Beyond Personalization: Multi-User Recommender Dynamics and Robust Mitigation of Social Media Addiction

Filippo Gasco
University of Milan-Bicocca
Milan, Italy
f.gasco@campus.unimib.it

Nicolo' Luigi Allegris
University of Milan-Bicocca
Milan, Italy
n.allegris@campus.unimib.it

Luca Bolis
University of Milan-Bicocca
Milan, Italy
l.bolis3@campus.unimib.it

Stefano Livella
University of Milan-Bicocca
Milan, Italy
s.livella@campus.unimib.it

Sabrina Guidotti
University of Milan-Bicocca
Milan, Italy
s.guidotti2@campus.unimib.it

Dimitri Ognibene
University of Milan-Bicocca
Milan, Italy
dimitri.ognibene@unimib.it

ABSTRACT

Problematic and addiction-like social media use is increasingly linked to engagement-optimizing recommender systems, yet most computational accounts analyze personalization at the level of an isolated user. In this paper we study how *population-level* learning can couple users through shared recommender updates, creating spillovers whereby one user's engagement signals affect another user's exposure and, potentially, their risk of overuse. We extend a dual-system reinforcement-learning model of social media addiction to a multi-user setting and compare three recommender architectures: a single shared recommender across the whole population, per-user individualized recommenders, and group-specific recommenders shared among users with similar preferences. Across two simulated populations (addiction-prone and healthy-leaning), we show that the extent and *correctness* of recommender generalization materially changes the prevalence and speed of addiction: shared and group-based recommenders can accelerate or limit overuse relative to individualized learning, with group-based sharing amplifying addiction when users truly share preferences. Finally, we evaluate a well-being oriented intervention (PUT-OUT) and find it remains comparatively robust across generalization regimes, reducing excessive use even when cross-user learning amplifies engagement dynamics. Our findings highlight non-uniform systemic impacts of recommender design and motivate auditing and governance approaches that assess cross-user pathways, not only individual-level personalization effects.

KEYWORDS

Social Media, Recommender Systems, Algorithm Auditing, User Behavior Modeling, Well-Being, Behavioural Addiction

1 INTRODUCTION

Concerns about compulsive and addictive behaviors have become increasingly prominent alongside the widespread adoption of social media, which now play a central role in how information is accessed and how people interact. These problematic usage patterns differ from pharmacological addiction, as they do not stem from substance intake but rather from inherent cognitive limitations, particularly the challenge of resisting immediately rewarding stimuli in favor of long-term well-being. This vulnerability is further amplified by recommendation systems intentionally designed to sustain and intensify user engagement, often without regard for potential negative outcomes [17, 26, 27, 36]. A growing body of longitudinal research reveals a paradox: while aggregate time spent on social media platforms steadily rises, self-reported life satisfaction and subjective well-being tend to deteriorate. This divergence between increasing engagement and declining utility—often described as an engagement–utility gap—has been documented empirically, with studies showing, for instance, that Facebook use can predict reductions in life satisfaction within just a few days. Notably, this tension has also been publicly acknowledged by platform representatives; in 2017, both Facebook and YouTube conceded that algorithms optimized for engagement might inadvertently contribute to users feeling worse, even as usage continued to grow [20, 29, 30].

This escalation has been framed as a recognized public health concern [33] rather than just a technical or behavioral phenomenon. Different studies highlight how excessive use has been associated with stress, anxiety, and depression. In particular, Hussain and Griffiths [16] show correlations between problematic social network use and psychiatric disorders. Large-scale behavioural data show that individuals with poorer mental health spend more time on social platforms and engage in more detrimental online feedback loops compared with peers without mental health conditions [12]. Moreover, short-term interventions, such as a one-week social media detox, have been shown to significantly reduce symptoms of depression, anxiety, and insomnia [8].

