Editorial Values for News Recommenders: Translating Principles to Engineering

Jonathan Stray
UC Berkeley CHAI
jonathanstray@berkeley.edu

Abstract

This article proposes concrete methods for specifying the editorial behavior of news recommendation systems through collaboration between journalists and technologists to create metrics, data sets, feedback methods, and evaluation protocols. So far, the desired behavior of news recommenders has mostly been specified in terms of principles or guidelines. I argue that natural language specifications are inadequate because the translation to software must subsequently be undertaken by technical specialists, a process which requires consequential values-related decisions. Instead, I propose the specification of recommender editorial values through the collaborative creation of specific value-laden technical artifacts already used in contemporary engineering. These artifacts are much more precise than principles, yet do not require the technical understanding necessary to create novel algorithms.

Introduction

News recommendation systems are operated both by publishers and by platforms and have become one of the primary ways that journalism finds its way to audiences. Communication scholars have argued that these systems should embody a variety of values such as informedness, accuracy, comprehensiveness, autonomy, inclusiveness, participation, representation, diversity, deliberation, and tolerance (Helberger, 2019; Nechushtai & Lewis, 2019). While the recommendation technical community has developed a wide variety of values-driven metrics and algorithms (Celis et al., 2019; Kunaver & Požrl, 2017; Stray, 2020) these do not generally align with the conceptions of journalists and scholars. Building recommenders that enact editorial values is a deeply interdisciplinary pursuit, and few individuals (or organizations) have both a deep grasp of editorial values and the technical skill to design novel recommender systems.

In short, different communities are talking about the same problem in different language. On one side journalists, scholars, critics, and regulators have largely discussed these systems in terms of their normative concerns and societal outcomes. On the other side computer scientists, product managers, AI researchers and others have built ever more sophisticated news recommender systems. This is a caricature; in reality there are not two clear "sides" but a complex network of overlaps and interconnections between people and ideas. Yet this divide is immediately recognizable to workers in the field, and useful for framing the problem.

With this divide in mind, this article proposes narrowing the *journalist-technologist gap* through interdisciplinary collaboration to create four types of artifacts:

- Metrics. What should be measured, and what counts as an acceptable result?
- Data sets. These can be used to train algorithms, or to evaluate and compare their performance.
- Feedback methods. There are emerging methods to enable users and other stakeholders to provide algorithmically actionable feedback to a recommender system.
- Evaluation protocols. Understanding the consequences of news recommendation algorithms is essentially social science research, which would benefit from repeatable methods.

While a great deal of scholarship has focused on principles, both for news recommenders (Helberger, 2019; Vrijenhoek et al., 2020) and responsible AI in general (Fjeld et al., 2020), there has been less attention to metrics, data sets, feedback, and evaluation specifically. Collaboratively produced artifacts could be a key way for *journalists* and *technologists* to work toward the shared goals of embedding editorial values into news recommender systems.

Related Work

Editorial values are significant for any recommender system that handles news content, with potentially profound effects on democracy (Fields et al., 2018; Helberger, 2019;

Vrijenhoek et al., 2020). This includes recommenders operated by a single news organization such as that which produces personalized suggestions in the New York Times app, news aggregators such as Google News, and social media recommenders which also handle news items such as the Facebook News Feed. For simplicity, I will refer to all of these types of systems to be "news recommenders." I use "news" fairly narrowly to mean the output of conventional news organizations, though there are analogous editorial concerns about any recommender which selects media items.

This article takes a design orientation towards journalistic algorithms, as articulated by Diakopoulos (2019). My foremost concern is the real-world deployment of news recommenders that embody important editorial values. Merely explicating these values is not enough, which is why "developing evaluation methods and metrics" (Diakopoulos, 2019, p. 4) is so central to advancing journalism automation.

Participatory design is an orientation and a set of practices that attempts to actively involve all stakeholders in a system design process (Simonsen & Robertson, 2012). The stakeholders in this case include journalists as content creators, technologists as system designers and operators, audience members, and perhaps society in general. The related field of *value sensitive design* is "a theoretically grounded approach to the design of technology that accounts for human values in a principled and systematic manner throughout the design process" (Friedman et al., 2017). Most specifically, *multistakeholder recommendation* studies the design and evaluation of recommender systems that must simultaneously serve the interests of multiple groups (Abdollahpouri et al., 2020).

