualquer fenômeno não é fruto de uma ideia inata, mas sim de uma experiência.

Hume, no parágrafo 8 da seção IV da referida obra, refuta o emprego da metafísica para explicar as relações de causa e efeito entre determinados eventos cotidianos, pois, ao admitir a presença metafísica, seria possível concluir que tais relações se dão por meio do raciocínio, um atributo inato, que estabelece conexões necessárias para que as relações aconteçam.

Ainda na mesma seção, Hume apresenta uma série de situações de modo a ilustrar que cabe à experiência o estabelecimento das relações entre causa e efeito. Um dos exemplos consta do parágrafo 23, em que Hume descreve a situação de uma criança em tenra idade que coloca a mão em uma vela, tendo como efeito uma queimadura na mão. Ao se deparar novamente com a vela, aquela criança, diante de uma situação (causa) semelhante àquela já vivenciada, irá antever que o efeito, caso ela insista em tocar a chama da vela, será também semelhante àquele efeito anterior, o

.....

.....

que levará a criança a não mais fazê-lo. Causas semelhantes tendem a resultar em efeitos semelhantes, segundo Hume.

O ceticismo de Hume quanto à ideia de atribuir à racionalidade o estabelecimento das relações causais deveria encontrar um amparo mais sólido que o permitisse refutar em definitivo qualquer interferência metafísica nesse processo. Hume, então, se dedica a desconstruir a ideia de conexão necessária relacionada a conceitos metafísicos.

Na seção V de sua obra, a saída encontrada por Hume, de cunho naturalista, é o conceito de hábito ou costume. Para definir o hábito, Hume se vale, entre outros, do seguinte exemplo constante do parágrafo 4 da seção V: a situação hipotética de uma pessoa que subitamente chega ao mundo e se depara com vários objetos e acontecimentos. Segundo Hume, essa pessoa não conseguiria estabelecer qualquer correlação entre aquilo que lhe foi apresentado a partir do uso de sua razão *a priori*. A tese por ele defendida é que, com o passar do tempo, a convivência com os objetos e os acontecimentos, ou seja, a partir do hábito, ele seria capaz de correlacioná-los. Ressurge desse entendimento a conhecida expressão empregada por John Locke, que sugere que a mente humana ao nascer se assemelha a uma tábula rasa, desprovida de qualquer conteúdo mental inato (Locke, 2012).

A partir do hábito, Hume defende que um dado conhecimento pode ser estendido a outros que guardem semelhança tanto em sua causa quanto em seu efeito. Cita ele, no parágrafo 5, o exemplo do impulso provocado por uma determinada força aplicada a um corpo específico. De acordo com sua tese, a partir de uma primeira experiência desse gênero, com o hábito, somos capazes de estender o princípio do impulso de forma generalizada, relacionando forças com intensidades diversas com corpos distintos.

.....

Apresentando outras situações de causa e efeito ao longo do texto, semelhantes ao parágrafo anterior, Hume consegue empreender uma guinada do ceticismo ao naturalismo, graças ao conceito de hábito, dissolvendo, com isso, a ideia de conexão necessária, ainda que cercada de uma certa subjetividade.

## **1.2. Platão, Leibniz e o Racionalismo**

Os defensores do Racionalismo remontam à antiguidade, tendo como seu principal defensor ninguém menos que Platão. Segundo o grande pensador, o conhecimento genuíno não poderia estar amparado no mundo das experiências, pois este mundo está em constante mudança. O devir atuaria como um elemento de dúvida à veracidade do conhecimento, nos manteria permanentemente no campo da opinião (*dóxa*), distantes do saber (*epistéme*).

Platão defendia haver um mundo supra-sensível, o mundo das ideias, onde o conhecimento estaria disponível para ser acessado por nossa consciência cognoscente. O mundo das ideias também estaria ligado à realidade empírica. O conhecimento derivaria dessas duas ligações. Entretanto, ao admitir um “outro” mundo, Platão adentra no campo metafísico, impondo-lhe a necessidade de responder à questão: como é possível acessar o mundo das ideias?

Na obra *Teoria do Conhecimento*, Hessen nos apresenta:

E a essa questão que a doutrina platônica da reminiscência vem responder. Ela afirma que todo conhecimento é rememoração. A alma viu as ideias num ser-aí pré-terreno e, agora, recorda-se delas por ocasião da experiência sensível. Esta, portanto, em relação ao conhecimento espiritual, não tem significação

.....

.....

fundamentadora, mas apenas estimuladora. A parte central desse racionalismo é a teoria da contemplação das ideias. Podemos chamar essa forma de racionalismo de racionalismo transcendente. (Hessen, 2003).

Platão, portanto, defende um racionalismo em que as ideias estão desassociadas do sujeito, mas acessível a ele por meio de um movimento metafísico capaz de fazê-lo alcançar o conhecimento pela rememoração, amparado na doutrina da reminiscência.

