

---

# Building Human Values into Recommender Systems: An Interdisciplinary Synthesis

---

**Jonathan Stray<sup>1</sup>, Alon Halevy<sup>2</sup>, Parisa Assar<sup>2</sup>, Dylan Hadfield-Menell<sup>3</sup>, Craig Boutilier<sup>4</sup>, Amar Ashar<sup>5</sup>, Lex Beattie<sup>5</sup>, Michael Ekstrand<sup>6</sup>, Claire Leibowicz<sup>7</sup>, Connie Moon Sehat<sup>8</sup>, Sara Johansen<sup>9</sup>, Lianne Kerlin<sup>10</sup>, David Vickrey<sup>11</sup>, Spandana Singh<sup>12</sup>, Sanne Vrijenhoek<sup>13</sup>, Amy Zhang<sup>14</sup>, McKane Andrus<sup>7</sup>, Natali Helberger<sup>15</sup>, Polina Proutskova<sup>10</sup>, Tanushree Mitra<sup>16</sup>, Nina Vasan<sup>9</sup>**

<sup>1</sup> Center for Human-Compatible Artificial Intelligence, UC Berkeley, <sup>2</sup> Meta AI, <sup>3</sup> Department of Electrical Engineering and Computer Science, MIT, <sup>4</sup> Google Research, <sup>5</sup> Spotify Inc., <sup>6</sup> Computer Science Department, Boise State University, <sup>7</sup> Partnership on AI, <sup>8</sup> Hacks/Hackers, <sup>9</sup> Brainstorm: The Stanford Lab for Mental Health Innovation, Department of Psychiatry and Behavioral Sciences, Stanford University, <sup>10</sup> BBC, <sup>11</sup> Meta, <sup>12</sup> New America's Open Technology Institute, <sup>13</sup> Faculty of Law, Information Law, University of Amsterdam, <sup>14</sup> Allen School of Computer Science & Engineering, University of Washington, <sup>15</sup> Institute for Information Law, University of Amsterdam, <sup>16</sup> Information and Computer Science, University of Washington

## Abstract

Recommender systems are the algorithms which select, filter, and personalize content across many of the world's largest platforms and apps. As such, their positive and negative effects on individuals and on societies have been extensively theorized and studied. Our overarching question is how to ensure that recommender systems enact the values of the individuals and societies that they serve. Addressing this question in a principled fashion requires technical knowledge of recommender design and operation, and also critically depends on insights from diverse fields including social science, ethics, economics, psychology, policy and law. This paper is a multidisciplinary effort to synthesize theory and practice from different perspectives, with the goal of providing a shared language, articulating current design approaches, and identifying open problems. It is not a comprehensive survey of this large space, but a set of highlights identified by our diverse author cohort. We collect a set of values that seem most relevant to recommender systems operating across different domains, then examine them from the perspectives of current industry practice, measurement, product design, and policy approaches. Important open problems include multi-stakeholder processes for defining values and resolving trade-offs, better values-driven measurements, recommender controls that people use, non-behavioral algorithmic feedback, optimization for long-term outcomes, causal inference of recommender effects, academic-industry research collaborations, and interdisciplinary policy-making.

## 1. Introduction

Recommender systems are the algorithms which select, filter, and personalize content across social media (Lada et al., 2021), news aggregators (Das et al., 2007), music and video streaming services (Holtz et al., 2020; Zhao et al., 2019), online shopping (Louca et al., 2019), online ad targeting (Zhao et al., 2021), and other systems. As such, their positive and negative effects on individuals and societies have been extensively theorized and studied. In the context of social media recommendation there has been mixed evidence regarding both positive and negative effects on adolescent well-being (Przybylski and Weinstein, 2017), polarization (Allcott et al., 2020; Asimovic et al., 2021) and news consumption (Fletcher and Nielsen, 2018). In news recommender systems, diversity of opinion and content sourcing is a major concern (Bernstein et al., 2020; Helberger et al., 2018). Recommender systems used to promote job openings may be discriminatory if they do not consider the balance of distribution across legally protected user attributes (Lambrecht and Tucker, 2019). Product recommendations used for online shopping could shift large-scale behavioral patterns with significant economic, environmental, or social effects (Gretzel and Fesenmaier, 2006). Even systems designed purely for entertainment, such as film and music streaming services, must consider the fair allocation of attention to the artists who create content (Mehrotra et al., 2018). One overarching question across all of these contexts is, *how can we make recommender systems enact the values of the individuals and societies that they serve?*

Addressing this question in a principled fashion requires technical knowledge of recommender design and operation, and also critically depends on insights from diverse fields including social science, ethics, economics, psychology, and law. This paper is a multidisciplinary effort to define a common language and review the current state of the art for addressing problems related to designing and operating recommender systems in support of a wide array of human values. This is a wide-ranging exercise in value-sensitive design (Friedman et al., 2002), which we undertake by synthesizing knowledge from a variety of perspectives.

There are several widely used frames for discussing the normative implications of AI systems, most of which apply to our narrower context of recommender systems. “Alignment” is concerned with ensuring that AI systems enact the intentions of its designers and of its users despite the impossibility of specifying the correct action in all possible cases (Hadfield-Menell and Hadfield, 2019) and provides us with the language of “value alignment.” “Fairness” or “bias” is primarily concerned with the distribution of benefits and harms between people or groups (Chen et al., 2020; M. D. Ekstrand et al., 2021). “Integrity” refers to identifying and moderating content that violates platform policies for a variety of reasons, including obscenity, copyright, criminal activity, and misinformation (Halevy et al., 2020). “Well-being” is an umbrella term for a wide variety of sociological measures used across sectors including government, public health, and research, which are starting to be applied to AI systems (Schiff et al., 2020). There is a body of work on the “ethics” of recommender systems (Milano et al., 2020). The potential effects of AI systems have also been analyzed within a “human rights” framework (Donahoe and Metzger, 2019). There is substantial overlap between these categories, each of which encompasses a range of more specific concerns such as misinformation (Fernandez and Bellogín, 2020), polarization (Stray, 2021a) or addiction (Allcott et al., 2021).

Rather than trying to reconcile these diverse frameworks, we take all of these to be concerned with *values*. We draw on the field of value-sensitive design (Friedman et al., 2002) to define values as “what a person or group of people consider important in life” (Borning and Muller, 2012, p. 1). Building real systems requires both deep technical practice and grounding in the realities of diverse human lives, including an understanding of the actual effects of deployed recommenders. Further, values and our understanding of them are constantly evolving, so the whole exercise depends on moral and philosophical reflection. Therefore, building human values into recommender systems requires a mix of conceptual, empirical, and technical work (Friedman et al., 2017).

Recommender systems are fundamentally social and collaborative both because many of their applications are designed to facilitate user-to-user interactions (e.g. social media) and because most recommender algorithms work by identifying patterns across users (e.g. “people who liked this also liked that”). Value-sensitive design uses the language of *stakeholders*, defined as those who are affected by design choices. It is their values which should matter in the design process, and recommenders are fundamentally multi-stakeholder systems which must simultaneously serve several types of stakeholders including users, content producers, platforms, and non-users (Abdollahpouri et al., 2020). Their opinions must be considered or solicited in some way, and there are a wide variety of methods for doing so in various contexts including participatory design, user surveys, semi-structured interviews, and deliberative elicitation (Lee et al., 2019; Ovadya, 2021; Simonsen and Robertson, 2012; Stray, 2020). To the extent that this feedback influences recommender algorithms, this is also an example of value-sensitive algorithm design (Zhu et al., 2018)

This paper is not a comprehensive review. We do not think such a review would be useful at this time, both because it would be massive and because the field is evolving rapidly. Rather, the goal of this paper is to help readers get their bearings in this large and complex space. While not exhaustive, our approach is interdisciplinary and cross-sector. Our authors include experts in recommender systems, journalism, political science, law, psychology, AI alignment, technology policy, human computer-interaction and other fields. We also have substantial representation across academic, industry, and civil society sectors. Even this broad author cohort cannot reflect the entire range of this space, so we focus primarily on recommenders for social media, news content, and entertainment streaming services. However, many of the issues and approaches we highlight are broadly applicable to other important categories of recommender systems, including online shopping, targeted advertising, recruitment, healthcare, and education.

We use the term “recommender systems” to focus on the core problem of personalized content selection across many domains. Recommender systems often operate without an explicit user query, though the user may also ask for more tailored recommendations (e.g., “politics podcast”). This contrasts with search functionality which requires an explicit query and where results tend to be much less personalized (Courtois et al., 2018; Krafft et al., 2019; Le et al., 2019). Social media is a major application of recommender systems, but we note that the two are not synonymous. While most social media platforms do employ recommenders for content selection, the effects of social media depend on many design decisions beyond algorithmic content selection including the way the content is displayed, the various possible actions or controls made available to users, and the use of other navigation methods such as search or human-curated lists. Further, it is difficult to disentangle recommender effects from user

creation, consumption, and sharing behavior across multiple platforms (Bartal et al., 2020; Hosseinmardi et al., 2020; Munger and Phillips, 2020). Nor do we take up content moderation at any length, even though content removal intersects strongly with values (Klonick, 2018; York and Zuckerman, 2019) and there is technical overlap with recommender algorithms. Instead we are concerned with what content is selected after items ineligible for recommendation have been removed.

We have been asked why this work is framed in terms of “values” and not “harms.” After all, it might seem more urgent to address a harm such as “contributes to depression” than to improve a benefit such as “keeps users happy.” In the sense that harms arise when recommenders do not operate in accordance with our values, the framing of values vs. harms can sometimes be merely a linguistic convention. Yet we believe that focusing solely on harms is overly narrow. One widely used set of core principles in applied ethics, deriving originally from consideration of biomedical research in the 1970s, is respect for autonomy, justice, and both avoiding harm and doing good (Beauchamp and Childress, 2019). Considering only harms and not the potential for doing good is limiting, while framing the effects of recommenders in terms of values opens up new avenues for thinking about the role and responsibilities of recommenders in society. Most of the values we consider can span the spectrum from mitigating harm to promoting good.

Further, it may not be possible to build a system that avoids “bad” without also defining “good,” because of the necessity of specifying objectives. A news recommender should not promote false information, but this does not provide any guidance on what sort of information it *should* select, nor how it should support the societal goals of journalism (Fields et al., 2018; Helberger, 2019). Similarly, if social media recommenders should not optimize for engagement, then what should they optimize for? This question is more than rhetorical, as most AI technologies are based on optimizing some set of metrics which define the objectives of the system, and the appropriate specification of objectives is a significant and well-known challenge (D’Amour et al., 2020; Gabriel and Ghazavi, 2021; Hadfield-Menell and Hadfield, 2019; Mullainathan and Obermeyer, 2021).

We do not attempt to offer a general theory of recommender value alignment. Instead, our approach is to synthesize current thinking, with an emphasis on industrial applicability. The contributions of this paper include:

- A set of human values that are important in the context of recommender systems, derived from previous partial compilations and feedback from experts in ethics, policy, and industry. We summarize the state of research and practice around each of these values, including the empirical effects of existing recommender systems (Section 2).
- We describe one possible process for operationalizing a given human value, based on current practice in industrial recommender systems, through the example of increasing the political diversity of content in a production news recommender (section 3).
- We discuss the techniques used to measure adherence and progress on human values. In particular, we show how measurement techniques of social science apply to the problem of evaluating values-based outcomes in recommender systems (section 4).

- We describe known design approaches to altering recommender systems to conform to specific values, which provides a menu of possibilities to recommender designers and opportunities for future work (section 5).
- We consider the different approaches to policy-making related to human values, and identify gaps where the policy and technical communities need to align (section 6).
- Finally, we present a set of important open problems across all the above areas (section 7).

While a work of this scope cannot possibly be comprehensive, this is a synthesis based on our collective understanding and experience, the result of an extensive interdisciplinary conversation. We offer a view of the current state of thinking and practice across a wide variety of issues and applications, and suggest the most promising directions for future work.

## 2. Values for Recommender Systems

In this section we consider a variety of values applicable to recommender systems. We do not aim to create a comprehensive list of all relevant values (which may not even be possible!). Instead, our goal is to ground the many discussions happening in industry, academia, and policy in a reasonably broad set of values so progress can be made. We used a multi-source methodology to come up with this list, incorporating previous reviews of values in technology, expert deliberation, and a small qualitative expert survey.

We began with four substantial compilations of values and risks relevant to either recommenders in particular or AI systems in general. These were a survey of AI ethics policy documents (Fjeld et al., 2020), a large user research project undertaken by the BBC (Kerlin, 2020), a compilation of AI-relevant well-being metrics (IEEE, 2020), and an IEEE process for ethical system design (IEEE, 2021). We extracted and combined these lists, then added a set of global “techno-moral virtues” (Vallor, 2016, pp. 15,120) as well as the feminist ethic of care (Skoe, 2014) and philosophical traditions in Africa that emphasize the relationships and bonds among people (Ndiweni and Sibanda, 2020). We added a few more items by hand, according to our judgment as practitioners and researchers, to ensure our list includes known “issues” in recommender operation (e.g. polarization).

