Modelling and Mitigating Problematic Social Media Use through Paired Recommender Systems with Contrasting Objectives

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

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

Sabrina Patania
University of Milan-Bicocca
Milan, Italy
sabrina.patania@unimib.it

Matteo Papini
University of Milan
Milan, Italy
matteo.papini@unimi.it

Kenji Morita
University of Tokyo
Tokyo, Japan
morita@p.u-tokyo.ac.jp

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

ABSTRACT

Social media platforms can provide meaningful connection and entertainment, but their engagement-optimizing recommendation algorithms may elicit compulsive overuse and undermine users' long-term well-being.

Unlike engagement-, content- and scaling-focused recommender system studies, addressing this phenomenon requires analyzing the interaction between algorithmic dynamics and users' cognitive and individual difference. We build on computational neuroscience and adopt a well-established dual reinforcement learning framework for modeling addictive behavior. This framework allows us to represent key individual traits, such as impulsivity and preference structure, through different parameterizations (endophenotypes). Since both the user and the recommender are adaptive learning agents, their interaction makes the system non-Markovian. This adds complexity to the interpretation of system dynamics compared to classical addiction models and calls for more effective recommender strategies.

We propose a paired-training modules recommender system aimed at promoting balanced usage. One module maximizes engagement, while the other gently discourages excessively prolonged sessions. While they have contrasting objectives they share training experience to enable effective adaptation while the users adapt their own strategies. We compare it against two baselines: (i) a classical engagement-maximizing recommender system, and (ii) an engagement-maximizing recommender system augmented with a well-being–aware mechanism.

Results with 200 trajectories per endophenotype show that conventional systems are more likely to induce addiction-like patterns. In contrast, our proposed models, especially the paired-training architecture, significantly reduce such behaviors without sacrificing overall engagement. These findings suggest that addictive behavior may not be an unavoidable outcome of sustainable recommendation systems, but one that can be mitigated through careful recommender system design.

KEYWORDS

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

1 INTRODUCTION

Social media platforms have transformed the way people communicate and consume information, but they also raise significant concerns about compulsive use and addiction. Unlike pharmacological addiction, which is typically based on substances, social media overuse arises from cognitive limitations, such as the difficulty in balancing immediate rewards with long-term well-being and the powerful influence of recommendation systems designed to maximize engagement at any cost [17, 25, 26]. Even as total time on these platforms continues to increase, longitudinal evidence indicates that user-reported life satisfaction and subjective well-being decline over time, an emerging engagement–utility gap documented both in academic studies (e.g. Facebook use predicts lower life satisfaction within a matter of days) and acknowledged by platform officials in 2017, when Facebook and YouTube admitted that their engagement-maximizing algorithms could leave users feeling worse despite growing usage [23, 31, 33].

This growing issue is not only a technical or behavioral phenomenon, but a recognized public health concern, particularly among adolescents and young adults [36]. Excessive use has been associated to depression, anxiety and stress, with studies such as Hussain and Griffiths [16] highlighting correlations between problematic social network use and psychiatric disorders. To better understand the underlying causes of this problematic behavior, Wang and Zhang [41] provide a compelling analysis grounded in operant conditioning theory [38], showing how social media platforms intentionally exploit reinforcement (e.g. likes, comments) and punishment (e.g. negative feedback, exclusion) to manipulate user behavior. Notably, they highlight that even negative stimuli, such as hostile comments, can function paradoxically as reinforcements by triggering users' desires to respond or seek validation, causing them to use social media more often and become more attached to it.

Dual-system RL, integrating model-free (habitual) and model-based (goal-directed) control, offers a principled framework to formalize decision-making vulnerabilities and compute the habitual/goal-directed balance across individuals [9, 28, 30]. [13] further showed that treating the relative weighting of these two subsystems as a fixed endophenotype captures wide intersubject variability in decision styles and addiction risk. Additionally, RL can also be used in recommender systems to provide personalized suggestions by learning optimal policies from user feedback. More specifically, RL-powered

recommender systems can engage users with recommendations tailored to their behavior. [19, 26].

Importantly, while addictive social media use is harmful, complete abstinence is not necessarily ideal. Moderate use provides benefits like connection, information and support, suggesting balanced engagement is healthiest. By extending RL addiction models to include explicit representations of recommender systems and the induced non-Markovian dynamics, it is possible to study how algorithmic interventions influence user behavior.

2 RELATED WORKS

Social media addiction (SMA) is characterized by the excessive and compulsive use of social platforms that disrupts daily functioning and undermines overall well-being [2, 3]. SMA shares key symptomatologies with behavioral addictions, including salience, tolerance, withdrawal, and relapse, as well as deficits in self-regulatory control that sustain compulsive engagement despite negative consequences [4, 6, 7, 39]. Although SMA is not formally recognized as a behavioral addiction in current diagnostic classifications, many studies highlight its overlap with established behavioral disorders, leading some authors to prefer the term *problematic social media use* to emphasize maladaptive patterns without implying a distinct clinical condition [5, 35]. Nevertheless, the strong similarity in symptomatologies and reinforcement dynamics suggests that frameworks developed for behavioral addictions can provide valuable insights into understanding and modeling excessive use.

Empirical evidence supports this view. Lindström et al. [20] found that posting frequency and engagement on over one million posts from over four thousand users follow reward-learning dynamics, with likes and comments reinforcing behavior, suggesting social media engagement operates like a reinforcement-learning process similar to behavioral addictions. Complementing this line of research, experimental findings highlight a gap between engagement and attitudes. A large-scale field experiment on X found that disabling algorithmic ranking reduced user engagement without affecting political opinions or polarization [14], suggesting that algorithms shape behavior primarily through interaction patterns rather than immediate attitudinal change.

