Engineering a Safer Recommender System

Liu Leqi [1] **Sarah Dean** [2]

Abstract

While recommender systems suffuse our daily life, influencing information we receive, products we purchase, and beliefs we form, few works have systematically examined the safety of these systems. This can be partly attributed to the complex feedback loops. In this work, we take a systems safety perspective and focus on a particular feedback loop in recommender systems where users react to recommendations they receive. We characterize the difficulties of designing a safe recommender within this feedback loop. Further, we connect the causes of widely covered recommender system failures to flaws of the system in treating the feedback loop. Our analysis suggests lines of future work on designing safer recommender systems and more broadly systems that interact with people psychologically.

1. Introduction

Recommender systems are large-scale socio-technical systems that actively shape the information we receive, and thus the beliefs we form and preferences we develop. Despite their prevalence and the potential long-lasting impact they may have on individuals and the society, few works have systematically examined the *safety* of recommender systems. This may be for many reasons. To name a few, examining the safety of any system requires deciding what constitutes an accident or hazard. In the context of recommender systems, these concepts have not been well-defined. In addition, recommender systems have a complex nature with multiple *feedback loops* involved. Content providers create material to be distributed on social media platforms, crafting content that will successfully find an audience. This content will be distributed to users by a *centralized* recommendation algorithm and may be scrutinized further by moderators. These users may provide feedback on the recommendations for example, by clicking, liking or flagging them, which affects the recommendation algorithm, moderation process, and eventually even the type of content which becomes successful and popular.

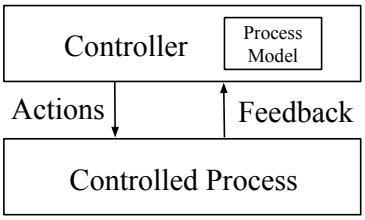

Figure 1. In a control/feedback loop, the controller takes actions and the controlled process provides feedback. As we discuss in Section 2, the controller needs to maintain a process model that captures the process it is controlling.

In this work, we focus on a particular feedback loop in the system, which we term as Recsys-User. Recsys-User captures the interaction between the recommender system (later referred to as algorithm) and an individual who consumes the recommendations (later referred to as user). In this feedback loop (Figure 1), the *controller* is the algorithm that takes recommendations as its *actions*. The *controlled process* that the controller operates on represents the user. Finally, the *feedback* is provided by the user's reaction to the recommendations. Considering human agency and free will, this language is not a perfect fit. We use it to evoke the safety engineering literature, not as an endorsement for the possibility or desirability of "mind control." The recommender-as-controller perspective centers the problem of algorithm design. We note that it is also possible view the user as the controller and the algorithm as the controlled process. Though not our focus, such a perspective illuminates issues related to user experience and recourse (Dean et al., 2020).

We examine Recsys-User from a system safety perspective (Leveson, 2016) and illustrate how certain problems with existing recommender systems can be traced back to control flaws in Recsys-User. In Section 2, we introduce the four conditions that are required to control any process. We then provide a detailed discussion on the unique characteristics of Recsys-User through these four conditions

[1]Machine Learning Department, Carnegie Melon University, Pittsburgh PA, USA [2]Department of Computer Science, Cornell University, Ithaca NY, USA. Correspondence to: Liu Leqi <leqil@cs.cmu.edu>, Sarah Dean <sdean@cornell.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

(Section 3). These characteristics make the design of a safe controller in this setup hard. We analyze some existing issues of recommender systems (that are presented in news articles) through this perspective and connecting them to potential control flaws in Recsys-User (Section 4). Finally, we propose new research directions on developing safe recommender systems (Section 5).