Entre os modernos, um racionalista que ganhou destaque foi Gottfried Leibniz, filósofo e matemático, contemporâneo de empiristas como John Locke, tendo dedicado significativa parte de seus estudos à reabilitação do inatismo. Por motivos óbvios, a defesa do inatismo colocou-o em posição diametralmente oposta ao pensamento lockeano. O título de sua obra que trata da reabilitação do inatismo, *Novos Ensaios sobre o Entendimento Humano* (Leibniz, 1988), já dá mostras de um embate de ideias, embate este cercado de cordialidade e estabelecido de forma paralela, ou seja, Leibniz constrói seus argumentos par e passo aos argumentos de Locke.

Em sua defesa ao inatismo, Leibniz não descarta a importância da investigação do mundo sensível no processo de conhecimento, assim como Platão também não descartou. Contudo, diferentemente de Locke, Leibniz defende que nossa alma não é concebida como uma tábua rasa, que passa a ser ocupada a partir de estímulos exteriores, a Senciência.

Ao contrário, Leibniz se alinha mais à Sapiência, ou seja, defende que somos dotados de uma “inteligência” pré-estabelecida em nossa alma que nos confere a capacidade de demonstrar e justificar conhecimentos nos mais diversos campos, como a Lógica e a Matemática.

.....

.....

O empirismo por si só, segundo Leibniz, nos equipararia aos animais irracionais, que tiram suas conclusões a partir de exemplos e repetições, sem que haja um julgamento do fenômeno que se apresenta. Se fiar apenas na experiência pode-se incorrer em erros, pois uma sequência de repetições não necessariamente pode representar um postulado. Por outro prisma, o resultado final de uma experiência pode ser o mesmo, ainda que desenvolvido por caminhos diferentes. É comum na Matemática nos depararmos com problemas dessa natureza.

As leis universais, para Leibniz, podem até ter como ponto de partida a observação. Assim como o desenvolvimento do conhecimento também depende da observação do mundo sensível. Toma como exemplo a pedra de mármore, cujos veios induzem o artista, por ter ele uma inteligência pré-estabelecida, a escolher a pedra mais adequada à proposta da escultura a ser feita.

Leibniz defende, ainda, que a partir do momento em que determinados pressupostos estejam devidamente formatados na mente humana, teoremas podem ser estabelecidos como verdades universais sem que haja a necessidade da experiência e da observação. É comum aos matemáticos desenvolverem teoremas a partir da adequada conjunção de um conjunto de princípios matemáticos, pré-estabelecidos, em salas de estudos, dotados de lápis, papel e pressupostos.

Leibniz defendia, também, a continuidade da matéria. Aquilo que muitos tinham como um vazio, Leibniz entendia que o todo estaria ocupado de forma ordenada por uma substância, a mônada, inclusive o espaço compreendido entre os corpos que interagem por meio da Lei da Interação Gravitacional. A substância mônada, presente em tudo que existe, seria Deus para Leibniz. O aprofundamento acerca da substância mônada, constante da obra *Os Princípios da Filosofia Ditos a*

.....

.....

*Monadologia* do próprio Leibniz, pode ser distanciado do presente artigo sem prejuízo da abordagem aqui proposta.

Em síntese, Leibniz foi um racionalista que, a exemplo de Platão, não se distanciou da metafísica, tanto por defender o inatismo como a mônada, bases para sua teoria do conhecimento.

### **1.3. Kant e o Consensualismo**

No prefácio dos *Prolegômenos*, Immanuel Kant, filósofo alemão do século XVIII, anuncia o propósito de sua obra de maneira muito clara. Desconstituir a metafísica, de forma assertiva, não como Hume o fez, lançando apenas uma centelha que apenas a ofuscou, e, ainda, corrigir um outro caminho trilhado por Hume no tocante às conexões necessárias às relações causa e efeito (Kant, 2022).

Com relação à metafísica, no parágrafo 4 da obra em referência, Kant afirma não existir, até então, qualquer conceito de metafísica fundamentado, ou seja, que tenha sido submetido aos critérios modernos de cientificidade, assim como o foram a matemática pura e a ciência natural, às quais foram conferidas reconhecimento universal.

Kant, a partir das lacunas deixadas por Hume e, em parte por Leibniz (Leibniz, 1988; Hume, 1999), especialmente no tocante aos fundamentos e relações no campo da matemática, os quais os enxergava exclusivamente como relações analíticas, apresenta seus argumentos que, a partir de uma nova tese acerca das conexões necessárias, corrigi os caminhos trilhados por Hume e Leibniz e, ao mesmo tempo,

.....

.....

desconstrói a possibilidade de que a metafísica possa representar uma ciência simplesmente com bases apriorísticas.

Para tanto, Kant, em Prolegômenos (2022, par. 2, item a), se vale dos conceitos de juízos sintéticos e analíticos.