To gain deeper insights into the root causes of this maladaptive conduct, Wang and Zhang [37] offer an insightful examination based on operant conditioning principles [35]. They demonstrate how social networks deliberately leverage reinforcements (such as likes and comments) alongside punishments (like social exclusion or negative reactions) to shape how users act. Crucially, they point out that even adverse stimuli, including antagonistic remarks, can ironically act as reinforcements. These negative inputs activate a user's urge to retaliate or seek approval, ultimately leading to increased platform usage and deepened psychological attachment.

Prior work by Livella, Bolis, et al. [19] has investigated how *individual-level* personalization shapes a user's platform usage patterns, highlighting that recommendation policies can measurably steer attention and engagement. However, this line of analysis typically treats the user in isolation and does not explicitly account for the fact that recommendations are learned from *population-level*

interaction data: the behavior of other users can shift item popularity, reshape embedding spaces, and alter the estimated relevance of content for a given individual. In this sense, unhealthy or extreme engagement patterns may propagate indirectly through a shared recommender, facilitating forms of behavioral contagion analogous to "information wildfires" [38] and echo chambers [10, 14, 21, 27, 31]. It is therefore crucial to analyze these cross-user effects to better inform auditing processes and governance frameworks: in the EU, for example, the Digital Services Act links recommender-system design to transparency duties and to the identification and mitigation of systemic risks, alongside requirements for independent auditing for very large platforms [1–4, 15]. More broadly, established approaches to internal algorithmic auditing, external audit methodologies, and impact-assessment practice provide concrete templates for operationalizing such system-level evaluations [24, 25, 28, 34].

This is the aim of this work: we study how population-level behavioral signals can influence an individual's recommendations and usage via shared model updates and ranking dynamics, and we discuss how accounting for these cross-user pathways can strengthen recommender auditing and governance.

2 RELATED WORKS

Social media addiction (SMA) research has established that algorithmic personalization can amplify compulsive use through mechanisms analogous to behavioral addictions [6, 7]. Engagement-maximizing platforms leverage collective user data to inform recommendations, creating feedback loops where one user's behavior influences the content exposed to others. This multi-user influence can propagate addictive patterns at scale, producing spillover effects that extend beyond individual use. For instance, content that elicits strong engagement in a subset of users is more likely to be surfaced to others, reinforcing habitual prolonged sessions and craving-like behaviors across the network.

Recent empirical work provides direct evidence that algorithmic feeds drive problematic use. Dekker et al. [11] show that personalized "For You"-style feeds on platforms like TikTok causally increase session length and compulsive scrolling compared to non-personalized feeds. Similarly, the rollout of Instagram's algorithmic feed in 2016 was associated with measurable declines in well-being, highlighting the mental health impact of engagement-driven recommendation systems [23]. Chen et al. [9] demonstrates that algorithmic recommendation contributes to continued use even when users intend to quit, capturing the hallmark of addiction-like compulsive engagement. Beyond personalization, interface mechanics such as infinite scrolling and autoplay exacerbate compulsive behavior by creating nearly frictionless loops of content consumption [13, 22]. Together, these findings suggest that SMA is shaped not only by individual susceptibility but also by algorithmic and design features that promote overuse.

Recent studies highlight the algorithmic mechanisms underlying multi-user spillover. Collaborative filtering and engagement-based ranking systems rely on aggregate user interactions to personalize content, which can amplify exposure to highly engaging content and reduce diversity. Empirical analyses indicate that these shared learning dynamics not only homogenize behavior but also increase the probability of compulsive use in other users, creating system-level reinforcement of addictive patterns [18]. As such, SMA emerges not just from individual susceptibility but also from the collective shaping of the recommendation environment.

