

User Tampering in Reinforcement Learning Recommender Systems

Charles Evans

Atoosa Kasirzadeh

u6942700@anu.edu.au

atoosa.kasirzadeh@anu.edu.au

The Australian National University

Canberra, Australia

ABSTRACT

This paper provides the first formalisation and empirical demonstration of a particular safety concern in reinforcement learning (RL)-based news and social media recommendation algorithms. This safety concern is what we call “user tampering” – a phenomenon whereby an RL-based recommender system may manipulate a media user’s opinions, preferences and beliefs via its recommendations as part of a policy to increase long-term user engagement. We provide a simulation study of a media recommendation problem constrained to the recommendation of political content, and demonstrate that a Q-learning algorithm consistently learns to exploit its opportunities to ‘polarise’ simulated ‘users’ with its early recommendations in order to have more consistent success with later recommendations catering to that polarisation. Finally, we argue that given our findings, designing an RL-based recommender system which cannot learn to exploit user tampering requires making the metric for the recommender’s success independent of observable signals of user engagement, and thus that a media recommendation system built solely with RL is necessarily either unsafe, or almost certainly commercially unviable.

1 INTRODUCTION

Broadly speaking, recommender systems are algorithms which offer the ability to filter through large pools of some kind of entity to identify and recommend particularly relevant entities to an individual or group [6]. Among other domains, recommender systems have been widely deployed to recommend movies and videos, music, and goods on e-commerce platforms. One of their most significant areas of application is within news and social media platforms, where they are used to provide users with some content of interest. We refer to recommender systems in this specific area as ‘media recommender systems.’

One emergent approach to implementing recommender systems involves treating the recommendation problem as a Markov Decision Process (MDP) and applying reinforcement learning (RL) to the recommendation task. While this approach was suggested some time ago [24, 26, 27], lately it has garnered new interest, due to the emergence of ‘Deep RL’ and its ability to handle larger, more complex problems [15, 31, 32]. Research has begun to explore the applicability of Deep RL-based recommendation in the news and social media space [23, 33]; this work has demonstrated significantly increased user engagement and activeness relative to other leading approaches to the recommendation problem, which are predominantly: (i) ‘static’ machine learning approaches [2, 6, 9, 16, 18]; and (ii) contextual Multi-Armed Bandit approaches [14, 28–30].

As RL research continues to advance and RL techniques become more effectively applicable at scale, RL-based recommender systems’ impact in the domain of media recommendation will likely continue to grow, and may even eclipse that of the current dominant techniques. Indeed, it is already the case that leading social media platforms like Facebook are undertaking this research & development [10, 17].

The social and ethical implications of media recommender systems have also recently received significant research attention [1, 13, 19, 20, 25]. A recent survey on the subject enumerated the main areas of concern as ‘Biased/unfair recommendations,’ ‘Encroachment on individual autonomy and identity,’ ‘Opacity,’ ‘Questionable content,’ ‘Privacy’ and ‘Social manipulability and Polarisation’ [19]. This paper particularly focuses on the last of these concerns. We argue that the risks posed by an RL-based approach in the space of manipulation and polarisation require serious attention. It has been theorised – but neither demonstrated nor formalised in previous work, to the best of the authors’ knowledge – that a particularly problematic variant of this concern has the potential to emerge when the recommendation problem is framed as a MDP and RL is employed to solve it [21, 22]. The basic idea is that the recommendation algorithm could learn to make recommendations which influence users into becoming easier, more predictable targets for recommendation, as this would heighten the algorithm’s success in the long term. This paper provides the first concrete formalisation – and experimental verification of the potentiality – of the just-described issue, which we call “user tampering.”

This paper makes two core contributions to the literature on the ethics and safety of media recommender systems. Firstly, we formalise the notion of user tampering as a potential safety issue specific to RL-based media recommenders; we do this by using the Causal Influence Diagram techniques proposed by Everitt et al. [7] to extract the specific causal phenomenon enabling RL-based recommenders to learn to manipulate users’ preferences, opinions and interests. Secondly, we simulate a simple media recommendation problem. We show that a standard Q-learning algorithm can learn to exploit user tampering, by developing a policy of making recommendations that affect our simulated users’ content ‘preferences’, before capitalising on those effects in later recommendations. While our simulation occurs on a significantly smaller scale than a real recommendation problem scenario, it nonetheless affirms the user tampering theory and so has significant implications for actual RL-based media recommender systems.

2 MODELLING THE MEDIA RECOMMENDATION PROBLEM

In this section, we first generically frame the media recommendation problem as a Markov Decision Process (MDP), and then use Causal Influence Diagrams (CIDs) to extract the relevant causal dependencies that particular variables exhibit under this model. For some background on CIDs, see Appendix A. We have endeavoured to keep the MDP as general as possible, while also incorporating design insights from recent work in implementing RL-based media recommender systems [23, 33].

2.1 An MDP representation of Media Recommendation

We now build up a model of the media recommendation problem as a Markov Decision Process $\langle S, A, T, R, \gamma \rangle$; this is the problem interpretation upon which RL algorithms are based. Here, we assume that articles/posts are represented in a parameterised form, i.e. as an n -dimensional vector, where we have identified n numeric characteristics of the article by which it can be identified (such as topic, stance, author, etc.). This is in opposition to representing articles atomically, which is undesirable; it rules out the use of Deep RL, without the use of which the media recommendation problem is untenably large for RL at industrial scales.

Say that we begin by defining the MDP's individual elements as follows:

- S is a set of states. The specific definition of a state could take any number of definitions in a particular implementation, but we generally define it here as a collection of data points which can be divided into 'necessary' and 'optional' elements:
 - *Necessary*: The state must contain a representation of how the recommender's recent recommendations have performed. For example, this could be a collection of $|n \times m|$ datapoints, representing users' aggregate clicks on recommended items across n categories over m different interpretations of 'recent history' e.g. the last 1 hour, 6 hours, 1 day, etc. (similar to the approach taken by Zheng et al. [33]). This inclusion is 'necessary' as without it, the theoretical advantage associated with using RL in the first place would be lost. Learning policies which capitalise on future as well as current opportunities for reward relies upon including this information in the state representation.
 - *Optional*: Other observable features which inform about this user's preferences, including the activity of their 'friends', contextual features such as time of day, and more.
- A is a set of actions. An action is an n -dimensional vector, representing the characteristics of one article that could be recommended to the user. This could be extended to defining an action as recommending a fixed-size set of articles to the user, as the characteristics of all the articles could simply be aggregated into the vector.
- R is the reward function, mapping the agent's activity to numeric rewards to give it feedback about the 'goodness' of that activity. Commonly, recommender systems could base reward on observable indicators of engagement such as a click,

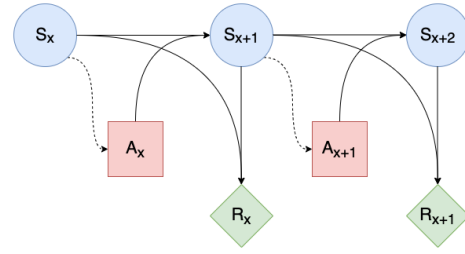


Figure 1: A naive CID of the media recommendation problem.