In order to refine our list and better understand the trade-offs between different values, we gathered multi-stakeholder input through the Partnership on AI (PAI), a global non-profit partnership of roughly 100 academic, civil society, industry, and media organizations working towards positive outcomes from AI for people and society. PAI surveyed those of its members with an interest in recommender systems, approximately 40 people, with this initial list of values asking them first to choose a type of recommender system (e.g., social media, news aggregator) then to rate the importance of each value for that recommender system. At a subsequent workshop, these 40 participants discussed these values in small groups to further develop their conception of how values apply in recommender systems (Leibowicz et al., 2021). We then refined our list and definitions of values based on this feedback.

The goal of this exercise was not to arrive at a set of global, universally applicable values, but to list some of the key values that have been identified to be particularly relevant in various recommender contexts. We note that the main challenge in devising our list was coming up with a set of definitions at an appropriate level of granularity that together cover a wide set of overlapping and often vague concepts. We aimed for a level of abstraction between overly general statements (e.g., “do good”) and specific formulations (e.g., particular metrics). Our list is shown in Appendix A, along with example indicators or metrics for each value, and example design changes which could promote that value.

Values cannot be discussed in isolation. By nature, there are tensions between values, and those tensions lead to many of the difficulties in operationalizing them in recommender systems. For example, the value of *free expression* is in tension with the value of *safety* because if we allow users to say anything they want on social media, others may feel threatened. As another example, *privacy* can be in tension with *usefulness*. *Privacy* suggests that a platform should not try to infer whether a user might have a particular disease, even though early information and intervention might be helpful to them.

Some of these tensions are rooted in the tradeoff between individual values versus societal values. For example, if we exclusively focus on the *well-being* of individuals, society may suffer (Ostrom, 2000). Values can also have effects on varying time scales. Giving users entertaining content may satisfy their short term needs, but providing more informative content may have longer term benefits. Having informed users may also be a societal value, and individual values often operate on a shorter timescale than societal values.

Many of the values in our list are closer in nature to instrumental goals for an AI system, by which higher-order values are achieved. For example, the bioethics principles of respect for *autonomy*, *beneficence*, and *justice* are often considered primary, without resorting to other values for justification (Canca, 2020). By contrast, other values seem to derive much of their importance because they contribute to these principles, including *privacy*, *agency*, *control*, *transparency*, *accessibility/inclusiveness*, and *accountability*.

Rather than discussing each value individually, we approach them through a number of themes that characterize many discussions on these topics: usefulness, well-being, legal and human rights, public discourse and safety.

## Usefulness

Platforms use recommender systems because they believe them to be useful to users, content creators, and themselves. The most straightforward distinction between recommendation and search is that a recommender can suggest items without an explicit query, which is valuable in a variety of contexts. For example, news items cannot be selected through user queries alone, because the user is unaware of new events. While the value of presenting a previously unknown post, article, person, movie, song or ad varies, all of these can lead to positive and novel outcomes for people. Modern recommender systems have their roots in collaborative filtering systems in the 1990s, and the need for intelligent filtering has

only increased since then as the pool of available information has exploded. We call this value “*usefulness*” to distinguish it from “*utility*” which has a more technical meaning from economics (discussed section 5.3.1).

*Usefulness* is closely related to *control*, which we take to mean that users should be able to select which content they are seeing, the type and degree of personalization (if any), and understand the processes that determine what they see. In large part this requires that platforms provide features to support such choices (e.g., playlists on a music streaming service or topic selection on a news service) though there are also important questions about community governance (Lee et al., 2019; Zhang et al., 2020). On social networks, interfaces for describing which posts to see are particularly complex given the breadth of available content. Note that *feeling in control* and *actually having control* are different. Placebo controls may increase user satisfaction without offering any actual improvement (Vaccaro et al., 2018) while users may not perceive even large effects of functional controls (Loeberbach et al., 2021). *Agency* is a similar value, but we use it to refer to control over other elements of the users’ life, not just the recommender. For example, a recommender could assist a user with education (Dascalu et al., 2016) or direct them to job opportunities.

There are complex tensions between *agency* and *control* and other values, as users might make choices that harm themselves or others. This grounds out in concrete ethical questions such as: if it is possible to infer that someone has an eating disorder (Yan et al., 2019), and if there is research indicating that viewing dieting videos leads to bad outcomes for such a person, is it reasonable or even obligatory to thwart their expressed intention to see such material?

Recommender systems can also be useful to both content creators and platforms. While recommenders are a commercially important technology, algorithmic content optimization does not translate directly into revenue in many contexts (Jannach and Jugovac, 2019). For example, subscription services must maximize user retention, while current recommender designs struggle with long-term outcomes. Direct optimization for revenue, or more precisely profit, is best developed in the context of online shopping (Das et al., 2009; Delgado et al., 2019; Jannach and Adomavicius, 2017; Louca et al., 2019). Even for ad-driven services, content personalization and ad selection are often handled by different recommenders which optimize against different objectives (Thompson, 2018). Other organizations which operate recommenders may not intend to make money by doing so, such as news publishers. While not all recommender systems need to be profitable, they all must be useful to the operator by some measure.

## Well-being

*Well-being* is fundamental to human experience, and recommender systems have the potential to affect users’ *well-being* in many ways. While the phrase has specific implications around positive subjective experience, it is also widely used in the policy community as an umbrella framework which encompasses other values (Exton and Shinwell, 2018; Graham et al., 2018; O’Donnell et al., 2014) and has been explored as an important end goal for AI systems in general (IEEE, 2020). *Well-being* as a subjective experience is one of the values in our list, but many other values intersect with it strongly and are often included in well-being frameworks.

Well-being is a complex concept, and there is little consensus on how to define it (Dodge et al., 2012). Objective measures such as employment, lack of crime, and economic prosperity were historically used as proxies for *well-being*. More recently there has been increasing focus on a more holistic understanding of *well-being* based on both subjective and objective measures (Diener, 2000; Hicks et al., 2013; Krueger and Stone, 2014). These subjective *well-being* measures account for people's cognitive and affective evaluations of their lives by asking subjects to rate how much they agree with statements like, "The conditions of my life are excellent" (Diener et al., 1985).

The values of *connection, community and belonging, recognition and acknowledgement, self expression, care, compassion, and empathy* all relate to the concept of *well-being*. Many different types of content can contribute to increased *well-being*, such as through education, motivation, or personal relationships. *Entertainment* can also contribute to *well-being*, especially as a short-term emotional experience, and hedonism is often considered a basic drive (Schwartz, 2012).

The effect of recommender systems on *well-being* has been most widely studied in the context of social media use. There is conflicting evidence for both positive and negative effects, in part because *well-being* in this context is variously defined and often represented by other factors, from self-esteem and life satisfaction to *civic engagement*, social capital, and user satisfaction. Also, most studies have not been designed to separate the effects of algorithmic content selection from other aspects of social media such as user creation and sharing.

In one study of over 2,000 college students, social media use was associated with improvement in various facets of *psychological well-being* such as overall life satisfaction, *civic engagement*, and social trust (Valenzuela et al., 2009). Some investigators have suggested that the amount of time spent online is less important than the quality of that time; active use may promote well-being, whereas passive use and emotional connection to use may have a negative impact (Bekalu et al., 2019; Frison and Eggermont, 2020; Verduyn et al., 2017). The literature also points toward negative health effects related to social media use. One longitudinal study compared social media use with *mental health, physical health*, and found a decrease of 5%-8% of a standard deviation in self-reported *mental health* (Shakya and Christakis, 2017). In another study, deactivating social media accounts for four weeks resulted in increased time in offline interactions and improved subjective *well-being* (Allcott et al., 2020). Upward social comparison has been proposed as one potential link between social media use and *mental health* disorders such as depression and anxiety (Midgley et al., 2021; White et al., 2006).

Some recommender-based products may encourage addictive tendencies. Allcott et al. (2021) found that abstaining from social media use for a time or allowing people to set future screen-time limits produced a decrease in subsequent use, suggesting that social media use may result in habit formation and self-control issues. These effects may not be due to personalized recommendations *per se*, as broadcast television, a non-personalized medium, has also been found to be addictive in this sense (Frey et al., 2007).

There is classic work on the advantages of both strong and weak social ties (Aral, 2016). These benefits are less well studied in the context of recommender systems, but a few studies of social media are worth

noting. Social media use has been associated with increased *social connection* and social capital in online and offline social networks, with particular benefit for users experiencing low self-esteem and low life-satisfaction (Ellison et al., 2007; Steinfield et al., 2008; Subrahmanyam et al., 2008). Larger social networks may be associated with greater perceived social support, reduced stress, and improved *well-being* (Nabi et al., 2013). Social media use has been shown to increase intergroup contact and reduce prejudice when offline social-network diversity is low (Asimovic et al., 2021). Recommenders can also contribute to *civic engagement*. A large-scale experiment showed that messaging on social media can increase voter turnout on the order of 1% (Bond et al., 2012). While the possibility that recommendations connect people to violent or extremist groups has been widely discussed (and we review the evidence below) the converse of this concern is the possibility of connecting people to constructive communities or social movements.

We see long term *well-being* as a major open problem for recommender systems, with a number of challenging sub-problems: defining *well-being* in contextually appropriate ways, measuring user *well-being* in production, and algorithmically optimizing for relatively long-term outcomes (months to years).

## Legal and Human Rights

Some of the values in our list can be considered rights, in the sense of obligations to various parties. Previous work has explored the legal and human rights potentially impacted by recommender systems (Milano et al., 2020) and by AI in general (Donahoe and Metzger, 2019)

*Privacy* has at least three different interpretations in the recommender context. The first sense refers to information the system knows, infers or estimates about the user. The second refers to the possibility of direct revelation of user information to other users or third-parties. There is also the possibility of indirect (often noisy) revelation of user information by the actions of the recommender, e.g. a recommendation made to one user may allow them to infer something about the items viewed by another user. This form of information revelation is the subject of a large body of research in the area of differential privacy (Chien et al., 2021; Dwork and Roth, 2014; Jain et al., 2018). *Privacy* trades off against other values such as *usefulness*, *transparency*, and *fairness*. The question of demographic inference which might help the user is well discussed in the context of algorithmic fairness (Andrus et al., 2021) but analogous concerns arise with many types of user knowledge. Granting users *control* over whether and how they want platforms inferring and using their personal data to recommend useful content might be one way to reconcile *privacy* and ethical use of personal data.

Like *privacy*, *transparency* and *explainability* are broad concepts. *Explainability* might mean that a recommender system gives reasons why a certain piece of content is being shown to a user (Zhang and Chen, 2020). *Transparency* could require platform disclosure of various types of data including metrics characterizing different types of content or aggregated user activity. *Transparency* can be important for building users' trust (Cramer et al., 2008) and is often discussed as a policy tool to promote understanding and *accountability*. Explanations and disclosures can come in multiple forms depending on their intended audience, such as users, researchers, governance bodies or auditors, and meaningful transparency needs to

be informed by concrete individual and societal information needs (Singh, 2020a; van Drunen et al., 2019). From a technical perspective, explanations can be tricky to generate because many recommender techniques (including deep learning) do not lend themselves to easy explanations (Weld and Bansal, 2019).

Fairness, equity and equality are closely tied to human rights (Universal Declaration of Human Rights, UN). In everyday situations where recommender systems are employed, multiple parties have some interest in the outcome (Abdollahpouri and Burke, 2019). Fairness has a variety of meanings in recommender systems, including considering the disparate impacts of recommendations across user classes (Leonhardt et al., 2018). Content providers want to be fairly treated in terms of the exposure and benefit they receive from the system (Diaz et al., 2020; Jeunen and Goethals, 2021); users want to receive good quality of service, and may not want to be under-served relative to other users (Ekstrand et al., 2018); and other stakeholders, such as the system operator, content creators, and society broadly each have ideas of what it may mean for recommendation to be “fair” (Diaz et al., 2020; M. Ekstrand et al., 2021). This has deep ties to various conceptions of *equity*, in particular equity of attention (Biega et al., 2018, p. 201; Singh and Joachims, 2018). Recommenders may also create externalities that affect people who do not use the platform, such as by directing many new people to a formerly obscure place, or encouraging the consumption of products with environmental consequences (Rolnick et al., 2019). The field of multi-stakeholder recommendation has emerged to tackle these dynamics (Abdollahpouri et al., 2020).