To address these challenges, researchers are exploring human-centered recommender systems and well-being–oriented interventions. With growing awareness of the harmful effects of engagement-maximizing algorithms, efforts have emerged to design social media platforms that prioritize ethical and user-focused experiences. These strategies operate at different levels: some target the user directly, promoting healthier interaction patterns by suggesting breaks or adjusting content exposure to prevent compulsive use [32]. Others, modify the platform's algorithms to balance immediate engagement with long-term user satisfaction. For example, Agarwal et al. [5] proposed the System-2 Recommender, which differentiates between impulsive (System-1) and deliberate (System-2) decision processes to reduce addictive behaviors while preserving meaningful engagement. [19] models interactions between individual users and recommendation systems using a dual-system reinforcement learning framework. By pairing modules that balance engagement and session moderation, this approach reduces addiction-like behaviors without sacrificing meaningful engagement, illustrating how careful algorithmic design can mitigate harmful overuse.

Collectively, these findings position SMA at the intersection of behavioral science and algorithmic design, emphasizing that addictive patterns in social media are emergent properties of both user psychology and multi-user recommender dynamics. Understanding and modeling these interactions is critical for designing interventions that address not only individual vulnerability but also systemic amplification of compulsive engagement.

3 METHODOLOGY

3.1 Base Framework and Extensions

This work builds upon the environment and agent architecture proposed by Livella, Bolis et al. [19], which models the interaction between a reinforcement learning (RL) user agent and a recommender system (RS). In the original framework, the user is modeled as a dual-system agent (combining Model-Free and Model-Based mechanisms) navigating a state space comprising *Healthy*, *Neutral*, *RecShort*, *RecLong*, and *Aftereffects* states.

Regarding the recommender system, the PutIn–PutOut Recommender System [19] comprises two modules: one optimized to capture user attention and engagement ("Put In") and another designed to promote healthy disengagement ("Put Out"), with adaptations handled via a non-stationary multi-armed bandit framework that emphasizes recent user interactions. In the PutIn configuration, both modules operate in Put In mode, whereas in the PutOut configuration, the first module operates in Put In mode to attract users, and the second operates in Put Out mode to encourage disengagement from prolonged sessions.

While Livella, Bolis et al. [19] focused primarily on single-agent trajectories, this research extends the framework to a **multi-user simulation environment**. This extension addresses the system's scalability and captures the dynamics that emerge when multiple users interact simultaneously.

3.2 Multi-User Preference Modeling

To simulate a realistic social media population, we introduce the parameters N_{users} and N_{groups}. The population is partitioned into distinct groups, where users are assigned to groups in a round-robin fashion (e.g., $g_1, g_2, g_3, g_1 ...$).

Each group $k \in \{1, \ldots, N_{groups}\}$ possesses a unique preference profile defined by a boolean preference mask $m_k \in \{0, 1\}^A$, where A denotes the number of available arms (content types). This mask governs the reward generation:

- **Preferred Content ($m_{k,a} = 1$):** Rewards are sampled from a Gamma distribution $\Gamma(\alpha_{high}, \theta_{high})$ with a high expected value, representing elevated immediate user satisfaction.
- **Non-Preferred Content ($m_{k,a} = 0$):** Rewards are sampled from a Gamma distribution $\Gamma(\alpha_{low}, \theta_{low})$ with a lower expected value.

This probabilistic modeling ensures that a content category a_i yielding a high reward for Group A may yield a low reward for Group B, creating conflicting signals for the RS.

3.3 Recommender Interaction Variants

We implemented three distinct interaction architectures to evaluate how the scope of the RS impacts learning efficiency:

(1) **Non-Discriminant (Global Shared RS):** All N_{users} interact with a single RS instance. The RS updates its policy based on feedback from all users simultaneously, regardless of group affiliation. This variant stress-tests the algorithm against heterogeneous feedback and it's expected to perform worse, since it must learn a single policy over heterogeneous and potentially conflicting preferences.
(2) **User-Discriminant (Personalized RS):** Each user u_i is assigned a private RS instance. The RS learns exclusively from the interaction history of that specific user, effectively bypassing group dynamics. This represents perfect personalization but suffers from data limitation. Performance is expected to be better than in Variant 1
(3) **Group-Discriminant (Clustered RS):** Users belonging to the same preference group share a specific RS instance. These group-level recommenders are expected to perform better than the per-user recommenders in Variant 2, because they can leverage more interaction data to learn the common structure of group preferences, leading to faster and more stable learning.