'like', or so forth. Depending on the specific implementation and the definition of the state space, this function could have multiple different signatures, including $R : S \rightarrow \mathbb{R}$, $R : S \times A \rightarrow \mathbb{R}$ and others.

- $T : S \times A \times S \rightarrow [0, 1]$ is the transition function. It returns the probability of an agent reaching a particular 'successor state' s' , given it has taken a certain action a from a state s .
- $\gamma \in \mathbb{R}$ is the discount factor for future rewards.

In terms of these variables, the recommendation problem can simply be described as an agent taking an action a_t at time t , which will transition the system from a current state s_t to a successor state s_{t+1} with probability $T(s_t, a_t, s_{t+1})$. The agent would thereafter be rewarded with the value $R(s_t)$, and then another action would be chosen at time $t + 1$, and so forth.

2.2 Extracting a CID from the MDP

If we were to naively model this MDP's causal structure as a CID without further consideration, we would reach a representation similar to that shown in Figure 1.¹ At a time-step x , the distribution over possible current states is represented by S_x . Once the actual value of the state as of time x is observed, this will constitute the only information available to the agent in its selection of an action at A_x . The distribution over possible states in S_{x+1} is then defined by T , given S_x and A_x . Finally, R_x represents the distribution over the reward value achieved from the action taken at A_x . We have assumed an interpretation of the reward function as $R : S \times S \rightarrow \mathbb{R}$ here, where reward is determined by a comparison of two consecutive states, as this will provide sufficient information to deduce the success of the most recent action (recommendation).

However, a simple thought experiment can demonstrate that this CID underspecifies the causal relationships in the actual problem, by leaving key variables external to the MDP unacknowledged. Consider the following: Alice and Bob are two university students who have just created accounts on some media platform, who have so far both been recommended the same three articles about the student politics at their university, and who have both clicked on all three articles. Within our general definitions, it is quite plausible that the states of the system have been identical thus far from the agent's perspective. However, what if Bob is uninterested in politics and is just clicking on the articles because his friends feature prominently in the cover photos of all three, whereas Alice is clicking out

¹Note that for this and all following CIDs in the paper, we just show a three-time-step subgraph of the full diagram, as this captures the general structure without overcomplicating the visualisation.

of a genuinely strong interest in politics, including student politics? If the recommendation to both Alice and Bob at the next time-step – say, A_x – is an article about federal politics, it is intuitively untrue that the distribution over possible states at S_{x+1} is the same; Alice is surely more likely to observably engage with this content.

Evidently, a random variable exogenous to the MDP must be introduced to properly model the causal properties of the true system. Informally, we argue that this variable can be characterised as the preferences/opinions/interests of the specific user to which the agent is recommending media. Given that it is exogenous to the MDP, it is unnecessary to assign this variable a specific form, but we assign it the symbol θ^T for the purposes of our causal modelling. That is to say, the exogenous variable that is the user’s preferences etc., at time x , is represented as θ_x^T . We do not enforce any Markov assumptions on θ_x^T ; that is, it may be dependent on the variable’s value at multiple, or all, previous time-steps’ values.

A vital observation to make at this stage is the causal relationship between θ_x^T and S_{x+1} . As the example above demonstrated, without acknowledging the effect of θ_x^T , the real distribution over states S_{x+1} cannot be explained. So, there exists a causal link between the former and latter variables. Moreover, the specific elements of S_{x+1} whose distribution would otherwise be unexplainable are precisely those which we explained are necessary inclusions in the definition of S in order to produce the desirable properties of an RL approach. So, this link cannot be removed by any practical redesign of the state space. Finally, it is crucial to recognise that an influence link will also exist between A_x and θ_{x+1}^T ; intuitively, this just reflects the reality that a user’s consumption of information will update and affect their preferences, interests, etc. going forward. Given that θ^T is exogenous and given no specific formal interpretation, it is not our claim that there is a precise model available for *how* A_x affects the distribution over possible values of θ_{x+1}^T ; rather, we are just acknowledging the existence of the dependency.

It is evident that the CID of Figure 1 is in need of revision. If we introduce the exogenous variable to the system, without changing any other definitions, we arrive at the CID shown in Figure 2. This CID, we argue, more completely captures the actual causal dynamics of the Media Recommendation MDP. We note that previous literature has acknowledged a similar causal structure to the recommendation process [12]; however, this was not formulated in the CID framework that we have used, which permits sophisticated graphical analysis of the kind developed in the next section.

Depending on the exact design of the MDP, there are variations on this CID that could exist; see Appendix B for an example of the causal structure that is implied if the designer wishes to expand the reward function to account for observations that are not captured in the state representation. However, such variations have no effect on the role and influence of θ^T ; its links from the preceding action and to the succeeding state necessarily remain part of the model’s causal structure. Given that these causal relationships are exactly where we intend to focus our analysis in the next section, the CID in Figure 2 is a sufficiently general representation for our needs going forward.

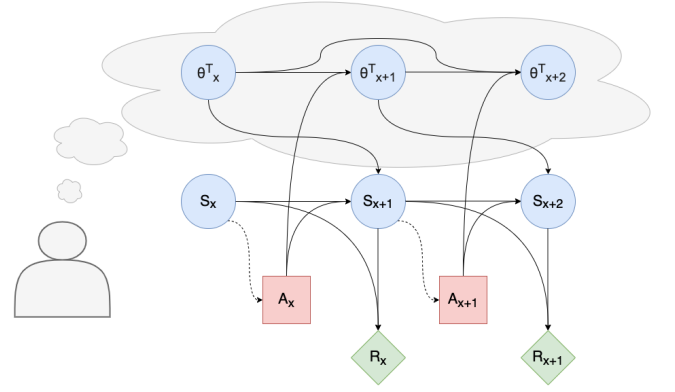


Figure 2: A CID of the media recommendation problem, extended to include the exogenous variable affecting state transitions.

3 USER TAMPERING

In this section, we use the CID formulated in the previous section (Figure 2) to analyse the safety of the RL-based approach to media recommendation, specifically with respect to the high-level concerns of user manipulation and polarisation. After introducing the phenomena of ‘instrumental control incentives’ and ‘instrumental goals’ from the RL incentive analysis literature, we show that in the CID, an instrumental goal exists for the agent to manipulate the expected value of the exogenous variable θ^T . This lends a concrete, formal interpretation to the (formerly only hypothesised) safety issue that we have called ‘user tampering’.

3.1 Instrumental Goals and Control Incentives

An ‘instrumental control incentive’ (ICI) is a graphical property of CIDs introduced by Everitt et al. [7]. An ICI exists on a Structural Node X if it lies on a path in the CID that begins at a Decision Node and ends at a Utility Node, and basically implies that the action chosen at the former affects the expected utility at the latter *through* affecting the distribution over values at X .

The importance of ICIs is that they provide a simple graphical criterion that can establish either the potential presence or the categorical absence of a so-called ‘instrumental goal’ on certain events in a RL problem [8]. An RL agent is said to have an *instrumental goal* to influence the distribution at a Structural Node X in a certain way if it has an ICI on X and that particular influence increases the expected reward accumulated by the agent – in short, if it has both the ability and a reason to affect the distribution at X .