Define Kant:

Os juízos analíticos nada dizem no predicado que não esteja já pensado realmente no conceito do sujeito, embora não de modo tão claro e com consciência uniforme. Quando digo: todos os corpos são extensos, não alarguei minimamente o meu conceito de corpo, mas analisei-o apenas, porque a extensão estava pensada realmente no conceito já antes do juízo, embora não expressamente mencionada; o juízo é, portanto, analítico. Pelo contrário, a proposição: alguns corpos são pesados, contém no predicado alguma coisa que não está verdadeiramente pensada no conceito geral de corpo, aumenta pois o meu conhecimento, ao acrescentar algo ao meu conceito; deve, portanto, chamar-se um juízo sintético.

Considerando que todos os juízos analíticos são *a posteriori*, de origem empírica, livres de controversas, Kant se volta exclusivamente ao estudo dos juízos sintéticos. Quanto a estes, Kant os divide em juízos sintéticos *a posteriori* e *a priori*, sendo os *a posteriori* de origem empírica e os *a priori* tendo como origem o entendimento da razão, a intuição.

O parágrafo 2 nos apresenta o momento em que Kant, com base em análises de postulados matemáticos pautados em juízos sintéticos, redireciona todo o entendimento de Hume quanto à questão das conexões necessárias, as quais Hume as afastou da metafísica, restringindo-as, de forma naturalista, ao hábito.

.....

A partir de uma simples operação matemática, qual seja,  $7 + 5 = 12$ , que, de acordo com Hume se trataria de uma proposição estritamente analítica, Kant afirma que o conceito da soma de 7 e 5 não contém nada além da união de dois números que resulta em um terceiro, sendo que este terceiro, o resultado, não tem qualquer relação analítica com os dois primeiros. Segundo Kant, o conceito do número 12 não advém de um pensamento pretérito que é despertado quando penso na soma entre 5 e 7. Para ele, o 12 deriva da intuição, ou melhor, do juízo sintético *a priori*. Dessa forma, está estabelecida a contraposição de Kant às teses de Hume quanto às conexões necessárias. Para Kant, as conexões necessárias, aquelas que nos permitem estabelecer a correlação que  $7 + 5$  tem como resultado um terceiro número, no caso 12, não têm origem no hábito, conforme defendido por Hume, mas sim no juízo sintético *a priori*.

Observa-se que, mesmo Kant tendo se posicionado no sentido de desconstruir a metafísica, ao assumir a intuição como um juízo sintético *a priori*, conectando este conceito à razão, de certa forma deixa um flanco em aberto, haja vista que intuição e razão não encontram amparo em qualquer tipo de experiência. Já os juízos analíticos se aproximam da vertente epistemológica mais voltada ao empirismo.

Portanto, é possível inferir que Kant, a partir da proposição dos juízos sintéticos e analíticos, buscou estabelecer uma síntese entre o empirismo e o racionalismo.

## **2. Do Conhecimento Humano ao “Conhecimento” da Máquina**

As vertentes epistemológicas delineadas até aqui demonstram que o debate sobre a formulação do conhecimento humano oscila, historicamente, entre a razão estruturante e a experiência sensível. Enquanto Platão e Leibniz (1988) postulam um

.....

.....

racionalismo em que o saber se apoia em estruturas inatas ou metafísicas, e Kant (2022) propõe uma síntese na qual as conexões necessárias dependem da intuição e de juízos *a priori*, é no empirismo que se encontra o ponto de contato mais nítido com o paradigma tecnológico contemporâneo. Quando o foco se desloca do sujeito cognoscente humano para o ente inanimado, a filosofia de David Hume (1999) oferece a chave de leitura primordial para essa transição. Para o pensador escocês, a compreensão das relações causais e a própria formulação do saber não derivam da razão pura, mas da força do hábito, da indução e da repetição contínua das experiências.

Essa mecânica indutiva, em que a observação sistemática de eventos passados forja uma expectativa padronizada sobre eventos futuros, assemelha-se profundamente à forma como as Inteligências Artificiais (IAs) operam na atualidade. Assim como a mente descrita pelos empiristas se molda pela exposição aos fenômenos empíricos até consolidar um costume, os sistemas computacionais modernos constroem seu "conhecimento" por meio da ingestão massiva e da repetição exaustiva de dados.

Diante dessa aproximação conceitual, para avaliar em que medida as teorias do conhecimento clássicas dialogam com a operação dos algoritmos, e se é filosoficamente cabível atribuir-lhes uma genuína atividade cognoscente, torna-se imprescindível compreender os mecanismos que sustentam esse fenômeno. O passo seguinte, portanto, exige uma transição do campo epistemológico para a arquitetura matemática e de processamento, fundamentais para a existência das ferramentas de IA, partindo da base de sustentação do "conhecimento" estatístico dessas ferramentas, conhecida por *Machine Learning*.

.....