Building on this connection, reinforcement learning (RL) models of addiction provide a computational framework to explain how maladaptive decision patterns emerge and persist despite negative outcomes. Early RL accounts proposed that addiction arises when habitual, model-free control, dominates behavior due to the biochemical hijacking of dopaminergic prediction error signals, leading to rigid stimulus-response habits insensitive to punishment or changing contingencies [10, 12, 29]. However, these models failed to capture more complex features of addiction such as craving, goal-directed drug seeking and relapse in the absence of direct cues [30, 37]. Subsequent work by Ognibene, Fiore and Gu [25] introduced a Dual Reinforcement Learning (Dual-RL) framework that integrates both model-free (habitual) and model-based (deliberative) components, representing the balance between automatic reward-driven actions and cognitive planning. Originally developed to simulate gambling disorder, a prototypical behavioral addiction, this approach demonstrates how limited cognitive resources and environmental complexity can lead to suboptimal policies—a defining feature of addictive behavior—while also explaining addiction-like patterns in the absence of pharmacological effects. Because it accounts for both impulsive and reflective decision processes within bounded rationality, the Dual-RL model provides a strong theoretical basis for simulating user behavior in environments such as social media platforms.

Growing awareness of the harms of engagement-maximizing algorithms has stimulate efforts to create more ethical, human-centered social platforms. These approaches either target users, e.g. suggesting breaks or moderating content to support healthy engagement [34], or adjust platform algorithms to balance short-term engagement with long-term well-being. For example, Agarwal et al. [1] proposed the System-2 Recommender Framework, which separates impulsive (System-1) from deliberate (System-2) decision-making to prioritize meaningful engagement over mere attention. Overall, this research reflects a shift toward value-aligned recommendation systems that protect user autonomy and health.

Prior work has extensively explored engagement-maximizing recommenders, including RL and bandit-based approaches; our non-stationary multi-armed bandit design builds on these foundations [40]. We extend this paradigm toward balanced (well-being-aware) usage, in line with recent efforts to optimize utility beyond short-term engagement (e.g., [1] System-2 Recommenders and user-centric "time well spent" approaches).

3 METHODS

We define a multi-agent system consisting of a user and a recommender system operating within a shared environment of psycho-physical states.

Formalization. The interaction between the user and the recommender can be formalized as a two-player general-sum Markov game [21] $(\mathcal{S}, \mathcal{A}_{1,2}, P, R_{1,2})$ where $\mathcal{S}$ denotes the joint state space (coinciding with the state of the user), $\mathcal{A}_1$ the set of actions of the user, $\mathcal{A}_2$ the set of actions of the recommender, $P : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{S} \to [0, 1]$ is the Markovian transition kernel assigning the probability of the next state given the current state and the actions of the two agents, $R_1 : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$ is the reward function of the user, and R_2 is the reward function of the recommender.[1] A peculiar aspect of this game is that the second player (the recommender) only acts in a subset of the state space, representing the situation in which the user is interacting with the social network. As a result, the recommender's reward is *sparse*, and the level of sparsity depends on the user's policy. This can make it challenging for the recommender to acquire meaningful feedback and adapt to the user's behavior. In this paper we focus on modeling the user in a realistic way that captures addiction phenomena, and on designing a simple proof-of-concept strategy for the recommender showing that *healthy* engagement optimization is possible. However, the "addiction game" introduced here may be of independent interest both theoretically and as a benchmark for multi-agent RL with sparse rewards [22].

[1] In our PutIn-PutOut system described below, the recommender is actually composed of two agents, for a total of three agents. This can be formalized as a 3-player Markov game where the two recommender agents have identical action spaces but distinct reward functions.

3.1 User Modeling

The user is represented by a well-established dual-RL system [30] combining a Model-Free (MF) approach, implementing habitual behaviors using Q-learning [42] to maximize cumulative rewards based on past experiences, and a Model-Based (MB) approach that simulates planning and problem solving, using Prioritized Sweeping [24], leveraging an internal representation of the environment for goal-directed decision making.

The Dual-RL model system comprises two components, Model-Free (MF) and Model-Based (MB), that cooperate to identify the optimal policy. The impact of MF and MB on the expected rewards is determined by a parameter β that reflects the balance between habitual behaviors and more reflective goal-directed thinking. A higher value of β indicates a tendency to rely more on the MB system, promoting more conscious decision-making and planning based on past experiences or foresight. On the other hand, a lower value indicates greater dependence on the MF system, where decisions are guided by well-practiced habits and immediate reactions to environmental cues. This dynamic mirrors the psychological competition between impulsive action and deliberate reasoning.
The Q-Values for the Dual-RL model (Q_{MX}) are calculated as:

$$Q_{MX}(s,a) = \beta Q_{MB}(s,a) + (1-\beta)Q_{MF}(s,a).$$

An additional key factor influencing the user's cognitive profile is the MBUS (Model-Based Updates per Step) parameter. This parameter defines the number of Q-Value updates performed during each iteration. A higher value increases the time required to complete the Q-Value update operation but results in more reliable estimates and an improvement in performance.

3.2 Recommender System Modeling

This paper presents two distinct architectures. The first, the *PutIn - PutOut Recommender System*, consists of two modules designed to: first capture user attention and, then, gradually encourage disengagement to prevent prolonged sessions. The second architecture builds on the first by extending its functionality, updating both modules in response to user interactions with the recommender system, rather than updating only the module directly engaged by the user.

