In the Mirror: Using Chess to Simulate Agency Loss in Feedback Loops[1]

Organisers: Catalin Mitelut, Esben Kran

Abstract

Machine learning is increasingly deployed in real world contexts, often with unexpected consequences. Recommender systems are commonly used to filter and suggest content, influencing the information we consume, the products we purchase, and the beliefs we form. Their continuous use forms a set of complex feedback loops, which few works seem to systematically examine. We frame this question and provide a brief overview of existing research in order to explore the ramifications for human agency of deploying machine learning models that include human feedback. We propose a research methodology for evaluating agency loss through human interaction with personalised behavioural chess models. Our analysis suggests a line of future work applying memetic models of human chess players - that we call *mirrors* - to evaluate changes in their strategic style. Our hypothesis is that we would see their existing strategies reinforced, highlighting the feedback loops inherent to many AI-based recommender systems. In a world where humans need ways to process exponentially growing amounts of data, recommender systems are proliferating - and we suggest that there is a gap in the literature when it comes to understanding when to use them, and in combination with what other methods, to deliver the best outcomes for users.

Keywords: Agency preservation, AI safety, memetic models, recommender systems, human-centred AI

[1] Research conducted at The Agency Challenge, 2023
(see https://alignmentjam.com/jam/agency)

1. Introduction

Recommender systems are used to present information to users in a variety of domains including music, video, dating, and advertising. Filtering content for predicted user preferences - based on past consumption - allows users to navigate the wealth of digital information available. However, with a system designed to invisibly narrow down the amount of information presented to users, comes selection bias - recommendation systems encode users' content preferences and biases (Dean et al., 2019) and by repeatedly presenting them with content that fits these criteria, also have the potential to reinforce those biases.

These systems operate on the basis of feedback loops where user preferences determine content suggestions. Initially, users' preference for certain content is partly driven by inherent biases, cultural influences, or the allure of sensational material, a phenomenon known as selective exposure (Thorburn et al., 2023). In response, the recommender system, seeking to maximise engagement, tailors its suggestions to mirror the user's past interactions - presenting more similar content. As users interact with this personalised content, it reinforces preexisting beliefs and preferences, aligning them more with the content they are consuming. This in turn results in users becoming increasingly selective in their choices, either consistently opting for familiar sources or avoiding others, creating a feedback loop.

This presents a number of challenges - both for individual outcomes and society as a whole. For the Chess learning case study proposed later in this paper, our suggestion is that if applied alone recommendation engines may reinforce flaws in players' approaches rather than helping them improve. Socially, recommendation engines may play a key role in hot button topics related to content addiction and polarisation.

From the perspective of addiction, it is possible that optimising for user engagement can lead to showing users content that is addictive to them - for example related to gambling, substance addiction, or pornographic content. This has the potential to exacerbate self-control issues and spill over into other areas of users' lives (Dean et al., 2019). A study that paid individuals to not use specific applications (including recommender driven social media) observed a long-term drop in subsequent use of these apps, indicating that social media use is habit-forming for some people (Allcot et al., 2021). Recommender systems may also lead to extreme and divisive material, and are often blamed in the media and public conversation for radicalising users, and showing teens harmful body-image related content (Bengani et al., 2022).

These demonstrated issues in the deployment of large scale machine learning-based systems in feedback with people highlight the need for analytical tools that anticipate and prevent problematic behaviours from being exacerbated by recommender systems (Leqi et al., 2023). We hope that future recommender systems will be able to measure the degree to which suggested content meets the information needs of its users, for which Dean et al. suggest the metrics of relevance, coverage, and diversity. Understanding the effects of recommendations on user agency requires treating humans as a component in the feedback loop, which necessitates modelling human behaviour.

We propose that existing modelling of individual human behaviour in chess provides a rich source for analysing human behaviour in an environment where agency can be clearly analysed, creating the opportunity for a novel agency simulation environment in which humans interact with their personalised AI models, which we refer to as *mirrors*.