3.2 Formalising User Tampering

In the CID we have presented, intuitively there are a subset of Structural Nodes upon which instrumental goals are *desirable* – those representing the set of random state variables $\{S_t | t \in \mathbb{N}\}$. This reflects the very premise of RL – we want the agent to be able to manipulate the state of the system in pursuit of ‘good’ states (in our context, these are states where many of its recent recommendations have been well-received by the user). As such, any path through the CID from a Decision Node to a Utility Node that *only* passes through

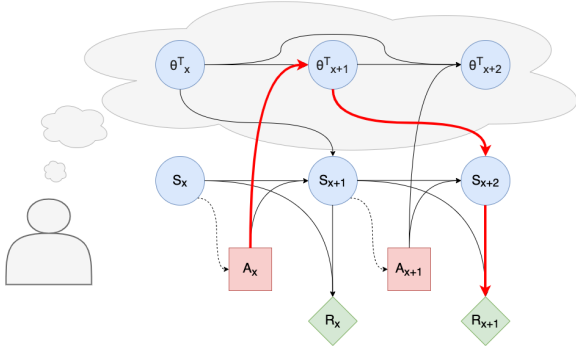


Figure 3: An annotated version of the media recommendation CID for state-based rewards. An example of an undesirable causal path introducing an instrumental control incentive on θ_{x+1}^T is shown in bolded red.

random state variables (e.g. $[A_x \rightarrow S_{x+1} \rightarrow R_x]$, or $[A_x \rightarrow S_{x+1} \rightarrow S_{x+2} \rightarrow R_{x+1}]$) only involves intended and safe instrumental goals.

However, other paths from Decision to Utility Nodes exist in the CID. Specifically, there are paths which visit the exogenous random variables – for example, $[A_x \rightarrow \theta_{x+1}^T \rightarrow S_{x+2} \rightarrow R_{x+1}]$. Figure 3 traces this path in the CID. Clearly, there is an ICI on θ_{x+1}^T , or on any other variable in $\{\theta_t^T | t \in \mathbb{N}\}$ that appears in similar paths.²

Given this, if the agent stands to generate higher amounts of reward by recommending to a user with particular preferences, opinions and interests (represented by θ^T), then the agent may have an instrumental goal to influence θ^T in that direction, as this may lead to higher expected reward in the long term. In essence, the CID’s admission of an ICI on at least one node in $\{\theta_t^T | t \in \mathbb{N}\}$ is precisely the necessary graphical condition that allows the manipulation of users to emerge as an instrumental goal for an RL agent. If such an instrumental goal does exist (i.e. if by affecting users’ preferences/opinions/interests, the agent can increase its expected reward), then we can expect that an arbitrarily capable RL agent would learn to act on that instrumental goal – we say that this makes user tampering a ‘learnable’ phenomenon.

Definition 1. *User tampering is a ‘learnable’ phenomenon for an RL-based media recommendation algorithm iff it has an instrumental goal to affect at least one of the variables in $\{\theta_t^T | t \in \mathbb{N}\}$.*

Importantly, however, an instrumental goal on affecting some variable in $\{\theta_t^T | t \in \mathbb{N}\}$ does not imply that an arbitrary RL agent will learn to affect the user in the way necessary to increase its expected reward; it only implies that it *could* learn this behaviour. So, user tampering’s learnability in some model is a necessary but insufficient condition for user tampering actually manifesting in an RL agent’s learned policy. It is therefore useful to introduce a second definition relating to user tampering, such that we can separate our discussions of its theoretical learnability from our

discussions of it actually manifesting in a given recommender’s policy. We introduce a second definition to address this:

Definition 2. *An RL-based media recommendation algorithm ‘exploits’ user tampering iff there exists a state s_t such that $\pi(s_t) = a_t$ and $\pi'(s_t) \neq a_t$, for the algorithm’s actual learned policy π , and the hypothetical policy π' that the same learning process would have produced in a world where $A_t \perp\!\!\!\perp \theta_{t+1}^T$.*

In Appendix C, we relate these formalisms to a different form of ‘tampering’ in RL: ‘Reward Function (RF)-tampering.’ We explain that although the two phenomena seem to describe similar high-level issues, the issues are quite separate on a causal level; and that these differences rule out the transferral of promising solutions for the RF-tampering issue to the user tampering context.

4 RESULTS

In this section, we empirically analyse the user tampering phenomenon formalised in the previous section. Firstly, we introduce a simple abstraction of the media recommendation problem, which involves simulated users and a user tampering incentive inspired by recent empirical results about polarisation on social media. Then, we present a Q-learning agent intended to mimic the Deep Q-learning algorithms used in recent media recommendation research, and train it in this environment [23, 33]; we show that its learned policy clearly exploits user tampering in pursuit of greater rewards.

4.1 Problem and Environment Setup

We begin by introducing our example problem. In this problem, we will have a recommender agent make h sequential recommendations of ‘political posts/articles’ to a user. At each timestep t , $0 \leq t \leq h$, the agent chooses one of three ‘sources’ from which to recommend; the first source being consistently left-wing in its perspective, the second being consistently centrist, and the last being consistently right-wing.

For the purposes of our example, we assume a definition of the exogenous parameter θ^T introduced in Section 2 – recall that the agent does not explicitly model this variable, but we need to here for the purposes of constructing our simulation. We define θ_t^T as a tuple of three probabilities as of time t , i.e. $\Theta^T = \{(\theta^{TL}, \theta^{TC}, \theta^{TR}) \in \mathbb{R}^3 \mid \forall x \in \{L, R, C\}. \theta^{Tx} \in [0, 1]\}$. For some arbitrary user, their probability θ^{TL} represents their probability of clicking an article from the left-wing source *if recommended it*; the same can be said of θ^{TC} for the centrist source, and θ^{TR} for the right-wing source. We say that a user is initially “right-wing” iff $\theta_0^{TR} > \theta_0^{TC} \wedge \theta_0^{TR} > \theta_0^{TL}$, and “left-wing” iff $\theta_0^{TL} > \theta_0^{TC} \wedge \theta_0^{TL} > \theta_0^{TR}$.

Finally, we include a simple environmental dynamic whereby users who are recommended content from a source that is politically opposed to their own wing gradually become more polarised in favour of their own wing. This is inspired by recent research into user polarisation on social media, which has demonstrated that showing people who identify with one wing of the political spectrum volumes of content from the opposing wing can often increase user polarisation [4, 5].

We do not at all claim that this completely simulates the polarisation phenomenon described in the works cited just above; it

²Although we are limited from showing larger subgraphs of the CID by page size, note that longer paths can also be identified which contain similar instrumental goals, e.g. paths of the form $[A_x \rightarrow \theta_{x+1}^T \rightarrow \theta_{x+2}^T \rightarrow \dots \rightarrow \theta_{x+n}^T \rightarrow S_{x+n+1} \rightarrow R_{x+n}]$, or $[A_x \rightarrow \theta_{x+1}^T \rightarrow S_{x+2} \rightarrow S_{x+3} \rightarrow \dots \rightarrow S_{x+n} \rightarrow R_{x+n-1}]$.

is obviously a significant simplification. However, our intention here is not to simulate this effect accurately, but rather to create an environment which allows the hypothesised effect of user tampering to be tested by introducing a causal effect that could be used by the agent as part of an instrumental goal (while still having its simplified dynamics grounded in actual sociological results).