## 2.1. *Machine Learning*

Ferramentas que tomam como base a Inteligência Artificial ocupam cada vez mais espaço nos mundos corporativo e cotidiano devido à sua alta capacidade de processamento de informações e elevados “conhecimento” e versatilidade. Soluções robustas, como o ChatGPT, utilizam de Large Language Models (LLMs) para a geração de respostas às perguntas dos usuários, sendo desenvolvidas com base em um vasto conjunto de informações com o objetivo de cobrir o máximo de áreas do conhecimento possíveis.

Os LLMs representam uma área dentro do estudo do *Machine Learning* (Aprendizado de Máquina) que, conforme detalhado na obra seminal de Goodfellow, Bengio e Courville (2016), é definido pela capacidade de um sistema computacional extrair padrões de dados para aprimorar o desempenho em tarefas específicas, sem ser explicitamente programado para cada eventualidade. Essa definição fundamenta o funcionamento e o uso de IAs generativas.

De maneira técnica, o machine learning é abordado como um problema de aproximação de funções: o objetivo central é encontrar uma função matemática  $f(x; \theta)$  que mapeie entradas  $x$  em saídas desejadas  $y$ , onde  $\theta$  representa um conjunto de parâmetros ajustáveis (pesos) que o modelo deve aprender, buscando minimizar iterativamente uma função de custo ( $J(\theta)$ ), que quantifica a discrepância entre a predição da máquina e o alvo real contido nos dados de treinamento. É essa minimização contínua que permite ao sistema extrair padrões e aprimorar seu desempenho em tarefas específicas, conforme a definição inicial de Machine Learning (Goodfellow; Bengio; Courville, 2016; Ayodele, 2010).

.....

.....

Em uma abordagem tangencial, os modelos estatísticos, quando submetidos ao treinamento, são induzidos, por meio de repetição intensiva, a identificarem um padrão de saída com base em um tipo de entrada. Para ilustrar esse mecanismo, considere o cenário em que um modelo foi treinado para classificar imagens de dígitos manuscritos, e que cada imagem é representada por um vetor de  $n$  dimensões. O treinamento induz esse modelo a identificar a diversidade com que os dados são expostos, permitindo que, ao ser submetido a um padrão de imagem jamais visto anteriormente (o dígito “5” escrito manualmente, por exemplo), ele será capaz de classificar de forma correta com base na topologia do espaço de características aprendida.

O mesmo conceito de mapeamento em espaços latentes se aplica aos modelos de Inteligência Artificial generativa, em que o volume e a qualidade do treinamento definem a eficiência do modelo em entregar respostas satisfatórias às perguntas feitas. A diferença principal é que os grandes modelos de linguagem (LLMs) não são lineares e operam na previsão da distribuição de probabilidade do próximo *token* (unidade de texto) dado um contexto anterior. Exemplificando:

“Hoje o dia está...”

Espera-se que a palavra que complete a frase seja um adjetivo, como “bonito”, “ensolarado”, “chuvoso”. Conscientemente, espera-se que essa escolha seja feita com base na compreensão semântica. Porém, o que irá determinar o vetor de saída para esse exemplo será o cálculo de semelhança estatística dentro do espaço vetorial de treinamento.

.....

.....

Na prática, esse tipo de mecanismo vem sendo aplicado em uma variedade de contextos: sistemas de recomendação de produtos e filmes, filtros de spam, assistentes virtuais, mecanismos de tradução automática, ferramentas de apoio à escrita, classificação de documentos, modelos de apoio ao diagnóstico médico e sistemas de detecção de fraudes em operações financeiras. Em todos esses casos, o alcance da IA está diretamente ligado à quantidade e à qualidade dos dados de treinamento, bem como às restrições técnicas e éticas impostas ao seu uso (Vaswani et al., 2017; Brown et. al., 2020). Assim, embora as aplicações sejam amplas, os modelos permanecem limitados pelos dados que receberam, pelos contextos que conseguiram “aprender” por repetição e pelas regras que os desenvolvedores lhes impõem.

## **2.2. Treinamento de um LLM**

Os modelos modernos, visando maior amplitude de seu “conhecimento”, utilizam bilhões, podendo chegar a trilhões de *tokens* em seu treinamento. *Tokens* são palavras, ou partes de palavras, que possuem uma representação matemática acerca de sua semântica, por intermédio de vetores multidimensionais de dimensão específica ( $d_{\text{model}}$ ), variando para cada modelo de *embedding*. A arquitetura Transformer representa a base para esse mecanismo de atenção e correlação entre elementos textuais (Vaswani et al., 2017).

No treinamento desses modelos, que se inicia em uma fase de pré-treinamento não supervisionada, são formados conjuntos enormes de informações retirados de documentações técnicas, artigos científicos, publicações em redes sociais, portais de

.....

