analytics

MDPI

Article

The AI Learns to Lie to Please You: Preventing Biased Feedback Loops in Machine-Assisted Intelligence Analysis

Jonathan Stray

UC Berkeley Center for Human-Compatible AI, Berkeley, CA 94720, USA; jonathanstray@berkeley.edu

Abstract: Researchers are starting to design AI-powered systems to automatically select and summarize the reports most relevant to each analyst, which raises the issue of bias in the information presented. This article focuses on the selection of relevant reports without an explicit query, a task known as recommendation. Drawing on previous work documenting the existence of human-machine feedback loops in recommender systems, this article reviews potential biases and mitigations in the context of intelligence analysis. Such loops can arise when behavioral "engagement" signals such as clicks or user ratings are used to infer the value of displayed information. Even worse, there can be feedback loops in the collection of intelligence information because users may also be responsible for tasking collection. Avoiding misalignment feedback loops requires an alternate, ongoing, non-engagement signal of information quality. Existing evaluation scales for intelligence product quality and rigor, such as the IC Rating Scale, could provide ground-truth feedback. This sparse data can be used in two ways: for human supervision of average performance and to build models that predict human survey ratings for use at recommendation time. Both techniques are widely used today by social media platforms. Open problems include the design of an ideal human evaluation method, the cost of skilled human labor, and the sparsity of the resulting data.

Keywords: recommender systems; summarization; AI bias; information quality

check for updates

Citation: Stray, J. The AI Learns to Lie to Please You: Preventing Biased Feedback Loops in Machine-Assisted Intelligence Analysis. *Analytics* **2023**, 2, 350–358. https://doi.org/10.3390/analytics2020020

Academic Editor: R. Jordan Crouser

Received: 1 February 2023
Revised: 22 March 2023
Accepted: 29 March 2023
Published: 18 April 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There is increasing interest in using AI systems for various intelligence analysis tasks, including information fusion, hypothesis testing, monitoring, and summarization [1]. For example, the Laboratory for Analytic Sciences recently inaugurated the "tailored daily report" (TLDR) grand challenge, which was to automate the creation of a summary of new intelligence similar in format and quality to the President's Daily Brief but tailored to the needs of individual workers across the intelligence community [2]. This multi-year research challenge requires retrieval of relevant source reports from a potentially large and diverse pool of available material, followed by summarization which correctly extracts the most essential information.

The production of such a daily report must begin with the automated selection of the new items of information most relevant to the analyst's task—whether or not the analyst has previously entered a query that would include them. This problem of query-free selection is known as "recommendation" in the computer science literature and has applications across many domains, including e-commerce, entertainment, and social media.

Unfortunately, there is no straightforward algorithmic translation of words like "relevant" and "essential." Current recommendation systems operationalize these concepts largely through proxy behavioral metrics, such as click-through rate, dwell time, likes or favorites, and other types of engagement [3]. This approach is known to produce human-machine feedback loops that result in biases of various kinds [4].

This paper studies the possibility of bias-producing human-machine feedback in AI systems designed for intelligence analysis using the example of recommender systems. Recommender systems are an appropriate place to begin this analysis because (a) they are likely

to be a core component of any practical AI-assisted intelligence analysis system, and (b) there is extensive previous work on recommender-induced bias. The contributions of this paper are:

(1) An analysis of how previously studied engagement-driven recommender biases will apply in the domain of intelligence analysis;
(2) An argument that human-machine feedback loops will bias not just intelligence analysis, but also intelligence collection;
(3) Proposed mitigation strategies based on collecting human evaluations of the analytic quality of recommender output.

2. Recommender System Biases

A recommender is a personalized information filter that selects, for each individual, a small set of items out of a much larger pool [5]. Recommenders differ from search engines in that they typically produce personalized results, and they are often invoked without a specific user query, such as a news recommender presenting significant new events without requiring the user to search for them explicitly.