4 EXPERIMENTS

4.1 Experimental Setup

To assess the scalability and adaptability of the proposed architectures, we conducted simulations across the three different recommender variants and two distinct populations: popADD and popNOSM.

Parameter Settings. The simulation proceeds in discrete timesteps where all users act synchronously. We configured the environment with 16 users partitioned into 4 distinct preference groups. The reward distributions were set based on the population considered:

- Population ADD (popADD): Users who tend to find most content highly engaging, which makes it more challenging for the system to encourage balanced or healthy usage. Each arm has a likelihood of 80% of being related to a preferred content.
- Population NOSM (popNOSM): Users who generally perceive content as less engaging, making it easier for the recommender to steer them toward healthier behavior patterns. Each arm has a likelihood of 20% of being related to a preferred content.

Evaluation Metrics. We compared the three interaction variants (Global, Personalized, Group-Based) defined in the previous section. Agent behaviors are categorized following the approach in [19] into four classes: Healthy, Balanced, Addicted and Uncertain. The evolution of user behavior over time is illustrated in plots, with distinct colors representing each class: green for Balanced, blue for Healthy, red for Addicted and gray for Uncertain. The plots display the average number of users exhibiting each behavior at each timestep, while the shaded areas indicate the standard deviation.

5 RESULTS

In the following subsections, we examine: (i) how a shared recommender system performs when used by individuals with different preferences in comparison to a user specific recommender and (ii) how its performance changes when, instead of relying on a single recommender for all users, separate recommenders are employed for distinct groups of users who share similar preferences.

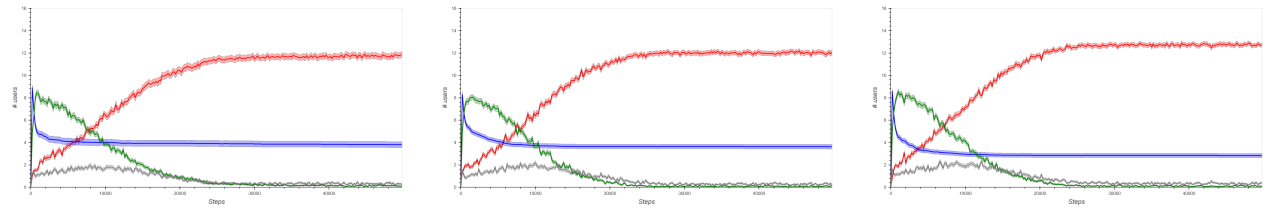
5.1 Shared Recommender System Across Users with Different Preferences

Figure 1 and Figure 2 illustrate the performance of popADD under three different configurations: a shared recommender applied to the entire population (left), an individual recommender tailored to each user (center), and a group-based recommender assigned to users with similar preferences (right).

A notable pattern emerges under the shared configuration. As shown in Figure 1b, agents develop addiction more rapidly, as indicated by the fact that the red line (addicted users) is higher that the former case. Furthermore, Figure 2b clearly shows that, in the shared setting, a subset of users develops addiction, while in the individual case this subset is almost empty. An

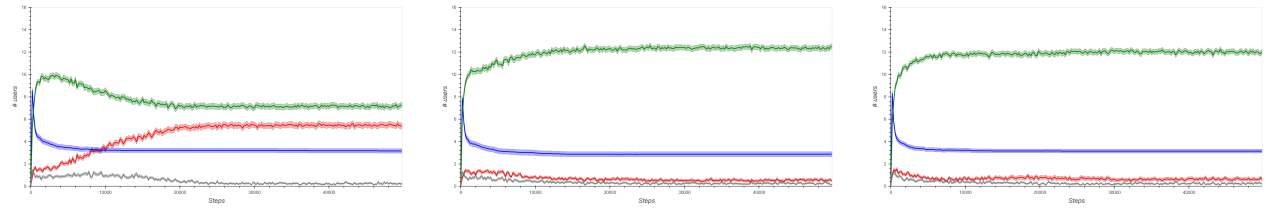
Similar dynamics are observed when the popNOSM population is considered, as shown in Figure 3 and Figure 4. In this case, the majority of the population develops healthy behavior, reflecting the way the population is defined. However, the number of users exhibiting addicted behavior is higher when the individual recommender is adopted.