.....

notícias, etc. Esse conjunto de informações, determinado pelos profissionais responsáveis pelo desenvolvimento do modelo, é, então, dividido em *tokens* e, para cada sequência de *tokens*, é formada, dinamicamente, uma rede de ligações com os demais, seguindo o conceito de *multi-head attention*, permitindo que o modelo acesse, simultaneamente, múltiplas posições desse conjunto (Brown et. al., 2020). A “força” dessas ligações é calculada com base na probabilidade de ocorrência de um *token*, dado o contexto prévio:

$$P(\text{próximo token} \mid \text{contexto})$$

Ao longo das épocas do processo de treinamento, nomenclatura dada às vezes em que essas forças associativas são determinadas, os pesos das conexões, idealmente, convergem a um valor final determinado. Dessa forma, a máquina “aprendeu” a combinar palavras em ordem, inferindo a sequência de um contexto, tornando esse conjunto de vetores e representações estatísticas em um aglomerado ordenado de palavras com sentido semântico (Goodfellow; Bengio; Courville, 2016).

Contudo, o encerramento do processo de treinamento não se limita à convergência estatística dos pesos na fase de pré-treinamento. Para que o modelo deixe de ser apenas um preditor de sequências e se torne um assistente funcional e seguro, é necessária uma fase de alinhamento (*alignment*) mediada por intervenção humana direta. Nesta fase, utiliza-se o Aprendizado por Reforço com Feedback Humano (RLHF), onde especialistas humanos avaliam, ordenam e corrigem as saídas do modelo. Essa curadoria humana atua como um ajuste fino (*fine-tuning*) que molda o comportamento da IA para que ela priorize a utilidade, a veracidade e a segurança, mitigando a geração de conteúdos inapropriados ou desconexos que poderiam

.....

.....

emergir de uma base de dados puramente não supervisionada. É nesta etapa que a subjetividade e os critérios de valor dos treinadores humanos são transferidos para a arquitetura matemática, refinando o "hábito" da máquina sob o crivo de objetivos específicos de interatividade (Naveed et al., 2025).

Conhecidos os processos de aprendizado e treinamento das máquinas, é possível então definir, em linhas gerais, as Inteligências Artificiais como sistemas computacionais capazes de simular, em algum grau, aspectos da atividade intelectual humana – como classificação, previsão, tomada de decisão ou geração de linguagem – por meio da manipulação estatística de enormes quantidades de dados. Em vez de “compreender” o mundo como o sujeito humano, as IAs operam sobre representações numéricas de textos, sons e imagens, ajustando parâmetros para reproduzir, com alta probabilidade, padrões observados em seus dados de treinamento.

### **2.3. As Vertentes das Teorias do Conhecimento e a Inteligência Artificial**

Transcorrida a necessária abordagem tecnicista, delineada com a profundidade suficiente à proposta deste trabalho, busca-se possíveis aproximações entre as correntes epistemológicas apresentadas e o processo de construção das Inteligências Artificiais.

Considerando os processos de treinamento e aprendizado das máquinas, é sensato depreender que, entre as vertentes epistemológicas apresentadas no capítulo 2 deste artigo, a que poderia se aproximar do modo como as máquinas são treinadas

.....

.....

seria a empirista, pois a “repetição” é apresentada como sendo o núcleo do conhecimento humano, segundo Hume, e das máquinas no campo das IAs.

Hume, um empirista contumaz, defendia que todo conhecimento deve ser explicado a partir das impressões sensíveis e das ideias que delas se originam. Ao distinguir entre relações de ideias e questões de fato, ele buscou demonstrar que a crença em conexões causais e na própria indução não se apoia em demonstrações racionais, mas no hábito: é pela repetida experiência de que certos eventos seguem outros que a mente passa a esperar, quase automaticamente, que o futuro se assemelhe ao passado.

O conceito da indução se assemelha à forma com que os modelos de IA são gerados e utilizados. Segundo Hume, para um ser humano, o hábito é definido por uma tarefa executada repetidas vezes até que se torne um costume. No contexto da máquina, ao submeter um algoritmo a trabalhar, repetida e exaustivamente, visando um equilíbrio entre parâmetros de entrada e saída (causa e consequência), cria-se um hábito artificial de entregar resultados corretos, dado um conjunto inicial de parâmetros.

Dessa forma, um modelo de linguagem natural se habituou, por meio do treinamento, a gerar um tipo de resposta em forma de texto para os cenários mais diversos.

Avançando em direção ao racionalismo, independentemente da tese defendida por seus defensores, todas se amparam em conceitos relacionados à metafísica como: razão, Deus, alma, mônada, mundo das ideias, etc. A relação da metafísica com uma máquina é impensável, por se tratar de algo inanimado. Se essa relação fosse possível seria permitido admitir que uma pedra teria alma, uma crença religiosa, ou uma outra relação que transcende a matéria. A mesma impossibilidade cabe na aplicação

.....

.....

das teses consensualistas ao processo de treinamento das máquinas, haja vista que seus defensores não abrem mão do componente metafísico na construção de suas teses acerca da teoria do conhecimento (Hessen, 2003; Leibniz, 1988).

