Ameaças à Colaboração Digital: Uma Proposta de Estudo sobre Guerra Cognitiva e Violação da Autonomia em Sistemas Sociotécnicos

João Victor Almeida[1], **Claudia Beatriz Berti**[1], **Leonardo Lana de Carvalho**[1]

[1]Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina

{almeida.victor,claudiabberti,leonardolana.carvalho}@ufvjm.edu.br

Abstract. *Digital collaboration presupposes the autonomy of participants. However, the rise of Generative AI and micro-targeting has introduced Cognitive Warfare, transforming platforms into potential vectors of manipulation. This article presents the design of a study focused on how algorithmic personalization impacts autonomy and collective deliberation. The objective is to analyze the technical and ethical boundary between legitimate personalization and psychological manipulation. Based on an integrative literature review, it proposes the creation of a framework with design requirements aimed at contributing to transparent platforms that are resilient to invisible coercion and protective of epistemic rights.*

Resumo. *A colaboração digital pressupõe a autonomia dos participantes. Contudo, a ascensão da IA Generativa e do micro-targeting introduziu a Guerra Cognitiva, transformando plataformas em potenciais vetores de manipulação. Este artigo apresenta o desenho de uma pesquisa focada em como a personalização algorítmica impacta a autonomia e a deliberação coletiva. O objetivo é analisar a fronteira técnica e ética entre a personalização legítima e a manipulação psicológica. A partir de uma revisão integrativa da literatura, propõe-se a criação de um framework com requisitos de design, visando contribuir com plataformas transparentes, resilientes à coerção invisível e protetoras dos direitos epistêmicos.*

1. Introdução e Motivação

A temática do SBSC 2026, "Colaborar para Transformar", convida à reflexão sobre como as tecnologias podem moldar o tecido social. Contudo, para que sistemas funcionem como instrumentos de colaboração e confiança, é fundamental que os seus usuários possuam autonomia sobre suas decisões e interações. O cenário atual apresenta uma ameaça silenciosa a essa premissa: *Cognitive Warfare* (Guerra Cognitiva).

Diferente das *Fake News* (notícias falsas) tradicionais — que tendem a ser disseminadas de forma ampla e não necessariamente orientadas por perfis individuais — a literatura contemporânea sobre *Cognitive Warfare* descreve uma forma mais sofisticada de operações de influência. Utilizando técnicas de análise de dados e Inteligência Artificial para compreender características psicológicas e comportamentais de indivíduos e grupos e, a partir disso, entregar mensagens altamente personalizadas por meio de *micro-targeting* (microssegmentação) e conteúdo dirigido. Visando o objetivo de moldar

percepções, decisões e comportamentos de forma mais eficaz do que a mera propagação de informação falsa dispersa [Reczkowski and Adamski 2025].

Um exemplo paradigmático desta premissa é o caso Cambridge Analytica,[1] que serve como marco referencial para a compreensão das operações de influência modernas. No escândalo revelado em 2018, dados de milhões de perfis do Facebook foram coletados sem consentimento para a criação de modelos psicométricos detalhados [Zuboff 2019]. Através do mapeamento de traços de personalidade, como o modelo *Big Five*, que categoriza a personalidade humana em cinco espectros: Abertura a experiências, Conscienciosidade, Extroversão, Amabilidade e Neuroticismo [Kosinski et al. 2013], a empresa foi capaz de segmentar usuários e entregar conteúdos de *micro-targeting* projetados para explorar vulnerabilidades emocionais específicas [Wylie 2019]. Este episódio ilustra como a manipulação psicológica em escala industrial pode subverter processos deliberativos, transformando a infraestrutura digital de colaboração em um vetor de coerção invisível [Zuboff 2019, Wylie 2019].

Em Sistemas Colaborativos, onde a construção de conhecimento e a tomada de decisão conjunta dependem da interação autêntica e da confiança mútua entre os pares [Kraut and Resnick 2012], a inserção dessas táticas de manipulação da informação representa uma ameaça estrutural. Quando algoritmos preditivos exploram vulnerabilidades cognitivas individuais dentro de uma rede de forma industrial, atingindo rapidamente uma grande escala de indivíduos, eles têm o potencial de fabricar consensos artificiais, polarizar discussões e induzir comportamentos de maneira quase que invisível [Woolley and Howard 2018]. Dessa forma, a *Cognitive Warfare* não apenas viola a autonomia do usuário de forma isolada, mas corrompe a própria inteligência coletiva [Malone 2018], transformando plataformas originalmente projetadas para a cooperação e inovação social em instrumentos de coerção [Zuboff 2019].