2. Background

Learning Models of Individual Behavior in Chess (McIlroy-Young et al., 2023), demonstrates the ability to develop machine learning models that closely mimic the individual decision-making style and agency of specific human players in the game of chess. The authors leverage a large public dataset of human chess games from LiChess to train personalised models, each targeting the moves of a single player. Through a combination of transfer learning techniques, including fine-tuning and adaptations of the AlphaZero reinforcement learning architecture, they are able to significantly enhance the prediction accuracy for a given player's moves compared to prior chess AI systems and models trained only on generalised human play at a certain skill level. Importantly, these personalised models maintain their superiority in predicting the target player's moves regardless of the move quality, accurately modelling both good moves and blunders.

The high specificity of the models to individual style is indicated by their ability to distinctly identify which player made a series of moves with 98% accuracy, successfully performing player attribution when provided with a sample of games without being pre-trained for this task. The authors posit that this approach could enable personalised AI systems that better align with human agency in general, rather than just exhibiting human-like behaviour in aggregate. McIlroy-Young et al. indicate that by tightly coupling human actions with AI recommendations, it is possible to develop intelligence systems that enhance rather than supersede human agency.

While chess serves as a more controlled initial domain to develop such techniques, the underlying methods could eventually translate to systems that preserve agency in higher-stakes human-AI interactions, such as healthcare. Overall, this work illustrates that, by tightly coupling what humans do, what AI recommends, and how humans think within a feedback loop, it is possible to develop intelligent systems that enhance rather than supersede human agency. The research lays the groundwork for further investigating this approach and its implications across different applications.

3. Proposal

A. Study Overview

We propose a study to investigate how repeated matches between humans and their mirrors could reinforce predictable flaws and reduce strategic diversity in human chess play. This is motivated by a desire to apply the concept of feedback loops as described in social media to an environment that can be more easily measured, and thus where we can assess human-AI interaction in the context of human agency loss.

We hypothesise that competing repeatedly against an AI encoded with one's own limitations will reinforce predictable bad habits in human participants, amplifying the player's inherent tendencies in decision-making, possibly narrowing their strategic horizon. The study is designed to test this by quantifying changes in performance, move quality, playing style, and predictions of the mirror model. Specifically, we will analyse ELO, move quality as judged by chess engines, changes in opening choices, and model calibrations before and after the training period. Significant decrease in ELO rating, increase in blunders, and convergence to a narrower band of strategies would demonstrate the risks of personalised AI models potentially reinforcing flaws through feedback loops without sufficient *diversity* in an informational diet.

Having created an individual player model, we can create a relatively closed simulation environment, where players will compete against their mirrors within specific timeframes. Experiments should be conducted to determine the ideal period after which to fine-tune the mirror with further updates regarding the games played between the model and the player. As more games are played, the mirror will continue to track subtle shifts in player strategies and tendencies if the player behaviour evolves. This simulated feedback loop could be measured using a number of benchmarks, considering chess as a statistical environment. Key indicators include the amplification rate of specific tendencies, gauging the frequency of specific mistakes that the player makes, especially when echoed by their AI counterpart. Another significant marker is the gradual narrowing of strategic diversity as captured by the number of possibilities proposed by the model for each move. Central to this analysis is the creation of an initial player model which can be used as a baseline from which deviations could be calculated. In addition, initial benchmarks of both the player and their mirror competing in a series of games of escalating difficulty in conventional settings against human players over a set ELO range would allow for skill changes to be better measured upon completion of the human-mirror playing period.

B. Evaluating Change in Player Behaviour

As part of understanding the effects of the mirrors on chess player’s strategies, we propose specific metrics aimed at evaluating the nuances in their stylometric evolution.

Mistake Frequency: McIlroy-Young et al calculate the quality of moves by the change in estimated win probability before and after each move, using Stockfish evaluations. Moves that decrease win probability by over 10% are classified as blunders. The frequency of blunders over a series of games could be calculated and compared before vs after competing against the mirror; an increase in frequency blunders would indicate a reinforcement of mistakes (McIlroy-Young et al., 2023).