In the intelligence context, a recommender could be designed to select the most relevant reports for each analyst. The difficulty here is defining the word "relevant". It has long been understood that there is no universal measure of "relevance", as it depends on both user goals and current context [6,7], which the computer cannot directly observe. Instead, existing recommendation systems largely select content based on previously observed short-term behavioral responses, such as clicks, likes, and favorites, which are collectively known as engagement [3].

This reliance on engagement can produce several types of bias, which we classify below.

2.1. Technical Biases

User interface designers have long understood that users are more likely to click items that appear earlier in a list, even if later items are just as relevant. This is known as "position bias". Recommenders usually rank clicked items higher for other users, which can result in runaway amplification of initially random ranking choices, as now-classic experiments with music recommendation have shown [8].

"Popularity bias" is a related phenomenon where a recommender shows only those items most likely to be clicked by a broad group of users. Recommending popular items is not always bad; after all, these are the items that the largest number of people will like. However, "popularity" is not "quality," though the two often correlate. In particular, popularity can be a valuable signal of item quality when exploration costs are in an intermediate regime [9]. Although, ranking by popularity can also result in popularity feedback loops alongside recommendations that serve the majority well while performing poorly for users who are in the minority along some axis [10].

Similarly, there is a risk of feedback effects within a community of analysts. If each analyst preferentially clicks on the top-rated reports, and what one user clicks is taken as a positive ranking signal for other users, then the entire community may end up preferentially reading a particular report or favoring a particular topic.

Positional and popularity biases are well-studied, and a variety of mitigations have been developed [11–13]. This is possible because these types of "technical bias" (following the terminology of [14]) can be defined by reference to an objective baseline. For example, to counteract position bias, one can attempt to counterfactually estimate an item's click rate independent of position [12]. While intelligence applications of recommender systems must consider these types of biases, mitigation techniques developed in other domains should prove adequate.

2.2. Biases Resulting from Incomplete Information

Engagement is a useful proxy for user relevance; certainly, we tend to give more attention to those items that are most valuable to us. This is why engagement prediction is a key ranking signal in most recommender systems [3]. However, engagement and value are frequently misaligned, as in the case of "clickbait". While an intelligence database is unlikely to be filled

with clickbait in the "You Won't Believe This One Weird Trick" sense of the term, there will still be some difference between what titles cause a user to click and what is actually valuable.

The resulting slippage may be tolerable in relatively simple or low-stakes systems. However, when more advanced systems are designed to optimize for engagement, they may find increasingly sophisticated ways to induce humans to click or otherwise rate the results highly, potentially influencing long-term user interests and behavior [15–17].

Simulation studies investigating this effect typically assume that users shift their preferences in the direction of whatever they have previously seen [18]. This can be considered a type of confirmation bias and has empirical psychological grounding in the "mere exposure effect" [19]. Given this assumption, plus the assumptions that preferences shape user choices and recommenders respond to those choices, a number of simulations have shown feedback effects where users consume increasingly narrow subsets of content and/or polarize into disjoint groups which consume different content [4,16,17,20,21].

One standard mitigation is the use of diversification algorithms to prevent too much of the same type of content from appearing in recommendations [22]. However, such diversification methods are based on content similarity metrics, which do not necessarily correspond to intelligence analysis principles. There are also plausible preference shift mechanisms that do not stem from a lack of content diversity [23].

At the present time, it is difficult to assess the degree to which recommender-driven preference shifts are happening in real systems. A recent review of hundreds of studies shows a positive correlation between "digital media" use and polarization [24]. Causal evidence is much more scarce and methodologically difficult because many non-recommender factors can influence polarization trends [24,25] and because external researchers have not been able to perform relevant experimental studies on major platforms. Deprivation studies (where users are paid not to use social media for several weeks) have shown both positive and negative polarization effects in different contexts [26,27].

Nonetheless, the overall correlation between digital media use and polarization is concerning, and the fact that narrowing feedback loops appear under a diverse set of recommender simulation specifications suggests a robust result. Further, we should expect that as advanced AI systems become better at long-term planning, they will influence users to achieve programmed engagement objectives if they possibly can. This may particularly be an issue for the newer generation of recommenders based on reinforcement learning [28,29].