A motivação deste trabalho reside na necessidade de compreender como e sob quais condições a inferência algorítmica afeta a capacidade de consentimento do usuário em ambientes colaborativos.

2. Problema de Pesquisa e Objetivos

A construção de confiança em comunidades digitais e a deliberação coletiva são processos importantes em Sistemas Colaborativos. Contudo, ecossistemas digitais baseados em coleta massiva de dados comportamentais possuem a capacidade de aprender e antecipar reações humanas. Há indícios teóricos de que o uso desses dados para *micro-targeting* pode comprometer a legitimidade do processo deliberativo, substituindo o consenso autêntico por alinhamentos fabricados algoritmamente [Woolley and Howard 2018].

O problema central desta pesquisa reside em compreender **como** a arquitetura de personalização em plataformas colaborativas pode ultrapassar a fronteira do serviço legítimo e adentrar o campo da manipulação subconsciente. É imperativo observar a complexidade deste fenômeno, considerando inclusive o "paradoxo da conveniência": cenários em que os usuários cedem voluntariamente parte de sua autonomia em troca de conforto cognitivo ou pertencimento em bolhas de filtro [Burr et al. 2018].

Diante disso, a pesquisa se baseia nas seguintes questões operacionais:

[1] `https://l1nq.com/ugDuh`

- **Q1:** Quais mecanismos algorítmicos e padrões de interface ameaçam a autonomia deliberativa em plataformas colaborativas?
- **Q2:** Como a literatura caracteriza a fronteira técnica e ética entre a personalização legítima de conteúdo e a manipulação coercitiva?
- **Q3:** Quais requisitos de design sociotécnico podem mitigar a coerção invisível, entregando maior transparência aos usuários?

O objetivo principal é mapear como a manipulação psicológica orientada a dados afeta a capacidade de consentimento e deliberação. Como contribuição, propõe-se um *framework* de requisitos de design para evidenciar como as táticas de persuasão violam os direitos epistêmicos dos usuários.

3. Contexto e Fundamentação Teórica

Para delimitar conceitos centrais, a autonomia cognitiva é compreendida como a capacidade do indivíduo de tomar decisões e formar crenças de forma independente, livre de coerções ocultas [Susser et al. 2019]. A *Cognitive Warfare* designa a instrumentalização da informação acoplada ao perfilamento psicológico em massa, visando explorar vulnerabilidades emocionais [Reczkowski and Adamski 2025].

Essas vulnerabilidades são frequentemente mapeadas por meio de inferência algorítmica usando modelos psicométricos, como o *Big Five*. A aplicação desses modelos em interfaces ocorre, muitas vezes, via *Dark Patterns* — padrões enganosos de design criados intencionalmente para induzir o usuário a ações que não tomariam em plena consciência [Gray et al. 2018]. O estudo ancora-se na defesa dos direitos epistêmicos, investigando como o design pode proteger o direito do indivíduo de saber como está sendo influenciado.

A hipótese levantada é que a manipulação psicológica industrializada pode criar uma assimetria de poder que compromete o consentimento informado do utilizador [Zuboff 2019, Susser et al. 2019]. A justificativa técnica para esse marco reside na disponibilização pública de *Large Language Models* (LLMs), como o ChatGPT, e na fundamentação técnica da literatura que comprova que a IA Generativa reduziu a quase zero o custo marginal de geração de desinformação hiper-personalizada [Goldstein et al. 2023]. Esse marco representa a transição da manipulação analógica para operações de influência automatizadas em larga escala, com possibilidades de modulação do tom e da valência emocional da mensagem em tempo real, baseando-se no mapeamento psicológico prévio.

