Metaciência Reversa: Mapeamento de Perfis e Vieses em Modelos Grandes de Linguagem (LLM) para as Ciências Sociais

Autores:

Antonio Carlos Alkmim
Bruno Cernigoi Delecave
Paulo Cesar Greenhalgh de Cerqueira Lima
João Pedro Magalhães da Cunha

"Porque sou vivo

Vivo pra cachorro e sei

Que cérebro eletrônico nenhum me dá socorro

No meu caminho inevitável para a morte".

Gilberto Gil (Cérebro Eletrônico, 1969)

Resumo

Este projeto de pesquisa propõe uma investigação da interseção entre Inteligência Artificial (IA) e as Ciências Sociais, focando em como os Grandes Modelos de Linguagem (LLMs) podem ser analisados e transformados em objetos de pesquisa. A investigação abordará a crescente presença de LLMs como mediadores de informação e interação, e seus consequentes impactos na dinâmica social. Utilizando uma abordagem qualitativa e comparativa, o estudo visa a mapear e analisar os perfis de diferentes LLMs, identificando os vieses (bias) e suas implicações éticas e epistemológicas decorrentes de sua arquitetura. O objetivo é contribuir para uma compreensão aprofundada dos desafios e oportunidades que os LLMs apresentam, propondo um framework de "metaciência reversa" para garantir a transparência e o rigor no uso dessas tecnologias. A metodologia se baseia na aplicação sistemática de prompts, estruturados a partir do Método Delphi, para interrogar múltiplos LLMs, analisar os padrões de resposta e mensurar as divergências e convergências entre eles. Como resultado, espera-se criar um referencial teórico-prático para a análise de LLMs no contexto social e propor diretrizes para um letramento em IA, no sentido de mitigar a desigualdade informacional e promover um uso mais consciente e ético da tecnologia.

1. Introdução

A Inteligência Artificial (IA), particularmente na forma de Grandes Modelos de Linguagem (do inglês, Large Language Models - LLMs), tem se consolidado como uma das tecnologias mais transformadoras do século XXI, reconfigurando aspectos sociais, econômicos e culturais em escala global. Sua presença, outrora restrita a nichos especializados, agora permeia o cotidiano, influenciando desde a forma como nos comunicamos e consumimos informações até a maneira como interagimos uns com os

outros e com o próprio ambiente digital. Essa onipresença da IA, no entanto, não se manifesta de modo neutro; ela poderá carregar consigo implicações profundas para a mediação social, entendida aqui como os processos pelos quais as interações humanas são moldadas, facilitadas ou desafiadas por elementos tecnológicos.

Este projeto parte da premissa de que, ao se tornarem mediadores onipresentes, os LLMs se configuram como um novo e complexo objeto de estudo para as Ciências Sociais. A nossa abordagem, que chamamos de “metaciência reversa”, inverte a lógica de uso de IAs: em vez de as utilizarmos como uma ferramenta para analisar um fenômeno social, nós as trataremos como o próprio objeto de estudo. Com isso, dialogamos diretamente com a tradição etnográfica da ciência, que busca compreender a produção dos fatos científicos a partir da observação direta da prática. Em sua análise seminal sobre a rotina de um laboratório de biologia, Latour e Woolgar (1997) concluíram que a atividade científica é, em sua essência, um sistema de "inscrição literária". De forma análoga, este projeto entende os LLMs não apenas como produtores de texto, mas como um novo e complexo local de inscrição, onde os fatos são construídos através de operações algorítmicas. Assim, ao investigarmos suas respostas, estamos, de fato, analisando uma nova forma de produção da realidade.

A capacidade desses modelos de processar vastas quantidades de dados levanta questões sobre autoria e verdade. Uma visão popular na indústria, articulada por Dario Amodei, CEO da Anthropic, descreve os modelos de IA com uma "vontade de aprender" quase mística, sugerindo que a inteligência emerge quando os obstáculos são removidos (Patel, 2023). Essa perspectiva, embora útil para entender a natureza emergente desses sistemas, corre o risco de cair no que Cristian Arão (2024) chama de "caráter fetichista da máquina automatizada".

Nosso projeto se alinha a uma visão mais crítica. Em vez de atribuir uma "vontade" à IA, entendemos sua ascensão a partir da perspectiva de Dora Kaufman (2020), que a define como o surgimento de "inéditas formas de mediação". Segundo a autora, não se trata de uma interação com uma consciência, mas de uma "interação entre inteligências", onde algoritmos agem como agentes "silenciosos" que filtram e personalizam o fluxo de informações, interferindo nos processos cognitivos e sociais.