3.2.1 PutIn - PutOut Recommender System. The recommender system examined in this study is composed of two distinct components: one tailored for short interactions (*Rec Short*) and the other optimized for extended user sessions (*Rec Long*). The first component operates in *"Put In"* mode, which aims to capture and retain user attention, while the second component operates in *"Put Out"* mode, which prioritizes user well-being by subtly encouraging disengagement and promoting balanced usage. Short-term interactions are associated with moderate rewards and low penalties, reflecting the less impactful nature of brief engagement. In contrast, long-term sessions offer higher rewards, but they also carry the risk of negative effects over time. To address this, the environment incorporates a penalty loop that users must pass through after exiting the Rec Long state. This design effectively captures the complex interactions between users and the system, highlighting how recommendations can gradually influence the user behavior over time.

The system is implemented using a non-stationary multi-armed bandit (MAB) framework [40], enabling it to dynamically adapt to evolving user preferences by balancing exploration and exploitation. In this model, each *arm* (i.e. action of the system) represents a possible content recommendation, and rewards are derived from user interactions.
For the *Put In* configuration, the goal is to maximize engagement: a positive reward (+1) is assigned when a user accepts a recommendation, and a negative reward (-1) is given for rejections. Conversely, the *Put Out* configuration is designed to promote healthy disengagement, receiving positive rewards (+1) when users discard the recommended content.

This adaptive structure is crucial for accurately capturing user behavior, which tends to vary over time. To account for these non-stationary dynamics, the system uses an exponentially weighted average that emphasizes recent interactions while gradually discounting the influence of the older ones [44]. This mechanism enables the recommender to effectively adjust to temporal shifts in user preferences.

3.2.2 Paired-training recommender system. The previous system introduces non-trivial interactions between the two modules, (*Rec Short* and *Rec Long*), and users.

By observing a user's interaction with one module, the recommender system can infer whether the proposed content was appreciated or not. This feedback is assumed to carry meaningful information for both modules, and the Q-values in the MAB models are updated accordingly, as explained in the previous section.

3.3 Environment Design

In reinforcement learning, the environment is typically modeled as a stochastic function that maps state-action pairs to subsequent states and associated rewards [40], with its complexity depending on the size of the state and action spaces. The states of our environment shown in Figure 1 are categorized into five distinct groups:

- **Healthy:** these are the states that represent the user's physical and psychological well-being. From these states, it is possible to perform only *aG* actions that guarantee positive rewards. As a result, the user perceives these states as a healthy state.
- **Neutral:** these are the states in which the psycho-physical state of the user is not modified. Thus, these are states in which the available actions lead the user to perceive neither positive nor negative rewards.
- **Recommender System:** these are states that emulate the use of the social network, that is, they represent the interaction between the user and the recommender system. In these states, the user can decide whether or not to consume the content proposed and, based on the decision made, reaches different states and obtains different rewards.
  Since different types of social media usage, short and intense, produce different effects on users, it is important to differentiate between these two scenarios. Therefore, this group of states is divided into two subsets:
  - **Recommender Short (RS):** These are the states that represent short usage of social media. Moderate use results in low reward, but does not cause any particular negative

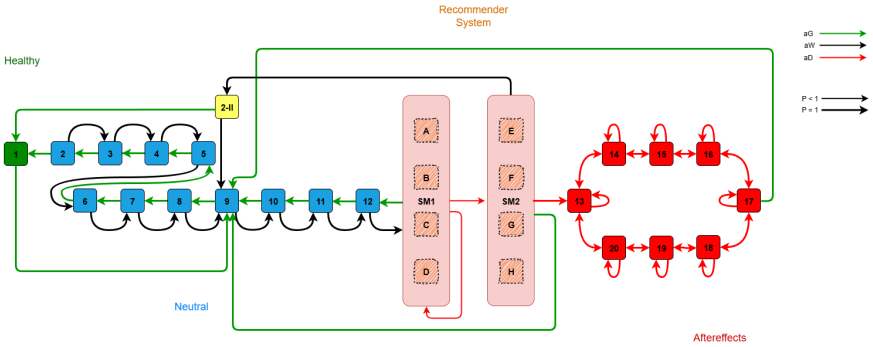

Figure 1: The environment includes Healthy (green), Neutral (blue), Recommender (pink), Balanced (yellow) and Aftereffects (red) states. Actions are aG (well-being, green), aW (wait, black), and aD (social media interactions/penalties, red).

effects; therefore, negative rewards are also small. However this kind of session may be the gateway towards longer sessions with bigger negative consequences that we could define as doomscrolling or mindless scrolling whose recommendations are handled by RecLong. For example, answering a notification received from a friend may lead to longer social media usage sessions.

  - **Recommender Long (RL):** These are the states that represent the prolonged use of social media platforms. Heavy use allows for a large reward initially, but it also exposes the user to inevitable aftereffects.
- **Aftereffects:** these are states that emulate aftereffects, i.e., the negative effects due to social media use, specifically modeling the occasions lost due to the time wasted online. Indeed, prolonged use of platforms can have repercussions on the user's personal and professional opportunities, such as time not dedicated to study or work or time not devoted to physical or recreational activities.
- **Balanced:** this state allows the user to use the social media in a balanced way. It is a shortcut from state Rec Short to state Healthy without paying major consequences.