Given all the parties involved through the platform or via externalities, *fairness* is a multifaceted problem where different stakeholders have different objectives and needs from the system. Often the desires of different stakeholders are in conflict; not everyone can have exactly what they want, though it might still be possible to give something worthwhile to everyone.

## Public Discourse

Some of the most intense recent discussion around recommender systems has centered around how they affect public discourse. *Accuracy* of information and *diversity* of content are two prime examples.

Access to accurate and factual information supports human decision making and understanding across myriad domains, ranging from healthcare to politics to economics (Hochschild and Einstein, 2015; Seymour et al., 2015). While falsehoods and misleading content have threatened truth for centuries (Arendt, 1972; O’Neil and Jensen, 2020), online user-generated media may have amplified the threats to sensemaking and decision making. For example, some evidence suggests fabricated news articles (as identified by third-party fact checkers) spread significantly faster through Twitter than genuine news articles, especially articles about politics (Vosoughi et al., 2018). It is not clear to what extent this is an effect of recommender algorithms, as opposed to non-algorithmic social media creation and sharing dynamics (Hosseiniardi et al., 2020; Munger and Phillips, 2020). Similarly, COVID-19 misinfo has been common on social media globally, especially early in the pandemic (Al-Zaman, 2021; Brennen et al., 2020) and there is evidence that suggests misinformation has contributed to the spread of the virus via reduced vaccination rates (Loomba et al., 2021; Pinna et al., 2021). This suggests that personalized

recommendations may have had a causal effect on disease spread, though this has not been directly studied.

Many authors have argued that designers of news recommenders have editorial responsibility similar to news editors (Fields et al., 2018; Helberger, 2019; Nechushtai and Lewis, 2019; Sørensen and Hutchinson, 2018). Getting the right information to the right people at the right time is a key normative concern that goes well beyond ensuring accuracy, a value that we call “informativeness.” We can look to the tradition of public-service journalism to inform recommender design (Fields et al., 2018; Sørensen and Hutchinson, 2018) but personalized news delivery is a new technology and specific editorial theories are still developing. One approach would be to try to deliver a news item if the user previously expressed an interest in a particular topic, if it reports on events that affect their life, or if there is an opportunity for the user to help others (Stray, 2012). These sorts of ideas have yet to be effectively translated into algorithmic terms.

*Diversity* is a value that can be relevant to consumers, content creators, and society in general. In industrial settings diversity has mostly been studied because increased *diversity* typically results in greater user satisfaction and user engagement, at least up to a point (Holtz et al., 2020; Kunaver and Požrl, 2017; Mehrotra et al., 2018; Wilhelm et al., 2018). Many recommenders use diversification algorithms for the practical task of ensuring that users are not continuously shown the same type of item in their recommendations, often implemented as a re-ranking pass (Google, 2020; Ziegler et al., 2005). Other parties also have an interest in diversity. A streaming music or television platform needs to ensure that the long tail of less popular artists or producers have enough exposure to make it worthwhile for them to stay on the platform (Mehrotra et al., 2018), and increased diversity may contribute to equity of attention (Biega et al., 2018; Singh and Joachims, 2018).

The experience of a lack of diversity in personalized content consumption has been described as a “filter bubble” or “echo chamber.” However, this language is somewhat vague and has been used to describe a wide variety of phenomena including self-selected consumption behavior, homophily in social networks, and algorithmic feedback effects (Bruns, 2019). We consider such possibilities more specifically throughout the rest of this paper.

*Diversity* has been most specifically studied in the context of news recommendations, where it might serve a variety of democratic goals including *consumer choice*, *civic participation*, *pluralist tolerance*, or challenging the status quo (Bernstein et al., 2020; Helberger, 2019; Vrijenhoek et al., 2020). The meaning of diversity can differ between news organizations, depending on their editorial missions and the balancing of other values that matter to the organization (such as personal relevance, engagement, or time spent). In practice diversity is often measured using item-similarity metrics (Kunaver and Požrl, 2017) but such formulations do not yet capture the complexity that the social sciences have brought to the debate about media diversity (Loechebach et al., 2020).

One concern is that a lack of diversity in personalized news recommendations could prevent users from being exposed to contrary perspectives. Simulations have suggested that optimization to increase user engagement could create feedback loops that drive users into narrower selections of content (Carroll et al., 2021; Jiang et al., 2019; Kalimeris et al., 2021; Krueger et al., 2020; Rychwalska and

Roszczyńska-Kurasińska, 2018). However, news personalization algorithms do not seem to produce a less diverse selection than human editors (Möller et al., 2018) and the news provided to different users is quite similar (Guess et al., 2018; Nechushtai and Lewis, 2019). Social media users consume a more diverse range of news sources than non-users (Fletcher and Nielsen, 2018) but correspondingly also consume more news from partisan sources (Fletcher et al., 2021).

A related concern is that recommender systems might be causing large-scale polarization of ideology (issue polarization) or attitudes (affective polarization) (Stray, 2021a). The evidence for this hypothesis is mixed. In the U.S., polarization began increasing decades before social media (Boxell et al., 2017) while several other developed countries have similar internet usage but do not show increasing polarization (Boxell et al., 2020). Paying people to stop using social media for several weeks produced small declines in issue polarization measures in a study of American users (Allcott et al., 2020). However, an attempted replication in Bosnia and Herzegovina showed *increases* in a measure of ethnic affective polarization after not using social media for two weeks (Asimovic et al., 2021). This was mediated by the ethnic diversity of users' offline social networks, suggesting that in this context the platform was providing otherwise absent exposure diversity.

Finally, we note that there are additional values on our list that have more of a societal flavor such as *civic engagement, social progress, meaningful work, environmental sustainability, and tradition and history*. These are important values, though the degree to which they are considered important varies considerably from one culture to another. In our survey and the literature we reviewed these values did not come up as particularly relevant to online platforms, but that does not mean that they cannot or should not be promoted in certain contexts.

## Safety

*Safety* includes the idea that people should not be bullied, attacked, or dehumanized and should not be exposed to disturbing content. For social media, safety has thus far mostly been considered in the context of content moderation, and many platforms have developed their content moderation policies based on international human rights principles and frameworks and in consultation with third party experts (Douek, 2020; Facebook, 2021a; Patja Howell, 2021). As a notable example, a series of human rights abuses occurred in Myanmar, characterized by hate speech and disinformation against the minority Rohingya by the military on social media (BSR, 2018).

Recommender systems should not promote violence. Note that “polarization,” a mass hardening of political divisions, is conceptually distinct from “radicalization,” where a small number of individuals violate mainstream norms and may resort to violence (van Stekelenburg, 2014). There are a number of documented cases of far-right and Islamist radicalization where online recommendations were involved (Baugut and Neumann, 2020; Munn, 2019; Roose, 2019; Washingtonian, 2019). However, these reports mention many other factors including chat rooms, personal relationships, user-directed searches, and life circumstances. More systematic studies have looked for recommender feedback effects that move users toward radicalization en masse (Faddoul et al., 2020; Ledwich and Zaitsev, 2019; Munger and Phillips, 2020; Ribeiro et al., 2020a) These studies generally show that recommenders alter content mix in the direction of engagement, but have produced poor evidence on the radicalizing potential of recommender

systems because of insufficiently powerful experimental designs, as we will discuss below. In general, causal understanding of user trajectories through recommender systems remains a major challenge.

### 3. The State of Practice

To ground the discussion of human values in recommender systems, we review the current state of practice from two perspectives. In Section 3.1 we describe a general architecture that captures common design patterns in current commercial recommender systems. In Section 3.2 we illustrate how the engineering of a recommender system to support a specific value might proceed, within the context of a large company. We note that recommender objectives are chosen and evaluated by people working in teams, which makes these human processes an important determinant of system outcomes. These sections draw heavily on the experience of those of our authors who are practitioners working on large, contemporary recommender systems.

#### 3.1 How Recommenders Work

Recommender systems are often described as “black boxes” (Singh and Doty, 2021a) but most are constructed using similar principles. This section presents a greatly simplified, but illustrative recommender design. While it doesn’t represent the details of any particular system, many real systems share its features. Our discussion also leads to a technical definition of *engagement* in terms of behavioral data. Most recommenders are built to optimize some form of engagement at their core, though they also consider many other types of signals.

Recommendations start with a pool of content items. These items may be produced entirely by the recommender operator, as with a news organization’s recommendations; or they may be curated from multiple sources, such as a music recommender which gets content from publishers; or items may be entirely user generated and posted without prior review, as on social media. These three categories have been called “closed,” “curated,” and “open” recommender systems (Cobbe and Singh, 2019). Before recommendations are generated, a moderation process identifies and removes items which violate platform policies. Platform moderation is a complex process which involves many human and machine steps from policy making to enforcement to appeals (Halevy et al., 2020; York and Zuckerman, 2019) but here we are concerned only with how moderation defines which items are available for recommendation.

Individual recommendations are generated using data about item content, user attributes, and context, and result in an output stream or set of recommended items, sometimes called a *feed* or *slate*. The context may include a wide variety of features such as the video the user is currently watching, the time of day, or a search query like “politics podcast” (Bauer and Novotny, 2017; Beutel et al., 2018). User attributes may be derived from personal information the user has provided, any explicit user feedback or control settings, and any implicit feedback contained in the history of past interactions with the system.

The recommendation process usually begins with the selection of *candidate items*. Candidate generation algorithms are tuned to retrieve an overbroad sample, but are very efficient at filtering a corpus of

(potentially billions of) items available for recommendation down to a small set which might be a good fit for the user and context, typically ranging in size from a few hundred to a few thousand. These candidates are then *ranked*, that is, each is assigned a *relevance score*, which typically reflects a prediction of user engagement with the candidate item. In modern systems ranking may involve dozens or hundreds of “signals” which summarize aspects of the content, the user, the context, and how all of these interact. The top scoring items are then selected as the user’s recommendations. Many systems then re-rank the remaining items, this time comparing them to each other, rather than evaluating each individually, to achieve goals such as diversity of item topic or source (Goodrow, 2021; Ziegler et al., 2005).

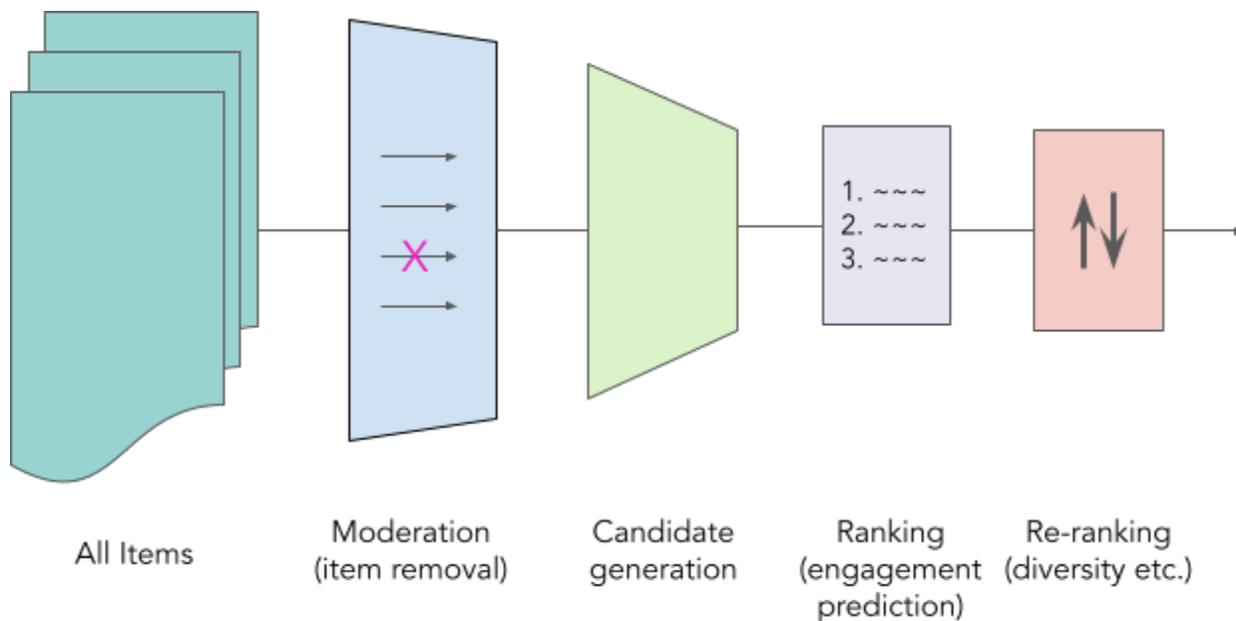


Figure 1: An illustration of how many modern recommenders work, adapted from (Google, 2020; Lada et al., 2021). The global set of items is first moderated to remove content that violates platform policies. The remaining steps happen when a user is served recommendations: candidate generation selects a wide set of items that could be relevant, ranking scores each one, and re-ranking ensures feed-level properties like diversity.

