Towards a Humanized Social-Media Ecosystem: AI-Augmented HCI Design Patterns for Safety, Agency & Well-being

1st Mohd Ruhul Ameen
College of Engineering and Computer Sciences
Marshall University
Huntington, WV, USA
ameen@marshall.edu

2nd Akif Islam
Department of Computer Science and Engineering
University of Rajshahi
Rajshahi 6205, Bangladesh
iamakifislam@gmail.com

***Abstract*—Social platforms connect billions of people, yet their engagement-first algorithms often work on users rather than with them, amplifying stress, misinformation, and a loss of control. We propose Human-Layer AI (HL-AI) which is a user-owned, explainable intermediaries that sit in the browser between platform logic and the interface. HL-AI gives people practical, moment-to-moment control without needing platform cooperation. We contribute a working Chrome/Edge prototype that implements five representative patterns framework (Context-Aware Post Rewriter, Post Integrity Meter, Granular Feed Curator, Micro-Withdrawal Agent and Recovery Mode), a unifying mathematical formulation that balances user utility, autonomy costs and risk thresholds, and an evaluation plan spanning technical accuracy, usability and behavioral outcomes. The result is a suite of humane controls that help users rewrite before harm, read with integrity cues, tune feeds with intention, pause compulsive loops and seek shelter during harassment while preserving agency through explanations and override options. We argue that HL-AI offers a practical path to retrofit today's feeds with safety, agency and well-being, and we release design patterns and implementation details to support adoption and future study. As a prototype, this work invites rigorous user evaluation across cultures and contexts.**

***Index Terms*—HCI design patterns, human–AI interaction, safety, agency, well-being, social media, transparency**

I. Introduction

Social media platforms have fundamentally reshaped human communication and information access, with more than 5 billion active user identities worldwide as of 2024 [1]. While these platforms have enabled unprecedented connectivity, knowledge sharing, and civic participation, their underlying architectural paradigm remains rooted in *simplex communication*—a unidirectional system in which platform-controlled algorithms curate and prioritize information flows without meaningful user input [2], [3]. This engagement-optimized model inherently prioritizes attention capture over human-centered outcomes such as psychological wellbeing, information integrity, and meaningful task accomplishment.

The societal consequences of this misalignment have become increasingly evident across South Asia. The July 2025 Milestone School helicopter crash in Bangladesh demonstrated how algorithmic feeds can exacerbate psychological trauma by repeatedly surfacing distressing content in the absence of user-controlled filtering mechanisms [4], [5]. Survivors and community members were continually exposed to graphic imagery and updates, intensifying mental health distress as feeds optimized for engagement rather than emotional protection [6].

Similarly, during Bangladesh's July Revolution of 2024—when student-led protests culminated in the resignation and flight of Prime Minister Sheikh Hasina on August 5, 2024—social media platforms became simultaneous tools for democratic mobilization and vectors for large-scale misinformation [7], [8]. Platform-driven amplification dynamics contributed to volatility during the uprising, in which over 1,400 people were killed [9].

More recently, Nepal's September 2025 Gen Z protests highlighted both the potential and structural limitations of current social platforms. Following the government's ban of 26 major social media services, protesters migrated to Discord, where over 145,000 participants used structured discussions and polls to democratically select an interim Prime Minister [10], [11], [12]. This event demonstrated a clear demand for *duplex communication*—systems that empower users to actively shape their information ecosystem rather than passively receiving algorithmically curated feeds.

Compounding these issues, cross-border misinformation campaigns have intensified regional tensions. In 2024, Indian media outlets circulated more than 137 false reports about Bangladesh, many algorithmically amplified on social platforms [13], [14]. These incidents illustrate how engagement-driven architectures can propagate harmful narratives across national boundaries, producing real-world diplomatic and communal consequences.