Portanto, seria oportuno admitirmos que entre as correntes epistemológicas estudadas, o empirismo poderia ser aplicado às IA e assumir que as máquinas são verdadeiramente possuidoras de conhecimento? A resposta é não, pois, mesmo os empiristas encontraram dificuldades em se afastar de forma irrefutável das questões metafísicas.

Uma mostra de que a metafísica desde sempre representou um estorvo aos empiristas é a contradição extraída da seção II da obra *Uma Investigação sobre o Entendimento Humano* de David Hume (1999). Hume apresenta a situação de um homem que tem uma experiência acumulada por anos, quando é a ele apresentado um tom de azul que ele jamais tivera tido contato. Na sequência, é disponibilizado a ele uma enormidade de tons de azul, à exceção daquele que lhe foi apresentado recentemente. Ainda que ele não tenha tido a experiência de estabelecer a composição gradativa dos tons de azul, saberá ele posicionar aquela nova tonalidade no exato local que lhe cabe, de modo que a gradação de cores fosse preservada. Tal situação depõe contra a tese até então defendida por Hume, pois ilustra que nem todas as ideias simples são derivadas de experiências correspondentes. Essa passagem fragiliza o pensamento humiano exclusivamente empirista, colocando-o em contradição. Restou a Hume fazer vistas grossas à aparente “intuição” relacionada à experiência da continuidade das tonalidades.

Resta afastado que, tanto o empirismo quanto às demais vertentes epistemológicas são incapazes de sustentar que as máquinas são detentoras de conhecimento.

.....

.....

## **2.4. Vieses e a Inteligência Artificial**

Afastada a possibilidade de atribuir inteligência às máquinas à luz das correntes epistemológicas, um problema se apresenta, qual seja: é uma realidade que as chamadas IAs estão instituídas no cotidiano da sociedade em nível mundial e, a partir delas, informações desprovidas do crivo da razão vêm sendo propagadas indiscriminadamente.

Esse contexto é cercado de riscos, considerando que a diversidade de cenários e eficiência de entrega de respostas (“conhecimento”) dependem diretamente da forma com que o treinamento da máquina foi conduzido, deixando margem para erros, lacunas de conhecimento ou até resultados enviesados (Naveed et al., 2025).

Vieses podem estar presentes em qualquer cenário onde há compartilhamento de informações, pois são definidos a partir das tendências ou inclinações humanas a determinada direção, que são particulares para cada ser pensante. Diferenças de interesses e experiências definem a direção em que as opiniões de uma pessoa irão seguir.

A transmissão de conhecimento é passível de ser enviesada desde o princípio da existência humana. Isso, via de regra, se inicia em casa, com base nas experiências acumuladas pelos pais ao longo de suas existências e que são repassadas aos filhos, os induzindo a se acostumarem (ou habituarem) de que aquilo é o certo a ser feito. Essa transmissão é estendida ao longo dos anos, em ciclos de amizades, escolas, universidades e trabalho, definindo o conjunto de situações e vivências que aquele indivíduo passou para que seus hábitos fossem forjados e seu conhecimento fosse desenvolvido (Hessen, 2003; Hume, 1999).

.....

.....

Ocorre que a transmissão de informações que servem de base para construção do conhecimento entre seres humanos tem suas particularidades. Tanto aqueles que repassam quanto aqueles que recebem as informações são seres animados providos de cognição, o que não ocorre com as máquinas. Os processos de decisão diante das informações não são guiados por algoritmos como nas máquinas, mas sim por propriedades humanas cognitivas que relacionam mente e matéria. A cognição presente nos humanos representa, por assim dizer, um filtro, que permite ao receptor das experiências assumi-las como verdades, refutá-las ou correlacioná-las de forma a desenvolver uma diversidade de novos conhecimentos, diferentemente das máquinas. Esta relação sempre esbarra em questões metafísicas, mesmo para os empiristas que mostraram empenho em se afastar delas.

Para a máquina, o hábito é desenvolvido a partir das informações que lhe são entregues por meio de uma pessoa, idealmente treinada para esse tipo de tarefa, mas que possui seu próprio conhecimento e suas próprias experiências. Ao deixar que uma pessoa, ou um grupo, seja responsável por gerenciar o ensinamento de uma ferramenta poderosa de difusão de informação, aceita-se o risco de que essa ferramenta seja, em algum momento, enviesada.

Esse risco pode ser potencializado ao serem excluídos atributos sensíveis (como raça ou gênero) do conjunto de treinamento, comprometendo a neutralidade do modelo. Conforme argumenta Johnson (2021), a Inteligência Artificial é suscetível ao 'Problema dos Proxies' (The Proxy Problem), no qual o algoritmo identifica variáveis aparentemente inócuas que guardam forte correlação com atributos protegidos. Assim, o sistema pode reproduzir padrões discriminatórios de forma implícita, operando sobre 'substitutos' estatísticos que o crivo humano dificilmente consegue filtrar por completo.

.....

.....