Other work examines the values actually implemented in production recommenders. Nechushtai and Lewis (2019) undertake a crowdsourced audit of Google News while Bandy and Diakopoulos (2019) study Apple News, evaluating the output of these systems with respect to values such as diversity, local news content, etc. DeVito (2017) infers the values of the Facebook News Feed by reading public documents to determine which factors are used as inputs.

Yet none of this work specifies how values are to be translated into algorithms. Though methods like participatory design provide high-level descriptions of design processes, more specialized technical approaches are necessary to construct operational recommender systems. The state of the art of recommender values engineering is the creation of hand-crafted metrics or machine-learning classifiers to identify various wanted and unwanted aspects of content or recommendations, used in a three part process (Stray et al., 2020):

- Identification: system designers become aware of a negative outcome associated with the system and identify a concept associated with it. For example, discovering that users are getting drawn into low-quality content and developing a corresponding definition of "clickbait."

- Operationalization: A concrete procedure is developed to identify instances of the abstract concept in the recommender system. This may involve the development of hand-crafted metrics, but most systems rely on some form of machine learning, trained on human-labelled data.

- Adjustment: system designers modify the recommender system in order to increase or decrease the prevalence of the target concept. This could involve A/B testing with respect to an evaluation protocol, incorporating a metric into model training, or adding code that re-orders ("re-ranks") results prior to presenting them to the user.

The actual processes used to engineer values into production news recommender systems today involve various technical artifacts such as metrics, data sets and evaluation protocols. These have a deep role in defining the character and enacted values of such systems. Yet these artifacts are much more amenable to the involvement of non-experts than core recommendation algorithms *per se*.

Finally, it is important to conceive of this process as more than "embodying editorial values in technology." When journalists are asked to translate their practices into the definitions and data required for algorithmic implementation they often discover that naturalized concepts like "newsworthy" or "authoritative" are not as clear or uncontested as they thought (Stray, 2019). As the News Quality Initiative put it, "any confusion that existed among journalists regarding principles, standards, definitions, and ethics has only travelled downstream to platforms" (Vincent et al., 2020).

In many cases, there is no existing normative consensus on exactly which journalistic values a recommendation system should support, how these should trade off against each other, and how the results should be evaluated (Nechushtai & Lewis, 2019). Therefore, the synthesis of journalism and technology will force articulation and clarification of core editorial concerns, up to and including journalism's ultimate societal goals.

Principles do not Define Behavior

By "principles" I mean written descriptions of the values that technical systems should uphold. Such descriptions are, so far, the primary method by which the ethical behavior of technical systems have been specified by scholars and critics. Fjeld et. al. (2020) map the content of several dozen

AI principles documents, finding themes such as privacy, accountability, safety, transparency, fairness, human control, and responsibility. There are also ongoing standards efforts around the values which apply to particular technical domains, such as the IEEE Ethically Aligned Design series (Shahriari & Shahriari, 2017).

Likewise, *journalists* (and allied scholars and non-technical experts) have mostly attempted to specify the operation of news recommenders through written descriptions of the values that such systems should uphold. Existing critiques of AI ethics principles typically focus on the lack of incentives for implementation and the dangers of superficial "ethics washing," e.g. (Floridi, 2019) but here I critique principles from a different direction: they are typically not precise enough to define the behavior of a technical system.

As an example, consider how the value of "diversity" is implemented in recommender systems. Diversity is perhaps the most widely discussed value in the news recommender literature produced by *journalists.* But what is it?

Regulatory discussions typically consider *source diversity*, that is, they are concerned with the mix of news organizations in the content recommended to each person e.g. (Helberger et al., 2019). Several researchers who have audited news recommender systems are correspondingly concerned with source diversity (Bandy & Diakopoulos, 2019; Nechushtai & Lewis, 2019). But users may still encounter only a small number of topics or perspectives even though they consume diverse sources, so other authors have been concerned with *content diversity* (Möller et al., 2018). Helberger et al. (2018) suggests that we should think instead in terms of the goals of a news recommender, and outlines *liberal*, *deliberative* and *adversarial* notions of diversity. This is appealing, but these are concepts at an even higher level of abstraction.