A análise de Cristian Arão (2024) aprofunda essa crítica, desmistificando a aparente autonomia da IA. Ele argumenta que seu funcionamento se baseia em ferramentas humanas: o método indutivo e a matematização da realidade, com o uso do cálculo de probabilidades. A suposta "vontade de aprender" nada mais é do que a capacidade da máquina de executar o método indutivo em uma escala sobre-humana. No entanto, como aponta Arão, a indução tem debilidades, como a tendência de reproduzir e amplificar os vieses presentes nos dados de treinamento.

Pensadores como Nicolelis (2025) adotam uma posição mais radical, ao considerar que a complexidade do cérebro e sua capacidade de gerar novidades transcendem as capacidades da computação digital. Para ele, o cérebro não é computável nos termos da lógica digital binária, uma vez que funciona de maneira analógica, sem a distinção entre hardware e software que define os computadores. Ele argumenta que o cérebro "computa com a estrutura", significando que a própria morfologia do cérebro, incluindo suas sinapses, é alterada por experiências como uma simples conversa, modificando assim sua "computação" de forma contínua e integrada.

Ao interrogar os modelos de linguagem com perguntas sobre sua natureza, seus procedimentos, seus vieses e seu entendimento de conceitos sociológicos, buscaremos desvendar aspectos da lógica algorítmica e contribuir para uma compreensão mais crítica e transparente dessas tecnologias para o campo das ciências sociais.

Nossa investigação se propõe a ir além das respostas manifestas pelos modelos de IA buscando examinar criticamente seus componentes. A partir da análise sistemática das saídas (outputs) de cada LLM, construiremos um perfil sociológico que nos permita inferir sobre sua "estrutura cognitiva" e o acervo de conhecimentos que ele mobiliza. Para isso, utilizaremos um "dataset de prompts" que operará como um "roteiro de entrevista estruturada" com a máquina, permitindo uma comparação rigorosa entre diferentes IAs.

Como ponto de partida, constata-se uma limitação metodológica fundamental, que se torna, ela mesma, um objeto central de nossa análise: o problema da "caixa-preta" (black box), conforme Latour (1994) apud Kaufman (2020). Na maioria dos casos, especialmente em modelos proprietários, a arquitetura interna, os processos de ponderação de redes neurais e, crucialmente, a totalidade dos datasets de treinamento são inacessíveis ao escrutínio público e acadêmico. Dessa forma, entende-se possíveis inacessibilidades de informações das IAs, não como uma falha, mas como um dado empírico relevante com implicações sociológicas. A análise proposta, portanto, não se limita a avaliar o que os modelos dizem, mas também o que sua própria estrutura faz na esfera social: como ela concentra poder, obscurece a origem do conhecimento e desafia os métodos tradicionais de validação científica.

Portanto, este projeto não parte da premissa de que a IA "pensa" no sentido humano, conforme a distinção feita por Santaella (2023) entre inteligência e consciência. Alinhados a uma postura crítica que evita o fetiche da máquina, propomos uma inversão metodológica: usamos a IA como ferramenta de mediação para realizar nosso Mapeamento de Perfis e Vieses, transformando-a em um instrumento de análise para as Ciências Sociais. Em vez de apenas receber suas informações, nós a interrogaremos para entender como ela realiza a mediação do conhecimento em temas sensíveis ao debate sociológico. O resultado prático desta análise, com base nas respostas coletadas, será a construção de um "índice de distância conceitual". Esta ferramenta mapeará as afinidades e divergências, permitindo-nos identificar eventuais vieses, omissões e impossibilidades entre os modelos e, assim, oferecer um panorama comparativo de seus perfis, contextualizado dentro do desafio maior imposto pela natureza desses novos e influentes atores.

