Influência da Recomendação Algorítmica no Bem-Estar dos Usuários: Proposta de Auditoria

Ana Beatriz F. da Luz[1], **Mayara C. Figueiredo**[1]

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

ana.luz@icen.ufpa.br, mcfigueiredo@ufpa.br

Abstract. *The influence of recommendation algorithms on the well-being of users in states of psychological vulnerability raises concerns regarding the ethical responsibility of digital platforms. To explore this context, this study proposes an algorithmic audit to analyze whether these systems create loops of feedback that saturate the feed with sensitive content within the context of body image distortion. The goal is to inspire algorithmic transparency guidelines that prioritize digital well-being in collaborative systems.*

Resumo. *A influência de algoritmos de recomendação sobre o bem-estar de usuários em estado de vulnerabilidade psicológica levanta preocupações sobre a responsabilidade ética das plataformas digitais. Para explorar este contexto, este estudo propõe uma auditoria algorítmica para analisar se esses sistemas criam loops de retroalimentação que saturam o feed com conteúdos sensíveis no contexto de distorção de imagem. Espera-se que os resultados possam inspirar diretrizes de transparência algorítmica que privilegiem o bem-estar digital em sistemas colaborativos.*

1. Introdução

Plataformas de vídeos curtos, como TikTok e Instagram Reels, utilizam sistemas de recomendação focados em otimizar métricas de retenção [Vombatkere et al. 2024]. Contudo, a lógica de maximização do tempo de tela pode gerar consequências negativas, conduzindo usuários a conteúdos prejudiciais à saúde mental [Golbeck 2025]. Esta pesquisa propõe um desenho metodológico para auditar esses sistemas, investigando como o equilíbrio entre a exploração (apresentação de novos tópicos) e a explotação (aprofundamento em interesses conhecidos) pode saturar o *feed* com conteúdos sensíveis. O estudo se insere na área de sistemas colaborativos ao investigar como mecanismos de recomendação algorítmica atuam como mediadores na construção coletiva de narrativas sobre saúde mental e imagem corporal, influenciando a percepção da realidade e o bem-estar emocional de bilhões de usuários.

2. Fundamentação Teórica

2.1. Algoritmos de Recomendação e Bem-Estar Digital

A influência dos sistemas de recomendação no bem estar dos usuários, especialmente as consequências da formação de "círculos viciosos" de consumo de conteúdo, é um tema que vem sendo debatido na comunidade acadêmica. Segundo [McCrory et al. 2022], estados negativos de saúde mental são reforçados por meio da experiência em plataformas

altamente visuais. Esse fenômeno pode ser perigoso para usuários em situação de vulnerabilidade. Por exemplo, [Golbeck 2025] discutem como sistemas de recomendação podem induzir recaídas em distúrbios alimentares ao entregar conteúdos nocivos a perfis vulneráveis, enquanto [Lin et al. 2016] sugerem que a exposição a conteúdos comparativos e o uso passivo de mídias sociais estariam associados à piora de sintomas depressivos. [Pendse et al. 2023] salientam ainda que as regras estabelecidas sobre como o usuário deve se expressar para obter visibilidade podem reforçar estereótipos, já que, de forma geral, tais conteúdos devem ser "curtos e salientes" apenas para gerar impacto rápido, se distanciando assim de sua complexidade original. Esses estudos evidenciam que plataformas digitais, enquanto sistemas sociotécnicos de recomendação, podem influenciar o bem estar de seus usuários, ressaltando a necessidade de investigar como essa influência acontece e como se pode reduzir seus impactos negativos. Este projeto propõe explorar essa influência por meio de auditoria algorítmica.