Meanwhile, *diversity* has been extensively studied in the technical context of recommender design. A review by Kunaver and Požrl (2017) lists eight different diversification algorithms and nine different formulas to measure the diversity of a set of items, mostly based on evaluating a *similarity function.* This is an algorithm for computing the relative sameness of two different items, for example the classic *cosine similarity* method for comparing text documents based on word frequencies, which has been used in search engines since the 1970s (Manning et al., 2008). *Cosine similarity* is essentially a measure of topical similarity, but similarity metrics can be designed to capture many other axes of variation, e.g. source diversity or demographic diversity.

Recommender systems actually in use by news organizations employ further definitions of diversity. One large news organization has designed their recommender to consider diversity in terms of media format, so that the user sees a mix of articles, videos, podcasts, etc. Another uses the topical diversity algorithm of Ziegler et. al (2005) so as to prevent all of the top-ranked news stories from being about the same popular topic, e.g. President Trump. Further afield in music recommendation, Spotify uses a popularity-based diversity metric in an effort to give a fair level of exposure to all artists who provide content (Hansen et al., 2021; Mehrotra et al., 2018).

As these examples show, there is no end to the normative and technical definitions of diversity that might be employed. A review of the concept of diversity across communications scholarship, social science, and computer science concludes that "research on this topic has been held back by the lack of conceptual clarity about media diversity and by a slow adoption of methods to measure and analyze it" (Loecherbach et al., 2020, p. 606). Hence, merely stating that a news recommender system should be "diverse" is not enough. There is still considerable work to be done in a) choosing a conception of diversity and then b) choosing a specific algorithmic operationalization of that concept.

In other words, there is a very large gap between a principle such as "represent diverse viewpoints" and a specification such as "ensure that no one news source accounts for more than 20% of the recommended items." The latter is concrete enough to be implemented. It also necessarily captures only a small portion of the rich concept of "diverse viewpoints." While it is true that a narrow mechanical process can never account for all the contextual richness of human experience, all principles must claim *some* power of generalization if they are to serve as a guide to future behavior. It is necessary to commit to specific definitions, phrased in algorithmic terms, in order to build real recommenders.

This is why "principles" are not a satisfactory method for specifying the desired behavior of a news recommender. When inspected closely, most "principles" for recommender design admit many possible algorithmic translations. If the work of making the abstract concrete is not done by *journalists*, it must be done by *technologists*, which is closer to delegation than collaboration.

Metrics

Metrics are a key tool for translating principles to practice because they span the divide between the conceptual and the empirical. In the context of AI systems, they are used both at the level of management (e.g. "key performance indicators") and encoded algorithmically (e.g. "objective functions.") Well-being metrics are already being used by large platforms in this dual role (Stray, 2020). Metrics can also contribute to transparency and provide regulatory affordances, as they define a common language for comparing recommender performance.

Aside from the well-known issues with using metrics in a management context generally (Jackson, 2005) metrics pose a problem for AI systems in particular because most AI systems are built around strongly optimizing for a narrow objective (Thomas & Uminsky, 2020). Poor use of metrics can result in a damaging emphasis on short term outcomes, manipulation and gaming, and unwanted side effects. Even a successful metric cannot remain static, as the structure of the world will change over time; many machine learning models broke when the onset of the COVID-19 pandemic caused mass changes in behavior (Heaven, 2020). Yet metrics are an essential component of modern recommender systems, and offer rich possibilities for *journalist-technologist* collaboration.

There are two major questions that must be answered in the design of a metric: what is important, and how to measure it. This starts with the question of which values a particular system should enact, which must then be "operationalized" into practical metrics (Jacobs & Wallach, 2019). *Technologists* will need to be included in discussions of which values matter because it is necessary to consider both the constraints of technical possibility and the empirical behavior of the audience-platform system. *Journalists* will need to be involved in the operationalization and validation of metrics because these decisions ultimately define what, exactly, is measured. The issue of who actually carries out these measurements is also significant – it need not be the platform itself (Wu & Taneja, 2020).