Strategic Diversity: Each possible move of the mirror is represented as a 1858-dimensional vector. The diversity of moves chosen could be quantified by metrics like the entropy or variance of these move vectors. A decrease in entropy/variance over a series of games would indicate reduction in strategic diversity, highlighting the importance of requiring a consistent regimen of games with consistent time controls.

Opening Choices: The distribution of openings played across games could also be quantified. Reduced variance in openings chosen before vs after training against the AI model would signify convergence on narrower opening preparation. Nibbler, the unofficial LeelaChessZero GUI provides methods to identify openings in large quantities of games, from changes in the frequency of use of specific openings could be calculated. This would require significant work on assembling a profile for individual players before

Model Uncertainty: The mirror’s output is a probability distribution over next moves. Comparing the entropy of this distribution for the human's moves before and after training would indicate if the model is becoming more certain of the human's predictable mistakes.

Centipawn Loss: Centipawn loss is used in the mirrors to evaluate move quality by approximating loss versus perfect play. The average centipawn loss per game could be compared before and after AI training to quantify changes. In summary, strategic diversity metrics combined with move quality benchmarks would allow quantifying potential harms like increased mistakes and narrowed play arising from the hypothesised human-AI feedback loop.

4. Practical & Future Work

The implementation of a pipeline for initially fine-tuning and iteratively updating personalised chess models is beyond the scope of this two-week project given its complexity and significant unknowns. We aim to apply this methodology in the coming months and will provide updates to the maia-individual repository which it currently requires; the data generation scripts only support Lichess game formats, which must be processed for the supervised learning component. Significant further effort is required to enable Chess.com game support, which we attempted in week 1 but were unable to complete. Chess.com integration is an important priority due to the larger player pool that would enable recruiting a more diverse sample of chess players. For the current project, we prioritised presenting a clear conceptual framework and rationale, alongside quantifiable metrics to evaluate changes in player strategy relative to their baseline model. While we hope to implement the full study soon, this project focused on establishing the motivation and methods for assessing potential harms arising from human-AI feedback loops.

5. Discussion and Conclusion

A. Ethical Implications

We propose a number of controls in order to ensure that study participants are respected. It is essential that participants provide informed consent, where the study's aims are clearly explained, as well as the hypothesised ramifications on their playstyles. Withdrawal should be made clear to all participants as a viable and recommended option in the case of increased stress or spillover in their personal lives. Anonymizing and securing the data is essential given the sensitive skill insights possible with the tools provided by McIlroy-Young et al., 2023. In addition, we hope to collaboratively engage participants through sharing results clearly and in detail after the study, including soliciting feedback.

We recognise that chess players are often passionate about the game, and often want to improve their play in order to win. Ensuring that we support their mental health if the hypothesised reduction in strategic diversity and increase in mistakes and affects their psych is an essential component of this study, and merits further research to propose mitigation strategies.

B. Limitations of the Mirror Approach

Chess constitutes a narrow, well-defined domain with clear optimal play and mistakes. Recommender systems instead suggest content across complex real-world topics like news and social media, where consensus on high quality or dangerous material is absent. The AI models in chess merely encode gameplay strategy. However, recommender systems in the wider world infer multidimensional user models encompassing interests, beliefs, and behavioural patterns. The risks they pose do not stem from less diversity of strategy, but of worldviews.

Additionally, chess models have a transparent set of incentives - winning the game by exploiting mistakes. By contrast, recommender systems have opaque engagement incentives users may not realise shape their information diet (Dean et al., 2019). Chess players voluntarily choose to play against the AI model, whereas users do not actively select the algorithms potentially manipulating their recommendations. The chess environment is fixed, while recommenders operate in dynamic environments with constantly emerging content. Therefore, while the proposed chess study offers an intriguing model for a controlled system, it does not replicate the complex behavioural dynamics emerging from recommender systems deployed at global scale. The risks arise less from strategic maximisation of a known objective, and more from inadvertent misalignments between system and user goals. Significant additional research is required to understand and address the challenges in designing recommender systems that benefit users and avoid unintended harms stemming from feedback loops.