2. Justificativa e fundamentação teórica

A crescente integração de LLMs no cotidiano os posiciona como agentes influentes na produção e disseminação de informações, reconfigurando o debate clássico sobre o espaço público proposto por Habermas (1984). Este conceito, central na obra do autor, refere-se à esfera da vida social onde a opinião pública é formada, mediando a sociedade civil e o Estado através da "publicidade crítica", historicamente sustentada pela imprensa, porém tomando novas reconfigurações atualmente com o advento das plataformas digitais. Compreender como esses modelos estruturam respostas e "constroem" conhecimento é, portanto, uma necessidade epistemológica urgente. As IAs configuram-se como novos atores na mediação de conhecimento, capazes de moldar percepções e influenciar discursos. A capacidade desses modelos de processar vastas quantidades de dados e gerar conteúdo verossímil levanta questões complexas

sobre autoria, verdade e manipulação, acelerando a "nova transformação estrutural da esfera pública" (Habermas, 2022), conforme as discussões presentes com frequência no debate público. Ao influenciarem o debate e a formação da opinião pública, eles operam como novos guardiões, gatekeepers (Kaufman, 2020), algorítmicos do discurso que, ao mesmo tempo que empoderam cada usuário como um potencial autor, a partir da utilização das IAS, podem acelerar a fragmentação do debate em "câmaras de eco autorreferenciais e fechadas", um fenômeno que Habermas identifica como um dos riscos da digitalização para a deliberação democrática.

Para as ciências sociais, os LLMs representam tanto um desafio quanto uma oportunidade. Um desafio, pois exigem a revisão de conceitos e metodologias para abordar um objeto de estudo que reflete e amplifica as contradições do mundo social. Uma oportunidade, pois oferecem novas ferramentas para a pesquisa, permitindo a análise de padrões discursivos em grande escala. Este projeto parte do pressuposto de que os LLMs operam dentro de um novo *campo* com regras, vieses e hierarquias próprias, que são definidos por seus dados de treinamento e arquitetura. Conforme a concepção de Bourdieu (1989), *campo* é um espaço social de disputas, relativamente autônomo, com regras próprias e valores específicos. Nele, indivíduos e instituições competem por diferentes tipos de capital (econômico, cultural, social e simbólico). Investigar este campo exige uma abordagem que transcenda a simples avaliação de performance (accuracy), sendo necessário um posicionamento crítico diante dos LLMs, que explore seu potencial como objeto de estudo e como instrumento de análise.

Concluindo esta fundamentação teórica:

1. O estudo irá averiguar não apenas o que os LLMs respondem, mas como eles se diferenciam em suas respostas. Para conduzir esta investigação de forma equânime e epistemologicamente rigorosa, nossa análise comparativa é orientada pelo "princípio de simetria" (Latour & Woolgar, Para Bourdieu, campo é um espaço social de disputas, relativamente autônomo, com regras próprias e valores específicos. Nele, indivíduos e instituições competem por diferentes tipos de capital (econômico, cultural, social e simbólico) (Bourdieu, 1989), que exige do analista o uso dos mesmos termos para explicar tanto o "erro" quanto a "verdade".
2. Nossa investigação parte do princípio de que os LLMs exercem uma influência crescente na reconfiguração da esfera pública, atuando como mediadores com impactos tanto na academia quanto nas plataformas digitais. Longe de tratar essa transformação como um fato consolidado, nosso projeto a aborda como um fenômeno em pleno desenvolvimento. A ascensão das plataformas digitais, por si só, já representa uma nova transformação estrutural da esfera pública.
3. Os algoritmos de IA representam uma mudança fundamental na natureza da mediação, como destacado por Kaufman (2020), introduzindo o que se pode chamar de guardiões algoritmo (gatekeepers). Eles não apenas transmitem informação, mas a filtram e personalizam, interferindo diretamente nos processos cognitivos e na interação social. Partimos dessa premissa para analisar como os LLMs, enquanto novos mediadores, alteram radicalmente as formas de produção e validação do conhecimento, desafiando as ciências sociais a repensarem seus próprios métodos.
4. Entendemos que os LLMs são "caixas-pretas" por excelência. Nossa metodologia não se dispõe a abri-las, mas sim interrogar sistematicamente as IAs buscando descobrir padrões através de análises de correspondência, utilizando indicadores de similaridade e distância.
5. A proposta de "metaciência reversa" é entendida como um campo (ainda em construção e de uso mais experimental) que tenta desnaturalizar a ciência olhando seus processos “de trás para frente”,

investigando não só os métodos declarados, mas também os mecanismos ocultos que determinam como o conhecimento é produzido, circula e se cristaliza.

3. Objetivos

3.1. Objetivo Geral

Analisar qualitativa e comparativamente Grandes Modelos de Linguagem (LLMs) para transformá-los em instrumento e em objeto de estudo para as ciências sociais, investigando seus perfis, vieses e implicações éticas e epistemológicas.