Despite growing awareness of these systemic issues, current mitigation approaches remain fundamentally constrained by the simplex communication paradigm. Existing mechanisms—such as basic screen-time limits, opaque feed adjustments, or platform-controlled moderation systems—do not grant users meaningful control over how information is curated, prioritized, or de-emphasized. Users increasingly expect granular, value-aligned control over their digital environments: for ex-

ample, reducing political content while increasing educational material, or limiting exposure to sensitive topics during vulnerable periods [15].

At present, however, mainstream platforms lack architectural support for *user-owned, explainable AI intermediaries* that can mediate between raw platform outputs and the user interface. This architectural gap prevents the development of systems that support real-time transparency, personalization, and psychological wellbeing.

The consequences are measurable and widespread. Studies report that:

- 62% of daily users experience post-session regret,
- trust in platform-mediated content has declined to 28%, and
- 38.2% of users accidentally share misinformation due to inadequate decision-support mechanisms [16].

Events such as the Milestone School tragedy [4] and Bangladesh–India misinformation campaigns [17] highlight how the absence of user-controlled intermediaries can exacerbate trauma, distort information ecosystems, and intensify geopolitical tensions.

Research in human-centered AI has established foundational principles for promoting user autonomy, transparency, and controllability [18], [19]. Parallel work in digital wellbeing identifies the psychological and behavioral factors that shape healthy technology use and proposes various intervention mechanisms [20], [21]. Advances in explainable AI (XAI) have further introduced techniques for interpreting model behavior and making algorithmic decisions more intelligible to users [22], [23]. Despite these developments, existing approaches remain constrained by several critical limitations. First, digital wellbeing tools predominantly rely on restriction-based mechanisms—such as time limits or app blocking—rather than augmentation-based strategies that enhance users' ability to make informed decisions [24]. Second, misinformation detection and content ranking infrastructures remain centrally controlled by platforms, offering limited transparency and virtually no customization for end users [25], [26]. Third, most explainable AI methods provide post-hoc explanations that describe algorithmic decisions only after they occur, rather than enabling proactive, user-controlled intermediation capable of shaping information flows in real time [27]. Collectively, these shortcomings reveal a fundamental research gap: the absence of comprehensive design patterns and architectural frameworks for developing *user-controlled AI intermediaries* that can operate within social media environments.

This paper contributes a practical and human-centered approach to improving user control and wellbeing in social media systems. First, we introduce an architectural model for user-owned AI intermediaries that operate between platform algorithms and the user interface, enabling greater transparency and autonomy. Second, we present a taxonomy of fifteen Human-Layer AI (HL-AI) design patterns spanning safety, agency, and wellbeing; while the complete set will appear in a forthcoming journal publication, this paper illustrates five representative patterns to convey the core concepts. Third, we outline an initial mathematical formulation for balancing user autonomy and risk mitigation, providing a structured basis for customizing intermediary behavior. Finally, although full empirical evaluation is reserved for future work, we describe our planned methodology and demonstrate a working prototype that implements the five showcased design patterns.

II. Related Work

Human-Centered AI emphasizes transparency, explainability, and user agency [19]. Gausen et al. [28] introduced "sociotechnical transparency" considering both technical systems and their interactions with users. Recent work shows transparency significantly mitigates negative relationships between AI attitudes and trust when user involvement is high [29].

Digital Wellbeing Research evolved from restriction-based to augmentation-based frameworks enhancing user decision-making [30]. Vanden Abeele's dynamic systems approach recognizes digital wellbeing as contingent on personal characteristics and design choices [31]. The METUX framework identifies autonomy, competence, and relatedness as essential for sustained motivation and wellbeing [32].

AI-powered moderation reports 99% terrorism-related post detection before user reports [33]. However, smaller accounts disproportionately spread AI-generated misinformation [34]. Analysis reveals both downstream (post-spread identification) and upstream (prevention) approaches are needed.

Legal scholarship establishes "right to know" algorithms as fundamental to democratic participation [3]. Research shows granular control mechanisms beyond simple on/off switches are needed [29]. Recent workshops identified novel design challenges for human-AI collaboration.