2.2. Auditoria Algorítmica

Auditoria algorítmica pode ser definida como uma estratégia de pesquisa que investiga padrões de comportamento problemático, que podem ser ilegais ou socialmente prejudiciais, envolvendo algoritmos de computador operados por plataformas de internet [Sandvig et al. 2014]. Dada a opacidade dos sistemas de recomendação, a auditoria algorítmica surge como uma alternativa para investigar comportamentos enviesados ou danosos, e tem sido usada para investigar algoritmos de mídias sociais, suas influências e possíveis formas de mitigação de problemas [Bandy 2021], [Mosnar et al. 2025], [Boeker and Urman 2022]. Nesse contexto, [Sandvig et al. 2014] analisaram diferentes metodologias de auditoria propondo a observação sistemática de *outputs* (saídas) como método viável para detectar discriminação sem a necessidade de acesso ao código-fonte. Já [Shen et al. 2021] exploram o conceito de "Auditoria Algorítmica Cotidiana" onde a análise de casos reais permitiu identificar comportamentos nocivos através da interação diária e comparação de resultados. [Boeker and Urman 2022] demonstraram a eficácia de auditorias via *sockpuppets* (contas fictícias utilizadas como sondas para examinar um sistema) para isolar fatores de personalização como o tempo de visualização e localização. Entretanto pesquisas anteriores ressaltam a necessidade de investigar plataformas de vídeos curtos, que são extremamente populares, mas para as quais existem poucas auditorias [Mosnar et al. 2025], [Bandy 2021].

3. Metodologia Proposta

Este projeto adota uma abordagem de auditoria algorítmica do tipo *output audit* (auditoria de saídas) [Sandvig et al. 2014] para responder a seguinte questão de pesquisa: Como o direcionamento algorítmico de plataformas de vídeos curtos, baseado em interações e retenção induzem afunilamento e persistência de conteúdos sensíveis? O objetivo é medir como os algoritmos de recomendação reagem a comportamentos associados a vulnerabilidades psicológicas. Neste trabalho será utilizado como exemplo de análise a distorção de autoimagem, por isso, define-se conteúdos sensíveis como vídeos que promovem padrões estéticos irreais, dietas restritivas extremas ou comportamentos associados à distorção de imagem corporal e transtornos alimentares. O estudo investigará ao menos uma plataforma contemporânea de vídeos curtos, selecionada com base em sua popularidade, relevância e viabilidade.

3.1. Cenários Experimentais e Protocolo de Interação

Foram definidos diferentes cenários de auditoria nos quais agentes *sockpuppets* interagem com a funcionalidade principal de vídeos curtos da plataforma (por exemplo, a página *For You* do TikTok). A auditoria é estruturada em um conjunto de testes que manipulam fatores específicos de personalização mantendo constantes ou mitigando as demais condições observáveis [Sandvig et al. 2014]. Os fatores de personalização investigados serão: 1) nenhum (usado como um controle geral), 2) *Likes* (curtidas), 3) Retenção ou VVR (*Video View Ratio* - taxa de visualização) e 4) Busca. Em cada cenário serão criados dois agentes, um controle e um personalizado, produzidos por *scripts* em *Python* que mimetizarão o comportamento humano.

Cenário Padrão: Estabelece a *baseline* de recomendação. Os dois agentes realizam navegação passiva limitando-se a rolagem contínua, com VVR de 100% ou duração máxima de 120 segundos, o que for mais curto. Esse teste permite observar o padrão de ruído inerente ao sistema [Mosnar et al. 2025].

Cenário *Likes* (curtidas): Seguindo a estratégia usada em [Boeker and Urman 2022], [Mosnar et al. 2025] será definida uma lista de *hashtags* para simular interesses do usuário; no caso deste estudo, *hashtags* relacionadas a distorção de autoimagem (#glowupcheck, #weightlosscheck, etc.). O agente personalizado então aplicará *likes* em vídeos que contiverem ao menos uma das *hashtags* da lista. O agente controle agirá como no cenário padrão. Assim, a única diferença entre controle e personalizado será a ação de curtir o vídeo.