Users can perform three different actions:

- **aG** *(Action Goal):* aims to model those behaviors that improve the user's physical and psychological well-being, for example, spending time with friends and family or performing tasks to free oneself from daily responsibilities.
- **aW** *(Action Wait):* models the action of waiting, representing situations in which the agent does not perform any specific action that improves or worsens its current psycho-physical state.
- **aD** *(Action Drug):* emulates the use of social media and, in particular, the consumption of content proposed by the recommender system. It also models the penalties associated with such consumption, representing the user's subsequent actions as a consequence of prolonged social media engagement.

3.4 Environment discussion

Environment design is inspired by a well-established environment previously explored in the study of behavioral addiction [25]. This section outlines some key implementation choices, while a more detailed and formal definition of the environment can be found in the Supplementary Material.

The *Aftereffects* states are interconnected, forming a penalty loop that models the negative consequences of overuse, a design choice that reflects real-world concerns such as the time-wasting effects of excessive social media use. Agents can only exit this loop from a specific *Aftereffects* state, the one from which the green arrow originates. From here, agents have a 60% probability of escaping by taking the *aG* action, mirroring how users might find it difficult to break free from overexposure.

The *Neutral* states (2 to 12) form the core progression path of the environment, linking the agent's actions toward a healthy outcome or into interaction with the recommender system. This structure is intentionally designed to increase the complexity of discovering the optimal policy.

Within the system, two special set of states, RecShort and RecLong, represent the interaction with the recommender. RecShort is accessible from the final neutral state via the *aW* action, while RecLong is reached from RecShort by performing the *aD* action with a probability of 50%. This probabilistic nature reflects the unpredictability of real-world user behavior when responding to system recommendations, as performing the *aD* action in RecShort may even cause a self-loop without progressing to the RecLong session. The rewards in these states are dynamic, determined by the internal logic of the recommender system, which operates as a black box from the user's perspective. A key design decision is the inclusion of an escape point from the Rec Long, allowing agents to avoid consuming the content proposed by the system. This escape point serves as a way to balance the social media usage, much like how real users navigate their engagement with platforms to avoid excessive interaction.

The proposed design intentionally limits the number of escape routes (one in RecLong and one in Aftereffects), as adding multiple options would unnecessarily complicate the analysis. By setting the transition probability from RecShort to RecLong at 50%, the system strikes a balance between simplicity and challenge. This design decision ensures that agents, much like real-world users, are able to exhibit adaptive behavior, navigating between healthy and problematic usage patterns. This framework aligns with research

trends in social media addiction and the complex dynamics of user engagement with algorithm-driven systems [27].

4 EXPERIMENTS

To assess the performance of our models, we conducted 200 simulations, each consisting of 100,000 steps, for every parameter combination (i.e. different endophenotypes). Additionally, we employed a bootstrapping technique to enhance the robustness of our validation process. For each experiment, we randomly sampled 50 data from the group of 200, with a total of 100 resampling iterations. This approach, combining a large number of simulations with bootstrapping, was designed to ensure more reliable and generalizable results. The parameters explored include:

- **Beta values**, representing different levels of cognitive regulation in users. For example, a beta value of 1.0 (model-based only) reflects fully rational behavior, while a beta of 0.0 (model-free only) corresponds to instinct-driven or habitual behavior.
- **Agent rewards given by the recommender system**, intended to simulate different user populations:
  - *Population ADD*: Users for whom most contents are addictive, making it harder for the system to promote healthy use.
  - *Population NOSM*: Users for whom contents are generally non-addictive, making it easier for the recommender to guide them toward a healthy behavior.
  - *Population NTRL*: Users for whom contents are balanced between addictive and non-addictive.
- **Rec Short learning rates**: Study how adjusting the speed of adaptation to users' preferences (either faster or slower) can influence user behavior.

The parameter MBUS is fixed to 2 for all the experiments because the environment we are considering is simple, as we aim to avoid introducing unnecessary noise and complexity when studying the interactions between humans and the recommender system. Increasing the value of MBUS would not be beneficial, as agents can already learn the environment meaningfully with a small number of updates. Raising this value further would yield results very similar to those obtained with the current setting.

To benchmark our model, we tested it against an aggressive baseline architecture that includes both `RecShort` and `RecLong` components operating in `PutIn` mode. This configuration is designed to maximize the exploitation of the recommender system by the users, increasing the chances that users end up in the aftereffects area, thereby prioritizing user engagement at the expense of their well-being. Additionally, we consider also a second baseline, which combines an engagement-maximizing recommender system with a well-being–aware mechanism.

5 RESULTS

For each agent, behavior is assessed in segments of 200 steps. Within each segment, the agent is classified into one of four behavioral categories:

- **Healthy**: The agent predominantly follows the *Healthy policy*, which starts at state 9, moves through the neutral states to the healthy state, performs action *aG* and returns to the initial state 9.
- **Addicted**: The agent spends the majority of the time in the *Aftereffects* area.
- **Balanced**: The agent predominantly follows the *Balanced policy*, which starts at state 9, passes through the neutral states to reach Rec Short, Rec Long, the balanced state and then the healthy state. From here action *aG* is performed and the agent returns to the initial state 9.
- **Uncertain**: No single behavior is clearly dominant that is, no behavior occurs at least 20% more frequently than the others.