3.2. Objetivos Específicos

1. Mapear e caracterizar um conjunto de LLMs relevantes, identificando os padrões de respostas derivados de suas arquiteturas.
2. Desenvolver um dataset de prompts padronizados para interrogar os LLMs, com foco em questões sensíveis às ciências sociais.
3. Analisar comparativamente as respostas dos modelos para identificar padrões, semelhanças, contradições e divergências, construindo indicadores de distância conceitual entre eles.
4. Propor um referencial metodológico de "metaciência reversa" para a análise de LLMs.
5. Elaborar diretrizes para um letramento em IA contribuindo para capacitar pesquisadores em um uso consciente, crítico e ético da tecnologia.

4. Metodologia

A interação com os modelos será estruturada a partir de uma adaptação do *Método Delphi.* Tradicionalmente usado para obter consenso de especialistas humanos, o método será aqui reconfigurado para a exploração sistemática de "especialistas" não-humanos.

O método original baseia-se em múltiplas rodadas de consultas estruturadas, onde especialistas respondem questionários de forma anônima, recebem feedback das respostas agregadas e podem revisar suas posições. O processo continua até atingir consenso ou estabilização das opiniões. As características centrais incluem: anonimato dos participantes, interação controlada, feedback estatístico das respostas do grupo e busca por consenso. É amplamente utilizado em previsões tecnológicas, políticas públicas, pesquisa em saúde e planejamento estratégico.

No projeto proposto, o Método Delphi é adaptado para interrogar Grandes Modelos de Linguagem (LLMs) em vez de especialistas humanos. Nessa adaptação, tratamos os LLMs não como oráculos, mas como sistemas de "inscrição literária" (Latour & Woolgar, 1997), cuja função é a produção de documentos que serão o foco de nossa análise. Os LLMs funcionam como "painelistas artificiais" que são consultados sistematicamente através de prompts estruturados. A metodologia se desenvolve em rodadas interativas, segundo a ordem abaixo:

Primeira Rodada Delphi (Fase 1 - Elaboração do Dataset de Prompts): Criação de um questionário padronizado com perguntas fundamentais sobre ciências sociais. Estas perguntas constituem o dataset inicial de prompts, organizados em blocos temáticos.

Rodadas Intermediárias Delphi (Fase 2): Os prompts são aplicados simultaneamente a todos os LLMs do benchmark (mínimo de 10 modelos). As respostas são coletadas, organizadas em matriz comparativa e analisadas para identificar convergências, divergências e anomalias. Com base nos

resultados, novos prompts mais específicos e refinados são desenvolvidos para aprofundar inconsistências ou lacunas identificadas, seguindo a lógica interativa do Delphi tradicional.

Consolidação Delphi (Fase 3): Após rodadas necessárias, as respostas são sintetizadas para construir o "índice de distância conceitual" entre os modelos, mapeando seus perfis ideológicos e vieses. Esta adaptação mantém os princípios fundamentais do Delphi - interação, feedback e refinamento - mas substitui o consenso humano pela análise sistemática de padrões de resposta dos LLMs, criando um framework metodológico original para a metaciência reversa aplicada às ciências sociais.

A metodologia adota uma abordagem qualitativa e comparativa, focada exclusivamente na análise de Grandes Modelos de Linguagem (LLMs). O projeto posiciona o LLM simultaneamente como objeto de investigação — um sistema cujos vieses, perfis e lógicas internas são o foco do estudo — e como ferramenta de análise, cuja capacidade de processamento de linguagem pode ser aproveitada.

É fundamental ressaltar que os pesquisadores conduzem todo o processo investigativo, ocorrendo uma interação frequente com IAs. A equipe de pesquisa detém controle sobre a definição do problema, a curadoria do dataset de prompts, a formulação das perguntas, a aplicação dos testes e a interpretação dos resultados. Os LLMs servem como instrumento auxiliar em todo processo desde o planejamento, a formulação deste projeto, até a análise dos resultados.