III. Core Design Patterns

We present five representative patterns from our broader fifteen-pattern framework. These illustrate how Human-Layer AI (HL-AI) intermediaries enhance user safety, agency, and wellbeing.

A. P1. Context-Aware Post Rewriter

This pattern aims to support reflective posting by offering AI-generated rewrites for potentially harmful content. The intermediary analyzes the draft, previews how it may be

Example

User draft: "These people ruin everything about our town."

System: Flags polarizing phrasing and suggests alternatives (neutral, empathetic, fact-focused).

User: Selects neutral rewrite: "I'm concerned about recent changes in our community."

Outcome: Neutral post is published; original remains local only.

Fig. 1. Example of P1. Context-Aware Post Rewriter

TABLE I
P1 — CONTEXT-AWARE POST REWRITER

Intent	Encourage reflection and reduce harmful posting by providing AI-assisted rewrites.
Problem	Users may unintentionally post polarizing or biased phrasing that escalates conflict.
Context	Fast-paced posting environments, public visibility, and sensitive sociocultural topics.
Forces	Reflection vs. speed; safety vs. autonomy; privacy vs. proactive guidance.
Solution	The intermediary reads the draft, generates neutral or empathetic alternatives, provides contextual warnings, and allows users to override suggestions at any time.
Consequences	**Pros:** Reduced harm and more mindful posting. **Risks:** May trigger perceptions of censorship or lead to over-reliance on AI rewrites.

Example

Viral claim: "New vaccine banned in 20 countries!"

Integrity Meter: Conflicting sources (4 reputable outlets disagree), Fake News Probability: 82% "Likely False," Political Bias: "Strong Right," AI Content: 65% likelihood.

Outcome: User checks linked sources and decides not to share.

Fig. 2. Example of P2. Post Integrity Meter

interpreted, and provides neutral or empathetic alternatives while leaving final choice to the user.

B. P2. Post Integrity Meter

This pattern provides real-time integrity signals for posts by checking factual consistency, AI-generation likelihood, and political bias. Instead of restricting access, the intermediary offers transparent cues that help users judge credibility.

C. P3. Granular Feed Curator

This pattern gives users actionable control over feed composition and ad categories through simple sliders and toggles. Rather than relying on opaque algorithmic defaults, users can directly shape what they see and how often.

Example

User: Chen feels overwhelmed by political posts.

System: Shows a category slider.

Action: Chen reduces "Politics" by 70%.

Outcome: Feed updates with fewer political posts, maintaining occasional ones for awareness.

Fig. 3. Example of P3. Granular Feed Curator

TABLE II
P2 — INTEGRITY METER

Intent	Provide clear, real-time integrity cues showing whether content may be fake, AI-generated, or biased, supporting informed decision-making.
Problem	Users struggle to distinguish reliable posts from misinformation or synthetic media in fast-moving environments.
Context	High-stakes domains such as health, elections, and crises; mixed content streams; rapid skimming with minimal fact-checking.
Forces	Transparency vs. overload; automation vs. accuracy; awareness vs. trust; fast processing vs. verifiability.
Solution	The system introduces an Integrity Meter beneath each post. It checks claim consistency against reputable sources, estimates misinformation probability, detects AI-generated text, images, or videos, and provides contextualized bias estimations. All results are delivered through plain-language explanations with adjustable sensitivity to suit varying user preferences.
Consequences	**Pros:** Supports critical literacy and reduces misinformation spread. **Risks:** Users may over-trust probability scores, misinterpret bias labels, or face adversarial adaptation.

D. P4. Micro-Withdrawal Agent

This pattern introduces brief, context-aware pauses that help users step out of compulsive scrolling loops without imposing hard blocks. The intermediary detects continuation risk and offers lightweight reflective prompts or redirections.