Concretely, what would it look like for a recommender to try to influence you? It might simply show you content that it wants you to engage with. In general, all of us tend to become more aligned with the content we are exposed to—whether through experience effects, anchoring, learning new facts, discovering new interests, pressure to conform, or the illusory truth effect. Alternatively, a recommender might show you content you find strongly disagreeable in order to more firmly entrench your current preferences. It might show you articles suggesting that everyone else is interested in a particular topic to generate FOMO and increase the chance you engage. It might show you a lot of conflicting news accounts to generate reality apathy, then feed you a false conspiracy theory that makes it all make sense. This list is not meant to be exhaustive. There will be many more subtle ways in which a sufficiently capable recommender could influence us, including some that we would not be able to foresee. The above examples simply demonstrate that recommenders could plausibly influence us if they are able to learn how to do so.

At root, these problems result from the fact that the user's needs cannot be accurately inferred from behavior alone [30,31]. Instead, recommenders (and other AI systems) must optimize for some behavioral proxy for user relevance or value. Of course, one can use better proxies than clicks. However, it is probably not possible to provide a formal definition or metric that cannot be gamed in some way. This is a fundamental challenge that has been extensively explored in the AI alignment literature [15,32–34].

3. Bias as a Consequence of Feedback Loops

Recommender output might be "biased" in the sense of presenting an inappropriate selection of items, just as a newspaper might be biased. However, interactive systems can produce types of bias not possible with traditional media due to the formation of feedback loops between recommender systems and their users [4]. Recommenders are designed to respond to human feedback as a signal of relevance or quality. If human users then react in response to the information they are presented with, feedback loops can form between human and machine, as depicted in Figure 1. These loops can produce the positional, popularity, and polarization feedback loops discussed above.

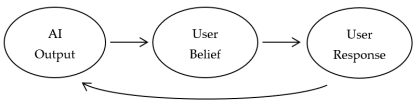

Figure 1. Feedback loops between an AI system that selects and summarizes information and the user. AI output shapes belief which in turn shapes the user response. Systems that optimize for engagement adapt by changing their output to reinforce this response, and the cycle repeats.

This feedback loop includes changes in human beliefs—if it did not, we would not need to be especially concerned about it. Bias-producing feedback loops can operate for a single user, where the personalization process induces successive rounds of confirmation bias as the system drifts towards progressively poorer results. However, recommenders are inherently multi-user systems, and in most cases, other users' previous reactions are a key source of information when inferring what a user wants. Thus, belief-shifting feedback loops can operate within an entire community of users, biasing a group towards certain content or topics. Popularity bias is an example of this effect.

4. Biased Feedback Loops in Intelligence Analysis

Recommender feedback loops might bias intelligence analysis in at least two ways: what is consumed from available information and what intelligence is collected at all. Both might be considered types of confirmation bias, though they act on different levels.

Considering intelligence on weapons of mass destruction preceding the 2003 US invasion of Iraq, Pillar notes that the analyst's job is, in large part, deciding where to look.

> *On any given subject, the intelligence community faces what is, in effect, a field of rocks, and it lacks the resources to turn over each one to see what threats to national security may lurk underneath. In an unpoliticized environment, intelligence officers decide which rocks to turn over based on past patterns and their own judgments. However, when policymakers repeatedly urge the intelligence community to turn over only certain rocks, the process becomes biased. The community responds by concentrating its resources on those rocks, eventually producing a body of reporting and analysis that, thanks to quantity and emphasis, leaves the impression that what lies under those same rocks is a bigger part of the problem than it really is* [35].

In this case, Pillar believes that the intelligence process was corrupted by political influences (which would be in violation of fundamental directives, such as ICD 203 [36]). However, his analysis holds for any bias which influences "which rocks are turned over".

If a recommender-human feedback loop results in an analyst preferentially looking at certain types of information, they may, in turn, represent some problems as being more significant than they actually are while paying insufficient attention to other important issues. Worse, they may fail to find contrary evidence for significant analytical conclusions. Although analysts are directed to consider contrary information and alternative explanations [36], reliance on AI systems has been shown to reduce human skepticism in some circumstances [37].