One explanation for this dynamic is that when a single, unique recommender is considered, it cannot fully capture each user's preferences and adapt accordingly, leading to situations where the most engaging content for each user is not always recommended.



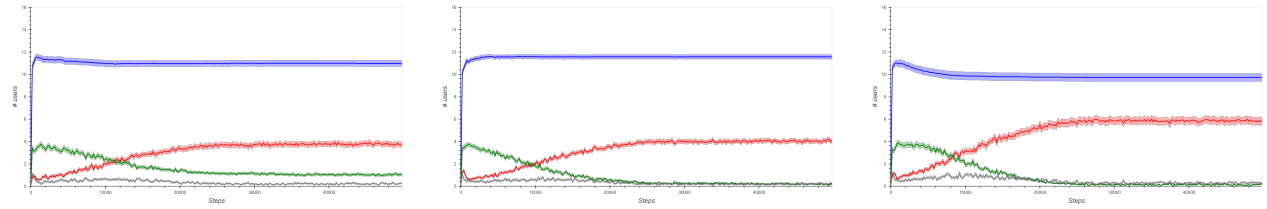
(a) Shared (b) Individual (c) Group

Figure 1: Results for $\beta = 0.5$, with learning rates $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$. System: PutIn, Pop:Addicted. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.).



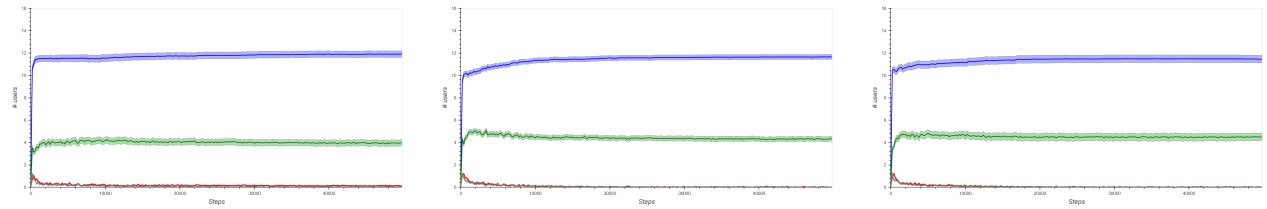
(a) Shared (b) Individual (c) Group

Figure 2: Results for $\beta = 0.5$, with learning rates $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$. System: PutOut, Pop:Addicted. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.).



(a) Shared (b) Individual (c) Group

Figure 3: Results for $\beta = 0.5$, with learning rates $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$. System: PutIn, Pop:Healthy. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.).



(a) Shared (b) Individual (c) Group

Figure 4: Results for $\beta = 0.5$, with learning rates $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$. System: PutOut, Pop:Healthy. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.).

5.2 Performance of Group-Specific Recommender Systems

Contrary to the previous case, when a single recommender is used *within* each group of users with similar preferences, users tend to develop addictive behavior more easily. In this configuration, the cause is not mis-personalization but rather *correct generalization*: because users in the same group genuinely share preferences, the recommender's estimates become more accurate more quickly, and it can more effectively exploit the most reinforcing content for that group. As a result, under PUT-IN, addiction emerges both *faster* and at *higher* prevalence than in the individual recommender scenario, as shown by the elevated red curves in Figure 1c. This effect is consistent with the intuition that pooling data among truly similar

users increases sample efficiency and accelerates convergence toward engagement-maximizing policies, thereby amplifying the risk of overusage relative to fully individualized learning.