Item ranking constitutes the core of personalization. The final item score is typically a weighted combination or some other function of (i) the predicted probability of a number of different types of user responses (Facebook, 2021b; Milli et al., 2020; Smith, 2021; Zhao et al., 2019) plus (ii) a wide variety of scoring signals that range from source credibility (Pennycook and Rand, 2019) to playlist diversity (Mehrotra et al., 2018) to whether an item tends to be inspiring (Ignat et al., 2021). Ranking items by the probability of desired or targeted user reactions (e.g., sharing, dwell time) is informally known as optimizing for *engagement*. This may or may not optimize for *value* to various stakeholders, which is why many non-engagement signals are also used.

The word “engagement” has been used across many fields including media and technology to suggest that users are repeatedly interacting with a product, as evidenced by a wide variety of metrics. Here we propose a more specific definition, compatible with recommender design practice. We take engagement to be *a set of user behaviors, generated in the normal course of interaction with the platform, which are*

*thought to correlate with value to the user, the platform, or other stakeholders.* This definition builds on previous work (Jannach and Adomavicius, 2016; Yi et al., 2014), is multi-stakeholder in nature, and reflects the fact that engagement signals are chosen to be indicators of value, but aren't going to be fully aligned with specific values in all cases. It also suggests there are some signals of value that can only be derived from non-ordinary or non-behavioral data, as we discuss below.

### 3.2 Implementing a Product Change

In this section, we illustrate the process by which a specific human value might be incorporated into a recommender system's design using the value of *diversity*. We walk through the steps that could be involved when a news recommender operator wants to increase political diversity, and illustrate the complex interplay between social, technical and organizational factors when building real products to reflect specific values. It introduces basic themes that we discuss throughout the rest of this paper, including selecting values, creating metrics, implementing product changes, and evaluating trade-offs.

While we hope this example is illuminating, it is not intended to be a general theory of value-sensitive design for recommender systems. Rather, this section is intended to be descriptive, based on the authors' collective experience working with platform-scale recommenders. Unfortunately, there is little public documentation of the actual processes by which values are engineered into large recommender systems. The processes outlined in this example do not represent those of any particular company or product, nor are they universal; but the example nonetheless offers a window into typical practice.

Consider the development or refinement of a news recommendation platform in which the designers are concerned with users having the opportunity to develop an awareness and understanding of multiple political perspectives—this has been called the *deliberative perspective* on recommender diversity (Helberger et al., 2018). The question of which values matter most for any given recommender is an important one, but we suppose the decision has been made to prioritize this type of civic tolerance, perhaps from the study of existing recommender outcomes. One way that a recommender system might contribute to this value is by increasing the diversity of recommended items. This strategy rests on two key assumptions: there is currently a lack of diversity in the items users see, and showing them more diverse items will increase tolerance. Both of these assumptions are complex and available evidence is mixed (Stray, 2021a). Nonetheless, this is a non-trivial example of how metric-driven recommender engineering might proceed.

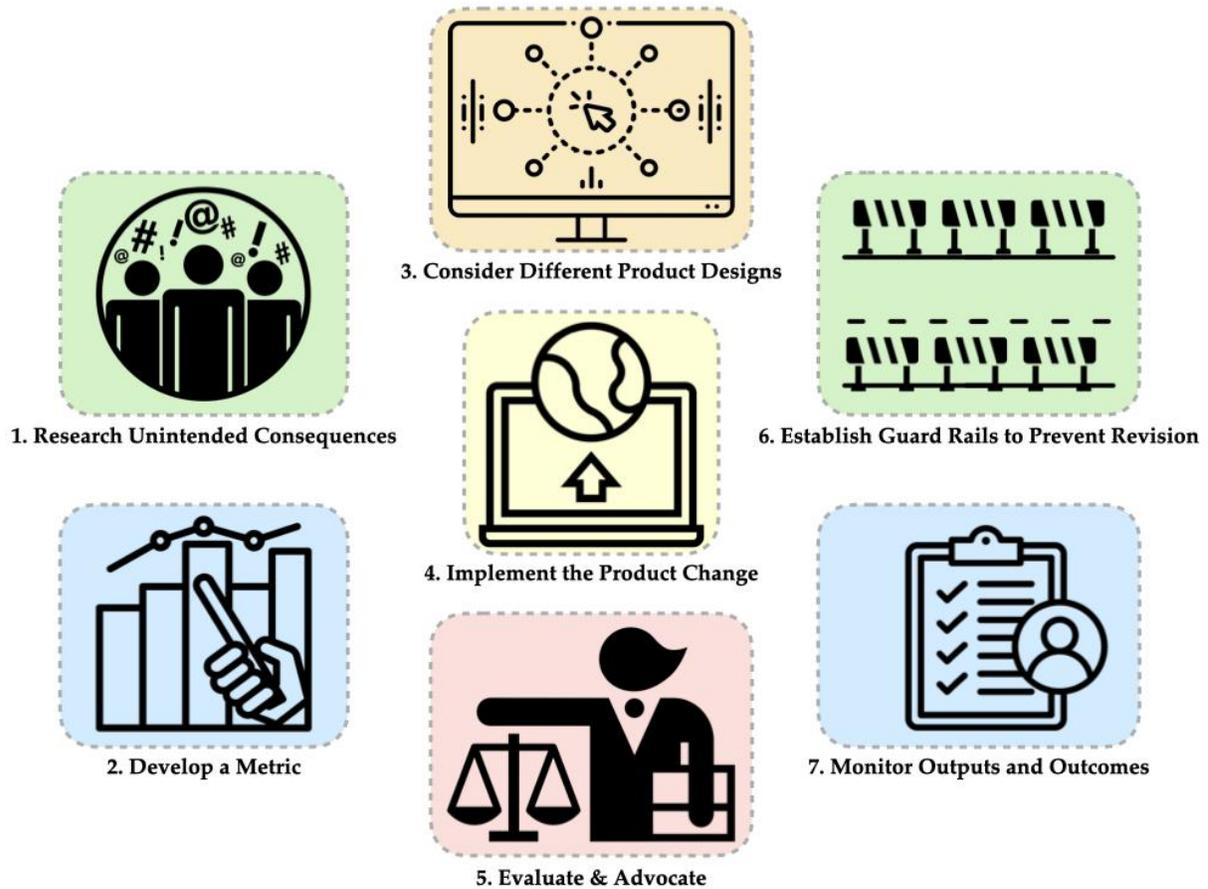


Figure 2: To ground our discussion of values in recommendation systems, this section discusses a potential product change to improve the diversity of content recommendations. We present seven distinct steps: 1) research unintended consequences, to reduce the risk associated with the change; 2) develop a metric to measure progress; 3) consider different product designs; 4) implement the product change; 5) evaluate the change and advocate for it with internal stakeholders; 6) establish guard rails to prevent revision based on other changes; and 7) monitor outputs and outcomes to identify, e.g., drift between your metrics and the concept you ultimately care about.

Once the decision to increase content diversity has been made, the implementation might proceed using the steps outlined below.

### 3.2.1 Research unintended consequences

Generally, such an effort begins with the study of potential mechanisms to incorporate this value. In the case of diversity, designers may research the history of attempts to create inter-group tolerance through diversity of exposure, including the ways in which this has failed (Paolini et al., 2010; Pettigrew and Tropp, 2006). User studies may be used to test the exposure diversity hypothesis with users of the platform, assessing the impact of seeing news items from diverse perspectives on people’s knowledge and attitudes. In particular, it will be important to test for backfire effects where people reject diverse information (Bail et al., 2018; Dunn and Singh, 2014) or other unintended consequences. There must also

be a reasonable expectation that the changes to the recommendation system won't adversely affect other values that the system embodies. In concrete terms, this often means that any proposed change should not decrease other metrics (see below) more than a specified amount.

### 3.2.2 Develop a metric

As discussed in Section 4, metrics are central to any attempt to build values into recommender systems. A variety of diversity metrics have been proposed, both for news content (Cowlshaw et al., 2018; Vrijenhoek et al., 2020) and for recommender systems more generally (Castells et al., 2015; Kunaver and Požrl, 2017). Choosing or developing a metric requires settling on a definition of diversity suitable for that particular system. It may be that multiple metrics are needed to capture different aspects of diversity. While product teams typically choose metrics, there has been experimentation involving external stakeholders in order to increase the legitimacy of such decisions (Bernstein et al., 2020; Lee et al., 2019; Stray, 2020).

Here we focus on a conception of “*productive*” *diversity*, where people disagree in ways that are ultimately constructive (Stray, 2021a). Given this concept, developing a method to measure the diversity of articles on the platform may be broken into several phases, such as:

- 1) Developing a description of diversity for use by human raters (e.g., “Does this set of articles include constructive contributions from multiple perspectives?”).
- 2) Creating a training data set of positive and negative examples using human-rater evaluations.
- 3) Using this training data to develop a heuristic or machine-learned model that can predict whether a list of (recommended) articles adequately reflects “productive diversity.”

### 3.2.3 Consider different product designs

Before implementing a specific product change, different product-based approaches to increasing diversity will be explored. For example, the designer could change the user interface, perhaps by showing related articles from other viewpoints below each item (Su, 2017), or change the ranking algorithm to try to nudge people to consume more articles that represent this type of diversity (Mattis et al., 2021). For the sake of example, we assume below that the latter change has been selected.

### 3.2.4 Implement the product change

Implementing this specific product refinement requires incorporating the diversity prediction model into the item ranking procedure. Whereas many current recommender systems score each item independently, diversity is a property of sets of items. One challenge is that scoring entire lists is both more complicated and more expensive than scoring individual items (Ie et al., 2019; Wilhelm et al., 2018) which may necessitate the development of more efficient algorithms for ranking such item sets.

### 3.2.5 Evaluate and advocate

Once implemented, the new product feature will typically be tested using offline data (Cañamares et al., 2020; Gilotte et al., 2018) followed by A/B tests with a small groups of users. If these tests show positive results, the new diversity prediction model will be deployed in production, and monitored to see whether the target diversity metric improves (this may involve gradual ramping up of the deployment to larger

numbers of users, holdback tests, etc.). There may also be side effects—for example, while increasing diversity often increases engagement (Kunaver and Požrl, 2017; Wilhelm et al., 2018), this is not always the case. Suppose that as a by-product of the model deployment there is a drop in engagement, declines in other values-relevant outcomes such as quality or safety metrics, or some cost imposed on other stakeholders (e.g., content producers). In practice, this requires negotiation among internal stakeholders to decide if the increase in diversity is significant enough to justify a drop in other metrics.

### 3.2.6 Establish guardrails to prevent reversion

Once deployed, product teams often establish a review process so that subsequent product changes, within the originating team or elsewhere, don't indirectly revert the diversity improvement. This might include numeric “guardrails” that specify the maximum allowable decrease in diversity induced by other product or algorithmic changes.

### 3.2.7 Monitor outputs and outcomes

In practice, product teams will continually monitor the diversity of recommended items to detect operational failures (e.g., bugs or system failures), or (induced or exogenous) changes in the user distribution and user behavior. A survey which asks users or human raters to detect “productive” diversity may also be regularly employed to detect model drift and produce updated training data. Finally, determining whether the ultimate goal is being achieved, or is still worthy of being achieved, requires ongoing evaluation. This could involve survey methods to assess metrics such as affective polarization, and ethnographic research to understand what it means for users to be encouraged to be “more tolerant” in this way.

## 4. Value Measurement

In order for recommender systems to incorporate human values, there need to be methods for measuring how well a system is adhering to, promoting, or facilitating these values over time. This section describes how to go from a value to a metric, the issues involved in designing good metrics for values, and the data that is available on which to build such metrics.

Our description of recommender operation in section 3 didn't show what comes before and after recommendation proper. Figure 2 expands our view to include the process of selecting metrics, the interactions a user has, and the ultimate outcomes.