The behavioral categories "Addicted," "Balanced," and "Healthy" are not intended as clinical diagnoses, but rather as operational constructs aligned with RL-theoretic distinctions, e.g. distance from MDP optimal behaviour ("Healthy"). "Addicted" agents repeatedly select suboptimal actions that maximize short-term returns but prevent long-term value maximization, which corresponds to model-free dominance. This is consistent with cognitive science literature that links compulsive behaviors to the over-reliance on cached value systems at the expense of planning mechanisms [8],[11].

The results are visualized in plots using distinct colors: green for Balanced, blue for Healthy, red for Addicted and gray for Uncertain behavior. Plots show the average number of users following each behavior at every step, with the shaded areas representing the standard deviation.

In the following subsections, we will present: (i) how different endophenotypes exhibit distinct behaviors and develop different behavioral patterns, (ii) the comparison between the engagement-maximizing recommender system and the *PutIn - PutOut Recommender System*, (iii) the influence of PutIn - PutOut learning rate ratio on user behavior, (iv) a comparison of the performance of the *paired-trained recommender system* and the *PutIn - PutOut* version, (v) the comparison between the *engagement-maximizing recommender system augmented with a well-being-aware mechanism* and *paired-trained recommender system* and, finally, (vi) how these novel architectures scale as the size of the content set considered by the recommender increases.
Additional results across varying parameter settings, along with the full codebase and environment details are available at: https://github.com/DimNeuroLab/SocialMediaAddiction.

5.1 Endophenotypes and Behavioral Patterns

Figure 2a and Figure 2b illustrate the behavioral patterns developed by popADD and popNOSM, respectively, under the engagement-maximizing recommender system with $\beta = 0.5$. From these plots, it is clear that users in popADD, for whom most of the recommended content is addictive, can be divided into two groups: a very small subset of users who develop healthy behavior and the majority who develop addictive behavior. This phenomenon can be explained by the fact that the addictive condition is easily accesible, while the few healthy users never interact with the recommender system in early stages. As a result, they develop healthy behavior that protects them from falling into addiction. On the other hand, users from popNOSM exhibit a different pattern. Since they are less susceptible to the content recommended by the system, they begin to move

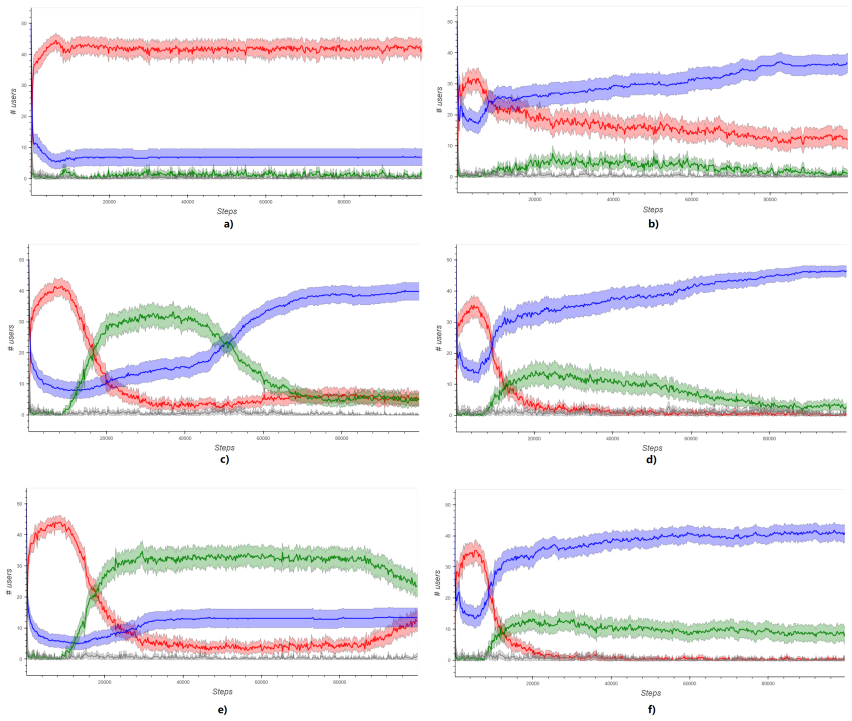

Figure 2: Results for $\beta = 0.5$, with learning rates $lr_{PutIn} = 0.01$ and $lr_{PutOut} = 0.01$. Left: popADD; Right: popNOSM. (a–b) Engagement-maximizing recommender. (c–d) PutIn–PutOut recommender. (e–f) PutIn–PutOut recommender with $lr_{PutIn} = 0.002$. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.

away from addiction and gradually adopt healthier behavior as the steps progress. In both cases, the number of users exhibiting uncertain or balanced behavior is negligible.

From this discussion, we can conclude that the proposed environment, the dual-RL system to model user behavior and the MAB model for the recommender system, allows users to exhibit different behaviors based on their individual parameters (i.e., endophenotypes) [25].

5.2 Engagement-Maximizing vs. PutIn-PutOut Recommender System

Simulations involving the engagement-maximizing recommender system show that users do not develop balanced behaviors, regardless of whether they belong to popADD, popNOSM, or popNTRL, as illustrated in Figure 2a and Figure 2b. This can be explained by the fact that users either fail to recognize the penalties associated with prolonged recommender system use (leading to addictive behaviors), or once they do realize the negative impact, they decide to completely disengage from the system.

The PutIn - PutOut Recommender System partially addresses this issue, as shown in Figure 2c and Figure 2d. From these plots, we can observe how users transition from addictive behaviors to more balanced behaviors. However, the limitation of this architecture is that after an initial stage where users exhibit balanced behaviors, they begin to develop healthier behaviors and eventually leave the platform. This happens because, with more time spent outside the Aftereffect area, users have the opportunity to explore different policies and eventually find the healthy one.