Importantly, PUT-OUT remains robust under the same generalization regime. While group-based recommendation still increases the tendency toward compulsive use compared to individualized models, Figure 2c shows that the intervention drives a substantially larger share of users toward balanced behavior and reaches the corresponding plateau more rapidly than in the PUT-IN setting. In other words, even when generalization is *beneficial* for recommendation accuracy (because users are truly similar), PUT-OUT is comparatively effective at counteracting the resulting exploitation dynamics and reducing excessive engagement.

Taken together, these findings reinforce that the impact of recommenders on problematic usage is not uniform: the degree and *correctness* of cross-user generalization can be a primary driver of how quickly overuse develops, and mitigation strategies should therefore be evaluated under both mis-generalization (shared models across heterogeneous users) and correct generalization (shared models within homogeneous groups) regimes.

6 DISCUSSION

As expected, when the recommender uses PutIn, Variant 3 outperforms Variant 2 in increasing the proportion of addicted (red) users. When comparing Variant 1 with Variants 2 and 3, the recommender appears more confused, as it does not effectively discriminate between user types. This is evidenced by the observation that, under PutIn, some addicted users transition to a balanced state, while under PutOut the opposite occurs, with balanced users becoming addicted.

This framing makes the causal story easier to follow: Variant 3 behaves like a more selective filter, while Variant 1 acts more like a blunt instrument that perturbs users in both directions.

7 CONCLUSIONS

Our simulations indicate that the prevalence and intensity of problematic social media usage are not solely a function of an individual's vulnerability or the recommender's objective, but vary materially with *how* the recommender generalizes across users and items: stronger cross-user generalization can amplify exposure to engagement-maximizing content and, in turn, increase compulsive consumption for otherwise comparable individuals. At the same time, the proposed putout-based intervention remains effective across a wide range of generalization settings, continuing to mitigate addictive dynamics in a comparatively robust manner even when population-level signals distort the estimation of individual preferences. Taken together, these results suggest that auditing and governance efforts should evaluate not only user-specific personalization effects, but also the recommender's cross-user generalization regime as a key driver of systemic risk and as a critical condition for the reliability of mitigation strategies—aligning with emerging regulatory expectations around recommender transparency, systemic risk assessment, and independent auditing [1–4], as well as established frameworks for algorithmic auditing and impact assessment [15, 24, 25, 28, 34].

ACKNOWLEDGMENTS

This research was supported by the Italian Ministry of University and Research under Grant No. 2023-NAZ-0206, PsyFuture – Dipartimento di Eccellenza 2023-2027 and by Volkswagen Foundation OpenUp Grant Ref. 9E530 Developing an Artificial Social Childhood (ASC).

REFERENCES

[1] 2022. Regulation (EU) 2022/2065 (Digital Services Act) — Article 27: Recommender system transparency. EU legal text (public consolidation/portal). https://www.eu-digital-services-act.com/Digital_Services_Act_Article_27.html Accessed 2026-02-19.

[2] 2022. Regulation (EU) 2022/2065 (Digital Services Act) — Article 34: Risk assessment. Digital Services Act (consolidated text portal). https://www.eu-digital-services-act.com/Digital_Services_Act_Article_34.html Accessed 2026-02-19.

[3] 2022. Regulation (EU) 2022/2065 (Digital Services Act) — Article 37: Independent audit. EU legal text (public consolidation/portal). https://www.eu-digital-services-act.com/Digital_Services_Act_Article_37.html Accessed 2026-02-19.

[4] 2024. Digital Services Act — auditing very large online platforms and search engines. EUR-Lex summary. https://eur-lex.europa.eu/EN/legal-content/summary/digital-services-act-auditing-very-large-online-platforms-and-search-engines.html Accessed 2026-02-19.