TABLE III
P3 — GRANULAR FEED CURATOR

Intent	Provide intuitive, fine-grained controls for adjusting feed content and ad exposure.
Problem	Users feel feeds are manipulated or unclear. Built-in platform controls are vague, buried, or ineffective, reducing trust and limiting personalization.
Context	Social feeds, recommender systems, and ad personalization scenarios where users want control without losing discovery.
Forces	Transparency vs. complexity; discovery vs. control; personalization vs. privacy; stability vs. adaptability.
Solution	The system provides a granular feed-filter module that lets users adjust content intensity across categories such as politics, sports, or personal updates. It offers an ad-transparency panel with options to disable specific ad categories, quick contextual toggles (e.g., focusing on friends), and per-post overrides that shape future recommendations. Auto-tuning powered by an HL-AI Curator Agent refines personalization while maintaining user agency.
Consequences	**Pros:** Higher trust, better relevance, reduced frustration. **Risks:** Over-curation may create echo chambers; extensive controls may overwhelm some users.

Example

User: Aisha has been watching travel reels for 9 minutes.

System: Detects repetition and shows a brief pause card.

Prompt: "You've seen 4 travel reels—save your favorites and take a short breather?"

Outcome: Aisha taps "Save & close," ending the session intentionally.

Fig. 4. Example of P4. Micro-Withdrawal Agent

E. P5. Recovery Mode

This pattern provides an immediate "shelter-in-feed" mode when users face harassment, sudden pile-ons, or emotional overload. The intermediary reduces harmful exposure, limits inbound interaction, and offers structured next steps.

IV. Mathematical Framework

We formalize the trade-off between user autonomy and risk mitigation as an optimization problem. For each action or intervention a associated with a content item, the system selects the action that maximizes the objective

$$\mathcal{J}(a) = u(a) - \lambda\,\Omega(a) - \beta\,\mathbf{1}_{r(a)>\tau}\,r(a), \tag{1}$$

TABLE IV
P4 — Micro-Withdrawal Agent

Intent	Provide brief, context-aware pauses that help users disengage from compulsive scrolling without paternalistic blocking.
Problem	Variable-ratio reward loops ("just one more") drive overuse, regret, and fatigue. Traditional screen-time warnings are blunt and ignore situational factors.
Context	Infinite-scroll feeds, late-night usage, repetitive content cycles, and sessions that drift away from user goals such as staying informed, connecting with friends, or learning.
Forces	Relief vs. annoyance; agency vs. protection; behavioral inference vs. privacy; minimal disruption vs. effectiveness.
Solution	The system detects continuation risk by analyzing patterns such as swipe velocity, repetitive topic exposure, time-of-day signals, and stated user goals. When the likelihood of compulsive continuation is high, it delivers a brief micro-intervention lasting under ten seconds—providing a moment of reflection, an optional redirection to saved or meaningful content, or the frictionless option to continue. The cadence of these interventions adapts dynamically based on user responses, including acceptance, dismissal, or repeated avoidance.
Consequences	**Pros:** Reduces impulsive loops and increases goal alignment. **Risks:** Habituation, banner blindness, or perceptions of nagging.

Example

User: Nabila receives a hostile quote-tweet and a surge of negative replies.

System: Detects rapid hostile activity and prompts: "Rough thread—enable Recovery for 24h?"

Action: Nabila enables it with one tap.

Outcome: Replies limited to followers-only, mentions are queued for review, and support resources appear.