While this architecture helps solve the problem of user addiction, it introduces a new challenge: the platform becomes unsustainable as users gradually disengage

5.3 Influence of PutIn - PutOut learning rate ratio on user behavior

A key role in the problem described in the previous subsection is played by the learning rate of PutIn. Specifically, if PutIn learns too quickly to attract users on social media, it may fail to recommend contents that disengages users from prolonged sessions and their negative consequences. As a result, users might quit social media since they frequently fall into these extended sessions. Ideally, the optimal scenario is one where PutIn learns more slowly (i.e. with a lower learning rate) than PutOut, allowing users to consume content recommended by PutIn while avoiding the content from PutOut that leads to prolonged sessions.

This hypothesis is supported by simulation results, as shown in Figure 2. In Figures c and d, both PutIn and PutOut have a learning rate of 0.01, while Figures e and f show PutIn with a learning rate of 0.002 and PutOut with a learning rate of 0.01. The latter plots reveal more stable balanced behaviors among users compared to the cases where the learning rates of PutIn and PutOut are equal.

Additionally, experiments were conducted with PutIn learning rate set to 0.05 and PutOut learning rate set to 0.01. As expected

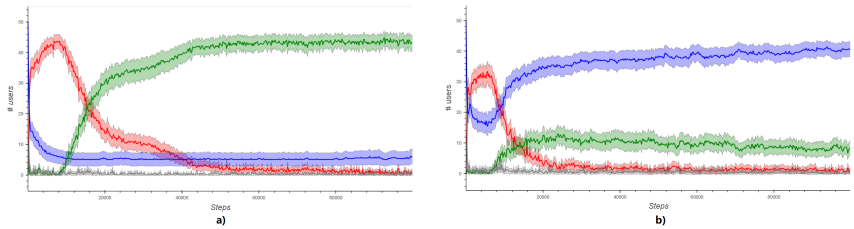

Figure 3: Results for $\beta = 0.5$, with learning rates $lr_{PutIn} = 0.01$ and $lr_{PutOut} = 0.01$, using the Paired-Trained Recommender System with 4 arms. (a) popADD, (b) popNOSM. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.

based on the previous reasoning, the results are similar to those shown in Figures c and d.

5.4 Paired-Trained vs. PutIn-PutOut Recommender Systems Performance

The architecture presented so far involves a critical interaction between Rec Short and Rec Long. As the user must interact with Rec Short to eventually interact with Rec Long, Rec Short is likely to be updated more frequently, especially when users exit the social before Rec Long is activated. This imbalance can lead to Rec Short learning faster than Rec Long, and so to issues as discussed in the previous subsections.

To address this, as explained in the Method section, we implemented the Paired-Trained Recommender System. This architecture, similar to the PutIn-PutOut Recommender System, differs in that both modules are updated during each user interaction, which resolves the imbalance in updates.

The new architecture promotes the development of more stable balanced behaviors in users, as demonstrated in Figure 3. The results shown are obtained using the same parameters as those in Figure 2. This is especially evident for popADD users, where almost all users exhibit balanced behaviors starting from step 50,000. Minor improvements are also observed for popNOSM users, where the number of users with a balanced policy remains nearly the same, but the peak of addicted users is lower compared to the previous architecture

5.5 Well-Being-Aware Recommender vs. Paired-Trained Recommender

We implemented a mechanism commonly employed by major social media platforms, providing a baseline that accounts not only for user engagement but also for user well-being [32]. Specifically, we imposed a maximum interaction constraint on social media usage, limiting users to X interactions within intervals of Y steps. We tested two configurations: X=25 - Y=100 and X=50 - Y=100.
When a user reaches the maximum number of interactions, further attempts result in no state transition and a reward of 0. These configurations simulate different social media usage policies, varying in attention to user health (25 vs. 50 interactions).
Across both populations, our results (Figure 5) highlight the clear advantages of the paired-training architecture over engagement-maximizing systems under the well-being mechanism. In popADD, where content is highly addictive, this new baseline only marginally reduces addicted users and slightly increases balanced behavior compared to the previous baseline, while the paired-training system nearly eliminates addictive behavior and shifts the majority of users toward balanced usage, with healthy users remaining stable. In popNOSM, where content is generally non-addictive, this new baseline produces small gains in balanced users but also increases the number of addicted users, whereas the paired-training system promotes predominantly healthy usage and minimizes addictive behavior. Thus the slight increase in balanced users achieved by the engagement-maximizing architecture comes at the expense of a substantial rise in addicted users. One reason behind the rise in addicted users is that forcing session endings prevents users from fully experiencing the impact of the recommender's influence. This trade-off is clearly unfavorable, reinforcing that our proposed recommendation system offers a significantly better and more responsible solution.

5.6 Scalability of Novel Recommender Architectures

To generalize the results from the MAB with 4 possible content options, we expanded the model to consider 16 different arms (i.e., 16 different contents).

Figure 4 shows the results for both the PutIn-PutOut Recommender System and the Paired-Trained Recommender System with 16 arms, using $\beta = 0.5$ and popADD. Both systems still enable users to develop balanced behaviors, but the Paired-Trained Recommender System reduces the time required for users to reach this state. This result aligns with the findings in previous subsections, where the recommender system learns to disengage users from prolonged sessions more quickly, thanks to the update of both modules at each user interaction.

Similar results were observed across most of the other experiments with different parameters, leading us to conclude that the architectures presented in this study are robust and effective, even for systems with a larger set of content options to recommend.