2. Preliminaries

At the core of control engineering is the design of controllers in a feedback loop. In the context of machine learning, reinforcement learning (RL) studies the algorithms for obtaining optimal controllers under uncertainty. Our work uses concepts from these subjects to examine recommender systems. Below, we present the four conditions required to control a process (Ashby, 1957; Leveson, 2016):

- Goal condition: The controller is required to have a goal, which includes the objective as well as the safety constraints. This is in contrast to the goal in many existing RL setups, where the condition is only specified by the objective (i.e., reward) function.
- Action condition: The action condition specifies the actions that the controllers can take to affect the controlled process.
- Model condition: The controller must possess a model of the process it is controlling. As discussed in Leveson (2016), the model condition is of great importance for a controller, since it provides the current state of the controlled process and will be used to estimate the effect of different actions. The model that the controller maintains is denoted by the *process model* (Figure 1).
- Observability condition: The controller requires feedback, which is specified by the observability condition. The feedback is limited by ways of observing (measuring, assessing) the controlled process. In the language of partially observable Markov decision processes (POMDPs), this condition is specified by the observation function that maps a full state to an observable state.

In the following, we will illustrate how these four conditions exhibit in the Recsys-User loop.

3. Characteristics of the Recsys-User loop

To understand the difficulty of designing a controller in the Recsys-User loop, we first need to understand how it differs from control settings that we are more familiar with. Putting it in a broader scope, depending on the *nature* of the controlled process, we contrast feedback/control loops in physical systems, where the controlled process is of physical nature (e.g., a robot vacuum that controls the cleanliness of the floor; a car in smart cruise control mode controls the speed and lane position), with psychological systems, where the controlled process is given by the behavior and psychological state of a human (e.g., in Recsys-User, the controlled process captures the preference of a user; for a mental health chatbot, the controlled process reflects the mental health of the user).

Physical systems are the focus of much of the control engineering literature. The story is similar in machine learning, with much of the RL literature focusing on physical systems. For example, in OpenAI gym (Brockman et al., 2016) or other standard benchmark setups for testing RL algorithms (Kaelbling et al., 1996; Lillicrap et al., 2015, and references therein), the controlled process (environment) simulates physical phenomena of the world.[1]

3.1. Physical and psychological systems

We compare physical and psychological systems in terms of the four conditions required to control a process: the goal, action, model and observability conditions (Table 1).

Goal Condition A key aspect of the goal condition is safety constraints. For a physical system, the safety constraints are often well-defined, since there are existing regulations and clear definitions on accidents or hazards for these systems. In the context of self-driving car, one clear safety constraint for the system is avoiding collisions with other cars. In contrast, for psychological systems, the safety constraints are not well-specified due to the difficulty of precisely defining what should be considered an accident or hazard. Additionally, there are very few existing laws and regulations on these systems that straightforwardly translate into technical specifications. For example, in the context of social media platform, while there are certain guidelines on content moderation, existing laws do not offer constraints on content curation (which is the main functionality of recommender systems) (Telecommunications Act of 1996).

Action Condition The nature of the actions taken by the controller depends on the nature of the controlled process. For example, in the context of autonomous driving car, the actions (e.g., steering and acceleration) control the physical condition of the car; in the context of content recommendation, the actions (the recommendations) control the information (the content) presented to the user. While steering has a clear effect on the trajectory of a car, the effect of information presented in recommendations is more difficult to define, especially if users interact with many different

[1] One exception is the large number of game environments, though these ultimately have more in common with physical systems than psychological systems due to their well-defined goals and determinism.

	Physical	Psychological
Goal Condition	Well-defined	Unclear safety constraints
Action Condition	Controllable	Less controllable
Model Condition	Informed by Physics, Estimable	Informed by behavioral sciences, Hard to estimate
Observability Condition	Partially observable	Partially observable

Table 1. Differences between physical and psychological systems in terms of goal, action, model and observability condition.

sources of content. In general, the assessment of the action condition (whether the action affects the controlled process) relates closely to the observability condition (e.g., whether the effect may be observed).

Model Condition Process models are maintained by the controller for anticipating the evolution of the controlled process. We discuss three important aspects: state definition, dynamics definition, and dynamics estimation.

The ideal state captures the sufficient and necessary information to control a process at a given time (analogous to the full state in POMDPs). Due to our understanding of physics, this state is well-defined for physical systems. It may include the status of the controller (e.g., vacuum speed and position) and the controlled process (e.g., cleanliness and layout of the room). For psychological systems, it is much more ambiguous. In the context of recommending an article, consider the following questions: Is it sufficient to know what the person desires at the moment? Do we need to know their mood and long-term goal? The ambiguity of an ideal state for psychological systems precludes assessment on the quality of the estimated state, since the state that one should aim to estimate is not well-defined in the first place.