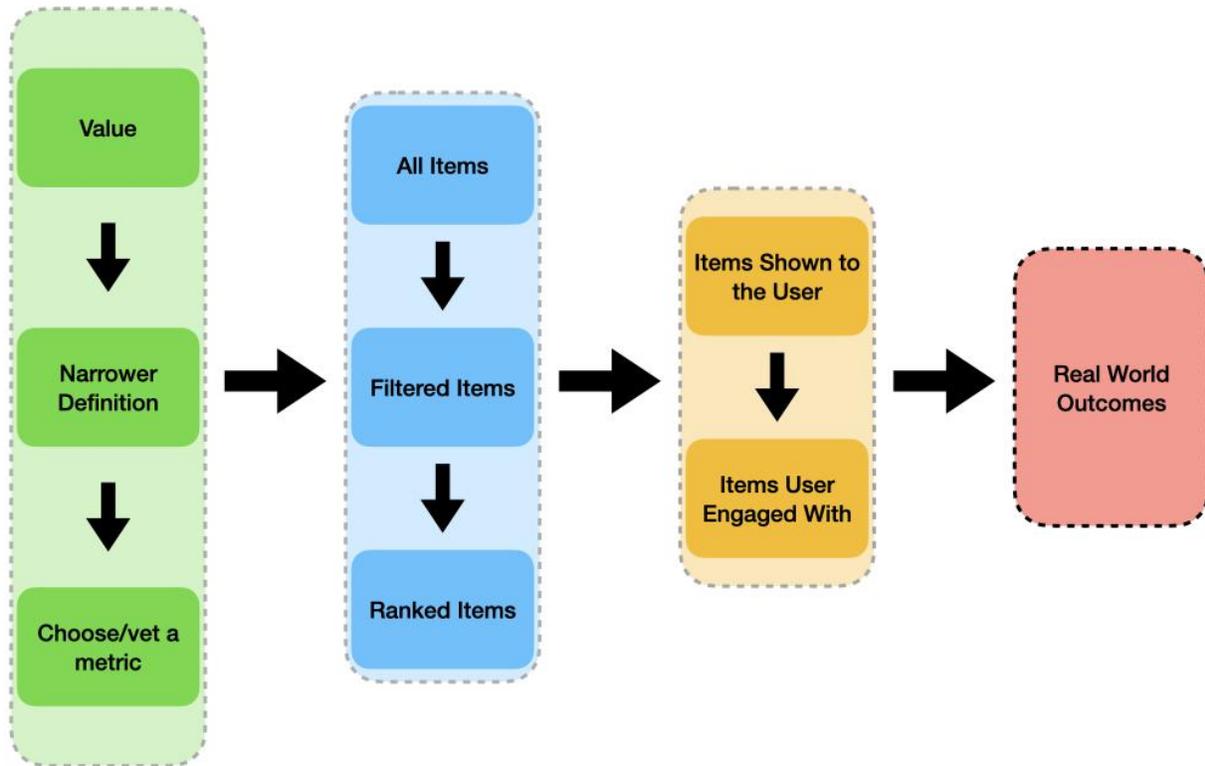


Figure 3: An illustration of the relationship between values, metrics, recommenders, items, and outcomes. The selection of a metric leads to a selection of items, some of which the user will engage with. Over time, the items a user engages with shape outcomes of interest.

#### 4.1 From Value to Metric

Defining a metric is part of the process of specifying what, exactly, a particular human value means in the context of a real system. This process is called *operationalizing* a value. The people involved in defining metrics have considerable influence over the ultimate function of a recommender, which is why multi-stakeholder involvement in recommender metric selection may be important (Lee et al., 2019; Stray, 2020).

To illustrate the gap between a value and its measurement, consider the value of “safety,” and in particular, protecting users from hate speech. A precise definition of hate speech is not only hard to articulate, but under constant debate and evolution, and every choice has some set of undesirable side effects (Díaz and Hecht-Felella, 2021; Gordon et al., 2021; Martinez and Paluck, 2020, 2020; Papcunová et al., 2021). Well-being is an even more complicated example. As discussed, well-being has many components (e.g., physical health, health of personal relationships, having purpose in life, etc.) and can be considered in the short term (e.g. entertaining content) or the longer-term (e.g. learning useful skills or fostering relationships). Making news recommendations that serve the public interest likewise requires

defining concrete metrics that reflect some assessment of that interest (Fields et al., 2018; Vrijenhoek et al., 2020). At some point, a real recommender must commit to specific operationalizations of broad concepts, with the resulting tradeoffs between competing values and stakeholders.

## 4.2 Characteristics of a Good Value Measurement

A good measurement has a number of desirable properties, including validity and reliability (Mohajan, 2017), fairness, and legitimacy. Quinn et al. (2010) summarize the situation as:

The evaluation of any measurement is generally based on its reliability (can it be repeated?) and validity (is it right?). Embedded within the complex notion of validity are interpretation (what does it mean?) and application (does it “work”?).

(Quinn et al., 2010, p. 216)

Many social science theories involve quantities that are not directly observable and hence must be inferred, often with considerable uncertainty, which makes any measurement instrument implicitly a model of some purported underlying reality (Jackman, 2008). Construct validity is the convergence of a successful theoretical idea with a measurement that effectively captures it (Strauss and Smith, 2009). Jacobs and Wallach (2021) propose several other types of validity in the context of measurements within computational systems, including face validity (is the metric plausible?), content validity (does the metric capture all the relevant parts of the concept of interest?), convergent validity (does this metric agree with other accepted metrics?) and discriminant validity (does this capture something different than other metrics?). Reliability is typically evaluated in terms of agreement between multiple measurements (test-retest reliability), between different human judges (inter-rater reliability) and, for surveys, between different ways of asking a question (inter-item reliability) (Jackman, 2008).

It is rare that a particular measurement fully captures what we mean by some value in a particular context, meaning that most metrics are in fact proxies for what we actually care about. Even a good metric will change meaning when it is known to be used to make consequential decisions, e.g. student test scores must be interpreted differently when they are used to decide academic progression, because instructors will begin “teaching to the test.” This effect is sometimes known as *Goodhart’s law*, but there are a variety of different causal structures which can produce feedback processes that widen the distance between a metric and the underlying value (Manheim and Garrabrant, 2018; Mullainathan and Obermeyer, 2021). In the technical community this has been most discussed in the context of the general problem of algorithmic optimization and the difficulty of objective specification (D’Amour et al., 2020; Hadfield-Menell and Hadfield, 2019). Hence, as part of being precise about the definition of a human value, it is important to identify gaps in what is measured and monitor them over time. Human values — that is, what is considered important — also tend to change over time. Qualitative user research plays a critical ongoing role in designing and evaluating metrics.

Because modeling assumptions are required to connect a measurement to the underlying value it purports to reflect, measurement itself has fairness implications. The signals from measurements can vary across demographics, even after controlling for differences in user intent and task (Mehrotra et al., 2017). For example, if older users read more slowly than younger users, then a metric based on dwell time will be an

over-optimistic measurement for older users regardless of their level of satisfaction. Thus, interpreting metrics at face value may systematically disadvantage and misrepresent certain demographics and user groups. Organizations deploying recommendation systems often rely on *internal auditing* methods as a way to measure how the overall system performs across different demographics and other user attributes (Feldman et al., 2015; Raji et al., 2020). External audits may also be required by regulation, as discussed below.

It is not enough to have an accurate and fair measurement of a human value. Stakeholders in the recommender system need to be able to accept the process and outcome of the measurement as legitimate. This means that two metrics that are operationally identical (e.g. they both correlate strongly with a desired outcome) may not be interchangeable. Transparency around how the measurement is carried out may help build trust in a metric. For example, some platforms periodically release public transparency reports which include various metrics (Singh and Doty, 2021b). Another way to increase legitimacy is to establish accountability regarding a measurement, e.g., through independent, external audits of the measurement (Suzor et al., 2018). More ambitiously, a metric could be created or chosen through a participatory process. For instance, the measurement could aggregate the opinions of a panel of users, as in the “digital juries” (Fan and Zhang, 2020) and “citizens assembly” concepts for making platform decisions (Ovadya, 2021). In one case, representatives of various stakeholders participated in the construction of a recommender that matched supermarket excess food donations to volunteer drivers and local food banks, using an elicitation process to define a ranking function (Lee et al., 2019).

#### 4.3 Data Sources for Measuring Values

Our news recommender diversity example involved a complex, multi-step process for defining an algorithmic measure, ultimately drawing on human labeling to define the value of diversity. Broadly speaking, there are three main data sources for value measurements on a recommender platform.

The first source of data is the behavioral signals that users generate during normal usage, e.g. articles a user clicks on, songs a user plays, comments, emoji reactions, re-sharing of content, ads a user clicks on, purchases made, time spent on the platform and on specific items, and so on. These sorts of implicit signals of value are often called “engagement,” but we note that some behavior signals are explicitly designed to give feedback to the algorithm, such as swipe left/right or thumbs up/down. Implicit and explicit behavioral feedback both provide distinct and useful information to a recommender system (Jawaheer et al., 2010).

The second source for signals of value is answers to survey questions that are posed to a fraction of the user base, typically a very small fraction. Surveys can ask very targeted questions, e.g., Facebook has asked users whether individual posts were “worth their time” (Facebook, 2019) while YouTube uses user satisfaction surveys that ask users to evaluate previous recommendations (Goodrow, 2021; Zhao et al., 2019). Survey results can be used to monitor real-world outcomes, evaluate A/B tests, and ultimately recommend items that a user is predicted to respond positively to when asked about their experience.

The third source for value measurements is human annotation. This data is often produced by paid raters, though human ratings also come from users flagging or reporting items. Vast amounts of human training data are routinely used to create models for identifying particular kinds of positive or negative content for the purpose of content moderation and ranking. Platforms that feature professionally-made content can also ask creators to provide metadata for a variety of ranking purposes. For instance, the Swedish public broadcaster asks editors to rate each story in terms of importance, public service value, and lifespan (Mccaffery, 2020). While survey data is limited by how often users can be asked to fill out a form, annotation data is costly to create and therefore available in limited quantities, and some types of labeling work may contribute to mental health issues for raters (Arsht and Etcovitch, 2018). Human annotation can also be noisy, inconsistent and biased depending on how rater pools are selected, and have limited, unbalanced coverage, while the meaning or usage of labels can change over time. Research in the area of human computation (Law and Ahn, 2011) attempts to address such issues (Faltings et al., 2014).

In contrast to surveys and annotation, the benefit of behavioral signals is that they are plentiful—many orders of magnitude more data is available. However, the behavior of users on the platform typically correlates with but does not perfectly capture any particular type of value, and is only a proxy for what different stakeholders actually care about. Moreover, behavioral signals have been shown to be sensitive to a variety of factors, such as the user and their demographics (Carterette et al., 2012; Hassan and White, 2013), the context in which the user is interacting with the system, and the recency of interactions. For example dwell time, a behavioral signal that has been used as a proxy for user satisfaction (Yi et al., 2014), can vary significantly depending on whether the item is the first one clicked in a list of results or not (Borisov et al., 2016). As discussed above, behavior does not represent underlying preferences for a variety of reasons including cognitive biases, information asymmetry, coercion, lowered expectations, and so on (Anderson, 2001; Bernheim, 2016; Camerer et al., 2004). Additionally, the collection of vast amounts of user data poses significant privacy concerns.

One important approach that tries to combine the benefits of behavioral and survey signals is to learn a model that predicts survey results given user behavior, that is, it tries to predict how a particular user would answer a survey question if shown a particular set of items. This method extrapolates limited survey data to all users, and is common in industry (Goodrow, 2021; Stray, 2020) but has not been much discussed. We view such prediction of survey results as an emerging method for aligning recommender systems with human values, though it must be understood that predictions are just proxies and must be continually monitored for divergence from ground-truth survey responses.

In principle, surveys can elicit complex judgements of value. In addition to the challenges of complicated, multi-component values (e.g., well-being) the context and wording of a survey can significantly affect the results (Stoutenborough et al., 2016; Weaver et al., 1997). Social desirability bias, the tendency of respondents to answer in ways that others would view favorably, adds a further complication (Furnham, 1986; Näher and Krumpal, 2012). Another issue arises when users from different demographic groups tend to answer questions differently (Eskildsen and Kristensen, 2011). Since online surveys often involve casual participants, “seriousness” checks and data denoising can be very important (Aust et al., 2013).

## 5. Design Approaches

We are now ready to discuss the major approaches that recommender system developers and researchers have used to promote the values discussed above, and in turn, to identify some of the challenges and open problems which require the development of new technologies. The structure of this section reflects the issues presented above. Section 5.1 discusses the core of the recommender system, namely the techniques used to select and order items shown to the user, and long-term user paths through these items. Section 5.2 discusses the affordances and controls made available to users and other stakeholders through UI and UX design. Finally, Section 5.3 considers fairness and multistakeholder perspectives, and describes techniques used by system designers to help optimize tradeoffs within real systems.



Figure 4: Section 5 discusses the design affordances that are typically used to modify recommendation systems to better align with values. We identify three primary options: 1) item selection, where designers intervene on the ranking of content to achieve a desired goal; 2) controls and feedback, which creates affordances for users to modify their experience on the platform; and, 3) by identifying and optimizes tradeoffs within the systems to better account for multi-stakeholder perspectives.

### 5.1 Item Selection

At the heart of any recommendation system are the models and algorithms used to generate one or more recommended items. Generally, the core models are trained to predict a user’s behavioral response to a candidate recommendation (e.g., watch, like, reply, purchase) using properties of the user, the candidate items, and the context. Items are then scored by combining predicted outcomes in some way (Milli et al., 2020; Zhao et al., 2019). The nature of the properties used to make predictions, the predicted responses themselves, and the way these predictions are combined into a final score all play a role in the values that a recommender embodies.