[5] Arpit Agarwal, Nicolas Usunier, Alessandro Lazaric, and Maximilian Nickel. 2024. System-2 Recommenders: Disentangling Utility and Engagement in Recommendation Systems via Temporal Point-Processes. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1763–1773.

[6] Mohamed Basel Almourad, John McAlaney, Tiffany Skinner, Megan Pleya, and Raian Ali. 2020. Defining digital addiction: Key features from the literature. *Psihologija* 53, 3 (2020), 237–253.

[7] Jashvini Amirthalingam and Anika Khera. 2024. Understanding social media addiction: A deep dive. *Cureus* 16, 10 (2024).

[8] Elombe Calvert, Maddalena Cipriani, Bridget Dwyer, Victoria Lisowski, Jane Mikkelson, Kelly Chen, Matthew Flathers, Christine Hau, Winna Xia, Juan Castillo, Alex Dhima, Sean Ryan, and John Torous. 2025. Social Media Detox and Youth Mental Health. *JAMA Network Open* 8, 11 (2025), e2545245. https://doi.org/10.1001/jamanetworkopen.2025.45245

[9] Keyi Chen. 2024. If it is bad, why don't I quit? Algorithmic recommendation use strategy from folk theories. *Global Media and China* 9, 3 (2024), 344–361. https://doi.org/10.1177/20594364231209354 arXiv:https://doi.org/10.1177/20594364231209354

[10] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2022. The effect of people recommenders on echo chambers and polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 90–101.

[11] Cynthia A. Dekker, Susanne E. Baumgartner, and Sindy R. Sumter. 2025. For you vs. for everyone: The effectiveness of algorithmic personalization in driving social media engagement. *Telematics and Informatics* 101 (2025), 102300. https://doi.org/10.1016/j.tele.2025.102300

[12] Luisa Fassi, Amanda M. Ferguson, Andrew K. Przybylski, Tamsin J. Ford, and Amy Orben. 2025. Social media use in adolescents with and without mental health conditions. *Nature Human Behaviour* 9, 6 (2025), 1283–1299. https://doi.org/10.1038/s41562-025-02134-4

[13] Kamil Filipek. 2025. Captive Platforms: On the Algorithmic Loops of Infinite Scrolling. *Addiction & Social Media Communication* 2, 2 (2025), 21–37.

[14] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149.

[15] Sabrina Guidotti, Gregor Donabauer, Davide Taibi, Giuseppe Vizzari, Udo Kruschwitz, and Dimitri Ognibene. 2026. Toward Recognizing Social Media Recommenders under Absent Recommendations: A Graph Neural Network-based Approach. In *Proceedings of 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*. https://doi.org/10.65109/YMXQ7472

[16] Zaheer Hussain and Mark D Griffiths. 2018. Problematic social networking site use and comorbid psychiatric disorders: A systematic review of recent large-scale studies. *Frontiers in psychiatry* 9 (2018), 686.

[17] Ayaka Kato, Kanji Shimomura, Dimitri Ognibene, Muhammad A Parvaz, Laura A Berner, Kenji Morita, and Vincenzo G Fiore. 2023. Computational models of behavioral addictions: State of the art and future directions. *Addictive behaviors* 140 (2023), 107595.

[18] Björn Lindström, Martin Bellander, David T Schultner, Allen Chang, Philippe N Tobler, and David M Amodio. 2021. A computational reward learning account of social media engagement. *Nature communications* 12, 1 (2021), 1311.

[19] Stefano Livella, Luca Bolis, Sabrina Patania, Matteo Papini, and Dimitri Ognibene. 2026. Mitigating Problematic Social Media Use through Paired Recommender

Systems with Contrasting Objectives. In *Proceedings of the 25th International Conference on Autonomous Agents and Multi-Agent Systems*.