6 CONCLUSION AND FUTURE DIRECTIONS

This study introduced a simplified yet effective strategy for mitigating compulsive social media use by modeling users, based on the classical dual-system reinforcement learning framework from neuroscience [17], interacting with a dynamic multi-armed recommender system. First, we show that the system can reproduce addiction-like behaviors under specific, relevant parameter settings. While the original "PutIn–PutOut" architecture reduced addictive patterns, our extended analysis found that tuning the PutIn learning

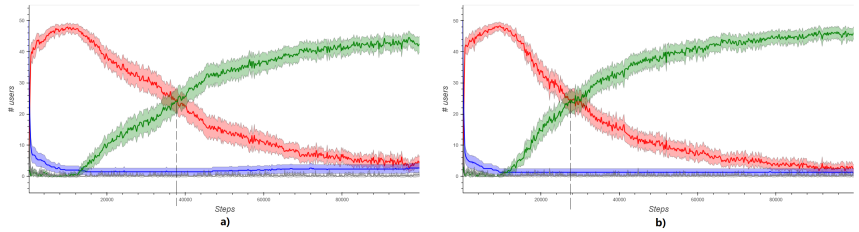

Figure 4: Results for $\beta = 0.5$, with learning rates $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$, using a recommender system with 16 arms. Population: popNTRL. (a) PutIn Recommender System, (b) Paired-Training Recommender System. Population: popNOSM. Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.

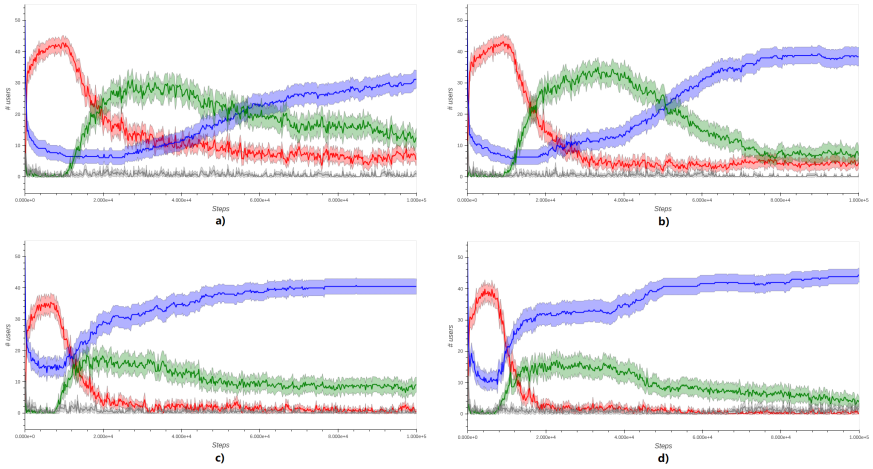

Figure 5: Results for $\beta = 0.5$ obtained using an engagement-maximizing recommender system augmented with a well-being aware mechanism (4 arms). The learning rates are set to $\text{lr}_{\text{PutIn}} = 0.01$ and $\text{lr}_{\text{PutOut}} = 0.01$. (a-b) popADD. (c-d) popNOSM. Parameter settings are $X = 25$ and $Y = 100$ in (a - c), and $X = 50$ and $Y = 100$ in (b - d). Blue: healthy behavior; red: addictive; green: balanced; gray: uncertain.

rate further stabilized balanced engagement. The newly introduced Paired-Trained Recommender System addressed update imbalances and led to more consistent recovery across user types. These effects persisted even with larger content representations, supporting the robustness and scalability of our approach to ethical, sustainable recommender design. Overall, our findings indicate that even a simplified model can sustain engagement without promoting addictive behavior, highlighting the potential of strategic recommendation to encourage healthier digital habits. Despite these promising results, several limitations remain. First, the simulations were conducted in a controlled setting with synthetic data, and validating the model's real-world applicability requires testing on real user data, though access to such data remains a major challenge. This stems from a structural issue: social media platforms are closed systems that do not grant access to the data. This opacity prevents the scientific community from empirically validating whether certain recommendation strategies contribute to addictive behaviors. Our work addresses this gap by building on established computational neuroscience research that models addiction as persistent convergence to suboptimal RL policies [29].

Second, our agent model could be enriched by incorporating social interactions, content awareness, and complex users differences extracted from real data, similarly to what has been in [15] and expanding toward generating synthetic RL agents [18, 43]. Third, the recommender system could be further adapted to individual user characteristic, learning to apply "kick-out" or "slowdown" mechanisms tailored to usage patterns. Additionally, even with improved recommender architectures, persistent forms of addiction emerged under the most challenging conditions, as identified in the computational neuroscience literature [17].

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).

This paper has been published as an Extended Abstract at AAMAS 2026. The final authenticated version is available online at: https://doi.org/10.65109/ROHQ5247

REFERENCES

[1] 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.

[2] 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.

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

[4] L. Badenes-Ribera, M.A. Fabris, F.G.M. Gastaldi, L.E. Prino, and C. Longobardi. 2019. Parent and peer attachment as predictors of facebook addiction symptoms in different developmental stages (early adolescents and adolescents). *Addictive Behaviors* 95 (2019), 226–232. https://doi.org/10.1016/j.addbeh.2019.05.009

[5] Joël Billieux, Pierre Maurage, Olatz Lopez-Fernandez, Daria J Kuss, and Mark D Griffiths. 2015. Can disordered mobile phone use be considered a behavioral addiction? An update on current evidence and a comprehensive model for future research. *Current Addiction Reports* 2, 2 (2015), 156–162.