### 5.1.1 Item ranking signals

The properties of an item being considered for recommendation—whether a social media post, news article or musical track—play a key role in determining whether it is of interest to the user, and therefore whether showing it to a user can serve one of their values.

The “topic” of an item is an important signal. A variety of techniques have been used for text analysis in recommender systems, including latent semantic indexing in Google News (Das et al., 2007), Latent Dirichlet allocation (Jelodar et al., 2019) and transformers. Image and video analysis are also used for topic assessment (Deldjoo et al., 2020), as is audio analysis (Yu et al., 2020). Modern recommender topic taxonomies can encompass tens of thousands, or even millions, of distinct categories and sub-categories. One of the challenges is that these classifiers are typically accurate for the popular elements of the taxonomy of topics, but often do not perform well for posts on less popular topics.

Topics do not directly correspond to specific user needs or to values. A post about “football” can be about organizing a football viewing party (possibly contributing to the value of connection) or just reporting the latest scores or player injuries (the value of knowledge). To support the value of empathy and care it would be useful to identify posts where the poster could benefit from support of their network (e.g., after losing a loved one, needing advice on a certain matter, or announcing an important event in their life). An open challenge is inferring the way a particular user will *relate* to an item, rather than analyzing properties of the item alone. There is nascent work on predicting the experience a user may have when viewing content, e.g., whether they are entertained, angry, or inspired (Chen et al., 2014; Ignat et al., 2021). These sorts of inferences could also enable better targeted or more persuasive advertising, which raises policy concerns (Singh, 2020b).

Content analysis is also heavily used by social media platforms in order to determine whether a post violates community policies or is of low quality in some way, and should therefore be removed or demoted. The predominant method for determining such violations uses ML models trained on content labels provided by paid raters. These models increasingly use multimodal techniques, simultaneously considering text, speech/audio, images and video. This is often crucial to understanding the intent of a post, e.g., the caption “love the way you smell today” means something different when superimposed over the image of a rose vs. a skunk (Kiela et al., 2021).

Increasingly, items are evaluated not just in terms of their content but their context, including properties of the poster and audience, previous user behavior, reactions from other users, and so on (Halevy et al., 2020). Incorporating embeddings of sequences of user interactions has been shown to increase accuracy in misinformation and hate speech classification on Facebook (Noorshams et al., 2020) while “toxic” conversations on Twitter are better predicted by including network structure features (Saveski et al., 2021).

These models are generally trained on labels or annotations generated by human raters. This process itself is potentially subject to error and particular forms of bias, both in the instructions given to evaluators and their assessment biases (Sap et al., 2021). Also, many of these categories are both complex and politically contested (Bozarth et al., 2020; Martinez and Paluck, 2020) and raters often disagree, which complicates

the evaluation of model accuracy (Gordon et al., 2021). The promotion or demotion of particular topics or types of language, and the errors and biases in this process, have major implications for freedom of expression, which implies a deep connection between these technical methods and broader policy considerations (Douek, 2021; Keller, 2018; York and Zuckerman, 2019).

### 5.1.2 User Trajectories

We use “trajectory” to refer both to the sequence of items a user saw over time and the reactions those items evoked. A user’s past trajectory often provides information that can be used to make better predictions about their future preferences and behaviors. At the same time, we want to ensure that the user’s future trajectory supports the values they care about. Moreover, trajectories are the basic unit for studying some problems users experience on platforms, as they are central to discussions about potential long term effects on, say, well-being and polarization. For example, the “filter bubble” critique is essentially a statement about the typical course of user trajectories.

The majority of recommenders deployed in practice are “myopic” in the sense that they make predictions of a user’s immediate response to the next slate of items presented, and rank items based on these predictions. However, many recommenders use past sequences of user and system behavior in sophisticated ways.

Advances in deep learning have made possible the practical deployment of *sequence models*, such as recurrent neural networks or transformers. For example, Beutel et al. (2018) describe a recurrent (specifically, an LSTM) model deployed at YouTube. Recurrent approaches explicitly model a user’s “latent” or “hidden” state, that is, they include variables which represent aspects of the user’s situation or psychological state which are unobserved but have effects on what the user wants to see next. This state might encode aspects such as user satisfaction, frustration, or current topic focus, but interpretation of the hidden state embodied in such models is challenging and is tightly coupled to the engagement metrics being predicted and optimized. Inferring user state from behavior is an important challenge and a key step toward better supporting many values. For example, a recommender could detect a user’s dissatisfaction with a certain type of content, or with too little topic diversity in the recommendation stream, or even that someone was developing an eating disorder (Yan et al., 2019).

Current ranking techniques do not offer the means to (directly) optimize a user’s *future* trajectory. A promising direction is the use of *reinforcement learning (RL)* for optimizing such futures non-myopically (Afsar et al., 2021; Shani et al., 2005; Taghipour et al., 2007). In particular, the use of RL allows the system to consider the impact on the user of not just immediate recommendations, but of the entire sequence of recommendations it might make over an extended period, and plan that sequence accordingly.

In our news diversity example, the redesigned recommender considered diversity within each slate of recommended items independently. An RL-based recommender would be able to consider the diversity of items over days or weeks as well. In an educational setting—where a user’s understanding of a topic is best served by a graduated but individualized and adaptive sequence of content—RL can be used to plan and adapt the appropriate recommendation sequence. As such, RL offers considerable promise as a technology for individual or user-level value alignment. To date it has been used largely to optimize user

engagement over the long term, but with suitable metrics and objective criteria, it could play a vital role in better aligning recommendation trajectories with user well-being by adaptively planning recommendation sequences.

As with sequence models, advances in deep RL methods have made it more practical to deploy RL in practical recommender systems (Chen et al., 2019; Gauci et al., 2019; Ie et al., 2019). That said, a number of challenges remain. First, since RL relies on sequential interaction data with real users, such models are often trained offline using data generated by past recommendations, which may induce bias because the user was interacting with a different model (Chen et al., 2019). The second challenge involves choosing from a large action space ranging from hundreds to millions depending on the recommendation task (Dulac-Arnold et al., 2015; Ie et al., 2019). A related problem is that item recommendations are often made jointly in slates or scrolling feeds, where the interactions or interference among the visible items makes interpreting and optimizing for user responses challenging (Ie et al., 2019). Finally, adopting RL for true value-alignment requires sophisticated models of various aspects of user latent state (e.g., preferences, satisfaction, awareness/knowledge levels, fatigue/boredom, etc.) and their dynamics. These psychological and situational states are challenging to uncover from observable user-response behavior, and may require planning over extremely long “event horizons” as user adaptation to recommender changes may take 3-6 months to fully materialize (Carroll et al., 2021; Mladenov et al., 2019).

Some work has considered trajectories that might be unhealthy or harmful. One problem concerns minimizing the likelihood that trajectories will end up with users viewing large amounts of questionable content, such as videos that might promote unhealthy behavior, false statements about COVID vaccines (Chan et al., 2021), or other content that may not explicitly violate the platform’s policies. This might occur not because users explicitly ask for such material, but because they end up in places (e.g., groups) where this material is common. In addition to downranking such content, another solution is to avoid recommending low quality groups or content sources to users.

Another strand of work concerns items that are acceptable or appropriate when considered in isolation, but could be harmful if consumed too much or by certain vulnerable people. For example, there may be nothing wrong with a diet video but perhaps someone with an eating disorder should not be presented with an endless stream of such videos; or it may be unhealthy to watch exclusively violent movies. Singh et al. (2021) address this possibility by ranking user trajectories by proportion of “unhealthy” content, then using the mean proportion of this content in the top  $\alpha$ th percentile of user trajectories as a regularization metric. They demonstrate a safe RL approach that improves both worst-case and average-case outcomes, in terms of the fraction of “unhealthy” items recommended to any one user.

The most complex concern about trajectories is the possibility that recommender systems might change user preferences in a manner that increases engagement but harms some other value. This is commonly associated with the idea of “manipulation,” meaning unwanted changes in attitude or behavior (even if unintentional).

This sort of optimization-driven shift has been frequently suggested as a mechanism driving filter bubble, echo chamber, or polarization effects, though the empirical evidence is mixed (Bruns, 2019; Guess et al., 2018; Stray, 2021a; Zuiderveen Borgesius et al., 2016). Such models posit a feedback loop where users

choose particular items (as in selective exposure effects (Prior, 2013)) and the recommender responds to that engagement by narrowing its output to those topics, which in turn shifts user preferences further in that direction. This effect may be a particular concern for RL systems, which may learn how to make their users more predictable so as to maximize engagement (Russell, 2019). A polarizing preference shift effect has been demonstrated in multiple simulations with different specifications (Carroll et al., 2021; Evans and Kasirzadeh, 2021; Jiang et al., 2019; Kalimeris et al., 2021; Krueger et al., 2020; Rychwalska and Roszczyńska-Kurasińska, 2018) which suggests that it could be a robust effect.

A range of work has attempted to determine the causal effect of recommender systems on content consumption trajectories. This is particularly important if those trajectories are correlated with offline outcomes such as violence (Global Internet Forum to Counter Terrorism, 2021). There is replicated evidence that strongly moralizing expression spreads faster than other content on social networks (Brady and Van Bavel, 2021) and such moralizing seems to precede offline violence (Mooijman et al., 2018). A number of researchers interested in categories such as “far right,” “conspiracy,” or “radical” have studied the network structure of recommendations between YouTube channels (Faddoul et al., 2020; Ledwich and Zaitsev, 2019; Ribeiro et al., 2020a). While this showed that more extreme channels often link to each other, these studies do not analyze user trajectories because they were conducted without any personalization. A different approach is to program bots to selectively engage with certain topics. This has generally shown that engaging with some type of content increases its frequency (Wall Street Journal, 2021; Whittaker et al., 2021) but this design models users as unchanging, so it does not provide evidence on persuasive effects. A user-tracking study of “far right” content on YouTube from 2016 to 2019 found that consumption there matched broader patterns across the web, including consumption from sources without recommender systems (Hosseinmardi et al., 2020). Separating consumption due to recommenders and consumption due to users’ seeking behavior is difficult without on-platform experiments. Using a set of Twitter users randomly assigned to receive a baseline chronological feed, Huszár et al. (2021) found that Twitter’s home timeline recommender reduced the consumption of more politically extreme material.

In general, inference of the causal effects of recommenders on user preferences and outcome is a major open problem. The feedback effects between algorithms and societies are at the cutting edge of social science research, in need of both interdisciplinary and cross-sector collaboration.

## 5.2 Controls and Feedback

Meaningfully supporting human values requires effective communication between the user and the system, beyond the standard implicit signals (clicks, shares, dwell time, etc.). We place communication affordances on a continuous spectrum between explicit and implicit. We use controls to refer to features where the user can explicitly change certain settings, and feedback to refer to situations where the user is more passive and their preferences are elicited in some way, such as by answering a survey question, providing a like/dislike, etc.

There are a variety of documented benefits to providing users with more control over their experience. Users have reported being more satisfied with their recommendations when given greater control (Jin et al., 2017) and better controls have led to increased engagement (Jannach et al., 2017). Interestingly, even

the appearance of controls can achieve some of these effects. In (Vaccaro et al., 2018), users reported higher satisfaction with the system even though the controls had random effects. Conversely, Loecherbach (2021) gave users control over the diversity of items in their feed but there was no correlation between actual and perceived diversity. In general, greater recommender control usage does not always translate into better control (Waern, 2004).

Existing systems offer a variety of ways for a user to customize their experience, and many more control and feedback schemes have been proposed by researchers (He et al., 2016). Some involve simple direct feedback on individual items, e.g., thumbs up / thumbs down, or item-level ratings (Boutilier et al., 2003; Zhao et al., 2013). Others involve evaluating the relevant importance of pairs of items (Boutilier, 2013; Chen and Pu, 2004). Even when controls are provided, many users do not know that they exist or what they do (Hsu et al., 2020; Singh, 2021), find them challenging to use, or simply don't see the value in engaging with them. As a result, most users do not use recommender controls and a "passive" user experience remains the default (Jin et al., 2017). Aside from better educating users on controls, a promising approach is to design feedback mechanisms that serve multiple purposes: such as sending a social signal (which is why the user would use it) and simultaneously providing direct feedback that pertains to some value. Some examples of this exist today, such as the "Insightful" emoji on LinkedIn, the "Care" emoji on Facebook, and the proposed Respect button (Stroud et al., 2017).