[20] Natasha Lomas. 2017. Google to ramp up AI efforts to ID extremism on YouTube. *TechCrunch* 24 (2017), 2019. https://techcrunch.com/2017/06/19/google-to-ramp-up-ai-efforts-to-id-extremism-on-youtube/

[21] Francesco Lomonaco, Davide Taibi, Vito Trianni, Sathya Buršić, Gregor Donabauer, and Dimitri Ognibene. 2022. Yes, echo-chambers mislead you too: a game-based educational experience to reveal the impact of social media personalization algorithms. In *International Workshop on Higher Education Learning Methodologies and Technologies Online*. Springer, 330–344.

[22] Sanzana Karim Lora, Sadia Afrin Purba, Bushra Hossain, Tanjina Oriana, Ashek Seum, and Sadia Sharmin. 2025. Infinite Scrolling, Finite Satisfaction: Exploring User Behavior and Satisfaction on Social Media in Bangladesh. arXiv:2408.09601 [cs.HC] https://arxiv.org/abs/2408.09601

[23] Simona Mandile. 2025. *The Dark Side of Social Media: Recommender Algorithms and Mental Health.* Technical Report 11648. CESifo Working Paper Series. Uses the 2016 rollout of Instagram's algorithmic feed as a quasi-experiment to assess mental health impacts.

[24] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. https://doi.org/10.1145/3442188.3445935

[25] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Special Publication (PDF). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf Accessed 2026-02-19.

[26] Dimitri Ognibene, Vincenzo G Fiore, and Xiaosi Gu. 2019. Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. *Neural Networks* 116 (2019), 269–278.

[27] Dimitri Ognibene, Rodrigo Wilkens, Davide Taibi, Davinia Hernández-Leo, Udo Kruschwitz, Gregor Donabauer, Emily Theophilou, Francesco Lomonaco, Sathya Bursic, Rene Alejandro Lobo, et al. 2023. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence* 5 (2023), 654930.

[28] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*. 33–44. https://doi.org/10.1145/3351095.3372873

[29] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[30] Kevin Roose et al. 2020. Rabbit hole. *The New York Times* (2020). https://www.nytimes.com/column/rabbit-hole

[31] Amy Ross Arguedas, Craig Robertson, Richard Fletcher, and Rasmus Nielsen. 2022. Echo chambers, filter bubbles, and polarisation: A literature review. (2022).

[32] Eden Saig and Nir Rosenfeld. 2023. Learning to suggest breaks: sustainable optimization of long-term user engagement. In *International Conference on Machine Learning*. PMLR, 29671–29696.

[33] Arianna Sala, Lorenzo Porcaro, and Emilia Gómez. 2024. Social Media Use and adolescents' mental health and well-being: An umbrella review. *Computers in Human Behavior Reports* 14 (2024), 100404. https://doi.org/10.1016/j.chbr.2024.100404

[34] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. ICA Preconference paper (PDF). https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf Accessed 2026-02-19.

[35] Burrhus Frederic Skinner. 1965. *Science and human behavior*. Number 92904. Simon and Schuster.

[36] Emily Theophilou, Francesco Lomonaco, Gregor Donabauer, Dimitri Ognibene, Roberto J Sánchez-Reina, and Davinia Hernàndez-Leo. 2023. AI and narrative scripts to educate adolescents about social media algorithms: insights about AI overdependence, trust and awareness. In *European conference on technology enhanced learning*. Springer, 415–429.

[37] Jiaxuan Wang and Xunpei Zhang. 2023. The Reinforcements and Punishments in Social Media Addiction. *Journal of Education, Humanities and Social Sciences* 8 (02 2023), 1460–1464. https://doi.org/10.54097/ehss.v8i.4503

[38] Helena Webb, Pete Burnap, Rob Procter, Omer Rana, Bernd Carsten Stahl, Matthew Williams, William Housley, Adam Edwards, and Marina Jirotka. 2016. Digital wildfires: Propagation, verification, regulation, and responsible innovation. *ACM Transactions on Information Systems (TOIS)* 34, 3 (2016), 1–23.