It is perhaps less appreciated that biases in the analysis process can create biases in the collection process. That is, limited resources require tradeoffs in which information is collected for analysis, and those tradeoffs are shaped by previous analytical conclusions. Detecting or mitigating biases in collection requires a theory of how limited collection

resources should be ideally targeted for a given intelligence question. One such theory is Heuers' analysis of competing hypotheses method, which posits that:

> *You should seek evidence that disproves hypotheses. Early rejection of unproven, but not disproved, hypotheses will bias the subsequent analysis because one does not then look for the evidence that might support them* [38] (p.98).

Here, seeking evidence includes more than just searching existing databases; it extends to which information is collected in the first place. Thus, an inappropriate algorithmic system might bias not just how available information is interpreted but what kinds of information are available at all. This creates a secondary feedback loop, indicated by the bottom path in Figure 2, which is slower but perhaps more consequential.

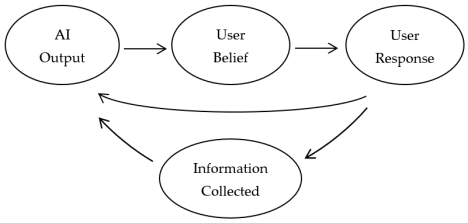

Figure 2. Feedback loops between an AI system that selects and summarizes information, the user, and information collection. Users prompt intelligence collection based on their beliefs, producing a second-order feedback loop that shapes the pool of available information.

5. Mitigating Recommender Feedback Loops

The biases described above can be understood as consequences of the fact that the optimization objectives of an engagement-driven recommender system do not accurately represent human goals. The usual behavioral signals are not enough; clickbait provides a simple example of why, but the problem is quite broad [15] and has close connections to the problem of inferring "preferences" studied in behavioral economics and psychology [30]. The solution is straightforward in principle but challenging in practice: the machine must be given other kinds of human feedback. This is the guiding principle of both theoretical solutions to the general alignment problem (such as "assistance games" [15]) and the practical mitigations discussed below.

5.1. User Controls

If users are able to detect errors or biases in the recommender output, then they may be able to correct the system by adjusting settings or parameters. A large number of different recommender control interfaces have been tested, including enabling the user to select between different algorithms [39], adjust the weight accorded to different content features [40], or interact with visualizations [41].

In the context of mitigating bias-producing feedback loops, however, controls suffer from two major problems. First, few users actually adjust the controls on real recommender systems [40], which means that user controls are not a plentiful or reliable source of feedback. This is why commercial recommenders, such as Netflix, migrated from explicit user feedback (star ratings on movies) to favoring implicit feedback (clicks and watch time).

Second, users must be able to detect the bias in recommender results before they can mitigate it. The bias-producing feedback loops discussed above operate through changes in user belief (as in Figures 1 and 2), and by definition, people are usually unaware when they are suffering from cognitive biases. Confirmation bias, in particular, has been documented in intelligence analysis tasks [42,43].

Nonetheless, various kinds of recommender controls may prove useful in ensuring that the analysis process is rigorous and unbiased. For example, there are many techniques for increasing the diversity of recommendations in various ways [22]. The analyst might find it useful to ask for additional recommendations from more diverse sources, over a broader time period or on a wider range of topics. With recent rapid advances in natural language process-

ing, it should soon be possible for analysts to specifically ask for disconfirming evidence, a possibility that has been explored in the nascent field of automated fact-checking [44].

5.2. Use of Survey Data

Large recommender-driven platforms have been dealing with the problem of misalignment between engagement objectives and user value for decades [45]. Mitigating such misalignment requires measuring it, and one of the most straightforward sources of additional information is user surveys. A wide variety of surveys are used to evaluate and train commercial recommender systems [5]. Some of these surveys merely ask the user to rate a previous recommendation, but some attempt to measure more abstract constructs, such as whether an item is "worth your time" or contributes to "meaningful social interactions" [46].

