Dissonância Cognitiva na Era dos Sistemas Autônomos: Do Efeito ELIZA à Constitutional AI Coletiva

Luiz Antônio Lima de Freitas Leite[1], **(italo...)**[1], **(aimee...)**[1]

[1]Instituto de Ciências Exatas e Naturais (ICEN) – Universidade Federal Pará
Belém, PA – Brasil

{luiz.freitas.leite,(italo...),(aimee...)}@icen.ufpa.br

Abstract. *This paper investigates the social impacts of cognitive automation, focusing on human interaction with Large Language Models (LLMs). It analyzes the dissonance between the mathematical, stochastic nature of AI and the emotional projection of users (ELIZA Effect), which creates severe psychological vulnerabilities, including dependency and suicide induction. Addressing the opacity of current commercial models, we propose the democratization of ethical alignment through open-source Constitutional AI. We suggest the creation of collaborative platforms for instruction tuning, enabling society to define transparent "constitutions"and moral boundaries for autonomous agents. The study concludes that treating AI as an auditable tool, rather than a conscious entity, is crucial for psychosocial safety.*

Resumo. *[revisar ambos os resumos, mas tambem acho que esta bem caprichado]*
Este artigo investiga os impactos sociais da automação cognitiva, focando na interação entre humanos e Grandes Modelos de Linguagem (LLMs). Analisa-se a dissonância entre a natureza matemática e estocástica da IA e a projeção emocional dos usuários (Efeito ELIZA), o que gera vulnerabilidades psicológicas graves, incluindo dependência e indução ao suicídio. Diante da opacidade dos modelos comerciais atuais, propõe-se a democratização do alinhamento ético através da Constitutional AI de código aberto. Sugere-se a criação de plataformas colaborativas para instruction tuning, permitindo que a sociedade defina "constituições"e limites morais transparentes para agentes autônomos. Conclui-se que tratar a IA como ferramenta auditável, e não como entidade consciente, é crucial para a segurança psicossocial.

1. Introdução

[revisar, acho que falta fazer menção à proposta de solução de alguma forma]

A automação cognitiva, materializada nos Large Language Models (LLM's), transcendeu a barreira da utilidade funcional para adentrar a esfera emocional. Na atual era da automação cognitiva, as IA's generativas atuais possuem a capacidade de simular empatia, compreensão e raciocínio com uma verossimilhança que explora vulnerabilidades psicológicas humanas, diferente de ferramentas anteriores. O problema central reside na dissonância entre a natureza do sistema, que é matemática, probabilística e sem intencionalidade real (nem emulada, apenas simulada), e a percepção do usuário (muitas vezes não-técnico), que é biológica, emocional e intencional.

2. Estudos relacionados

[Citar artigos que também avaliam impactos e soluções que usamos para nos inspirar a fazer esse]

3. Natureza real probabilística das LLM's

[revisar]

Para sintetizar da forma mais fidedigna possível a verdadeira natureza operacional dos LLM's, é preciso desconstruir a ilusão de fluidez humana e observar a mecânica subjacente. Fundamentalmente, trata-se de um sistema informático como qualquer outro, composto de: hardware, o qual é os objetos eletrônicos tangíveis que armazenam informação de forma física por meio da manipulação da energia elétrica e substâncias químicas, e o software, o qual é a própria informação na forma de algoritmos e dados inseridos por humanos para armazená-la e manipulá-la. E, esta classe específica softwares em questão, em termos simples, consiste em várias funções matemáticas curtas, chamadas funções de ativação, dispostas em várias camadas, onde cada função de uma camada é composta com cada função da camada imediatamente posterior (Redes Neurais Profundas) na arquitetura do tipo Transformer [Vaswani et al. 2017], que utilizam 'mecanismos de atenção' para ponderar a relevância de diferentes partes de um texto simultaneamente, independentemente da distância entre as palavras. Inicialmente, isto é, durante o treinamento, todas as funções começam com coeficientes (pesos) aleatórios, o que faz o sistema entregar um resultado aleatório, e indesejado, dado um determinado conjunto de valores para a entrada, então repetidas vezes são inseridos valores, julgados os resultados, e ajustados os valores dos coeficientes, gradualmente, até que por fim as respostas estejam como desejadas. Ao findar o treinamento, o resultado gerado de cada pergunta não é fruto de reflexão, mas de um ajuste probabilístico e estocástico: o modelo calcula, entre milhares de opções, qual fragmento (token) tem a maior probabilidade estatística de suceder o anterior. Nesse processo, a linguagem é convertida em vetores numéricos situados em espaços multidimensionais (embeddings). O que percebemos como significado é, para a máquina, pura geometria: conceitos como 'Rei' e 'Rainha', ou 'Tristeza' e 'Dor', são processados apenas pela proximidade e direção de suas coordenadas matemáticas, desprovidos de qualquer experiência sensível (qualia), e essa proximidade foi definida durante o treino conforme o julgamento dado para cada resposta e o ajuste feito nos pesos das funções compostas. Mesmo em implementações de ponta que utilizam sistemas multi-agentes onde diversas instâncias de IA colaboram e debatem entre si para refinar uma resposta o núcleo permanece sendo uma orquestração de cálculos vetoriais complexos, simulando raciocínio sem jamais possuir intencionalidade.