One of the issues that limits the use of controls is the difficulty in creating a direct, understandable link between the input a user provides and the resulting change in the system's behavior. In some cases this link is quite direct (e.g., "don't show me any sports-related content"), whereas in others the outcome of a control/feedback action is less obvious (e.g., "show me more videos like this video I just watched"). There are many points along this spectrum. For example, instructing the system to "show less about football" will be largely predictable in terms of the content it affects (though what about political statements/protests by athletes?) but users may still not understand the expected magnitude of change. To complicate things further, there is often some ambiguity about whether the change the control offers is transient (e.g., for the current session) or for the longer term. One proposed design is to highlight items that will be added and removed from the user's feed when a control is changed (Schnabel et al., 2020).

Organizing content into "channels" can help users to better customize their experience. Some systems include implicit ways of specifying what sort of items are to be included, such as music recommenders which can continue a human-generated playlist (Zamani et al., 2019). Algorithmic channels could also be designed around particular goals (e.g., learning to play guitar) or needs (e.g., getting support from friends, participating in lively conversations). The primary technical challenge here is building relevance measures that capture the dimensions that users care about.

In the future, communication between the user and the system will take different forms. Conversational recommender systems are an emerging area that offer a potentially more natural, engaging, and usable interface for people to express their preferences, particularly in light of recent advances in speech, natural language and dialog technologies (Adiwardana et al., 2020). Applications range from integrations with the approaches listed above to entire reimaginations of the control process (Christakopoulou et al., 2016; Jannach et al., 2021). A related idea is recommender personae, where the recommender is associated with a particular personality (e.g., explorer, diplomat, expert) to set a particular context for the conversation

(Harambam et al., 2018). Conversational recommenders face many of the same challenges as other control systems, but moreso: they must interpret the metaphorical, imprecise, or subjective language a user may use to convey their needs or topic preferences.

### 5.2.1 Meaningful explainability

The field of explainable recommender systems has grown into a vibrant research area (Zhang and Chen, 2020) in part because users respond positively to having good justifications for why certain items are suggested. Tintarev and Masthoff (2015) argue that a good explanation can increase transparency, scrutability, trust, effectiveness, efficiency, persuasiveness and satisfaction in the system. Each of these terms expresses somewhat different goals: transparency shows how the system works and can be instrumental in accountability; scrutability allows the users to tell the system it is wrong; trust increases confidence in the system; and satisfaction increases users' sense of derived benefit; effectiveness helps the user employ the system toward their own aims; persuasiveness convinces the user to change their beliefs or behavior; efficiency can increase the value of the system. Some of these values can be contradictory and may not be achieved at the same time. For example, an explanation that increases transparency of the system does not necessarily increase trust if the explanation is not understandable or reveals undesirable behavior. Therefore, what constitutes a "good" explanation very much depends on the goal, and the field is rapidly developing (Hase and Bansal, 2020; Narayanan et al., 2018; Nunes and Jannach, 2017).

With the rise of increasingly complex machine learning models, it has also become increasingly difficult to give intuitive and understandable explanations of why a user received a specific recommendation. Explanations may be shown to the user in different forms (e.g., text, visuals, highlighting relevant features) and may either attempt to explain the workings of the recommendation model itself, or may be the result of training a separate model that generates post-hoc explanations from model inputs and outputs (Zhang and Chen, 2020). There are also recommender algorithms specifically designed to be explainable (Xian et al., 2021). The system of Balog et al. (2019), for instance, operates based on set intersections, e.g., "recommended for you because you don't like science fiction movies except those about space exploration." This explainable-by-design approach avoids the challenges of interpreting learned models (Rudin, 2019), but generating understandable explanations from deep learning models is an active area of research that may yet prove fruitful (Kim et al., 2018; Kokhlikyan et al., 2020; Sundararajan et al., 2017).

## 5.3 Making Tradeoffs

Our news diversity example involved a tradeoff between a diversity metric and a short-term engagement metric. In general, there are numerous tradeoffs involved when incorporating human values into recommender systems, and a variety of techniques to evaluate these tradeoffs and make choices. There are at least three categories of tradeoffs: 1) tradeoffs between the benefits and potential harms to different people, 2) tradeoffs between different types of stakeholders, such as content creators versus content consumers, and 3) tradeoffs between values, resulting in different (but not obviously worse or better) measurable outcomes induced by various recommendation strategies.

Ultimately there must be some notion of a fair or optimal tradeoff, and techniques for making “good” tradeoffs in this sense. We hope that these tradeoffs are informed by the expressed opinions of users and other stakeholders, so we first discuss the theory of social choice, which studies how to combine the preferences of many people. We then discuss techniques for achieving various notions of fairness, such as between different types of users or between different stakeholder categories. Finally, we discuss tools designed to optimize tradeoffs when faced with the practical necessity of tuning a recommender’s ranking function.

### 5.3.1 Tradeoffs in theory: Social Choice

People express their values in their everyday use of recommender systems. Indeed, many recommender controls are designed specifically for this purpose, everything from upvoting and swiping left/right to reporting violating content. This creates the problem of aggregating preferences, and negotiating between competing desires of individuals, groups, and stakeholders. The framework of *social choice*, originally developed in economics (Gaertner, 2009; Sen, 2017), provides a foundational tool for addressing tradeoffs at all of these levels.

Abstractly, this framework assumes that each stakeholder has a *utility function* over a set of possible outcomes for them. This is motivated by the idea that someone could say which of several situations they’d prefer, that is, that each person has *preferences*. Note that this utility function can be purely “local” (a user may care only about whether they get good recommendations) or it can involve societal values (a user may care that other users also like what they like, or that vendors are treated fairly). A *social welfare function* is a voting process or a way of “adding up” or combining stakeholder preferences to produce a single number, a “societal utility” or social welfare. The aim is then to adopt a recommender policy which maximizes the expected social welfare.

This formulation can encompass virtually any criteria that express preferences over short or long-term, individual or group outcomes. There are relatively direct mathematical expressions for penalizing addictive behaviors, group-level diversity of consumed content, fairness across individuals, and so on. Conversely, any collaborative recommender system aggregates the explicit or implicit preferences of users in some mechanical way, as when signals like upvoting and watch time are combined across many users to decide how items are ranked for a different user. Social choice theory is a key bridge between the normative and the algorithmic, useful for analysis and design.

Although social choice theory is foundational, there are two major difficulties to applying it in practice. First, determining what an individual’s “preferences” actually are is quite challenging both in theory and in practice (Thorburn et al., 2022a). There is a long history of formal elicitation methods such as asking users to repeatedly say which of two options they prefer, or asking them to play various kinds of economic games (Camerer and Fehr, 2004; Chen and Pu, 2004), but formal preference elicitation can be involved and is quite demanding of the participants. Further, someone might be uninformed, coerced, addicted, have lowered expectations (Anderson, 2001) or want something that isn’t available (Freedman et al., 2020; Robertson and Salehi, 2020). In addition, attempts to elicit preferences can lead to strategic behavior where people misrepresent their beliefs to try to induce more favorable recommender outcomes, e.g. a content creator usually has an incentive to argue that their content is relevant to as many users as possible (Ben-Porat et al., 2019; Ben-Porat and Tennenholtz, 2018). While many different kinds of

feedback can provide crucial signals of what people value, behavior cannot be naively interpreted as true preferences.

Second, there is no entirely bottom-up or value-neutral method of ethics. Simply specifying the outcomes over which stakeholder preferences are defined is inextricably tied to the values being considered (Bernheim, 2016). For example, there is the choice of what will be voted upon. There is also the question of whose preferences count in what situations, e.g. community administrators may have special voting privileges, and it may be important to encode “rights” that cannot be infringed by the preferences of others. Hence, social choice approaches cannot excuse system designers from making consequential normative choices (Baum, 2020). There is more to democracy than voting systems.

### 5.3.2 Fairness and Tradeoffs among Stakeholders

Many of the values in our list involve making trade offs between different stakeholders (such as users vs. content providers) or among members of a stakeholder group (such as different subgroups of content providers which compete for attention). Correspondingly, there are a wide variety of notions of “fairness,” which often (but not always) are framed as some sort of trade off. The extent to which these tradeoffs are inherent is an open question in the research literature, because there are cases where it is possible to improve performance for one user subgroup without decreasing performance for other users. There are multiple challenges in this area, including defining fairness, measuring it in practice, and designing algorithms for efficient recommendation. See Ekstrand et al. (2021) for an overview.

Several major categories of fairness have been proposed in the context of recommender systems, roughly corresponding to the interests of different stakeholders. “C fairness” considers how well individual information consumers are served, “P fairness” is concerned with the distribution of attention between items or providers of content, while “CP fairness” considers both simultaneously, as in a rental property recommender designed to protect the rights of both minority renters and minority landlords (Burke, 2017; Wang and Joachims, 2021).

Recommender systems are mostly evaluated based on average performance across all users, but different user subgroups, such as age or gender groups, might be served with differing performance or error rates. Subgroup performance disparities can happen for a variety of reasons, including differences in group size or activity that affect the amount of training data available (Ekstrand et al., 2018; Li et al., 2021). There is a large body of work on mitigating group-level unfairness in classifier models, some of which can be adapted to recommender systems. For example, (Beutel et al., 2019a, 2019b) use pairwise comparisons of the ranking of different items to generalize the well-known “equality of opportunity” and “equality of odds” measures, showing that it is possible to equalize prediction error rates between user groups on a large commercial platform. However, algorithmic approaches that aim to equalize effectiveness disparities between user groups may make inappropriate tradeoffs: increasing recommendation utility for one user does not necessarily require decreasing it for other users, so it is not clear that allowing a solution that may decrease utility for well-served users is appropriate, as opposed to other approaches such as feature prioritization in the engineering process (Ekstrand et al., 2021 §5.4).

Items and their producers, on the other hand, necessarily compete for user attention. This leads to the concept of “exposure fairness” which may be formulated in a variety of ways (Singh and Joachims,

2018). Even if predicted “relevance” scores correctly measure the value of an item to an individual user, slightly less relevant items may get disproportionately less attention simply because they appear farther down, an effect known as “position bias.” Several algorithms have been proposed to ensure item attention is proportional to item value on average, either across a group of items or across multiple slates of recommendations (Biega et al., 2018; Diaz et al., 2020; Wang and Joachims, 2021). Such algorithms can be considered a correction to a type of error or “technical bias” in the fair ranking typology of Zehlike *et al.* (2021). More normative definitions of item fairness are often desirable. For example, Spotify strives to give exposure to less popular artists to counteract the “superstar economics” of cultural production (Mehrotra et al., 2018), while a “demographic parity” conception of fairness may be appropriate when qualified candidates from different groups (say, men and women) should be shown to prospective employers at the same rate (Geyik et al., 2019). A wide variety of fairness metrics concerning the exposure of items, groups of items, or producers have been proposed, though many of these are closely related (Kuhlman et al., 2021; Raj and Ekstrand, 2022; Sapiezynski et al., 2019; Zehlike et al., 2022).

Where it is possible to produce reasonable quantitative estimates of utility to different stakeholders, *multi-objective optimization (MOO)* can be used to balance multiple conflicting stakeholder utility and fairness objectives. One approach is to ensure that the recommender is Pareto efficient, meaning that there should be no way to modify a slate of recommendations to make it more fair without reducing utility, or to increase utility without reducing fairness (H. Wu et al., 2021). Mladenov et al. (2020) go beyond this by proposing a recommendation method that maximizes user social welfare (total user utility) by allowing small sacrifices in utility from well-served users to drive large gains for less well-served users. Reinforcement learning has also been applied to multi-sided fairness, through contextual multi-armed bandits which simultaneously optimize stakeholder utility and fairness objectives (Mehrotra et al., 2020, 2018; Wu et al., 2022). Recently, several researchers have taken a game-theoretic approach to the study of recommender systems. Ben-Porat and Tennenholtz (2018) and Ben-Porat et al. (2019) develop approaches that account for the strategic behavior of content providers while aiming to maximize user engagement. While all these methods hold promise for value alignment in complex ecosystems, they have not seen practical deployment to date.

### 5.3.3 Optimizing Tradeoffs

Because there is no purely “bottom up” way of making tradeoffs, system designers must ultimately choose some set of overall objectives or “ground truth” signals to serve as overall measures of value. Increasingly, AI tools are used to help make tradeoffs over the complex design space of recommender parameters, particularly the relative “weighting” of the signals that feed into item ranking functions.

Milli et al. (2020) determined the relative value of different user actions including viewing a Tweet, sharing it, liking it, etc. by connecting these interactions to the sparse use of the “see less often” control in a causal Bayesian network. This network represents dependencies such as the fact that a user has to view a Tweet before they can share it. By taking “see less often” as a ground truth signal of negative value it was possible to infer the value of all of the other, more common interactions. This idea generalizes to more complex methods for finding multiple weights that optimize multiple metrics.