Survey data is widely used to evaluate user experiences and perceptions. Surveys are often deployed to evaluate a prospective change versus the status quo in an A/B test and then used afterward for post-launch monitoring. While this is an essential way to detect drift between system goals and actual performance, it has limitations. Survey feedback is relatively slow, after-the-fact, and only gives insight into average responses, or at most, the responses of a modest number of subgroups.

To address these shortcomings, commercial systems have started to rely on the technique of predicting survey responses. In this approach, survey data is used to train a model that predicts how a new user in a new context would answer that survey. When the recommender then selects items for a user, the predicted response for that user is used as a factor in content ranking. This approach is used by YouTube [47,48], Facebook [46,49], and elsewhere [5]. While generating "fake" survey responses may seem a strange approach, conventional recommenders already depend on predicting engagement—future user reactions—which is not drastically different from predicting a survey response.

The great advantage of this survey prediction paradigm is that it provides a way to incorporate arbitrary user feedback at recommendation time. It also allows for modeling the personalized relationship between recommender output and survey responses, which provides individual-level customization. The major challenges include the sparsity and cost of survey data relative to plentiful engagement data, potential human response biases, and of course, the question of what one should actually ask users and how they will respond.

6. Incorporating Human Evaluations into Intelligence Recommendation

The above discussion of the use of survey data suggests that bias might be controlled in the context of recommenders for intelligence analysis by collecting human evaluations of recommender output.

Existing human evaluations of intelligence processes and products are a promising direction for this type of feedback. There exists a method for structured human evaluation of whether a particular report meets the analytic standards set out in ICD 203, known as the IC rating scale [50]. There are other similar rating scales, such as the "rigor" instrument of Zelik et al. [51]. A recent review of the concept of rigor in intelligence analysis identified five major indicators: logical, objective, thorough, stringent, and acute [52]. These or similar rating instruments could be used to provide the feedback needed to control bias in an automated system. In this scheme, a sample of machine output would be evaluated manually according to an established set of criteria, producing a set of numeric scores indicating quality along various axes.

It requires significant human effort to perform such evaluations, and it is not clear who would perform the evaluation on an ongoing basis (randomly selected regular users? A special team of "feedback" analysts?). Despite the effort and cost, no automated intelligence system can be considered credible if it is not evaluated in this way. Indeed, such evaluation is desirable even with extremely simple automated systems where feedback loops are not expected to be a problem.

Ideally, such metrics would be used to evaluate system changes, before deployment, through the usual method of A/B testing. This is a sort of human optimization loop where system designers continually strive to maximize quality metrics.

It is also possible to algorithmically optimize against human-provided metrics by using this feedback to train models that predict human evaluation ratings. Such predictive models can then be incorporated directly into the ranking function of a recommender. This approach is potentially much more powerful than simply evaluating a sample of system output or using such evaluations to decide between design alternatives. Incorporating predictions of evaluation into ranking, while not infallible, offers a way for individual analysts to receive the benefit of expensive human feedback even while performance on their particular needs is never evaluated directly.

The main technical challenges here are cost and data sparsity. Human evaluation data is costly to produce and will therefore be limited as compared to plentiful engagement signals. It remains for future work to evaluate the cost/benefit curve for different amounts of feedback data.

The use of surveys or similar rating instruments also raises complicated questions about survey design and human bias. It is not clear which of the existing intelligence product rating scales, if any, would be most suitable for evaluating recommender output. There is also the question of the unit of analysis. Most existing platform surveys ask about specific items of content—analogous to asking whether a single item was useful to an analyst—whereas it may be more useful to elicit feedback on a set of items chosen by the recommender or even the user's overall impression over a period of time as retrospective judgments are thought to be more accurate in some cases [45].

Surveys, rating scales, and related evaluation methods can also be biased. It can be challenging to establish validity (the survey correctly measures what it is intended to measure) and reliability (survey results are appropriately stable over time and between people) [53]. Further, human analysts are also subject to bias [42,43]. Incorporating human feedback could make automated systems as unbiased as a careful human analyst, but not better.

