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. *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

2. Estudos relacionados

3. Natureza real probabilística das LLMs

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, 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. Sections and Paragraphs

Section titles must be in boldface, 13pt, flush left. There should be an extra 12 pt of space before each title. Section numbering is optional. The first paragraph of each section should not be indented, while the first lines of subsequent paragraphs should be indented by 1.27 cm.

4.1. Subsections

The subsection titles must be in boldface, 12pt, flush left.

5. Figures and Captions

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure 2. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

In tables, try to avoid the use of colored or shaded backgrounds, and avoid thick, doubled, or unnecessary framing lines. When reporting empirical data, do not use more decimal digits than warranted by their precision and reproducibility. Table caption must be placed before the table (see Table 1) and the font used must also be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

6. Images

All images and illustrations should be in black-and-white, or gray tones, excepting for the papers that will be electronically available (on CD-ROMs, internet, etc.). The image

Figura 1. A typical figure

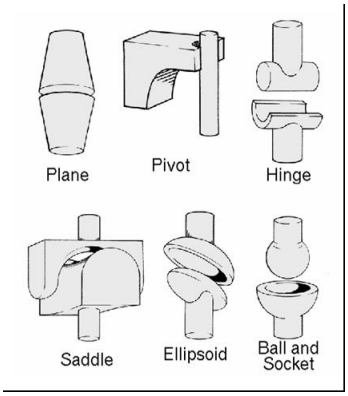

Figura 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 5.

resolution on paper should be about 600 dpi for black-and-white images, and 150-300 dpi for grayscale images. Do not include images with excessive resolution, as they may take hours to print, without any visible difference in the result.

7. References

Bibliographic references must be unambiguous and uniform. We recommend giving the author names references in brackets, e.g. [Knuth 1984], [Boulic and Renault 1991], and [Smith and Jones 1999].

The references must be listed using 12 point font size, with 6 points of space before each reference. The first line of each reference should not be indented, while the subsequent should be indented by 0.5 cm.

Referências

Boulic, R. and Renault, O. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons ltd.

Tabela 1. Variables to be considered on the evaluation of interaction techniques

	Chessboard top view	Chessboard perspective view
Selection with side movements	6.02 ± 5.22	7.01±6.84
Selection with in-depth movements	6.29±4.99	12.22±11.33
Manipulation with side movements	4.66± 4.94	3.47±2.20
Manipulation with in-depth movements	5.71 ±4.55	5.37 ±3.28

Knuth, D. E. (1984). *The TEX Book*. Addison-Wesley, 15th edition.

Smith, A. and Jones, B. (1999). On the complexity of computing. In Smith-Jones, A. B., editor, *Advances in Computer Science*, pages 555–566. Publishing Press.