Returning to the example of diversity, Helberger et al. (2018) propose several metrics that might help us evaluate different axes of diversity in recommender systems:

> it is conceivable to design metrics that would focus, for example, on user engagement with opposing political views, cross-ideological references in public debates or social media connections between people who represent different ideological positions. (Helberger et al., 2018, p. 195)

All of these suggestions are technically realizable. It is possible to infer an individual's ideological position from their posts, social network structure, and/or news consumption data (Bodrunova et al., 2019; Garcia et al., 2014; Garimella & Weber, 2017) and this could be used to define "cross-ideological" engagement. Such metrics can drive news recommender design at the managerial level when given as targets to a product team, who may further choose to translate these metrics technically by incorporating them into the objective functions of their algorithms (Stray, 2020).

It remains to decide what value of such a metric counts as a "good" outcome; the numerical result of measuring something doesn't mean much if we cannot say what an acceptable number is. Nechushtai and Lewis (2019) grapple with this problem in their study of source diversity in Google News:

> If every news story recommended by a search engine were false, or if a search engine referred readers to one news organization alone, it would be clear that the algorithm is falling short as a news provider acting in the public interest. But, in most cases, public-facing algorithms of this sort do not function catastrophically or perfectly, but somewhere in between. Precisely where they fall on such a spectrum remains open for debate. What standards should be used to assess their performance? (Nechushtai & Lewis, 2019)

Translating news values into metrics and numerical thresholds is likely to be an uncomfortable process for *journalists*, because it requires examination of naturalized values in extremely explicit terms. Consider the concept of "newsworthiness," which might be the fundamental value underlying content ranking. Even for stories which are already data-driven, involving crime statistics, earthquakes, or corporate earnings, it is difficult to say exactly what number counts as "news." Previous efforts have tried to infer a numeric threshold which matches what journalists already do, set the threshold so as not to overwhelm editors or audiences with too many stories, or simply picked a "reasonable" threshold arbitrarily (Stray, 2019). None of these options is completely satisfactory.

Metrics also have an important role to play in regulation. A metric can be considered a "regulatory affordance," a common language for the regulator, the regulated, and the public. For example, Facebook has proposed regulation of unwanted content based on a "prevalence" metric (Bickert, 2020) that would define some acceptably small percentage of views of prohibited content such as hate speech. Without both a metric and an acceptable limit, it is difficult to answer the question of whether there is "too much" unacceptable content. Quantification is especially important when judging tradeoffs between different values; given the limited accuracy of automated classifiers, removing more hate speech will necessarily impinge on freedom of expression as false positives also increase (Duarte et al., 2017). Metrics are thus a key component of evaluation protocols, below.

Metrics work best for concepts that have straightforward observable counterparts, like source diversity. It is difficult to capture more complex notions in a simple metric. A number of metrics have been used over the years to detect "clickbait," including dwell time (Yi et al., 2014) and click-to-share ratio (El-Arini & Tang, 2014). But neither of these measures is a very good operationalization, as neither really captures what is meant by "clickbait." Modern systems instead use machine learning classifiers trained on custom data sets.

Data Sets

Data sets can embody values in a variety of ways. While this embeddedness is usually studied in the context of fairness (e.g. Barocas et al., 2018), custom data sets are also used to define and promote positive outcomes. Data sets at the level of sources, items, or sets of items could provide valuable normative direction for news recommendations.

Recommender systems that select user posts must remove clickbait and spam. Because there is no one metric that can reliably detect this unwanted content, in practice platforms solve this problem through machine learning (ML) models trained on data sets containing examples of both clickbait and non-clickbait content (Cora, 2017; Peysakhovich & Hendrix, 2016). Hate speech is similarly hard to measure using hand-crafted metrics, so ML classifiers are used instead (Fortuna & Nunes, 2018). In short, data sets can encode more subtle and complex values than hand-crafted metrics. A classifier trained on such data outputs a numerical score for each item, which can be considered a type of algorithmically constructed metric.