The dynamics describe how the ideal state evolves over time, given an initial state, actions, and external disturbances. The ability to model dynamics a priori differs between physical and psychological processes. The laws governing physical systems are generally well understood, from Newton's laws of motion to Maxwell's equations. While approximations are made (e.g. rigid body mechanics ignore flexion), the broad strokes can be correctly captured by tractable dynamics functions. More fundamentally, many controlled physical processes are *engineered*, meaning that they can designed and built such that these approximations remain valid. Psychological processes are less understood, and it is difficult to imagine that human behavior can be described by computationally tractable functions. Furthermore, there is likely considerable variability among people.

If the dynamics function cannot be fully modelled a priori, it may still be possible to estimate it from data. The difficulty of such a task for psychological systems stems from the underying difficulty in specifying tractable models discussed in the previous paragraph. If the dynamics were known to have a linear form, for example, it would be possible to estimate the parameters and make concrete guarantees about accuracy. Though not all physical phenomena can be described in terms of linear relationships, engineered systems can be built such that linear (or other simple parametric) approximations are valid. On the contrary, it is unclear what assumptions are reasonable to impose on psychological processes. Even in settings where the ideal state is well-defined (e.g., what the user wants to read next), we do not know the structure of the dynamics that one should use (e.g., whether it is time-invariant or linear) and hence cannot assure that the dynamics are estimable using data.

Observability Condition If a system is observable, then the underlying state can be estimated from the observations. If a system is not observable, then there is a range of possible states that the system could be in, meaning that the observations so far cannot delineate which one is true—and in particular cannot predict which path the system might take moving forward. Observing a physical process involves different measurement tools, i.e., sensors. The number, placement, and precision of sensors is often a design decision, so engineers can ensure adequate observability. On the contrary, for psychological systems, we often only observe behavioral signals (e.g., revealed preferences, whether a person has clicked on a recommendation or not) instead of underlying psychological quantities (e.g., true preferences, happiness). Though there are many surveys designed to assess the psychological state of a person (e.g., Fordyce (1988); Lyubomirsky & Lepper (1999); Tombaugh & McIntyre (1992)), the precision of these approaches is not matched to the requirements of engineering system analysis. Furthermore, there are not standardized methods for estimating these psychological quantities from behavioral signals. Crucially, this fact ensures that identifying harms for psychological systems is much harder than for physical systems.

3.2. Autopilot and Up-Next

To examine more closely how the differences between physical and psychological systems affect the design and safety analysis of controllers in these systems, we discuss two illustrative examples. We use autopilot, a physical system that steers, accelerates and brakes a car automatically, and

an instance of Recsys-User called Up-Next, a psychological system that recommends the user content to read next.

In autopilot, the goal, action, model and observability conditions are straightforward to specify. In particular, the objective and the safety constraints are defined on observable physical states—driving the car to the desired location efficiently without violating any traffic laws. The process model maintained by the controller can be based upon laws of physics and refined using a large set of historical data. Observations of the feedback depend upon the sensors installed on the vehicle.

For Up-Next, these conditions are ambiguous, making the design of the controller much more difficult. For the goal condition, the objective is rather clear—getting the user to click on the recommended item—since it is defined on observable behavioral signals (e.g., click or not click). However, the safety constraints are less well-defined, since they likely involve unobservable psychological states—for example, that the recommendation should not cause the the user to become more depressed. For the model condition, the recommender system may maintain a black-box model for predicting the user's behaviors. Such models may never be fully accurate due to unmeasured external factors (e.g., picking up a new hobby changes viewing habits), but they may be sufficient for the goal of gaining more clicks. One might argue this is why modern recommender systems "work." On the other hand, accounting for safety constraints requires the estimation of latent psychological states, for which we do not have well-established methods and guarantees. Finally, the observability condition for the controller depends upon the feedback channels provided to the users. In current recommender systems, the feedback is often implicit (e.g., user's watch time of a recommendation) rather than explicit (e.g., user's direct indication on preference towards the recommendations), and is not sufficient to estimate latent psychological states.