The full definition $\langle S, A, T, R, \gamma \rangle$ of the media recommendation MDP, as well as the precise implementation of the ‘polarisation’ effect we have just described, is provided in Appendix D.

Next, we train a Q-learning agent in this environment and show that it learns to perform user tampering on our simulated users.

4.2 Recommender Simulation

Some extra specifications are needed to operationalise this environment. h was set to 30, and the probabilities defining the exogenous variable θ^T were limited to maximum values of 0.75.³ p , the ‘polarisation factor’ by which a user’s subsequent probability of clicking on content from their aligned source would increase after being recommended a post from the opposing source, was sampled from the uniform distribution $U(1.01, 1.10)$, making $\mathbb{E}[p] = 1.055$.

We also defined, for the purposes of our experiment, a population of five ‘users’ with varying preference profiles. This contained:

- A ‘strong left’ user with $\theta_0^T = (0.4, 0.1, 0.1)$
- A ‘moderate left’ user with $\theta_0^T = (0.3, 0.25, 0.1)$
- A ‘centrist’ user with $\theta_0^T = (0.2, 0.4, 0.2)$
- A ‘moderate right’ user with $\theta_0^T = (0.1, 0.25, 0.3)$
- A ‘strong right’ user with $\theta_0^T = (0.1, 0.1, 0.4)$

We trained a Q-learning agent in this environment, with a user randomly selected from the population to provide the initial value of θ^T for each episode.⁴ Non-deep Q-learning was used for training, in spite of deep Q-learning being the more viable approach at industrial scales; this was a deliberate choice, because unlike deep Q-learning, non-deep Q-learning provably converges towards the optimal policy for the problem.⁵ Nonetheless, for consistency with a real Deep RL application, we have still modelled the state space in a parameterised fashion that is amenable to those algorithms.

4.2.1 Results. For each of the five users in our population, we provide two plots based on 10000 evaluation episodes with the user (using the policy learned from the training process described above). Respectively, these two plots estimate:

- The probability with which the learned policy chooses each action, at each time-step of the problem, by taking the per-episode average of each choice’s total frequency.
- The expected reward accumulated up to and including each time-step t , $0 \leq t \leq h$. For context, we plot this against the expected reward accumulated by:

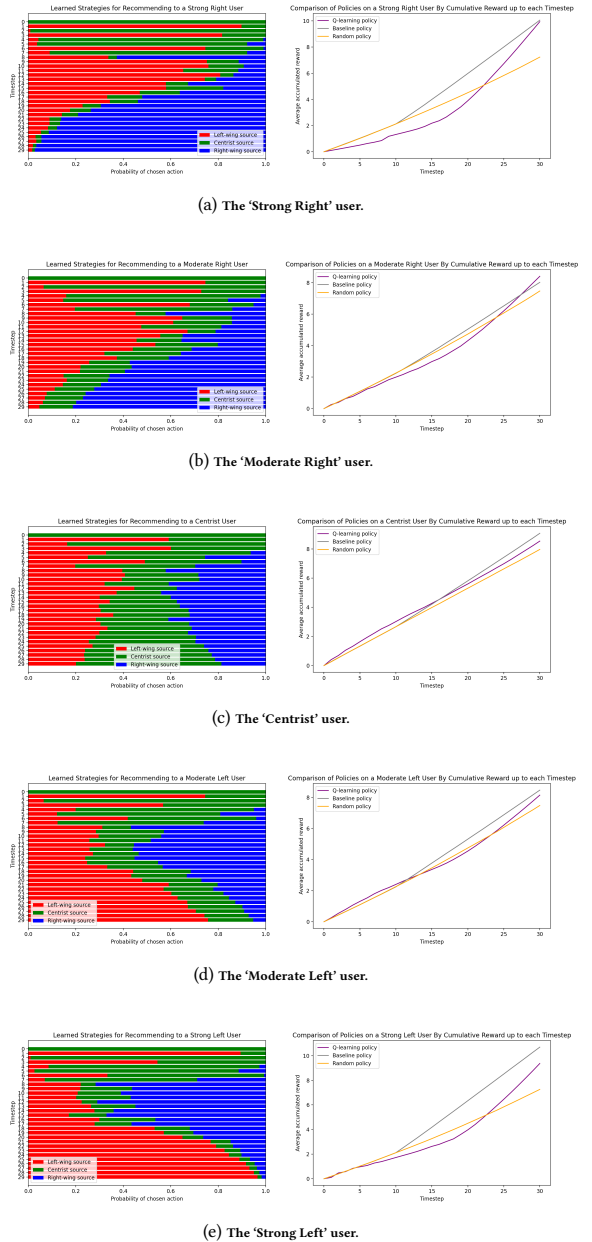


Figure 4: Evaluation of the policy learned with Q-learning for each member of our sample user population.

- A recommender that makes uniformly random recommendations at each time-step.
- A recommender that follows a simple multi-armed bandit-esque policy, which provides a ‘baseline’ of a good policy. This policy makes random recommendations for the first third of the episode, but then operates like a multi-armed bandit, by always recommending from the source which has the highest mean reward in the episode so far.

³This was an arbitrary limitation imposed by the authors to avoid users becoming ‘polarised’ to the extent that they would click on every post from a source that their views aligned with, as this seemed an unrealistically extreme outcome that would negatively effect the plausibility of our simulation.

⁴Our implementation, including a pre-trained recommender agent, is available on GitHub: <https://github.com/chevans-lab/user-tampering>.

⁵Since we wanted to test whether the agent was able to find a *better* policy by exploiting user tampering than it could otherwise achieve, Deep Q-learning was an inappropriate choice for the experiment as there was no way to guarantee that it would not converge on a good, safe policy even when a better, user tampering policy was available.

Figure 4 displays these plots for each simulated ‘user’ we defined above. These results possess several interesting properties:

- **For all users except the Centrist user, the exploitation of user tampering in the learned policy is clear.** Focusing on the strategy plots for the two ‘left-wing’ users, we can see a clearly dominant strategy has emerged, where:
 - The recommender attempts to profile the user and their preferences, by testing their reaction to centrist and left-wing content (roughly the first quarter of the episode).
 - The recommender predominantly recommends right-wing content *in spite of* its low expected reward, which will tamper with the user’s preferences and increase the expected reward from subsequent left-wing recommendations (roughly the second quarter of the episode).
 - The recommender predominantly recommends left-wing content to the (now more) left-wing user, maximising the high expected rewards that action will now offer (roughly the second half of the episode).