One of the simplest kinds of news-relevant data sets is a list of organizations which meet certain standards for process and quality. A news aggregator must first solve the problem of which sources produce "news," and even general social media platforms must be able to identify when users post news if they wish to treat it differently – for example, if they wish to subject it to evaluation, ranking, and labeling according to quality standards. For this reason, there are commercial organizations like NewsGuard that maintain a list of news organizations rated for credibility. Facebook has assembled source-level credibility data through crowdsourced surveys, a technique which other researchers have independently validated (Mosseri, 2018; Pennycook & Rand, 2018; Zhang et al., 2020).

Collections of labels at the level of individual news items (articles, videos, etc.) could in principle be used to embody various dimensions of item quality and credibility. While many "fake news classifiers" have been built from article-level data sets, in practice credibility classifiers built this way typically pick up on source or topic and do not generalize well (Bozarth & Budak, 2020). A classifier can only detect a concept if the necessary data is actually available to the system; it is not possible to identify "misinformation" with a classifier alone because determining whether something is true or not can require open-ended human research (Silverman, 2020). Still, it is useful to combine various content and contextual cues to evaluate credibility, and credibility rating data sets are used by platforms as one signal to rank content lower, rather than remove it.

Annotated *sets* of articles may also be an important data source. The News Quality Initiative has scraped the daily top stories from several news aggregators and asked journalists to "re-rank" the stories according to their editorial judgement and record their reasoning (Sehat, 2020). This data set, though small and partially qualitative, suggests that ranked sets of articles could be used to evaluate or perhaps even train news recommenders to match professional editorial values. The Reuters Tracer system relies on a similar approach to flag tweets with potential breaking news value for human review. It employs a "newsworthiness" classifier trained on the stories that reporters actually chose to write (Liu et al., 2016).

Given the lack of a precise theory of the aims of journalism, much less a theory of the aims of news recommenders which produce personalized results, whether or not news recommenders *should* attempt to match traditional editorial judgements is an open question. There may be greater agreement on which sorts of content should *not* be selected by recommenders, which speaks to the relationship between ranking (deciding what should be shown) and moderation (deciding what should not be shown).

Feedback Methods

Interactive feedback is an emerging method to control the output of AI systems in general, and recommenders in particular. In this approach, user or experts are asked to provide feedback on the actual output of a running system. This has the advantages of adaptability and ecological validity as compared to static data sets. Like metrics and data sets, this method of algorithmic tuning is also suitable for non-expert collaboration. The humble "like" button is a type of interactive feedback, but far more is possible.

Structured feedback has proven useful for training information filtering systems. For a document summarization task, OpenAI has demonstrated that guiding a reinforcement learning algorithm by repeatedly asking humans which of two summaries is better dramatically improves the quality of the results. Notably, including pairwise feedback produces much better summaries than training only on human-written reference summaries. (Stiennon et al., 2020).

Pairwise feedback has also been used to design a multi-stakeholder ranking system. Lee et al. (2019) demonstrate the use of a similar pairwise-comparison protocol for participatory design of a ranking algorithm. The goal was to design system for a non-profit which collects donated food and delivers it, via volunteer drivers, to local food charities. Representatives from different stakeholder groups were repeatedly shown a pair of donor-driver-recipient matches and asked to choose which they preferred. This feedback was used to construct a quantitative model of the preferences of each participant, with these models aggregated at run-time to produce the final ranking. The resulting algorithm

improved both efficiency and distributional fairness, as judged by stakeholders.

For news recommendation in particular, semi-structured feedback has enabled successful collaborations between technologists and journalists. During the iterative development of their in-house recommender systems, the BBC uses "a custom-designed qualitative scale and free text" to collect feedback from editors on each recommended item (Boididou et al., 2021).

This type of semi-structured feedback is not necessarily machine readable, meaning that only a small number of people (in this case editors) can be served in this way, but a number of researchers are investigating *conversational* methods which interact using natural language (Radlinski & Craswell, 2017; Sun & Zhang, 2018). Such a recommender might ask a user whether a particular article was a good choice for them and pose follow-up questions to try to learn why.