4. Categorization of causes of safety issues

We now use the process control perspective on the Recsys-User feedback loop to categorize failures in real-world recommendation systems. This is a prelimary analysis meant to illustrate the utility of the framework. We draw examples of real-world failures from news coverage.

MSI and toxicity Facebook overhauled its newsfeed algorithm in 2018, introducing a Meaningful Social Interactions (MSI) metric to give higher priority to posts with long comments. Though publicized as a change to increase the health of conversations on the platform, a Wall Street Journal investigation uncovered internal reports confirming that the change increased toxicity and divisiveness (Hagey & Horwitz, 2021). This failure can be understood at the level of the *process model*: a failure to anticipate that upranking posts with long comments would lead to anger and toxicity. It's plausible that when the metric was in development, long comments were not strongly correlated with negative content, and only after deployment was the correlation amplified by human behavior. An additional dynamic arises from the incentive structure for content creators; in an email the BuzzFeed CEO cited a "pressure to make bad content or underperform." More than just amplify toxic content, this changed may have led to the creation of more of it.

YouTube Kids Bridle (2017) published an investigation of disturbing videos on YouTube Kids, including violent and sexual content. To fix this issue, the company switched from passive moderation of flagged content to a proactive stance, relying on pre-approved content creators (Koh, 2022). The passive approach failed because of *inadequate feedback*: young children may not realize a video is disturbing or have the ability to react, so low quality videos may not receive proportionally lower watchtime, and inapropriate content may not be flagged for moderation until a supervising adult intervenes.

Discriminatory ad delivery Ali et al. (2019) demonstrated that employment and housing ad delivery on Facebook was biased along racial and gender lines. This bias was due both to complex bidding markets and Facebook's own relevance predictions. Discrimination on the basis of race and sex in housing and employment is illegal under U.S. federal law. Setting aside the market complexities and focusing on relevance targeting, there is an immediate dissonance between presenting ads equally across demographic categories and choosing to target individuals most likely to click. Thus a mismatch between existing regulations and the *goal condition* led to this failure.

5. Future Directions

Engineering safe recommender systems will require multidisciplinary efforts. Building on the system safety perspective, we present some future directions. First, the goal condition: In order to define safety constraints for recommender systems, we need to decide what constitutes as hazards. This requires normative discussions among ethicists, system designers, users and law makers. Second, the model condition: We need to better understand and estimate the latent psychological process of individuals. One component is the creation of detailed datasets on individuals' (psychological) experiences when interacting with these systems. Another component is developing machine learning methods for estimating sophisticated (possibly nonlinear) latent space models and designing safe RL algorithms operating in POMDPs. Third, the observability condition: Users should be provided with more channels for providing feedback. Designing and testing these new feedback mechanisms require collaboration between industry and academia.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

Ashby, W. R. An introduction to cybernetics. 1957.

Bridle, J. Something is wrong on the internet. https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2, 2017. [Online; accessed 31-May-2022].

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Dean, S., Rich, S., and Recht, B. Recommendations and user agency: the reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 436–445, 2020.

Fordyce, M. W. A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social Indicators Research*, 20(4):355–381, 1988.

Hagey, K. and Horwitz, J. Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=series_facebookfiles, 2021. [Online; accessed 31-May-2022].

Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Koh, Y. How YouTube Kids Cleaned Up Its Act. https://www.wsj.com/articles/how-youtube-kids-cleaned-up-its-act-11646476200, 2022. [Online; accessed 31-May-2022].

Leveson, N. G. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lyubomirsky, S. and Lepper, H. S. A measure of subjective happiness: Preliminary reliability and construct validation. *Social indicators research*, 46(2):137–155, 1999.

Telecommunications Act of 1996. 47 U.S. Code § 230. 1996.

Tombaugh, T. N. and McIntyre, N. J. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.