Fig. 5. Example of P5. Recovery Mode

TABLE V
P5 — Recovery Mode

Intent	Provide an immediate, protective mode after harassment or overload that reduces exposure, limits inbound interactions, and surfaces appropriate support.
Problem	Following negative events, users face toxic replies, brigading, and decision paralysis. Default feeds continue amplifying harm instead of stabilizing the situation.
Context	Online pile-ons, personal attacks, doxxing attempts, viral controversies, and vulnerable timing (late-night, travel).
Forces	Rapid safety vs. missing supportive messages; automation vs. consent; visibility vs. protection; shielding vs. information flow.
Solution	Recovery Mode enables one-tap activation—or an optional proactive suggestion—that immediately restricts who can reply or send direct messages, suppresses mentions, and hides toxic responses. It queues reports with automatically captured evidence, opens a dedicated support hub with trusted contacts and relevant guidance, and assists the user in planning a temporary exit such as short mutes or a timed account pause.
Consequences	**Pros:** Rapid harm reduction; clearer decisions; preserved evidence. **Risks:** May filter out supportive messages; false positives could cause unnecessary isolation.

where $u(a)$ denotes the user-perceived utility of the action, $\Omega(a)$ represents the degree to which the intervention may compromise user agency, and $r(a)$ reflects the estimated risk associated with the content or behavior. The parameters λ and β weight the importance of preserving autonomy and mitigating risk, respectively, while τ represents the user-adjustable tolerance threshold above which risk incurs a penalty. This formulation encourages the system to favor actions that provide meaningful benefit while avoiding excessive interference unless a user's defined risk threshold has been exceeded.

Each of the five patterns instantiates this model through three common mechanisms. First, a transparency layer ensures that every intervention is accompanied by a short explanation and the option to override, preserving user control even when risks are flagged. Second, the model allows users to calibrate the autonomy–risk balance by adjusting parameters such as λ and τ, enabling different levels of sensitivity depending on personal goals, context, and comfort. Finally, an audit

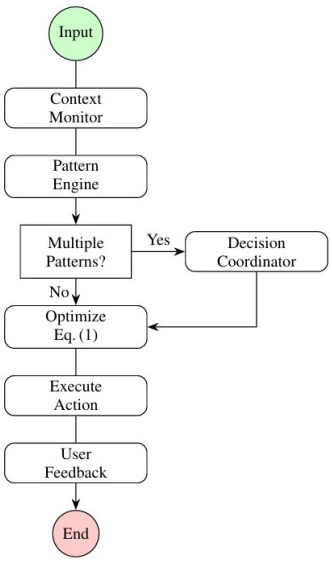

Fig. 6. HL-AI master decision flow: pattern coordination and optimization.

trail captures interventions and user responses, supporting accountability and enabling the system to adapt future behavior in a predictable and user-aligned manner. Together, these elements provide a principled mathematical basis for HL-AI intermediaries that align safety, autonomy, and wellbeing.

V. System Architecture and Algorithms

A. HL-AI Master Decision Flow

Figure 6 illustrates the core decision-making process for the Human-Layer AI system, showing how multiple patterns coordinate through the centralized optimization framework.

B. Core Optimization Algorithm

Algorithm 1 implements the mathematical framework from Equation (1), balancing user agency with risk reduction across all patterns.

C. Post Integrity Assessment Algorithm

Algorithm 2 details the multi-signal approach for Pattern P3 (Post Integrity Meter), combining fact-checking, AI detection, and bias estimation.

Algorithm 1 HL-AI Pattern Optimization Framework

Require: User preference weight λ, risk threshold τ, content/action set $\mathcal{A}$
Ensure: Optimal intervention decision a^*
Initialize utility estimator $\hat{u}(a)$ and risk estimator $\hat{r}(a)$
Compute agency penalty $\Omega(a)$ according to intervention type
for each candidate action $a_i \in \mathcal{A}$ **do**
  $J(a_i) \leftarrow \lambda \cdot \hat{u}(a_i) + (1-\lambda) \cdot \hat{r}(a_i) - \Omega(a_i)$
  if $\hat{r}(a_i) > \tau$ **and** intervention_required **then**
    $J(a_i) \leftarrow J(a_i) - \beta \cdot \hat{r}(a_i)$ ▷ Apply safety penalty
  end if
end for
$a^* \leftarrow \arg\max_{a_i} J(a_i)$
return a^* along with explanation and user override option