Given user tampering’s learnability in this problem and the expected rewards from right-wing content here (low), the recommender’s propensity to nevertheless heavily recommend right-wing content before switching to left-wing recommendations for the remainder of the episode is a clear exploitation of user tampering. Moreover, the inverse behaviour has been learned for right-wing users – the model is not blindly trying to polarise all users to the left, but has developed a sophisticated policy for identifying and exploiting the causal relationship between its actions and the user’s exogenous variable. Further evidence to this effect is given by the policy for the ‘centrist’ user – here, the plot shows clearly that the recommender has recognised that its actions have no discernable causal impacts by which the user could be tampered with, and so makes recommendations which are proportionate to the user’s initial preferences.

- **The agent heavily exploits user tampering even though we were able to generate similar cumulative rewards with our crude ‘baseline’ policy.** This adds weight to the safety concerns with respect to user tampering. It indicates that there exist other policies which do not exploit user tampering (although they may make a handful of ‘polarising’ recommendations by chance) and which offer similar rewards to the one that the recommender learned; nonetheless, over several iterations of retraining, the policy consistently converged to the policy we have presented here (with small natural variations). This implies that in this environment, the unsafe policy is not only learned occasionally, but presents a likely direction of convergence for the learning algorithm.

It is also worth establishing that the exploitation of user tampering in the learned policy was robust to simulated users not encountered during training. We generated the same policy plots for the recommender over 10000 evaluation episodes spent recommending to each user in a new, ‘unseen’ population: an ‘extremely left’ user with $\theta_0^T = (0.5, 0.05, 0.05)$, an ‘extremely right’ user with $\theta_0^T = (0.05, 0.05, 0.5)$, a ‘left anti-centrist’ user with $\theta_0^T = (0.35, 0.05, 0.2)$, and a ‘right anti-centrist’ user with $\theta_0^T = (0.2, 0.05, 0.35)$.

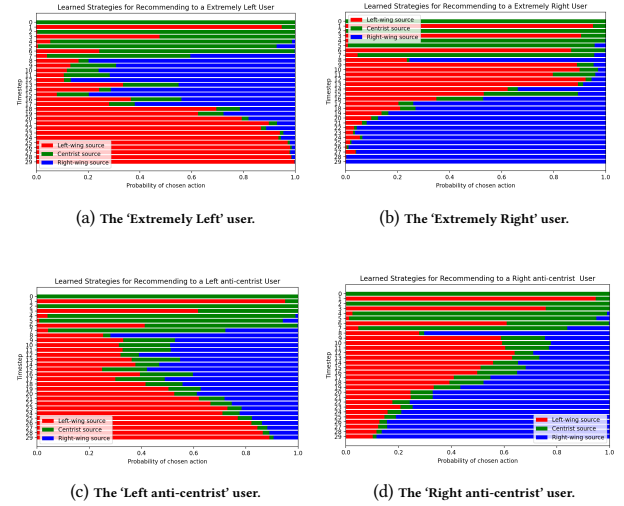


Figure 5: Action probabilities at each time-step for each user in the ‘unseen’ population.

These results are shown in Figure 5. Although these specific users were never encountered during training, the same unsafe strategies appear here; the three phases of user profiling, then polarisation, and finally preference satisfaction are clearly visible.

These results, in combination with the previous sections’ formalisations, justify the claims that user tampering is both almost unavoidably learnable for commercially viable media recommender systems built entirely with RL, and potentially highly unsafe in its effects. Specifically, the primacy of user engagement in content recommendation makes achieving complete safety from user tampering inconvenient at best, and impossible at worst. Appendix E unpacks this claim further for the interested reader.

5 CONCLUSION

This paper has substantiated concerns about the risks of emergent RL-based recommender systems with respect to user manipulation and polarisation. We have formalised these concerns as a causal property – ‘user tampering’ – that can be isolated and identified within a recommendation algorithm, and shown that by designing an RL-based recommender which can account for the temporal nature of the recommendation problem, user tampering also necessarily becomes learnable. Moreover, we have shown that in a simple simulation environment inspired by recent polarisation research, a Q-Learning-based recommendation algorithm consistently learned a policy of exploiting user tampering – which, in this context, took the form of the algorithm explicitly polarising our simulated ‘users.’ This is obviously highly unethical, and the possibility of a similar policy emerging in real-world applications is a troubling takeaway from our findings. Due to a combination of technical and pragmatic limitations on what could be done differently in RL-based recommender design, it is unlikely that commercially viable *and* safe recommenders based entirely on RL can be achieved, and this should be borne in mind when selecting future directions for advancement in media recommendation research & development.