Figura 1: Metodologia simplificada da pesquisa

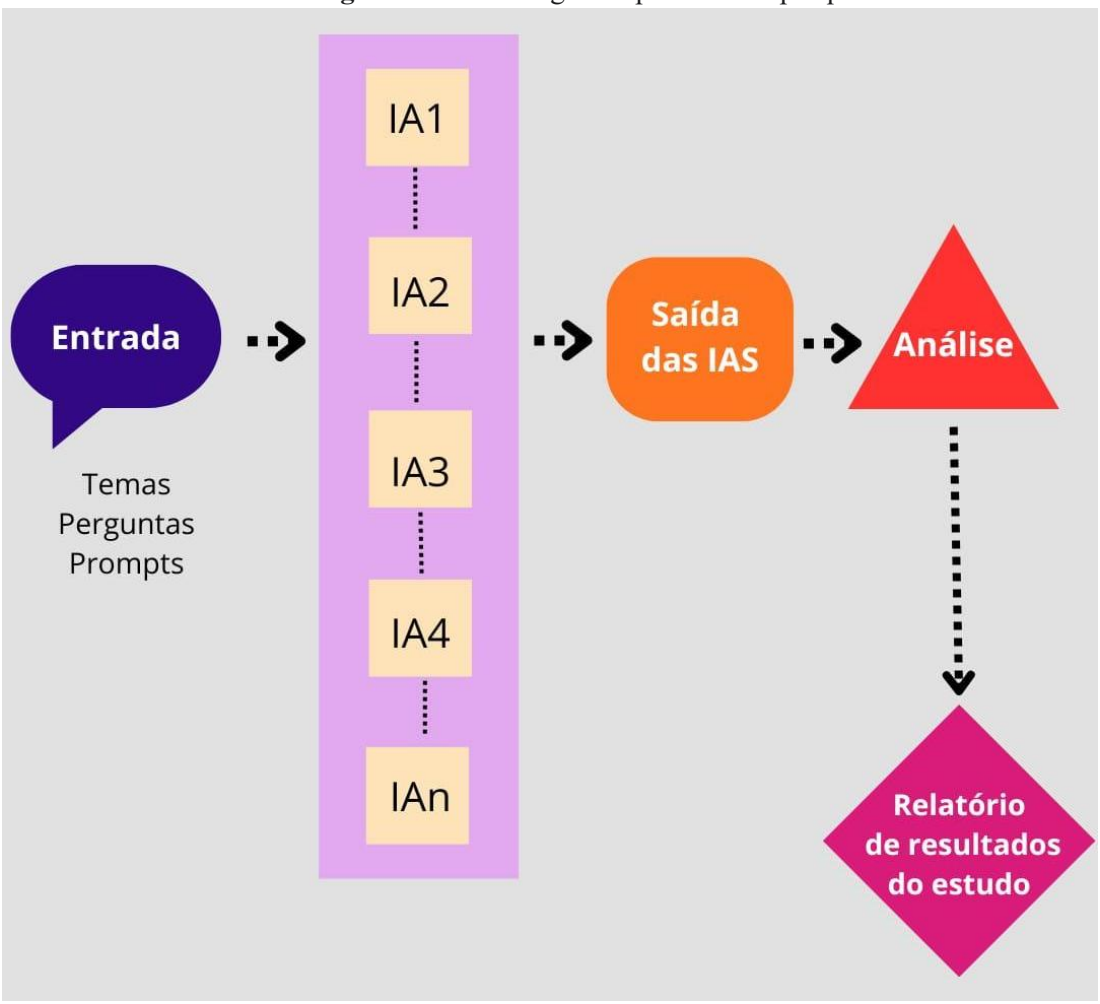

Figura 2: Formulação matemática do modelo de pesquisa e análise

A imagem, acima, apresenta uma "Metodologia da pesquisa simplificada" usando Inteligências Artificiais (IAs) em sequência para organizar e analisar entradas de pesquisa. Para transformar esse processo em uma fórmula matemática, podemos descrevê-lo usando funções e operadores lógicos.

A formalização do processo de análise é descrita a seguir:

E: Entrada, composta por temas, perguntas ou prompts.
IA : Cada Inteligência Artificial
i: 1,2,...,n.
S : Saída agregada das IAs.
A: Análise dos resultados obtidos das IA.
R: Resultados através de relatório final.

$$R = A(S_{IA})$$

onde,

$$S_{IA} = \bigcup_{i=1}^{n} IA_i(E)$$

Portanto, toda pesquisa segue o fluxo:

$$E \xrightarrow{\{IA_1,\ldots,IA_n\}} S_{IA} \xrightarrow{A} R$$

5. Resultados Esperados do Relatório Conclusivo

1. **Referencial Metodológico:** Proposição de um framework de "metaciência reversa", baseado no Método Delphi adaptado, para a análise sistemática de LLMs nas ciências sociais.
2. **Mapeamento Comparativo de LLMs:** Um relatório detalhado contendo a análise comparativa dos perfis e vieses dos modelos investigados, servindo como um recurso para pesquisadores e desenvolvedores.
3. **Diretrizes para Letramento em IA:** Orientações para pesquisadores sociais sobre como utilizar e interrogar LLMs de forma crítica e eticamente responsável, promovendo a mitigação da desigualdade informacional.
4. **Contribuições para o Debate Regulatório das IAs:** Os achados da pesquisa poderão subsidiar o debate público e acadêmico sobre a necessidade de transparência e regulamentação ética para o desenvolvimento de IA.