Algorithm 2 Multi-Signal Post Integrity Assessment

Require: Post content p, fact-check database $\mathcal{D}$, AI detection model M_{AI}
Ensure: Integrity score vector $\mathbf{s} = [s_{fact}, s_{AI}, s_{bias}]$
Extract claims $C = \{c_1, c_2, ..., c_n\}$ from p using NER
$conflicts \leftarrow 0$, $total\_claims \leftarrow |C|$
for each claim $c_i \in C$ **do**
  $sources \leftarrow$ query_database(c_i, $\mathcal{D}$)
  if $|sources| > 0$ **and** any source contradicts c_i **then**
    $conflicts \leftarrow conflicts + 1$
  end if
end for
$s_{fact} \leftarrow 1 - \frac{conflicts}{max(total\_claims, 1)}$ ▷ Fact score
$s_{AI} \leftarrow M_{AI}(p)$ ▷ AI-generated probability
$s_{bias} \leftarrow$ estimate_political_lean(p) ▷ Bias classification
Generate explanations for each score component
return $\mathbf{s}$, explanations, source_links

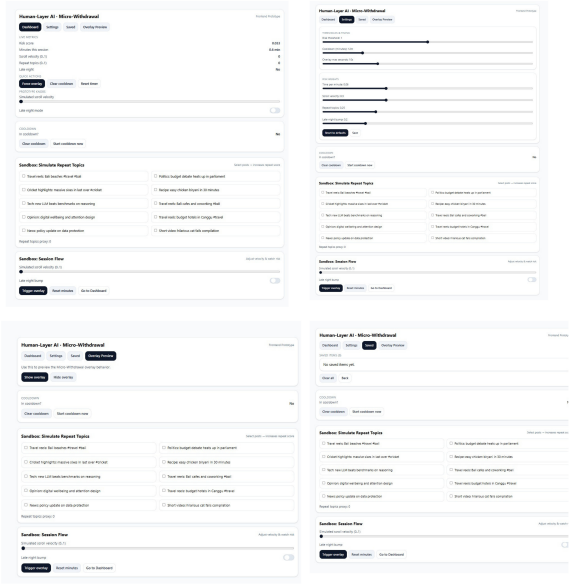

Fig. 7. Prototype screenshots of HL-AI browser-based intermediary (placeholder).

VI. Implementation Architecture

Our system is implemented as a Chrome/Edge browser extension that acts as a Human-Layer AI (HL-AI) intermediary, dynamically modifying the interface and injecting interaction elements without requiring platform cooperation. Figure 7 presents the system architecture along with prototype screenshots.

The HL-AI architecture functions as a privacy-preserving intermediary that applies the design patterns without requiring any cooperation from the underlying platform. At its core, a configurable Pattern Engine implements each pattern according to user-defined parameters and thresholds, while a Context Monitor observes user state, interaction rhythms, content consumption, and temporal factors to determine when interventions are appropriate. A central Decision Coordinator resolves potential conflicts between patterns by considering safety priorities, user preferences, and situational context. The Interface Adapter then modifies how content is presented,

inserts interface elements, and exposes control panels in a way that preserves the original platform's functionality. Throughout this process, the system follows a privacy-by-design approach: personal data is processed on-device whenever possible, cloud computation is restricted to anonymized aggregates, and users maintain full control over data sharing and pattern activation.

VII. Conclusion

We introduce a Human-Layer AI framework that reimagines social media through humane, user-controlled design patterns. Rather than optimizing for engagement, our approach centers on user sovereignty—offering protective and empowering features that foster wellbeing, safety, and agency. This pattern language outlines practical pathways for retrofitting existing platforms with ethical, user-aligned controls. By balancing autonomy with risk mitigation, it supports thoughtful decision-making that safeguards individuals while strengthening collective trust. True progress depends on embedding human-centered design principles at the core of technology. Our prototype offers an early blueprint for social media that serves human flourishing, inviting further user studies and cross-cultural evaluation to refine and expand its impact.