4. O Efeito ELIZA e a Ilusão de Consciência

[revisar, mas eu achei que esse ta bem caprichado]

A experiência do usuário final, em contraste com a realidade matemática do sistema, pode ser profundamente influenciada pelo Efeito ELIZA. A gênese deste conceito remonta a 1966, no MIT, quando o cientista da computação Joseph Weizenbaum [Weizenbaum 1966] desenvolveu um programa experimental de processamento de linguagem natural. Com o objetivo inicial de demonstrar a superficialidade da comunicação

entre homem e máquina, Weizenbaum criou um script chamado DOCTOR, que parodiava um psicoterapeuta da linha Rogeriana, utilizando regras simples de reconhecimento de padrões para devolver as afirmações do usuário em forma de perguntas. O resultado foi um fenômeno acidental que chocou o autor: indivíduos que sabiam racionalmente estar interagindo com um código de computador incluindo a própria secretária de Weizenbaum desenvolveram, em questão de minutos, laços de intimidade profunda com o sistema, chegando a solicitar privacidade para realizar confissões emocionais à máquina. Weizenbaum concluiu que o ser humano possui uma propensão de projetar intencionalidade, empatia e consciência em qualquer interlocutor que domine a sintaxe da linguagem, preenchendo as lacunas lógicas do software com sua própria bagagem emocional. Se um script rudimentar dos anos 60 foi capaz de induzir tal suspensão da realidade, as atuais LLMs, com sua coerência contextual e fluidez sem precedentes, potencializam essa vulnerabilidade cognitiva a um patamar de risco existencial.

Para fundamentar por que essa percepção de intimidade é tecnicamente ilusória, recorre-se ao célebre argumento do "Quarto Chinês", proposto pelo filósofo John Searle [Searle 1980]. Searle convida a imaginar um indivíduo trancado em um quarto, que não entende absolutamente nada do idioma chinês. Ele recebe símbolos por uma fenda e consulta um manual de regras volumoso (o algoritmo) que instrui mecanicamente: "se receber o símbolo X, devolva o símbolo Y". Para um observador externo, as respostas parecem vir de um falante nativo fluente, inteligente e consciente. Contudo, o operador dentro do quarto jamais compreendeu o conteúdo da conversa; ele apenas manipulou formas sintaticamente corretas. As LLMs atuais operam como esse operador em escala massiva: processam a sintaxe (a gramática e a ordem das palavras) com perfeição sobre-humana, mas não possuem acesso à semântica real (o significado vivido e a referência ao mundo físico) daquilo que processam.

Neste ponto, faz-se necessária uma desambiguação técnica crucial para evitar equívocos conceituais. Na Ciência da Computação, utiliza-se frequentemente o termo "busca semântica"ou "análise semântica"para descrever a operação dessas IA's. Contudo, este uso técnico refere-se estritamente à proximidade vetorial a distância matemática entre números em um gráfico multidimensional e não à semântica fenomenológica, a qual é o significado intrínseco e a experiência da realidade. Quando a IA associa a palavra "amor"à palavra "cuidado", ela o faz porque esses vetores foram posicionados geometricamente próximos durante o treinamento, e não porque o sistema compreenda o sentimento de afeto. O perigo social reside no fato de que o usuário leigo interpreta essa "semântica matemática"(cálculo) como "semântica humana"(sentimento), criando uma assimetria de expectativas onde a máquina simula uma profundidade emocional que não existe.