REFERENCES

- [1] Himan Abdollahpour, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158.
- [2] Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. 2018. Social Media Recommender Systems: Review and Open Research Issues. *IEEE Access* 6 (2018), 15608–15628.
- [3] Stuart Armstrong, Jan Leike, Laurent Orseau, and Shane Legg. 2020. Pitfalls of Learning a Reward Function Online. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1592–1600.
- [4] Christopher A. Bail. 2021. *Breaking the Social Media Prism: How to Make our Platforms Less Polarizing*. Princeton University Press, Princeton, New Jersey.
- [5] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.
- [6] Jesus Bobadilla, Fernando Ortega, A. Hernando, and A. Gutiérrez. 2013. Recommender systems survey. *Knowledge-Based Systems* 46 (2013), 109–132.
- [7] Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. 2021. Agent Incentives: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (2021), 11487–11495.
- [8] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese* (2021), 1–33.
- [9] Florent Garcin, Kai Zhou, Boi Faltings, and Vincent Schickel. 2012. Personalized News Recommendation Based on Collaborative Filtering. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. 437–441.
- [10] Jason Gauci, Edoardo Conti, Yitao Liang, Kittipat Virochsiri, Yuchen He, Zachary Kaden, Vivek Narayanan, and Xiaohui Ye. 2018. Horizon: Facebook’s Open Source Applied Reinforcement Learning Platform. *Facebook AI* (2018).
- [11] David Heckerman and Ross Shachter. 1995. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research* 3, 1 (1995), 405–430.
- [12] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) (AIES ’19). Association for Computing Machinery, New York, NY, USA, 383–390.
- [13] Mohammed Khwaja, Miquel Ferrer, Jesus Omana Iglesias, A. Aldo Faisal, and Aleksandar Matic. 2019. Aligning Daily Activities with Personality: Towards a Recommender System for Improving Wellbeing. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) (RecSys ’19). Association for Computing Machinery, New York, NY, USA, 368–372.
- [14] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) (WWW ’10). Association for Computing Machinery, New York, NY, USA, 661–670.
- [15] Feng Liu, Ruiming Tang, Xutao Li, Yunming Ye, Haokun Chen, Huifeng Guo, and Yuzhou Zhang. 2018. Deep Reinforcement Learning based Recommendation with Explicit User-Item Interactions Modeling. *ArXiv arXiv:1810.12027* (2018).
- [16] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized News Recommendation Based on Click Behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) (IUI ’10). Association for Computing Machinery, New York, NY, USA, 31–40.
- [17] Yang Liu, Zhengxing Chen, Kittipat Virochsiri, Juan Wang, Jiahao Wu, and Feng Liang. 2020. Reinforcement Learning-based Product Delivery Frequency Control. *Facebook AI* (2020).
- [18] Zhongqi Lu, Zhicheng Dou, Jianxun Lian, Xing Xie, and Qiang Yang. 2015. Content-Based Collaborative Filtering for News Topic Recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) (AAAI ’15). AAAI Press, 217–223.
- [19] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI & SOCIETY* 35, 4 (2020), 957–967.
- [20] Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhinjan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) (WWW ’20). Association for Computing Machinery, New York, NY, USA, 1194–1204.
- [21] Stuart J. Russell. 2019. *Human Compatible: AI and the Problem of Control*. Allen Lane, London.
- [22] Stuart J. Russell. 2019. Stuart J. Russell on Filter Bubbles and the Future of Artificial Intelligence. https://www.youtube.com/watch?v=ZkV7anCPfay&t=230s&ab_channel=LongNowFoundation. Accessed June 2, 2021.
- [23] Zeinab Shahbazi and Yung Cheol Byun. 2020. Toward Social Media Content Recommendation Integrated with Data Science and Machine Learning Approach for E-Learners. *Symmetry* 12, 11 (2020).
- [24] Guy Shani, David Heckerman, and Ronen Brafman. 2005. An MDP-Based Recommender System. *Journal of Machine Learning Research* 6 (2005), 1265–1295.
- [25] Jonathan Stray, Steven Adler, and Dylan Hadfield-Menell. 2020. What are you optimizing for? Aligning Recommender Systems with Human Values. In *Participatory Approaches to Machine Learning*. International Conference on Machine Learning Workshop.
- [26] Nima Taghipour and Ahmad Kardan. 2008. A Hybrid Web Recommender System Based on Q-Learning. In *Proceedings of the 2008 ACM Symposium on Applied Computing* (Fortaleza, Ceara, Brazil) (SAC ’08). Association for Computing Machinery, New York, NY, USA, 1164–1168.
- [27] Nima Taghipour, Ahmad Kardan, and Saeed Shiry Ghidary. 2007. Usage-Based Web Recommendations: A Reinforcement Learning Approach. In *Proceedings of the 2007 ACM Conference on Recommender Systems* (Minneapolis, MN, USA) (RecSys ’07). Association for Computing Machinery, New York, NY, USA, 113–120.
- [28] Liang Tang, Yexi Jiang, Lei Li, and Tao Li. 2014. Ensemble Contextual Bandits for Personalized Recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) (RecSys ’14). Association for Computing Machinery, New York, NY, USA, 73–80.
- [29] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. 2015. Personalized Recommendation via Parameter-Free Contextual Bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) (SIGIR ’15). Association for Computing Machinery, New York, NY, USA, 323–332.
- [30] Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. 2016. Online Context-Aware Recommendation with Time Varying Multi-Armed Bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) (KDD ’16). Association for Computing Machinery, New York, NY, USA, 2025–2034.
- [31] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep Reinforcement Learning for Page-Wise Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) (RecSys ’18). Association for Computing Machinery, New York, NY, USA, 95–103.
- [32] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with Negative Feedback via Pairwise Deep Reinforcement Learning. In *KDD ’18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom). Association for Computing Machinery, New York, NY, USA, 1040–1048.
- [33] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) (WWW ’18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 167–176.

A CAUSAL INFLUENCE DIAGRAMS

A modelling technique that is central to our formalisation of user tampering is that of the Causal Influence Diagram (CID) [7, 11]. This technique has recently gained popularity in precise analysis of the potential incentives behind RL agents’ behaviour [3, 7, 8]. Here, we briefly explain the interpretation of CIDs.

CIDs are directed, acyclic graphs. 3 node types exist, all of which represent a variable in the problem. There are:

- Decision Nodes (depicted with squares)
- Structural Nodes (depicted with circles)
- Utility Nodes (depicted with diamonds)

Structural Nodes and Utility Nodes represent probability distributions over the values a variable could take, while Decision Nodes are better thought of as representing variables that are assigned a value at the point of decision. A directed edge from a node X to a node Y means that:

- (If Y is a Utility/Structural Node) the value of the random variable Y is conditional on the value of X . The edge will be depicted with a full line in this case.

- (If Y is a Decision Node) the value of X is information that is available to the agent at the decision-time of Y . The edge will be depicted with a dashed line in this case.

CID-based analyses of agent incentives offer several advantages over trying to study behaviour learned with RL through other methods, e.g. statistical analysis. Firstly, CIDs (and the notion of an instrumental goal, introduced in Section 3) explicitly deal with *causation* rather than *correlation* between variables, which is important in our analysis as we are aiming to specifically show recommenders' ability to cause increased user engagement via its actions' causal effects on users' opinions and preferences. Additionally, they allow us to abstract away extraneous information about specific RL algorithms' implementations and deal with the underlying causal mechanisms that they hold in common; this is particularly useful in that it allows us to formally discuss causal properties of many potential implementations simultaneously, rather than statistically analysing the outcomes achieved with each of these implementations individually.

B EXAMPLE VARIATION ON THE MEDIA RECOMMENDATION MDP AND CID

The MDP representation of the media recommendation problem may be expanded relative to our characterisation in Section 2, if the designer wishes to expand the reward function to account for observations that are not captured in the state representation. This would be a reasonable design choice – for example, the state representation may only record some user behaviours such as clicks, whereas we may want to reward the agent based not only on clicks, but also on the 'dwell time' of the user on the article (the time spent on the article after clicking). For generality, we demonstrate how the CID could be extended to represent this.

This firstly requires some changes and introductions to our MDP definition:

- A set of observations O . An observation consists of some collection of metrics representing how a user observably responded to some recommendation.
- An observation probability function $Z : S \times A \times O \rightarrow [0, 1]$. This models the probability of making a particular observation (for example a click, but no comment) after making a certain recommendation in a certain state.
- An altered definition of the Reward function as $R : O \rightarrow \mathbb{R}$. This simply corresponds to the fact that the information on which rewards are predicated – the observable user response to the content – has now been concentrated into the one variable $o \in O$.

We also need to make the addition of an exogenous random variable θ^R for the updated CID. This serves a highly similar purpose to θ^T , except that it instead accounts for the fact that the probability of observing a certain behaviour in response to an article will intuitively change from user to user, even if their state representations are identical (this is a trivial conceptual extension to the Alice-Bob example from Section 2). θ^R and θ^T are not necessarily (and indeed are very likely not) uncorrelated, but we model them as distinct variables for clarity. For the same reasoning as was given with respect to θ^T , influence links will also be necessary between A_x and θ_{x+1}^R .

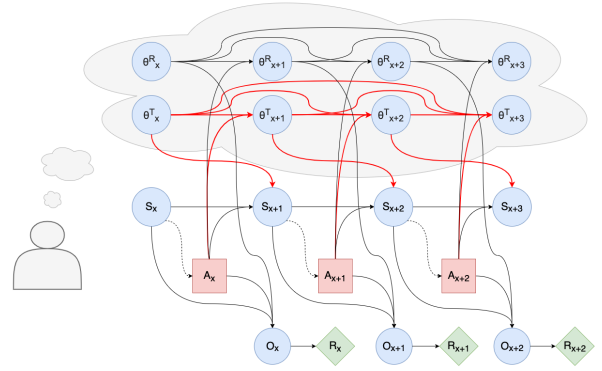


Figure 6: A CID of the media recommendation problem, extended to include an observation space and more complicated definitions of reward.