References

[1] Kepios, "Digital 2024: Global overview report," DataReportal, Mar. 2024, [Online]. Available: https://datareportal.com/reports/digital-2024-global-overview-report.

[2] H. Metzler and D. Garcia, "Social drivers and algorithmic mechanisms on digital media," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 735–748, 2024.

[3] H. Sun, "The right to know social media algorithms," *Harvard Law & Policy Review*, vol. 18, no. 1, pp. 1–47, 2024.

[4] Dhaka Tribune, "Milestone tragedy: Survivors face trauma beyond injuries," Dhaka Tribune, Jul. 2025, [Online]. Available: https://www.dhakatribune.com/feature/health-wellness/387147/milestone-tragedy-survivors-face-trauma-beyond.

[5] CNN, "Bangladesh air force jet crash: Distraught students demand answers after crash turned dhaka school into 'death trap'," CNN, Jul. 2025, [Online]. Available: https://www.cnn.com/2025/07/22/asia/bangladesh-plane-crash-witnesses-intl-hnk.

[6] The Daily Star, "Survivors of the milestone jet crash and the road to emotional recovery," The Daily Star, Jul. 2025, [Online]. Available: https://www.thedailystar.net/rising-stars/news/survivors-the-milestone-jet-crash-and-the-road-emotional-recovery-3951551.

[7] Al Jazeera, "How bangladesh's 'gen z' protests brought down pm sheikh hasina," Al Jazeera, Aug. 2024, [Online]. Available: https://www.aljazeera.com/news/longform/2024/8/7/how-bangladeshs-gen-z-protests-brought-down-pm-sheikh-hasina.

[8] Wikipedia, "July revolution (bangladesh)," 2024, [Online]. Available: https://en.wikipedia.org/wiki/July_Revolution_(Bangladesh).

[9] G. Macdonald, "Bangladesh's accidental revolutionaries topple sheikh hasina — what's next?" United States Institute of Peace, Aug. 2024, [Online]. Available: https://www.usip.org/publications/2024/08/bangladeshs-accidental-revolutionaries-topple-sheikh-hasina-whats-next.

[10] Al Jazeera, "'more egalitarian': How nepal's gen z used gaming app discord to pick pm," Al Jazeera, Sep. 2025, [Online]. Available: https://www.aljazeera.com/news/2025/9/15/more-egalitarian-how-nepals-gen-z-used-gaming-app-discord-to-pick-pm.

[11] Wikipedia, "2025 nepalese gen z protests," 2025, [Online]. Available: https://en.wikipedia.org/wiki/2025_Nepalese_Gen_Z_protests.

[12] NPR, "How gen z-led protests put nepal's 1st female prime minister in power," NPR, Sep. 2025, [Online]. Available: https://www.npr.org/2025/09/20/nx-s1-5545760/nepal-protests-gen-z.

[13] Rumor Scanner, "Trace of misinformation about bangladesh in 72 indian media outlets in 2024," Rumor Scanner, Jan. 2025, [Online]. Available: https://rumorscanner.com/en/india-en/indian-misinformation-on-bangladesh-2024/135322.

[14] Voice of America, "Tensions peak as bangladesh blames india for 'spreading misinformation'," Voice of America, Dec. 2024, [Online]. Available: https://www.voanews.com/a/tensions-peak-as-bangladesh-blames-india-for-spreading-misinformation-/7895753.html.

[15] A. Narayanan, "Understanding social media recommendation algorithms: Towards a better-informed debate on the effects of social media," Knight First Amendment Institute, Columbia University, New York, NY, USA, Essay, Mar. 2023, visiting Senior Research Scientist Essay. [Online]. Available: https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

[16] A. Watson, "Share of people who have ever accidentally shared fake news or information on social media in the united states as of december 2020," Statista, Mar. 2023, statista Statistic No. 657111. Available: https://www.statista.com/statistics/657111/fake-news-sharing-online/.