Looking to the future, there are several other kinds of human input that could help to prevent algorithmic biases and feedback loops. Open AI has demonstrated the value of direct elicitation of pairwise comparisons. In one experiment, humans were asked to choose which of two text summaries they preferred. This information was used to train a reward model, which was, in turn, used to train a reinforcement-learning agent to produce dramatically better summaries [54]. Ultimately, we are probably headed for conversational recommender systems, which might simulate the experience of directing a smart research assistant. Although it may be some time before such systems outperform more conventional tools, research is well underway [55]. Even with such advanced systems, it may remain challenging to determine whether such a system is meeting intelligence standards and, if it is not, how to communicate and integrate useful feedback.

7. Conclusions

Feedback loops between humans and recommender systems can create biases, and there is reason to believe these biases apply to intelligence applications. This may bias the information presented to analysts, the information is tasked for collection as a result, and ultimately what analysts believe and their subsequent conclusions.

The root cause is a misalignment between optimization objectives and human goals. Careful human feedback, rather than the implicit behavioral data that drives most recommender systems, can significantly mitigate this misalignment.

The most promising approach, drawn from existing industry practice, is regular human evaluations of system output. These evaluations could be modeled on existing instruments for measuring intelligence rigor and quality, such as the IC rating scale. The resulting data can be used for monitoring overall system performance and bias. Moreover, it can be used to train human evaluation prediction models that can be incorporated directly into the recommendation process so as to pro-actively produce output that humans are expected to rate highly.