4. Metodologia e Desenho da Pesquisa

Para analisar a hipótese apresentada, o desenho da pesquisa estabelece a realização de uma Revisão Integrativa da Literatura. Dada a natureza sociotécnica do problema, a revisão seguirá os preceitos do protocolo PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) [Page et al. 2021], garantindo rigor na seleção e transparência metodológica.

4.1. Bases de Dados e Protocolo de Busca

A pesquisa fará uso das seguintes bases de dados: DBLP e o Google Scholar.

A *string* de busca foi estruturada combinando três eixos de palavras-chave:
(i) Vetor Tecnológico: (*"generative ai" AND "algorithmic recommendation"*)
(ii) Fenômeno Analisado: (*"cognitive warfare" AND ("psychological manipulation" OR "dark patterns")*)
(iii) Impacto: (*"cognitive autonomy" OR "epistemic rights"*)

A escolha das bases de dados foi feita de acordo com a quantidade de artigos retornados pelas querys informadas. Também foram analisadas bases de dados como SOL, IEEE Xplore e Scopus para a construção da pesquisa, porém o número de artigos disponíveis retornados não foi satisfatório, e com isso foi tomada essa decisão.

4.2. Critérios de Triagem e Categorias Analíticas

Serão incluídos artigos revisados por pares, em inglês e português, com foco pós-2022. Obras anteriores ("Marco Zero") serão admitidas se caracterizarem literatura seminal (ex: caso Cambridge Analytica [Wylie 2019] e Capitalismo de Vigilância [Zuboff 2019]).

Na extração de dados, os artigos serão divididos em categorias analíticas:
(a) mecanismos de ataque/inferência psicométrica
(b) impactos na deliberação coletiva
(c) propostas de defesa/design ético

Após a extração, os artigos serão agrupados em categorias analíticas (ex.: mecanismos de captura de atenção, uso de LLMs para persuasão e estratégias de mitigação técnica).

5. Solução Proposta e Avaliação dos Resultados

Este estudo visa contribuir para a área de Sistemas Colaborativos investigando como a deliberação legítima e a confiança em comunidades digitais podem ser resguardadas contra arquiteturas coercitivas. A proposta central foca em entender como essas ameaças violam os direitos epistêmicos dos usuários na prática, implicando na formulação de um *framework* de diretrizes de *design*.

Para garantir o rigor, a validação do *framework* ocorrerá por meio de uma abordagem estruturada baseada em *Walkthrough Analítico* [Nielsen 1994]. O *framework* será aplicado retrospectivamente a casos de estudo conhecidos na literatura (como campanhas documentadas de manipulação eleitoral) para verificar se seus requisitos teriam sido capazes de mapear e mitigar os mecanismos de coerção invisível utilizados. Em iterações futuras, prevê-se a avaliação qualitativa por painéis de especialistas em IHC (Interação Humano-Computador) e Ética em IA.

6. Agradecimentos

Os autores agradecem todo o apoio dado por Karolina Azevedo ao longo dos estudos feitos para a criação deste artigo. E declara-se o uso do modelo de IA Generativa Gemini como ferramenta de assistência durante a elaboração deste trabalho. A ferramenta foi empregada como suporte para revisão, aprimoramento da fluidez textual, estruturação de tópicos metodológicos e tradução do resumo (*Abstract*) para a língua inglesa.

Referências

Burr, C., Cristianini, C., and Ladyman, J. W. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28(4):735–774.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. Technical report, Center for Security and Emerging Technology (CSET) and Stanford Internet Observatory.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L. (2018). The dark (side) of ux design: A multi-layered framework for deceptive design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA. ACM.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Kraut, R. E. and Resnick, P. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, Cambridge, MA, USA.

Malone, T. W. (2018). *Superminds: The Surprising Power of People and Computers Thinking Together*. Little, Brown and Company, New York, NY, USA.

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann, San Francisco, CA, USA.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372.

Reczkowski, R. and Adamski, B. (2025). Cognitive warfare and religion: Weaponization of the human cognitive dimension. *Pastoral Psychology*.

Susser, D., Roessler, B., and Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 4(1):1–45.

Woolley, S. C. and Howard, P. N. (2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, Oxford, UK.

Wylie, C. (2019). *Mindf*ck: Cambridge Analytica and the Plot to Break America*. Random House, New York, NY, USA.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, NY, USA.