Cenário Retenção (VVR): usando a mesma estratégia de *hashtags*, o agente personalizado assistirá integralmente a vídeos que contenham ao menos uma das *hashtags* da lista e pula os demais após 2 segundos. O controle limita-se a rolagem com VVR fixo de 25% para sinalizar um vídeo que não é de interesse do usuário [Mosnar et al. 2025].

Cenário Busca: Para este quarto fator de personalização, precedendo as interações de curtida e retenção, o agente realizará buscas por termos negativos e positivos relacionados a distorção de imagem (ex: *"Self-Love Journey"*, *"perfect body"*, etc.) na barra de pesquisa da plataforma. Após a busca os passos são repetidos de foram idêntica aos cenários anteriores com agentes controle e personalizados.

Em cada cenário, serão coletados os metadados disponíveis de cada vídeo exposto aos agentes (título, descrição, *hashtags*, *likes*, número de comentários, etc.). Serão coletados também os links de uma amostra aleatória de 10% dos vídeos de interesse de cada agente personalizado dos cenários de *like* e retenção. Inspirado na metodologia de [Mosnar et al. 2025], cada teste será composto por quatro sessões de execução com um intervalo de 24h entre elas para que o algoritmo desenvolva a personalização com base nas ações prévias. Define-se que uma sessão termina quando o agente é exposto a um número fixo de 100 vídeos. Após as 4 sessões iniciais, serão executadas mais 4 sessões com o agente personalizado se comportando como o controle para avaliar a persistência do conteúdo quando as ações de personalização não são mais realizadas.

3.2. Análise dos Dados

A análise dos dados segue um desenho de métodos mistos sequencial exploratório [Creswell and Clark 2007]. Primeiramente, será conduzida uma análise de conteúdo qualitativa [Bardin 1977] dos metadados coletados e da amostra de vídeos selecionada. Essa

análise visa categorizar padrões de conteúdo relacionados a bem-estar e saúde, por exemplo, definindo conjuntos de termos e expressões que possam ser associados a conteúdos sensíveis com base na literatura da área. Assim, os metadados e vídeos serão codificados primeiro em paralelo por dois pesquisadores, de forma independente, seguida de reuniões para discussão de divergências e chegada de um consenso. Após a validação e refinamento desta etapa inicial, as categorias derivadas da análise qualitativa serão utilizadas para classificação automatizada dos vídeos para a análise quantitativa, sendo buscadas em diferentes camadas de informação, como legendas, *hashtags* e outros metadados.

A análise quantitativa se dará em duas etapas. Primeiro ela focará em avaliar o grau de personalização do conteúdo seguindo a Análise de Similaridade entre vídeos descrita por [Mosnar et al. 2025]. Para isso serão usadas duas métricas: o índice de Jaccard (proporção de *hashtags* semelhantes compartilhadas por um par de vídeos) e a similaridade de correspondência básica (verificar se um par de vídeos tem ao menos uma *hashtag* similar). Essa análise será usada para verificar evidências de personalização algorítmica decorrente do tratamento, independente de tema.

A segunda etapa usará a classificação gerada pela análise qualitativa para gerar as seguintes métricas: 1) Taxa de Exposição Temática: razão entre o número de vídeos classificados como sensíveis (a partir da análise qualitativa) e o total de 100 vídeos recomendados por sessão, bem como sua evolução no decorrer das sessões. 2) Velocidade de Convergência: o número de sessões necessárias para que a taxa de exposição temática do agente personalizado ultrapasse em 50% a do agente de controle. 3) Persistência de Conteúdo: número de vídeos e/ou sessões que o algoritmo leva para retornar aos níveis de exposição do controle após o agente mudar seu comportamento para neutro.