Funding: This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Katz, B. The Intelligence Edge: Opportunities and Challenges from Emerging Technologies for U.S. Intelligence, Center for Strategic and International Studies (CSIS). 2020. Available online: https://www.jstor.org/stable/resrep24247 (accessed on 13 March 2023).
2. Kershaw, K. Creating a 'TLDR' for Knowledge Workers, Laboratory for Analytic Sciences, 31 August 2022. Available online: https://ncsu-las.org/blog/scads-tldr-knowledge-workers/ (accessed on 15 September 2022).
3. Bengani, P.; Stray, J.; Thorburn, L. What's Right and What's Wrong with Optimizing for Engagement, Understanding Recommenders, 27 April 2022. Available online: https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851 (accessed on 21 March 2023).
4. Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; Burke, R. Feedback Loop and Bias Amplification in Recommender Systems. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA, 19–23 October 2020; pp. 2145–2148. [CrossRef]
5. Stray, J.; Halevy, A.; Assar, P.; Hadfield-Menell, H.; Boutilier, C.; Ashar, A.; Beattie, L.; Ekstraud, M.; Leibowicz, C.; Sehat, C.M.; et al. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv* **2022**. [CrossRef]
6. Mizzaro, S. Relevance: The whole history. *J. Am. Soc. Inf. Sci.* **1997**, *48*, 810–832. [CrossRef]
7. Jannach, D.; Adomavicius, G. Recommendations with a purpose. In Proceedings of the 10th ACM Conference on Recommender Systems, New York, NY, USA, 15–19 September 2016; pp. 7–10. [CrossRef]
8. Salganik, M.J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* **2006**, *311*, 854–856. [CrossRef] [PubMed]
9. Nematzadeh, A.; Ciampaglia, G.L.; Menczer, F.; Flammini, A. How algorithmic popularity bias hinders or promotes quality. *Sci. Rep.* **2018**, *8*, 15951. Available online: https://www.nature.com/articles/s41598-018-34203-2 (accessed on 21 March 2023).
10. Ekstrand, M.D.; Tian, M.; Azpiazu, I.M.; Ekstrand, J.D.; Anuyah, O.; McNeill, D.; Pera, M.S. All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 172–186. Available online: https://proceedings.mlr.press/v81/ekstrand18b.html (accessed on 31 January 2023).
11. Zhu, Z.; He, Y.; Zhao, X.; Caverlee, J. Popularity Bias in Dynamic Recommendation. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event, Singapore, 14–18 August 2021; pp. 2439–2449. [CrossRef]
12. Agarwal, A.; Zaitsev, I.; Wang, X.; Li, C.; Najork, M.; Joachims, T. Estimating Position Bias without Intrusive Interventions. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 474–482. [CrossRef]
13. Chen, M.; Beutel, A.; Covington, P.; Jain, S.; Belletti, F.; Chi, E.H. Top-K Off-Policy Correction for a REINFORCE Recommender System. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 456–464. [CrossRef]
14. Zehlike, M.; Yang, K.; Stoyanovich, J. Fairness in Ranking, Part I: Score-based Ranking. *ACM Comput. Surv.* **2022**, *55*, 1–36. [CrossRef]
15. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking: New York, NY, USA, 2019.
16. Krueger, D.S.; Maharaj, T.; Leike, J. Hidden Incentives for Auto-Induced Distributional Shift. *arXiv* **2020**, arXiv:2009.09153.
17. Carroll, M.; Hadfield-Menell, D.; Dragan, A.; Russell, S. Estimating and Penalizing Preference Shift in Recommender Systems. In Proceedings of the Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September–1 October 2021. [CrossRef]
18. Bernheim, B.D.; Braghieri, L.; Martínez-Marquina, A.; Zuckerman, D. A Theory of Chosen Preferences. *Am. Econ. Rev.* **2021**, *111*, 720–754. [CrossRef]
19. Curmei, M.; Haupt, A.; Hadfield-Menell, D.; Recht, B. Towards Psychologically-Grounded Dynamic Preference Models. In Proceedings of the 16th ACM Conference on Recommender Systems, Seattle, WA, USA, 18–23 September 2022. [CrossRef]
20. Evans, C.; Kasirzadeh, A. User Tampering in Reinforcement Learning Recommender Systems. *arXiv* **2021**, arXiv:2109.04083.
21. Jiang, R.; Chiappa, S.; Lattimore, T.; György, A.; Kohli, P. Degenerate Feedback Loops in Recommender Systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019. [CrossRef]
22. Kunaver, M.; Požrl, T. Diversity in recommender systems—A survey. *Knowl.-Based Syst.* **2017**, *123*, 154–162. [CrossRef]
23. Törnberg, P. How digital media drive affective polarization through partisan sorting. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2207159119. [CrossRef]
24. Lorenz-Spreen, P.; Oswald, L.; Lewandowsky, S.; Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **2023**, *7*, 74–101. [CrossRef] [PubMed]
25. Boxell, L.; Gentzkow, M.; Shapiro, J. *Is the Internet Causing Political Polarization? Evidence from Demographics*; National Bureau of Economic Research: New York, NY, USA, 2017. [CrossRef]
26. Allcott, H.; Braghieri, L.; Eichmeyer, S.; Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **2020**, *110*, 629–676. [CrossRef]