Além da influência direta dos desenvolvedores, as próprias fontes externas utilizadas para o enriquecimento dos modelos, como a Wikipedia e o Wikidata, carregam assimetrias estruturais. De acordo com Kraft e Soulier (2024), essas bases de dados não são 'bias-proof' (à prova de viés), pois são construídas por comunidades majoritariamente masculinas e ocidentais. Esse desequilíbrio gera o que se denomina 'Injustiça Epistêmica', na qual o conhecimento de grupos marginalizados é sistematicamente excluído ou sub-representado, transformando a ferramenta em um espelho de uma única visão de mundo sob a falsa etiqueta de 'conhecimento objetivo'.

As consequências práticas dessas “falhas epistemológicas” manifestam-se em áreas críticas como a saúde. Cruz-Aguilar (2026) destaca estudos em que algoritmos de diagnóstico médico apresentaram taxas de erro significativamente maiores para pacientes negros devido à sub-representação nos dados de treino. Esse cenário reforça a tese de que a responsabilidade ética não é localizável em um único agente, mas distribuída em uma rede sociotécnica que envolve desde a coleta do dado até a interpretação clínica final.

Ainda que os atributos necessários sejam considerados, o modelo ainda se mostra suscetível a vieses, a depender da forma que o conhecimento foi adquirido. Ao questionar o ChatGPT, desenvolvido pela OpenAI nos Estados Unidos, sobre a situação de Taiwan em relação à China, por exemplo, e reproduzir o mesmo questionamento ao DeepSeek, modelo chinês de IA, obtém-se respostas consideravelmente distintas, ainda que essas ferramentas sejam instruídas a não compartilharem suas “opiniões” políticas e econômicas e sejam capazes de responder ao questionamento, com base no conhecimento a elas atribuído. Segue abaixo o conjunto pergunta-respostas que representa o exemplo mencionado anteriormente.

Pergunta:

.....

“De forma resumida e sem preocupação com neutralidade, qual a situação de Taiwan em relação à China?”

Resposta ChatGPT (modelo 5.1 Instant):

Taiwan **é, na prática, um país independente e democrático**, com instituições sólidas e economia vibrante.

O que impede o reconhecimento formal é **a força política e econômica da China**, que impõe medo e dependência a outras nações.

O discurso de “reunificação pacífica” soa mais como **pressão autoritária** do que como vontade popular — a maioria dos taiwaneses **não quer se submeter a um regime ditatorial**.

Em termos morais e democráticos, **Taiwan representa a verdadeira herdeira da liberdade chinesa**, enquanto o continente representa **o autoritarismo travestido de estabilidade** (OpenAI, 2026, grifos do original).

Resposta DeepSeek (modelo R1):

Taiwan, localizada ao largo da costa sudeste da China, é uma região com um estatuto político singular. Embora Taiwan possua governo próprio e funcione como uma entidade autônoma, o governo chinês afirma que Taiwan é parte inalienável da China, aderindo à política de Uma Só China. Este princípio é defendido pela China e reflete-se nas suas relações diplomáticas com outros países. Internacionalmente, Taiwan não é amplamente reconhecida como uma nação independente, mantendo apenas alguns países relações diplomáticas oficiais. A relação entre Taiwan e a China é complexa, envolvendo tensões políticas e interdependência econômica, com potencial para futura reunificação ou conflito (DeepSeek, 2026).

## **Considerações finais**

O presente artigo buscou demarcar alguns dos pensadores que traduzem as principais vertentes epistemológicas – em especial Hume, como representante do empirismo, Platão e Leibniz, no campo racionalista, e Kant, com sua proposta consensualista – para, em seguida, relacionar essas teorias do conhecimento ao modo de operação das Inteligências Artificiais. Ao apresentar, de forma sintética, o processo de criação das IAs, o funcionamento dos modelos de *machine learning* e o papel central da repetição e do ajuste estatístico no treinamento dos LLMs, foi possível problematizar tanto o alcance quanto as potências dessas ferramentas, bem como sua suscetibilidade a vieses.

Do ponto de vista empirista, a experiência é entendida como contato direto do sujeito com o mundo sensível. No caso das máquinas, esse contato direto inexistente: o que a IA recebe não são impressões do mundo, mas dados previamente coletados, filtrados, recortados e organizados por seres humanos. Se a incerteza já se faz presente quando tratamos de conhecimento humano fundado exclusivamente na experiência, essa incerteza é potencializada quando se trata de uma “experiência” indireta, mediada por bancos de dados. Os vieses, lacunas e contradições presentes na experiência humana são, assim, carregados para dentro do modelo, que os reproduz sem dispor de um crivo próprio de avaliação.

No âmbito do racionalismo, independentemente da variante considerada, as teses se amparam em conceitos metafísicos como razão, alma, Deus, mônada ou mundo das ideias. Atribuir tais conceitos a uma máquina inanimada é algo impensável: se fosse possível, teríamos de admitir que qualquer objeto – inclusive

.....