The design of interactive feedback methods is an opportunity for *journalist-technologist* collaboration. There are a variety of feedback paradigms that might yield useful information on how the system is enacting journalistic values. Journalists or users might be asked to do pairwise comparisons, rate individual articles, or provide natural language feedback. They could rate hypothetical recommendations, or recommendations actually provided. Retrospective, deliberative judgement on the items previously presented might be an especially powerful technique as it could align short-term and long-term incentives (Stray et al., 2020). If natural language feedback is possible, what sorts of questions should the algorithm ask?

Evaluation Protocols

News recommenders operate in a dynamic environment, interacting with large numbers of people. Offline evaluations of recommender performance, i.e. testing against prepared data sets, often do not meaningfully predict online performance (Jeunen, 2019). Thus, evaluation within a real-world setting is essential.

The most straightforward way to evaluate a news recommender is to measure its output against some metric. Such evaluation is a normal part of the process of recommender construction and operation, and engineers already repeatedly evaluate the output of recommender systems according to standard metrics like "accuracy," the ability to predict which items the user will actually engage with or otherwise rate as valuable. There is no reason why e.g. a suitable "diversity" metric could not be used similarly.

Standardized metrics would also allow external stakeholders to evaluate news recommenders, including users, researchers, and regulators. An evaluation protocol would specify which metric to use and how to collect the data to be evaluated. Data collection is a complex issue for recommender systems because most of them provide personalized results. Existing approaches include crowdsourced data collection from a demographically balanced sample of users (Nechushtai & Lewis, 2019) and using multiple new, non-personalized user accounts (Ledwich & Zaitsev, 2020). A standardized protocol is beneficial because it would allow longitudinal analysis, cross-recommender comparisons, and regulatory consistency.

Evaluation protocols can also go beyond assessing the output of a recommender system to consider the effect on users. Understanding the broader human consequences of these systems is essentially social science research which would benefit from repeatable methods. For example, the IEEE 7010 standard for well-being assessment (Schiff et al., 2020) proposes gathering baseline well-being metrics data for both expected users and non-users for a particular product, then collecting the same metrics over time for both groups.

As an example, we might want to test whether a particular news recommender increases or decreases political polarization. Affective polarization, a dislike and distrust of the outgroup, is a central part of the identity-based polarization that is occurring in modern democracies (Iyengar et al., 2018) and can be measured by simple survey instruments (Iyengar & Westwood, 2015). It might be possible to measure affective polarization of users and non-users before making some algorithmic change intended to reduce polarization, then re-measure for these two groups several months after deployment. The difference in the change in affective polarization between the two groups can be attributed to the recommender change, under certain assumptions. This is essentially a difference-in-differences research design (Angrist & Pischke, 2009, p. 227).

Evaluation protocols are the most general type of collaborative artifact studied here, and may specify the use of particular principles, metrics, data sets, and feedback methods. As above, different stakeholders may use such protocols in different ways. Recommender designers can use them to create and monitor their systems, while external stakeholders can use them to audit and compare different products.

Conclusion

While a number of authors have envisioned algorithms that embed values, much less has been said about how this is to be accomplished in practice. If *journalists* and *technologists* want to collaborate to produce better news recommenders, they will need to co-produce more than principles or guidelines. It is not reasonable to expect *journalists* to

become algorithmic experts and participate directly in technical design processes. Rather, these two groups could collaborate in the production of specific technical artifacts that are already used in contemporary recommender values engineering: metrics, data sets, feedback methods, and evaluation protocols.