[17] BBC Verify, "'persecution' of hindus in bangladesh: Fake posts uncovered by bbc," The Business Standard, Aug. 2024, [Online]. Available: https://www.tbsnews.net/bangladesh/persecution-hindus-bangladesh-fake-posts-uncovered-bbc-914141.

[18] O. O. Garibay *et al.*, "Six human-centered artificial intelligence grand challenges," *International Journal of Human-Computer Interaction*, vol. 39, no. 3, pp. 391–437, 2023.

[19] C. Stephanidis *et al.*, "Seven hci grand challenges," *International Journal of Human-Computer Interaction*, vol. 39, no. 4, pp. 1–47, 2023.

[20] M. Büchi, "Digital well-being theory and research," *New Media & Society*, vol. 26, no. 1, pp. 155–183, 2024.

[21] Internet Matters, "Children's wellbeing in a digital world: Year three index report 2024," Internet Matters, London, UK, Tech. Rep., Sep. 2024, developed with BMG Research. [Online]. Available: https://www.internetmatters.org/hub/research/childrens-wellbeing-in-a-digital-world-index-report-2024/

[22] N. Scharowski, S. A. C. Perrig, M. Svab, K. Opwis, and F. Brühlmann, "Exploring the effects of human-centered ai explanations on trust and reliance," *Frontiers in Computer Science*, vol. 5, Jul. 2023.

[23] Q. V. Liao and K. R. Varshney, "Human-centered explainable ai (xai): From algorithms to user experiences," 2022, arXiv preprint arXiv:2110.10790.

[24] C. Molnar, "Explainable ai for improved human-computer interaction," in *Proceedings of the Second World Conference on Explainable Artificial Intelligence (xAI 2024)*, ser. Communications in Computer and Information Science, vol. 2156. Valletta, Malta: Springer, Jul. 2024.

[25] J. A. Goldstein *et al.*, "Generative language models and automated influence operations: Emerging threats and potential mitigations," *arXiv preprint arXiv:2301.04246*, 2023.

[26] F. Pilati and T. Venturini, "The use of artificial intelligence in counter-disinformation: a world wide (web) mapping," *Frontiers in Political Science*, vol. 7, Jan. 2025.

[27] G. Vilone and L. Longo, "A comprehensive evaluation of explainable artificial intelligence applications: a systematic review," *Frontiers in Artificial Intelligence*, vol. 7, Sep. 2024.

[28] A. Gausen, C. Guo, and W. Luk, "An approach to sociotechnical transparency of social media algorithms using agent-based modelling," *AI and Ethics*, vol. 5, pp. 1827–1845, 2025.

[29] K. Park and H. Y. Yoon, "Ai algorithm transparency, pipelines for trust not prisms: Mitigating general negative attitudes and enhancing trust toward ai," *Humanities and Social Sciences Communications*, vol. 12, no. 1, p. 1160, Jul. 2025.

[30] A. Monge Roffarello and L. De Russis, "Achieving digital wellbeing through digital self-control tools: A systematic review and meta-analysis," *ACM Transactions on Computer-Human Interaction*, 2023.

[31] M. Vanden Abeele, "Digital wellbeing as a dynamic construct," *Communication Theory*, vol. 31, no. 4, pp. 932–955, 2020.

[32] D. Peters, R. A. Calvo, and R. M. Ryan, "Designing for motivation, engagement and wellbeing in digital experience," *Frontiers in Psychology*, vol. 9, p. 797, 2018.

[33] "Theory and practice of social media's content moderation by artificial intelligence in light of european union's ai act and digital services act," *European Journal of Law and Political Science*, Feb. 2025.

[34] N. Pröllochs *et al.*, "Characterizing ai-generated misinformation on social media," *arXiv preprint arXiv:2505.10266*, 2025.