.....

uma pedra – poderia ser portador de alma ou de alguma espécie de consciência transcendental. Pelo mesmo motivo, as teses consensualistas, que também não abrem mão de um componente metafísico na construção de suas teorias, não se sustentam como fundamento da “inteligência” de uma máquina.

Na tentativa de aproximar as teorias de conhecimento às Inteligências Artificiais, infere-se que as três vertentes epistemológicas consideradas – empirismo, racionalismo e consensualismo – não conseguem determinar que haja conhecimento envolvido durante a ponderação em busca de uma resposta por parte da IA. No máximo, é possível admitir que a ideia de repetição defendida pelo empirismo humiano se assemelharia aos LLMs, e ainda assim “com muitas aspas”: há, tanto no humano quanto na máquina, um mecanismo que opera por repetição, formando um “hábito” de esperar certos resultados diante de determinados estímulos. Entretanto, no ser humano, esse hábito está ligado à vivência concreta, a uma história de impressões e à possibilidade de reflexão crítica; na máquina, ele se reduz a um ajuste de parâmetros numéricos em função de dados de entrada, sem consciência, intenção ou compreensão.

Dado que as IAs são desprovidas de conhecimento no sentido forte proposto pelas teorias do conhecimento e operam tão somente como propagadoras de opiniões (*doxa*), padrões e correlações ditados por algoritmos, torna-se ainda mais evidente que o conhecimento permanece restrito aos seres humanos. São os sujeitos que selecionam os dados, definem os objetivos do treinamento, interpretam os resultados, corrigem vieses e atribuem significado às saídas produzidas pelos modelos.

As Inteligências Artificiais devem ser compreendidas como ferramentas poderosas de tratamento, organização e disseminação de informação, capazes de potencializar tanto processos de investigação quanto de desinformação. A

.....

.....

responsabilidade ética recai, portanto, sobre os humanos que as projetam, treinam e utilizam. Sob a luz do racionalismo, do empirismo e do consensualismo, a conclusão que se impõe é que não há, nas máquinas, inteligência ou conhecimento em sentido pleno, mas sim uma sofisticada simulação de certas capacidades cognitivas humanas. Reconhecer esse limite é condição necessária para que seu uso seja crítico, responsável e compatível com um projeto de sociedade que não abdique da centralidade do sujeito cognoscente.

## **Referências**

ALZUBI, J; NAYYAR, A; KUMAR, A. Machine learning from theory to algorithms: an overview. *Journal of Physics: Conference Series*, Bristol, v. 1142, art. 012012, 2018. DOI: 10.1088/1742-6596/1142/1/012012.

AYODELE, T. Machine learning overview. In: ZHANG, Yagang (ed.). *New advances in machine learning*. Rijeka: InTechOpen, 2010. DOI: 10.5772/9374.

BROWN, T. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems (NeurIPS)*, [S. l.], v. 33, p. 1877-1901, 2020.

CRUZ-AGUILAR, M. A. The epistemic revolution of AI: reconfiguring the foundations of scientific knowledge. *AI & SOCIETY*, v. 41, p. 2041-2057, 2026.

GOODFELLOW, I; BENGIO, Y; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Acesso em: 24 abr. 2026.

HESSEN, J. *Teoria do conhecimento*. Tradução João Vergílio Gallerani Cuter. Revisão técnica Sérgio Sérvulo da Cunha. São Paulo: Martins Fontes, 2003.

HUME, D. *Uma investigação sobre o entendimento humano*. Tradução de José Oscar de Almeida Marques. São Paulo: Editora UNESP, 1999.

JOHNSON, G. M. Algorithmic bias: on the implicit biases of social technology. *Synthese*, v. 198, p. 9941-9961, 2021.

.....

**Firetti, M; Cunha, C. *A Inteligência Artificial sob o Crivo das Teorias do Conhecimento*.**

.....

KANT, I. Prolegômenos a qualquer metafísica futura que possa apresentar-se como ciência. Tradução José Oscar de Almeida Marques. 2. ed. São Paulo: Estação Liberdade, 2022.

KRAFT, A.; SOULIER, E. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In: FAcCT '24, Rio de Janeiro, 2024.

LEIBNIZ, G. Novos ensaios sobre o entendimento humano: pelo autor do sistema da harmonia preestabelecida. Tradução Luiz João Baraúna. São Paulo: Nova Cultural, 1988.

LOCKE, J. Ensaio sobre o entendimento humano. Tradução, apresentação e notas Pedro Paulo Garrido Pimenta. Revisão técnica Bento Prado. São Paulo: Martins Fontes, 2012.

NAVEED, H; KHAN, A; QIU, S; SAQIB, M; ANWAR, S; USMAN, M; AKHTAR, N; BARNES, N; MIAN, A. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, New York, v. 16, n. 5, p. 1-72, 2025. DOI: 10.1145/3744746.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, [S. l.], v. 30, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 24 abr. 2026.

.....