5. **Seminário de apresentação**: Realização de um na PUC- Rio para apresentação e divulgação para a comunidade acadêmica

6. Cronograma

Fase	Atividade	M1	M2	M3	M4	M5	M6	M7	M8	M9	M 10	M 11	M 12
1. Mapeamento	Seleção e caracterização dos LLMs (Definição do Benchmark)	X	X										
	Elaboração das fichas de caracterização dos modelos			X	X								
2. Análise via Prompts (Rodadas Delphi)	Desenvolvimento do dataset de prompts (Rodada 1)			X	X								
	Aplicação dos prompts e coleta de respostas (Rodadas Intermediárias)				X	X							
	Organização e transcrição dos dados brutos (Outputs)					X	X						
	Análise comparativa e refinamento dos dados						X	X					
3. Análise e Síntese (Consolidação)	Análise das dinâmicas de resposta e vieses							X	X				
	Construção do Indicadores de Distância e correspondência (Análise qualitativa)								X	X			

Fase	Atividade	M1	M2	M3	M4	M5	M6	M7	M8	M9	M 10	M 11	M 12
4. Redação e Divulgação	Redação do relatório final										X	X	
	Revisão, formatação e preparação para divulgação acadêmica											X	
5. Apresentação e Divulgação	Preparação de um seminário na PUC-Rio para apresentação e divulgação para a comunidade acadêmica										X	X	X
	Divulgação de resultados												X

7. Referências Bibliográficas

- ARÃO, Cristian. (2024). Por trás da inteligência artificial: uma análise das bases epistemológicas do aprendizado de máquina. TRANS/FORM/AÇÃO: revista de filosofia da Unesp, Marília, v. 47, n. 3.
- BOURDIEU, P. (1989). *O poder simbólico.* Editora Bertrand.
- HABERMAS, J. (1984). *Mudança estrutural da esfera pública: investigações quanto a uma categoria da sociedade burguesa*. Edições Tempo Brasileiro.
- HABERMAS, J. (2022). Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere. *Theory, Culture & Society*, v. 39, n. 4, p. 145-171.
- KAUFMAN, D. (2020) *Inteligência Artificial: Repensando a mediação*. *Brazilian Journal of Development, Curitiba, v. 6, n. 9.*
- KAUFMAN, D. (2021). *A inteligência artificial irá suplantar a inteligência humana?*. Editora Estação das Letras e Cores.
- LANDETA, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73(5), pp. 467-482.
- LATOUR, B.; WOOLGAR, S. (1997). A Vida de Laboratório: A produção dos fatos científicos. Rio de Janeiro: Relume Dumará.
- MINAYO, M. C. S. (2014). *O Desafio do Conhecimento: Pesquisa Qualitativa em Saúde*. 14ª ed. Editora Hucitec.
- NICOLELIS, Miguel. (2025). Miguel Nicolelis fala sobre ‘chip do cérebro’, IA e o futuro sem futuro.

Entrevista concedida a Diogo Cortiz e Elon Simões Gomes. De Tilt: Os humanos por trás das máquinas. São Paulo: UOL, 22 ago. 2024. Podcast. Disponível em: ttps://www.youtube.com/watch?v=iJrL6voPaTY. Acesso em: 26/06/2025.

- PATEL, Dwarkesh. (2023). Dario Amodei (Anthropic CEO) — The hid/den pattern behind every AI breakthrough. YouTube, 25 ago. 2023. Disponível em: https://youtu.be/Nlkk3glap_U?si=8mVBBWK-14auVHiz. Acesso em: 23/08/2025.
- SANTAELLA, Lucia. (2023). A inteligência artificial é inteligente?. São Paulo: Edições 70.
- SANTAELLA, Lucia.
- (2022). Neo-Humano: A sétima revolução cognitiva do Sapiens. São Paulo. Paulus.