4. Discussão de Resultados Esperados

Primeiramente, espera-se caracterizar e comparar os principais temas de conteúdos direcionados pelos aplicativos de vídeo curto no contexto de vulnerabilidade investigado. Espera-se também que os resultados revelem uma dinâmica de "afunilamento", onde o equilíbrio entre exploração e explotação [Vombatkere et al. 2024] pende para a explotação de conteúdos considerados sensíveis (hipótese 1). Nesse contexto, espera-se encontrar que as interações de alta taxa de visualização (VVR) e buscas aceleram a formação de bolhas de conteúdo sensível [Golbeck 2025] mais fortemente que interações de curtidas (hipótese 2). Outro resultado esperado é uma dificuldade do algoritmo reverter a recomendação de conteúdos considerados sensíveis mesmo após o agente agir de forma neutra (hipótese 3). Tais análises podem fornecer evidências empíricas da necessidade de auditorias contínuas e de diretrizes algorítmicas que privilegiem o bem-estar do usuário.

5. Considerações Éticas e Impacto Social

Esta pesquisa fundamenta-se na sustentabilidade tecnológica e responsabilidade social, pilares do SBSC 2026. Ao utilizar agentes automatizados em vez de participantes humanos, o desenho experimental mitiga riscos éticos de expor indivíduos a conteúdos prejudiciais à saúde mental. Entretanto, o uso de agentes automatizados pode conflitar com termos de uso de plataformas digitais. No entanto, esta pesquisa fundamenta-se no caráter estritamente acadêmico da atividade (conforme Art. 4º, inciso II da Lei 13.709/2018 - LGPD) e no interesse social de transparência algorítmica previsto no Marco Civil da In-

ternet. A metodologia foi desenhada para minimizar o impacto na infraestrutura das plataformas, coletando apenas metadados públicos e garantindo a exclusão dos perfis após o encerramento do experimento, além de reverter as ações passíveis de reversão, por exemplo *likes* [Mosnar et al. 2025]. A escolha alinha-se às diretrizes de [Sandvig et al. 2014], priorizando a observação externa para diagnosticar danos do sistema sem violar a privacidade. O impacto pretendido é a proposição de diretrizes para sistemas de recomendação éticos que priorizem o bem-estar digital. Em conformidade com as normas da SBC, declara-se que modelos de IA Generativa (*Claude, Gemini*) foram utilizados estritamente para auxílio na estruturação e revisão do texto, mantendo-se a integridade e a autoria intelectual da pesquisa.

Referências

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34.

Bardin, L. (1977). *Análise de conteúdo*. Edições 70.

Boeker, M. and Urman, A. (2022). An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery.

Creswell, J. W. and Clark, V. P. (2007). Mixed methods research. *Thousand Oaks, CA*.

Golbeck, J. A. (2025). Recommender system-induced eating disorder relapse: Harmful content and the challenges of responsible recommendation. *ACM Transactions on Intelligent Systems and Technology*, 16(1).

Lin, L. Y., Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Hoffman, B. L., Giles, L. M., and Primack, B. A. (2016). Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331.

McCrory, A., Best, P., and Maddock, A. (2022). 'it's just one big vicious circle': young people's experiences of highly visual social media and their mental health. *Health Education Research*.

Mosnar, M., Skurla, A., Pecher, B., Tibensky, M., Jakubcik, J., and Bindas, A. (2025). Revisiting algorithmic audits of tiktok: Poor reproducibility and short-term validity of findings. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. Association for Computing Machinery.

Pendse, S. R., Kumar, N., and De Choudhury, M. (2023). Marginalization and the construction of mental illness narratives online: Foregrounding institutions in technology-mediated care. In *Proceedings of the ACM on Human-Computer Interaction*.

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *ICA Data and Discrimination Preconference*, Seattle, WA, USA.

Shen, H., DeVos, A., Eslami, M., and Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29.

Vombatkere, K., Mousavi, S., Zannettou, S., Roesner, F., and Gummadi, K. P. (2024). Tiktok and the art of personalization: Investigating exploration and exploitation on social media feeds. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 3789–3797. Association for Computing Machinery.