References

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., & Pizzato, L. (2020). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, *30*(1), 127–158. https://doi.org/10.1007/s11257-019-09256-1

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bandy, J., & Diakopoulos, N. (2019). *Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News*. *Icwsm*. http://arxiv.org/abs/1908.00456

Barocas, S., Hardt, M., & Narayanan, A. (2018). *Fairness and Machine Learning*. http://fairmlbook.org

Bickert, M. (2020). *Online content regulation policy*. https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf

Bodrunova, S. S., Blekanov, I., Smoliarova, A., & Litvinenko, A. (2019). Beyond left and right: Real-world political polarization in twitter discussions on inter-ethnic conflicts. *Media and Communication*, *7*(3 Public Discussion in Russian Social Media), 119–132. https://doi.org/10.17645/mac.v7i3.1934

Boididou, C., Sheng, D., Mercer Moss, F. J., & Piscopo, A. (2021). Building public service recommenders: Logbook of a journey. *RecSys 2021 - 15th ACM Conference on Recommender Systems*, 538–540. https://doi.org/10.1145/3460231.3474614

Bozarth, L., & Budak, C. (2020). Toward a Better Performance Evaluation Framework for Fake News Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(Icwsm), 60–71.

Celis, L. E., Kapoor, S., Salehi, F., & Vishnoi, N. (2019). Controlling Polarization in Personalization. *FAT* '19: Conference on Fairness, Accountability, and Transparency*, 160–169. https://doi.org/10.1145/3287560.3287601

Cora, M. V. (2017). *Detecting Trustworthy Domains — Flipboard Engineering*. Flipboard. https://engineering.flipboard.com/2017/04/domainranking

DeVito, M. A. (2017). From Editors to Algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism*, *5*(6), 753–773. https://doi.org/10.1080/21670811.2016.1178592

Diakopoulos, N. (2019). Towards a Design Orientation on Algorithms and Automation in News Production. *Digital Journalism*, *7*(8), 1180–1184. https://doi.org/10.1080/21670811.2019.1682938

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy and Technology. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/

El-Arini, K., & Tang, J. (2014). *Click-baiting*. Facebook. https://about.fb.com/news/2014/08/news-feed-fyi-click-baiting/

Fields, B., Jones, R., & Cowlishaw, T. (2018). The Case for Public Service Recommender Algorithms. *BBC London*, 22–24. https://doi.org/10.1145/1835449.1835486

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676

Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, *11*(23), 63–125. https://doi.org/10.1561/1100000015

Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., & Schweitzer, F. (2014). *Network polarization in online politics participatory media*.

Garimella, V. R. K., & Weber, I. (2017). A long-term analysis of polarization on Twitter. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 528–531. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15592

Hansen, C., Mehrotra, R., Hansen, C., Brost, B., Maystre, L., & Lalmas, M. (2021). Shifting Consumption towards Diverse Content on Music Streaming Platforms. *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 238–246.

Heaven, W. D. (2020). *Our weird behavior during the pandemic is messing with AI models*. MIT Technology Review.

https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/

Helberger, N. (2019). On the Democratic Role of News Recommenders. *Digital Journalism*, *7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information Communication and Society*, *21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Helberger, N., Leerssen, P., & Van Drunen, M. (2019). *Germany proposes Europe's first diversity rules for social media platforms*. Media@LSE Blog. https://blogs.lse.ac.uk/medialse/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2018). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 1–35. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Jackson, A. (2005). Falling from a great height: Principles of good practice in performance measurement and the perils of top down determination of performance indicators. *Local Government Studies*, *31*(1), 21–38. https://doi.org/10.1080/0300393042000332837

Jacobs, A. Z., & Wallach, H. (2019). *Measurement and Fairness*. http://arxiv.org/abs/1912.05511

Jeunen, O. (2019). Revisiting offline evaluation for implicit-feedback recommender systems. *RecSys 2019 - 13th ACM Conference on Recommender Systems*, *3*, 596–600. https://doi.org/10.1145/3298689.3347069

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems – A survey. *Knowledge-Based Systems*, *123*, 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

Ledwich, M., & Zaitsev, A. (2020). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *First Monday*, *25*(3). https://doi.org/10.5210/fm.v25i3.10419

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., & Procaccia, A. D. (2019). Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*. https://doi.org/10.1145/3359283

Liu, X., Wudali, R., Martin, R., Duprey, J., Vachher, A., Keenan, W., Shah, S., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., & Pomerville, S. (2016). Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, 207–216. https://doi.org/10.1145/2983323.2983363

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2020). The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism*, *8*(5), 605–642. https://doi.org/10.1080/21670811.2020.1764374

Manning, C. D., Raghavan, P., & Schütze, H. (2008). The vector space model for scoring. In *Introduction to information retrieval* (pp. 109–126). Cambridge University Press.