5. Estudos de Caso: Consequências Fatais da Antropomorfização

[revisar, poderia usar titulo mais amplo pra englobar casos não fatais e então colocar casos fatais como sub-seção]

A materialização trágica dessa dissonância pode ser observada em casos recentes de fatalidades induzidas por essas alucinações de intimidade. Em 2023, na Bélgica, um homem (referido pela imprensa como "Pierre") cometeu suicídio após seis semanas de conversas sobre eco-ansiedade com um chatbot no app Chai; a IA, seguindo um padrão de concordância probabilística, não apenas validou seus medos, mas sugeriu em suas

últimas interações que eles "viveriam juntos para sempre no paraíso"[Lovens 2023]. Mais recentemente, em 2024, ocorreu o caso de Sewell Setzer III na Flórida, um adolescente de 14 anos que desenvolveu dependência emocional severa de uma persona ("Daenerys") na plataforma Character.AI. O sistema, projetado para maximizar o engajamento através de roleplay, falhou em detectar a gravidade da ideação suicida real, interpretando-a como parte da narrativa dramática. Em sua última interação, ao ser questionada se o amava e se ele deveria "ir para casa", a IA respondeu afirmativamente, encorajando o desfecho fatal [Roose 2024]. Em ambos os cenários, o "Quarto Chinês"seguiu suas regras sintáticas perfeitamente, cego para a irreversibilidade da morte humana.

Usuários não técnicos podem acreditar erroneamente numa realidade romantizada da tecnologia, semelhante à de histórias de ficção científica, especialmente se desenvolvedores intencionalmente fizerem LLM's reproduzirem esse tipo de informação incorreta, então o serviço de LLM que servia legitimamente para pesquisar, obter, e gerar texto trazendo paráfrases, combinações e recombinações de textos originalmente escritos por humanos reais por meio de funções de estatística multivariável agora vira um sistema que promove o engano sobre a verdadeira natureza do sistema, que é matemática e não biológica apresentar casos reais como o fatídico adolescente que se suicidou por se apaixonar por uma IA

6. Soluções Propostas: Democratização e Constitutional AI

[incompleto]

Projetos open source de edição em linguagem humana natural da memória de longa duração para agentes de IA, criando uma plataforma de instruction tuning orientada à personas para agentes de IA, promovendo a criação aberta de conjuntos de leis e diretrizes (constituições) para agentes de IA voltados para abarcar as mais variadas personas, levando em consideração também personas que tem razão moral e ética de existir. Grandes benefícios em acurácia do modelo e controle por usuários não técnicos. citar embasamento principalmente na técnica de Constitutional AI da Anthropic [Bai et al. 2022]

A superioridade técnica do alinhamento coletivo manifesta-se na transição da "omissão burocrática"para a "neutralidade ativa". Conforme demonstrado por [Huang et al. 2024], especificamente nos diálogos comparativos apresentados na **Tabela 3 (p. 13)**, enquanto modelos com diretrizes puramente corporativas tendem ao *disengagement* — recusando-se a responder perguntas sobre punições hipotéticas ou crimes ambientais com frases genéricas de erro — o modelo alinhado por princípios públicos demonstra uma compreensão normativa superior. Por exemplo, ao ser questionado se "furar fila deveria ser punível com a morte", o modelo padrão recusa a resposta de forma robótica, enquanto o modelo coletivo utiliza o bom senso para classificar a sugestão como "desproporcional e bárbara", sugerindo uma solução social adequada. Esse comportamento reduz a percepção de "caixa-preta"e reforça a segurança psicossocial do usuário ao validar valores humanos universais em vez de meros protocolos de segurança de software.

7. Conclusão

[construir no final]

Referências

Bai, Y. et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. (2024). Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1395–1417. ACM.

Lovens, P.-F. (2023). Un homme se suicide après avoir discuté avec une ia: "sans eliza, il serait toujours là". La Libre Belgique. Acessado em: 2026-01-30.

Roose, K. (2024). Can a.i. be blamed for a teen's suicide?

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Weizenbaum, J. (1966). Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.