C. Opportunities and Advantages

In contrast, we are hopeful that this research into feedback loops could reveal some of the psychological mechanisms that make recommender systems dangerous, and launch further research into their impact. Beyond the domain of chess, this study constitutes an experimental test of how similar feedback loops that power recommender systems and personalised AI could negatively impact users if designed without care. The dynamics investigated here, arising from models tailored too closely to an individual without diversity, are relevant across many human-AI interaction contexts. Our study design provides a methodology for quantifying resulting harms.

Understanding feedback loops is crucial for several reasons. The framework helps pinpoint research areas such as the need to discern the type of content users engage with, the nature of the content displayed, and the subsequent beliefs the users develop. It is essential to establish a causal connection between these factors to validate the existence of feedback loops as described, considering that their underlying premise is that they significantly impact user agency in widely used online platforms that govern our information diet and patterns of media consumption. There is already significant attention from governance bodies such as the European Commission who describe the use of machine learning algorithms that “produce improved and refined nudges in a self-propelling cycle that is beneficial to [online platforms] but may be detrimental for consumers” (Directorate-General for Justice and Consumers et al., 2022).

Interdisciplinary studies bridging computer science and psychology, such as the proposed personalised chess AI experiment, hold value for enhancing scientific understanding of potential unintended consequences in human-AI interaction. While not an ideal analog, chess provides a controlled abstraction situated between minimal examples and the complexity of societal-scale systems. The proposed analysis of quantitative metrics like mistake frequency and strategic diversity can reveal how repeated exposure to an AI reflecting one's own limitations may negatively impact human performance and narrow cognitive diversity over time. From an AI safety perspective, such revelations improve considerations for human welfare factored into AI design, irrespective of domain specifics. While care is warranted when generalising findings, evidence within chess for how personalization can detrimentally influence strategy may have parallels to how recommender system personalization could lead to unintended effects like narrowed exposure.

Overall, the interdisciplinary nature of the study offers a worthwhile scientific effort to expand knowledge regarding risks of feedback loops on human psychology. Insights from such experiments blending computer science and the behavioural sciences are important for steering the ethical development of human-AI systems. Even imperfect analogs further illuminate the subtle influences personalised AI could exert on human behaviour, serving to inspire future studies across a breadth of domains. We believe that it is in the interests of users and providers to drive toward a future best practice for recommendation engines where users have more understanding of how their content is being selected, and are better protected from the effects of feedback loops. Chess is a zero sum game, but our Instagram feeds do not need to be.

6. References

Dean, S., Rich, S., & Recht, B. (2019). *Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information* [Review of *Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information*]. Department of EECS, University of California, Berkeley, Canopy Crest. https://arxiv.org/pdf/1912.10068.pdf

Thorburn, L. (2023, April 17). From "Filter Bubbles", "Echo Chambers", and "Rabbit Holes" to "Feedback Loops." Tech Policy Press. https://techpolicy.press/from-filter-bubbles-echo-chambers-and-rabbit-holes-to-feedback-loops/

Allcott, H., Gentzkow, M., & Song, L. (2021). Digital Addiction. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3870938

Bengani, P. (2022, April 27). What's Right and What's Wrong with Optimizing for Engagement. https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851

McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., & Anderson, A. (2022). Learning Models of Individual Behavior in Chess. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1253–1263. https://doi.org/10.1145/3534678.3539367

Leqi, L., & Dean, S. (n.d.). Engineering a Safer Recommender System. Retrieved September 24, 2023, from https://responsibledecisionmaking.github.io/assets/pdf/papers/30.pdf

McIlroy-Young, R., Sen, S., Kleinberg, J., & Anderson, A. (2020). Aligning Superhuman AI with Human Behavior. Proceedings of the 26th ACM SIGKDD

International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3394486.3403219

Directorate-General for Justice and Consumers (European Commission), Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., & Rodríguez de las Heras Ballell, T. (2022). Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation. In Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/606365bc-d58b-11ec-a95f-01aa75ed71a1/language-en#