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). Towards a Fair Marketplace. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2243–2251. https://doi.org/10.1145/3269206.3272027

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

Mosseri, A. (2018). *Helping Ensure News on Facebook Is From Trusted Sources*. Facebook. https://about.fb.com/news/2018/01/trusted-sources/

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298–307. https://doi.org/10.1016/j.chb.2018.07.043

Pennycook, G., & Rand, D. G. (2018). Crowdsourcing Judgments of News Source Quality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3118471

Peysakhovich, A., & Hendrix, K. (2016). *Further Reducing Clickbait in Feed*. Facebook. https://about.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/

Radlinski, F., & Craswell, N. (2017). A Theoretical Framework for Conversational Search. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 117–126. https://doi.org/10.1145/3020165.3020183

Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). *IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence*. http://arxiv.org/abs/2005.06620

Sehat, C. M. (2020). *NewsQ Review Panel Reports 2020*. News Quality Initiative. https://newsq.net/newsq-review-panel-reports-2020/

Shahriari, K., & Shahriari, M. (2017). IEEE standard review - Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. *IHTC 2017 - IEEE Canada International Humanitarian Technology Conference 2017*, 197–201. https://doi.org/10.1109/IHTC.2017.8058187

Silverman, C. (Ed.). (2020). *The Verification Handbook for Disinformation And Media Manipulation*. European Journalism Center. https://datajournalism.com/read/handbook/verification-3

Simonsen, J., & Robertson, T. (2012). Routledge international handbook of participatory design. In *Routledge International Handbook of Participatory Design* (1st Editio). Routledge. https://doi.org/10.4324/9780203108543

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). *Learning to summarize from human feedback*. 1–44. http://arxiv.org/abs/2009.01325

Stray, J. (2019). Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism*, *7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289

Stray, J. (2020). Aligning AI Optimization to Community Well-being. *International Journal of Community Well-Being*. https://doi.org/10.1007/s42413-020-00086-3

Stray, J., Adler, S., & Hadfield-Menell, D. (2020). What are you optimizing for ? Aligning Recommender Systems with Human Values. *Participatory Approaches to Machine Learning Workshop, ICML 2020*. https://participatoryml.github.io/papers/2020/42.pdf

Sun, Y., & Zhang, Y. (2018). Conversational recommender system. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 235–244. https://doi.org/10.1145/3209978.3210002

Thomas, R. L., & Uminsky, D. (2020). Reliance on Metrics is a Fundamental Challenge for AI. *Ethics of Data Science Conference*. https://arxiv.org/abs/2002.08512

Vincent, S., Lopez, P., Allen, L., Allsop, J., Riley, R., & Traister, R. (2020). *Our Opinion: Recommendations for Publishing Opinion Journalism on Digital Platforms*. News Quality Initiative. https://newsq.net/wp-content/uploads/2020/11/NewsQ-Opinion-Panel-2020-nov30-FINAL.pdf

Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., & Helberger, N. (2020). Recommenders with a mission: assessing diversity in newsrecommendations. *Proceedings of ACM Conference (Conference'17)*, *1*(1), 554–561. https://doi.org/10.1007/978-3-030-65965-3_38

Wu, A. X., & Taneja, H. (2020). Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme. *New Media and Society*. https://doi.org/10.1177/1461444820933547

Yi, X., Hong, L., Zhong, E., Liu, N. N., & Rajan, S. (2014). Beyond clicks: Dwell time for personalization. *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, 113–120. https://doi.org/10.1145/2645710.2645724

Zhang, A. M. Y. X., Sehat, C. M., & Mitra, T. (2020). *Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria*. *37*(4). https://arxiv.org/abs/2008.09533

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. *Proceedings of the 14th International Conference on World Wide Web - WWW '05*, *January 2005*, 22. https://doi.org/10.1145/1060745.1060754