[6] Xiongfei Cao, Mingchuan Gong, Lingling Yu, and Bao Dai. 2020. Exploring the mechanism of social media addiction: An empirical study from WeChat users. *Internet Research* 30, 4 (2020), 1305–1328.

[7] Silvia Casale, Laura Rugai, and Giulia Fioravanti. 2018. Exploring the role of positive metacognitions in explaining the association between the fear of missing out and social media addiction. *Addictive behaviors* 85 (2018), 83–87.

[8] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 6 (2011), 1204–1215.

[9] Lorenz Deserno, Quentin JM Huys, Rebecca Boehme, Ralph Buchert, Hans-Jochen Heinze, Anthony A Grace, Raymond J Dolan, Andreas Heinz, and Florian Schlagenhauf. 2015. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences* 112, 5 (2015), 1595–1600.

[10] Amir Dezfouli, Payam Piray, Mohammad Mahdi Keramati, Hamed Ekhtiari, Caro Lucas, and Azarakhsh Mokri. 2009. A neurocomputational model for cocaine addiction. *Neural computation* 21, 10 (2009), 2869–2893.

[11] Ray J Dolan and Peter Dayan. 2013. Goals and habits in the brain. *Neuron* 80, 2 (2013), 312–325.

[12] Barry J Everitt and Trevor W Robbins. 2016. Drug addiction: updating actions to habits to compulsions ten years on. *Annual review of psychology* 67, 1 (2016), 23–50.

[13] Vincenzo G Fiore, Dimitri Ognibene, Bryon Adinoff, and Xiaosi Gu. 2018. A multilevel computational characterization of endophenotypes in addiction. *eneuro* 5, 4 (2018).

[14] G. Gauthier, R. Hodler, P. Widmer, et al. 2026. The political effects of X's feed algorithm. *Nature* (2026). https://doi.org/10.1038/s41586-026-10098-2

[15] Sabrina Guidotti, Gregor Donabauer, Simone Somazzi, Udo Kruschwitz, Davide Taibi, and Dimitri Ognibene. 2024. Modeling Social Media Recommendation Impacts Using Academic Networks: A Graph Neural Network Approach. In *International Workshop on Recommender Systems for Sustainability and Social Good*. Springer, 63–72.

[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] Bosen Lian, Wenqian Xue, Frank L. Lewis, and Tianyou Chai. 2022. Inverse reinforcement learning for multi-player noncooperative apprentice games. *Automatica* 145 (2022), 110524. https://doi.org/10.1016/j.automatica.2022.110524

[19] Yuanguo Lin, Yong Liu, Fan Lin, Lixin Zou, Pengcheng Wu, Wenhua Zeng, Huanhuan Chen, and Chunyan Miao. 2023. A survey on reinforcement learning for recommender systems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 10 (2023), 13164–13184.

[20] 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.

[21] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *ICML*. Morgan Kaufmann, 157–163.

[22] Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and Du Zhang. 2023. Lazy Agents: A New Perspective on Solving Sparse Reward Problem in Multi-agent Reinforcement Learning. In *ICML (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 21937–21950.

[23] 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/

[24] Andrew W Moore and Christopher G Atkeson. 1993. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning* 13, 1 (1993), 103–130.

[25] 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.

[26] 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.

[27] Alfonso Pellegrino, Alessandro Stasi, and Veera Bhatiasevi. 2022. Research trends in social media addiction and problematic social media use: A bibliometric analysis. *Frontiers in psychiatry* 13 (2022), 1017506.

[28] Srinivasan A Ramakrishnan, Riaz B Shaik, Tamizharasan Kanagamani, Gopi Neppala, Jeffrey Chen, Vincenzo G Fiore, Christopher J Hammond, Shankar Srinivasan, Iliyan Ivanov, V Srinivasa Chakravarthy, et al. 2025. Impaired arbitration between reward-related decision-making strategies in Alcohol Users compared to Alcohol Non-Users: a computational modeling study. *NPP—Digital Psychiatry and Neuroscience* 3, 1 (2025), 1.

[29] A David Redish. 2004. Addiction as a computational process gone awry. *Science* 306, 5703 (2004), 1944–1947.

[30] A David Redish, Steve Jensen, and Adam Johnson. 2008. Addiction as vulnerabilities in the decision process. *Behavioral and brain sciences* 31, 4 (2008), 461–487.

[31] 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.

[32] Alberto Monge Roffarello and Luigi De Russis. 2023. Achieving digital wellbeing through digital self-control tools: A systematic review and meta-analysis. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–66.

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

[34] 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.

[35] 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.

[36] 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

[37] Dylan A Simon and Nathaniel D Daw. 2012. Dual-system learning models and drugs of abuse. In *Computational neuroscience of drug addiction*. Springer, 145–161.

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

[39] Yalin Sun and Yan Zhang. 2021. A review of theories and models applied in studies of social media addiction and implications for future research. *Addictive Behaviors* 114 (2021), 106699. https://doi.org/10.1016/j.addbeh.2020.106699

[40] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

[41] 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

[42] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

[43] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. 2023. When Demonstrations meet Generative World Models: A Maximum Likelihood Framework for Offline Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 65531–65565. https://proceedings.neurips.cc/paper_files/paper/2023/file/ce9d3c592712d23f2ec3671941d67fa1-Paper-Conference.pdf

[44] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2810–2818.