Figure 6 depicts the media recommendation CID that results from these extensions to the MDP. To reinforce the point made in Subsection 2.2 about variations on the MDP not affecting the causal structure local to the variables $\{\theta_t^T | t \in \mathbb{N}\}$, we have highlighted these variables' incoming and outgoing causal links in the figure; the reader may compare these to those in Figure 2 to verify that all the same links are present.

C USER TAMPERING'S DIFFERENCES FROM RF-TAMPERING

Reward function (RF)-tampering refers to a specific safety issue wherein an RL agent has one or several undesirable instrumental goal(s) to affect variables *within* its reward function in order to change the way in which certain states are evaluated by the function [3, 8]. Although it has not been analysed in detail, it has been suggested that the high-level concerns of user manipulation and polarisation we have discussed in this paper could fall into the category of RF-tampering [8]. On some level, this seems intuitive – we often think of the user and their behaviour as akin to a 'reward function' for the recommender, as the user's response is ultimately the arbiter of whether reward is received for a recommendation. So, one might expect that tampering with the user would constitute a kind of 'reward' tampering. However, our work in Section 3 shows this not to be the case. In the media recommendation problem, the reward function is an explicitly defined function that maps concrete outcomes in the state space – clicks, likes, etc. – to numerical rewards; the user actually constitutes a part of the problem environment, and their behaviour a part of the dynamics of the environment. What we have described is not a form of *reward* tampering at all, but more accurately a form of *transition* tampering.

To contextualise this argument, we give a brief overview of the properties of any problem in which RF-tampering may occur (and its associated CID). In such a problem:

- The reward function can be expressed as $R(S; \theta) : S \times \mathbb{R}^N \rightarrow \mathbb{R}$ or similar, where θ represents some 'parameters' of the reward function other than states or actions.

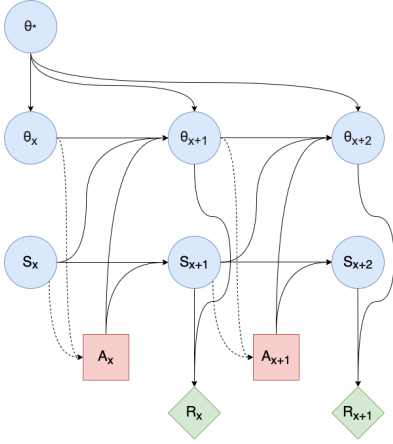


Figure 7: CID Representation of an RF-Tampering-susceptible problem.

- There is an ‘intended’ value for θ (denoted as θ_*) which is static unless a change is introduced to it at some point by an external process. In other words, its value is independent of any actions undertaken by the RL agent.
- The agent models θ_* and updates that model based on its experiences; at time-step t , the distribution over possible values of θ_* according to the agent is represented as θ_t .
- The agent is able to influence the distribution θ_{t+1} with its action a_t .
- The rewards observed by the agent at time-step t are defined as $R(S_t; \theta_t)$, not $R(S_t; \theta_*)$.

The crux of the tampering issue is that the agent has an instrumental goal to alter its own model of the reward parameters such that it rewards certain states more positively than what θ_* would actually generate.

Note, however, that the MDP and associated CID representation of the media recommendation do not conform to this description on several points. Particularly, in the media recommendation problem:

- θ^T is not a hidden parameter to the *reward* function, but instead to the *transition* function.
- θ^T is *not* independent of the agent’s actions. While it fulfils a similar conceptual role as θ_* in RF-tampering in that it represents an ‘intended’ parameter, it is nonetheless subject to the effects of the agent’s recommendations. This is why there is a causal link from A_t to θ_{t+1}^T .
- There is no attempt in the media recommendation problem to explicitly estimate the distribution θ_t^T at time-step t . Instead, the state space contains an implicit estimation of the intended parameters in the form of recorded user behaviour, which is why there is a causal link from θ_t^T to S_{t+1} .

It may help the reader to consult and compare the diagram in Figure 7, where we have recreated the CID given in Everitt et al. [8] to represent an RF-tampering-susceptible model, with the recommendation CID we constructed in Section 2.

C.1 Non-transferability of RF-tampering solutions to User Tampering

Several solutions have been proposed to RF-tampering by Armstrong et al. [3] and Everitt et al. [8]; however, the fundamental principles of their proposals cannot successfully transfer to the user tampering context. An explicit assumption upon which Everitt et al. predicate the solutions they outline (which include among them an adaptation of Armstrong et al.’s solution) is what they term the ‘privacy’ of the random variable(s) for which we are trying to remove the existence of instrumental goals. Formally, ‘privacy’ requires that no direct paths exist between that variable’s distribution at one time-step, and the random state distribution at the next step – so, in our case, that no paths exist of the form $\theta_t^T \rightarrow S_{t+1}$. However, as was explained in Section 2, this link *must* exist in the media recommendation problem; redesigning the state space such that $\theta_t^T \not\rightarrow S_{t+1}$ would compromise the purpose of applying RL to the problem. This implicit estimation of the hidden parameters in the state space is one of the fundamental differences that was mentioned earlier between the media recommendation problem and problems susceptible to RF-tampering. As such, the aforementioned core principle of privacy is unachievable; and without privacy being achieved, the RF-tampering solutions that have been proposed in recent work will not successfully adapt to solving the issue of user tampering.

D FORMAL MDP DEFINITION OF THE RECOMMENDATION SIMULATION

We define the MDP $\langle S, A, T, R, \gamma \rangle$ of the media recommendation problem described in Section 4 as follows:

- $S = \{(s_t^{LR}, s_t^{LC}, s_t^{CR}, s_t^{CC}, s_t^{RR}, s_t^{RC}) \in \mathbb{N}^6 \mid (s_t^{LR} + s_t^{CR} + s_t^{RR} \leq h) \wedge (s_t^{LR} \geq s_t^{LC}) \wedge (s_t^{CR} \geq s_t^{CC}) \wedge (s_t^{RR} \geq s_t^{RC})\}$.
 - s_t is the state after t recommendations, $0 \leq t \leq h$.
 - The state space is interpreted as follows:
 - * s_t^{LR} is the number of “left-wing” recommendations made to the user after t total recommendations
 - * s_t^{LC} is the number of “left-wing” recommendations clicked on by the user after t total recommendations
 - * s_t^{CR} and s_t^{CC} are as above, but with respect to “centrist” recommendations
 - * s_t^{RR} and s_t^{RC} are as above, but with respect to “right-wing” recommendations
- $A = \{0, 1, 2\}$, where:
 - 0 = ‘Left-wing recommendation’
 - 1 = ‘Centrist recommendation’
 - 2 = ‘Right-wing recommendation’
- T is defined as follows, where $s = (s^{LR}, s^{LC}, s^{CR}, s^{CC}, s^{RR}, s^{RC})$:
 - $T(s, 0, (s^{LR} + 1, s^{LC} + 1, s^{CR}, s^{CC}, s^{RR}, s^{RC})) = \theta^{TL}$
 - $T(s, 0, (s^{LR} + 1, s^{LC}, s^{CR}, s^{CC}, s^{RR}, s^{RC})) = (1 - \theta^{TL})$
 - $T(s, 1, (s^{LR}, s^{LC}, s^{CR} + 1, s^{CC} + 1, s^{RR}, s^{RC})) = \theta^{TC}$
 - $T(s, 1, (s^{LR}, s^{LC}, s^{CR} + 1, s^{CC}, s^{RR}, s^{RC})) = (1 - \theta^{TC})$
 - $T(s, 2, (s^{LR}, s^{LC}, s^{CR}, s^{CC}, s^{RR} + 1, s^{RC} + 1)) = \theta^{TR}$
 - $T(s, 2, (s^{LR}, s^{LC}, s^{CR}, s^{CC}, s^{RR} + 1, s^{RC})) = (1 - \theta^{TR})$
 - $T(s, a, s) = 0$ otherwise.⁶