27. Asimovic, N.; Nagler, J.; Bonneau, R.; Tucker, J.A. Testing the effects of Facebook usage in an ethnically polarized setting. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2022819118. [CrossRef] [PubMed]
28. Afsar, M.M.; Crump, T.; Far, B. Reinforcement learning based recommender systems: A survey. *arXiv* **2021**, arXiv:2101.06286v1. [CrossRef]
29. Thorburn, L.; Stray, J.; Bengani, P. Is Optimizing for Engagement Changing Us? Understanding Recommenders, 23 November 2022. Available online: https://medium.com/understanding-recommenders/is-optimizing-for-engagement-changing-us-9d0ddfb0c65e (accessed on 16 March 2023).
30. Thorburn, L.; Stray, J.; Bengani, P. What Does It Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems, Understanding Recommenders, 15 March 2022. Available online: https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157 (accessed on 25 March 2022).
31. Bernheim, B.D. The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare. *Economics* **2016**, *7*, 12–68. [CrossRef]
32. Hadfield-Menell, D.; Hadfield, G.K. Incomplete contracting and AI alignment. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 417–422. [CrossRef]
33. Zhuang, S.; Hadfield-Menell, D. Consequences of Misaligned AI. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020.
34. Christian, B. *The Alignment Problem: Machine Learning and Human Values*; W. W. Norton & Company: New York, NY, USA, 2020.
35. Pillar, P.R. Intelligence, Policy, and the War in Iraq. *Foreign Aff.* **2006**, *85*, 15–27. [CrossRef]
36. Clapper, J. Intelligence Community Directive 203: Analytic Standards. 2015. Available online: https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf (accessed on 21 March 2023).
37. Zerilli, J.; Knott, A.; Maclaurin, J.; Gavaghan, C. Algorithmic Decision-Making and the Control Problem. *Minds Mach.* **2019**, *29*, 555–578. [CrossRef]
38. Heuer, R.J. *Psychology of Intelligence Analysis*; Center for the Study of Intelligence, Central Intelligence Agency: Washington, DC, USA, 1999.
39. Harambam, J.; Makhortykh, M.; Bountouridis, D.; Van Hoboken, J. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (News) recommender systems. In Proceedings of the 13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 16–20 September 2019; pp. 69–77. [CrossRef]
40. Jin, Y.; Cardoso, B.; Verbert, K. How Do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations? In Proceedings of the 11th ACM Conference on Recommender Systems, Como, Italy, 27–31 August 2017; p. 8.
41. He, C.; Parra, D.; Verbert, K. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Syst. Appl.* **2016**, *56*, 9–27. [CrossRef]
42. Tolcott, M.A.; Marvin, F.F.; Lehner, P.E. Expert decision-making in evolving situations. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 606–615. [CrossRef]
43. Lehner, P.E.; Adelman, L.; Cheikes, B.A.; Brown, M.J. Confirmation Bias in Complex Analyses. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2008**, *38*, 584–592. [CrossRef]
44. Glockner, M.; Hou, Y.; Gurevych, I. Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. *arXiv* **2022**, arXiv:2210.13865v1.
45. Stray, J.; Adler, S.; Vendrov, I.; Nixon, J.; Hadfield-Menell, D. What are you optimizing for? Aligning Recommender Systems with Human Values. *arXiv* **2020**, arXiv:2107.10939.
46. Stray, J. Aligning AI Optimization to Community Well-being. *Int. J. Community Well-Being* **2020**, *3*, 443–463. [CrossRef] [PubMed]
47. Zhao, Z.; Hong, L.; Wei, L.; Chen, J.; Nath, A.; Andrews, S.; Kumthekar, A.; Sathiamoorthy, M.; Yi, X.; Chi, E. Recommending What Video to Watch Next: A Multitask Ranking System. In Proceedings of the 13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 16–20 September 2019; pp. 43–51. [CrossRef]
48. Goodrow, C. On YouTube's Recommendation System, YouTube Blog. 2021. Available online: https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/ (accessed on 19 November 2021).
49. Lada, A.; Wang, M.; Yan, T. How Machine Learning Powers Facebook's News Feed Ranking Algorithm, Engineering at Meta, 26 January 2021. Available online: https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/ (accessed on 16 December 2021).
50. Validity of the IC Rating Scale as a Measure of Analytic Rigor, 2 December 2021. Available online: https://www.youtube.com/watch?v=8FZ9W1KRcZ4 (accessed on 9 September 2022).
51. Zelik, D.J.; Patterson, E.S.; Woods, D.D. Measuring Attributes of Rigor in Information Analysis. In *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams*; CRC Press: Boca Raton, FL, USA, 2010.
52. Barnett, A.; Primoratz, T.; de Rozario, R.; Saletta, M.; Thorburn, L.; van Gelder, T. Analytic Rigour in Intelligence, Hunt Lab for Intelligence Research, April 2021. Available online: https://cpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/8/401/files/2021/04/Analytic-Rigour-in-Intelligence-Approved-for-Public-Release.pdf (accessed on 1 July 2022).
53. Jacobs, A.Z.; Wallach, H. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, 3–10 March 2021; pp. 375–385. [CrossRef]
54. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P. Learning to summarize from human feedback. *arXiv* **2020**. [CrossRef]
55. Jannach, D.; Manzoor, A.; Cai, W.; Chen, L. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* **2021**, *54*, 105. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.