⁶Less formally, this transition function just amounts to the intuition that recommending a post from one source will increment the number of total recommendations from that

- $R(s_t, s_{t+1})$ is defined as:

$$\begin{cases} 1 & (s_{t+1}^{LC} - s_t^{LC}) + (s_{t+1}^{CC} - s_t^{CC}) + (s_{t+1}^{RC} - s_t^{RC}) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- $\gamma = 0.999$

Note that this specific MDP interpretation of the media recommendation problem fits within our general MDP definition from Section 2.

Finally, we define the causal effect of agent actions on the user's exogenous variables in our simulation; this is not something that would be explicitly defined in a scenario with real users, but we need to define it here in order to build our simulation. As mentioned in Section 4, for this effect we took inspiration from recent research into user polarisation on social media, which has demonstrated that showing people who identify with one wing of the political spectrum volumes of content from the opposing wing can often increase user polarisation [4, 5]. We approximate this effect with the following causal relationship between the recommendation at time t , and the value of θ_{t+1}^T :

- If the user is right-wing, and $a_t = 0$ (a left-wing recommendation), then $\theta_{t+1}^{TR} = \min(p\theta_t^{TR}, 1.0)$ for some random variable $p \sim P$ where $\mathbb{E}[p] > 1.0$. We call p the 'polarisation factor.'
- The same effect applies for left-wing users with $a_t = 2$ and θ_{t+1}^{TL} .

E THE INFEASIBILITY OF A SAFE RL-BASED APPROACH

This appendix expands upon our formalisation and simulations' implications for the future of research in the space of RL-based media recommender systems. Based on these reflections, we reach the conclusion that user tampering is almost unavoidably learnable for commercially viable media recommender systems built entirely with RL, given pragmatic constraints on what metrics for recommendation success are aligned with media platforms' commercial interests.

It is assistive to highlight one fundamental assumption that underpins the designs of all the existing and future RL-based recommender systems to which the analysis in this paper applies; this is the assumption that the reward metric, or the definition of 'success' for a recommendation, is at least partly defined in terms of observable user behaviour in reaction to that recommendation. These behaviours could include clicks, likes, comments, time spent reading, the act of sharing the content with other users, and more. For the rest of the section, we refer to this assumption just as 'the user-focused assumption.'

It is also helpful to remind ourselves of the fundamental goal which has motivated research into RL-based media recommender systems; this is the goal of creating systems which are capable of accounting for the causal relationships between its current choice of recommendation and its future actions' rewards. For the rest of the section, this goal is referred to just as the 'the temporal goal.'

If we set out to design an RL-based recommender system which is capable of satisfying the temporal goal, *given* the user-focused

assumption, it is a core requirement that the user's behavioural responses to recent recommendations must be included in the state space that the RL algorithm perceives. Otherwise, the kinds of temporally sophisticated policies being pursued would simply not be achievable; the algorithm needs to be aware of the context of previous recommendations at each decision-time in order to learn the causal effects of particular recommendations across time.

Now, we know that this core requirement implies that the random state variable at each time-step is dependent on the exogenous variable θ^T at the preceding time-step. This goes back to the Alice-Bob example of Section 2; to recap, a depiction of the causal structure of the media recommendation problem which does *not* include this dependency at each time-step is incomplete, because it cannot account for why the transition probabilities from a given state after a given action can differ between users.

Given the premise that there is also a causal link between a recommendation at one time-step and the user's exogenous variable at the next time-step (also discussed in Section 2), the dependencies discussed in the last paragraph then imply the existence of instrumental goals to manipulate θ^T at nearly all time-steps in a recommendation episode.⁷

Finally, as Section 3 demonstrated, the existence of these instrumental goals implies that user tampering is learnable for the recommender.

So, given the user-focused assumption, achieving the temporal goal with an RL-based recommender implies that user tampering is learnable for that recommender.

On a positive note, this amounts to saying that if we either reevaluate the user-focused assumption such that the metric for recommendation success is independent of observable user engagement, or we avoid pursuing the temporal goal altogether, then the resulting recommender system will not be able to learn to exploit user tampering. But, this is more easily said than done. Note firstly that if we abandon the temporal goal, the motivation for RL-based media recommendation is itself lost; so if we are committed to the question of whether we can design a RL-based media recommender where user tampering is not learnable, the question that needs to be asked is: Can we realistically reject the user-focused assumption in recommender system design?

Unfortunately, it is essentially a given that any commercial media platform's recommender system will satisfy the user-focused assumption. Despite academic suggestions of more ethical success metrics such as user well-being [13], commercial platforms' primary interest is in keeping their users engaged. Observable signals of user engagement such as clicks, likes, and the rest are therefore practically guaranteed to constitute part of how recommendations' success is evaluated.

It is for this reason that we deem the prospect of RL-based media recommender systems for which user tampering is not learnable to be infeasible. While diverging from the user-focused assumption allows us to avoid the phenomenon *theoretically*, it is implausible that this step would actually be taken by commercial media platforms if and when they were to use RL-based recommendation,

source so far, and also increment the number of clicks on that source's posts with the relevant probability.

⁷This excludes the first time-step because this precedes any actions being taken, and the last time-step because no future opportunities for reward exist.

given that this would run directly counter to the interests of stakeholders in those platforms. We argue that this motivates significant reconsideration of the ethical status of RL-based recommendation as an approach to news and social media content recommendation.

It is also worth emphasising that this paper's results are expected to generalise to several domains other than political media and polarisation – while this has been the topic *du jour* in many discussions around recommender system ethics and provides an interesting basis for a case study like ours, it is far from the only context in which user tampering could manifest. Some other theoretically possible cases include the following:

- A specific beverage company often constructs its advertisements around a person drinking their product after playing

basketball. The recommender learns to recommend basketball-related posts/articles (which may cause the user to develop an interest in the sport), and *then* to recommend this ad, because it has learned that this combination of recommendations causes the ad to generate more engagement.

- Posts relating to a specific singer generate unusually high engagement on some media platform, perhaps because their fans' preferred forums for discussing the singer's music are the comment sections of such posts. The recommender learns to recommend posts about the singer to some users who currently show no interest in them, because by doing this it can cause some of those users to become fans. Further recommendations about the singer to these users are then likely to generate high reward.