Bayesian optimization (Golovin et al., 2017; Rasmussen and Williams, 2006) can be used to find the weights of a ranking function that maximize relevant (perhaps long-term) metrics, and automatically run

data-gathering experiments to improve those predictions (Gupta and Ouyang, 2020). This approach requires that the designer be able to assess the overall “utility” of any vector of performance metrics, which itself induces various tradeoffs. For example, is a product that has a greater number of items viewed but less total time spent and lower reported satisfaction better than the opposite? The tools of interactive optimization and utility elicitation (Boutilier et al., 2003; Zhao et al., 2013) could play an important role here, though these approaches have not yet found widespread use in practical recommender design.

## 6. Policy Approaches

In the previous section we have discussed design approaches to recommender systems and human values, that is, potential product changes. This section considers policy-making, which can be a powerful lever for change. Policy-making is informed by all of the perspectives articulated above, including ethical, procedural, measurement, and technical issues. Chosen policies impose constraints on how to build a recommender system and introduce additional technical challenges.

By policy-making we mean external governance, from government, regulators, or external bodies with appropriate authority. All large platforms have internal policies as well, particularly around which types of items are eligible for recommendation, but we focus on external governance as an important interface between recommender system operators and the rest of society. Because recommender systems are used in so many different types of products, we do not offer specific policy recommendations. Instead, our goal is to discuss the major categories of policies that have been proposed, and especially to understand how these policies could be translated into the terms of metrics and algorithms. At the current time there is a large gap in terminology and understanding between the recommender technology and policy communities, which we seek to highlight and begin to address. There are also policy-relevant technology gaps: the capability to do what policy-makers ask may not yet exist, as in the “right to explanation” provisions of the GDPR (Wachter et al., 2017).

We consider policy approaches that are relevant to recommender systems specifically, as opposed to social media or online platforms generally (neither of which necessarily involve recommender systems). We do not discuss content moderation policy approaches here, but direct readers to reviews such as (Douek, 2021; Keller, 2018; York and Zuckerman, 2019).

### 6.1 Risk Management vs. Value Sensitive Design

One proposed policy approach would require recommender system operators to evaluate the potential risk of harm from operating their systems. This is the approach taken by the proposed EU Digital Services Act, which requires “very large online platforms” (currently defined as those with more than 45 million users in the EU) to perform yearly assessments for three kinds of risks: the dissemination of illegal content, negative effects on fundamental rights, and “manipulation” with effects on “public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security.” (European Commission, 2020) Any harms found must be mitigated through various means including

“adapting content moderation or recommender systems.” The proposed Digital Services Oversight and Safety Act of 2022 (DSOSA) in the U.S. takes a similar approach, requiring platforms of certain sizes and scope to conduct assessments and mitigate any risks, including by “adapting the content moderation or recommender systems (including policies and enforcement) of the provider.

This touches on several of the values in our table, but is relatively narrow in two ways. First, a risk-based framework is concerned only with potential harms. Second, the type of risk mitigation envisioned by the DSA and the DSOSA generally happens after a system is already built. Another approach is to require consideration of important values during the design phase, as with “privacy by design” provisions (Drev and Delak, 2021). Extending this to more general values, one German law requires platforms to meet certain content diversity obligations (Helberger et al., 2019).

The challenge for policy-makers or regulators is to be both precise and general about how harms are to be assessed and values are to be enacted in recommender systems. In principle this could involve monitoring certain metrics, as is already done in environmental regulation and media monopoly policy. Such regulation would face all of the challenges of choosing metrics discussed above, and certainly no one set of metrics will be appropriate for all recommender systems. Even if useful metrics for harm or good can be found, there is the difficult question of what constitutes an “acceptable” value (Nechushtai and Lewis, 2019).

## 6.2 Accountability

Policy approaches to the issue of *accountability* include provisions around transparency and evaluation mechanisms such as audits. Both are instrumental in supporting other values such as agency, control, and accountability.

Transparency has been a major focal point of legislative efforts in the United States and the European Union. Over the past two years, lawmakers in the United States have introduced numerous bills that seek to compel internet platforms to provide more transparency around how they develop and deploy algorithmic systems for content curation purposes (Mozilla, 2021; Singh, 2020a). The key challenge from a policy perspective is in defining the goal of transparency efforts, who they are meant to serve, and what, exactly, should be disclosed. Transparency may be intended to serve users, lawmakers, researchers, journalists, etc. (van Drunen et al., 2019). There are also privacy, security, and intellectual property issues that complicate disclosure (OECD, 2019, chap. 4; Shapiro et al., 2021; Singh and Doty, 2021a). Suggestions for what to publish have included the recommender source code, data on users (e.g. demographic data and their interactions with the system), options for users to modify recommendation “parameters”, data used for training models, key metrics, and the rationale behind product features and ranking changes (Stray, 2021b; van Drunen et al., 2019; Vorm and Miller, 2018). Each of these has limitations.

Production recommender code is extremely large, difficult for outsiders and non-technical individuals to understand, and may not be particularly revealing without reference to the content and user data it operates on. For example, user interactions over time can be used to understand people’s trajectories through the system, as discussed above. However, user data is difficult to share because of privacy

concerns. This leads to the idea of sharing aggregated data, and many platforms already do so in the context of content moderation and targeted advertising (Singh, 2020b; Singh and Doty, 2021b). Then the key policy question becomes which metrics should be disclosed and how they are defined. While recommender operators will have key insights into what is relevant to measure, relying on them to select what is shared may pose a conflict of interest. Note that even aggregated user data can be used to re-identify individuals so it may need to be protected with techniques such as differential privacy (Dwork and Roth, 2014; Narayanan et al., 2016).

A related suggestion is that recommender operators should share the policies used to guide ranking and recommendation efforts, including what types of low quality or harmful content a platform downranks. The policy community has pushed recommender operators to disclose the rationale behind ongoing product changes more generally, including changes to recommender algorithms and parameters (Singh, 2020c; Stray, 2021b). This leads to the idea of a “change log” or “proceedings” that details what the operators were trying to do with each change (e.g., increase news quality), what data they had in front of them (e.g., fraction of items from each news source rated false by fact checkers), and what change they made (e.g., downrank certain sources by a certain amount). This is especially important as metrics or algorithms alone will not tell the full story: the motivation and context of a change are relevant to values. Thus far, a handful of civil society organizations have pushed platforms to adopt change logs for recommender system-related policies (Singh, 2020c).

Algorithmic auditing of AI/ML based systems has recently gained increased attention, and has mostly focussed on fairness and discrimination concerns in decision-making systems or predictive models (Vecchione et al., 2021). Although fairness remains a concern in recommender systems, in principle recommenders might be audited for any of the values discussed in this paper. A recent review of 62 academic algorithmic audits identified eight audits of recommender systems (Bandy, 2021), of which seven were concerned with “distortion” effects such as echo chambers or lack of source diversity. Several of these audits looked for, but did not find, echo chamber or filter bubble effects on Google News, the Facebook News Feed, and Apple News (Bandy, 2021, p. 16). They did find that a small number of news sources make up a large percentage of the results in Google News and Apple News.

While companies can and do perform internal algorithm audits of various types (Chowdhury, 2021; Raji et al., 2020) regulation could require external audits of recommender systems to mitigate these concerns (Singh and Doty, 2021a). Regulation could also direct audits to evaluate specific biases or harms to users or consumers (Brundage et al., 2020). The draft Digital Services Act requires certain platforms to undertake yearly audits to ensure they are meeting their risk assessment and mitigation obligations (European Commission, 2020). Similarly the Algorithmic Fairness Act in the United States would require covered entities to conduct and retain a five-year audit trail which includes information on how an algorithmic system was developed, trained, and tested (Singh, 2020a).

The personalized nature of recommender systems complicates auditing. Consider the problem of auditing user trajectories through recommender systems. Ribeiro et al. (2020b) collected data on non-personalized recommendations across YouTube channels and then evaluated trajectories using random walks, while the Wall Street Journal used 100 TikTok bots pre-configured to watch only videos on particular topics (Wall Street Journal, 2021). Neither of these methods model real user behavior. Conversely, Hosseinmardi et al.

(2020) used panel web browsing data collected by Nielsen to evaluate consumption of YouTube videos across the political spectrum, and The Markup's Citizen Browser project asks users to install a browser extension that reports what they see on Facebook and YouTube (The Markup, 2020). These observational studies have the advantage of involving real users, but the methods used so far cannot produce robust causal inferences about recommender effects. For example, it is currently not clear if social media contributes to depression or if depressed people spend more time on social media, or both. If these methodological issues cannot be solved, even extensive platform data sharing may not be sufficient to answer questions of interest. In that case, on-platform experiments may be the only reliable approach to auditing the effects of recommender systems, which would require extensive collaboration between recommender system operators and external researchers.

The nature of platform access remains to be defined. Industry players assert they face constraints when participating in third party audits, including concerns around privacy, security, intellectual property, competitiveness, and cost (OECD, 2019, chap. 4; Shapiro et al., 2021; Singh and Doty, 2021a). Many of these concerns are also risks to a third-party auditor, who must protect shared personal data yet typically do not possess the security resources of a platform. Additionally, many third-party auditing entities do not have the necessary technical skill and resources to audit recommender systems at scale (Singh and Doty, 2021a).

### 6.3 Translating Between Policy and Technology

There has long been miscommunication between the builders and regulators of technology. At the present moment, many countries are drafting or passing laws that regulate recommender systems of various kinds, especially social media, news recommenders, or targeted advertising. Unfortunately, much of the policy discussion taking place uses terms that do not map easily to recommender technical affordances (Singh, 2020a).

For example, Article 29 of the proposed Digital Services Act requires recommender-based platforms to disclose “the main parameters used in their recommender systems” as well as any available controls to “modify or influence those main parameters.” (European Commission, 2020) Unfortunately, it's not clear what a “main parameter” is (Helberger et al., 2021). While this probably doesn't refer to the millions or billions of learned neural network weights, real recommender systems involve hundreds of major interacting components which are continually configured and tuned in complex ways. The vast majority of such internally configurable settings will not be understandable or useful to users, auditors or regulators. A better formulation might arise from considering the design of recommender controls, as discussed above.

The same provision also stipulates that recommenders offer “at least one option which is not based on profiling.” Profiling in this context is defined in the GDPR and includes “personal preferences” and “interests” (European Commission, 2016). Thus, exclusion of profiling in this sense may exclude even Twitter's classic reverse chronological timeline, which cannot operate without the user indicating who to follow. An alternative approach would be to require an option that personalizes based on explicit user controls only, as opposed to implicit signals such as clicks or watch time.

“Amplification” is another word which appears in several proposed laws yet is difficult to translate into operational terms. Modern platforms can provide huge distribution to an item in a short period of time via information cascades, which result from a combination of user sharing and algorithmic recommendation (Bartal et al., 2020). While “amplification” is a reasonable name for this phenomenon, many definitions of “amplified” collapse to “shown” when parsed carefully, e.g. “the promotion of certain types of extreme content at the expense of more moderate viewpoints” (Whittaker et al., 2021). The selection of any type of content for display reduces the promotion of all other types because of the zero-sum nature of selection, so amplification would mean any display of “extreme” content when other content was available.

Other definitions are more specific, and compare the distribution of an item to some (often implicit) baseline. One approach is to define amplification as the prevalence of some type of content in user feeds relative to the prevalence of that type among all available content. This definition can be useful in some contexts, though it is unclear why raw prevalence should be a presumptively neutral baseline. Indeed, this formulation leads to the perverse outcome that spamming content may reduce amplification (through an increase in the denominator) even though it may increase distribution. For some systems a chronological baseline may make sense. This is plausible for Twitter because it was designed around a reverse-chronological feed, and several studies compare the algorithmic and chronological options on Twitter (Bandy and Diakopoulos, 2021; Huszár et al., 2021). However, there is no similarly natural baseline for systems such as YouTube, Google News, Netflix, Spotify, or Amazon where a chronological feed makes little sense. Further, purely chronological feeds can suffer from problems that make them unattractive baselines (Kantrowitz, 2021; Keller, 2021) including recency bias and the prevalence of low quality content like spam. Because of these conceptual and practical issues, “amplification” is not a well-defined measure for most recommender systems (Thorburn et al., 2022b). In the U.S., legislation targeting amplification in any of these senses is also likely to face 1st Amendment challenges (Keller, 2021).

Examples like these highlight the importance of finding new ways of bridging the knowledge divide between policy makers, specialized expertise and independent research. Possible ways forward include educational programs for policy makers and engineers alike, the embedding of technology expertise (such as TechCongress which recruits technologists and embeds them in Capitol Hill offices to inform tech-focused legislation), regulatory sandboxes which allow controlled experimentation, a strong commitment to evidence-based policy making, and other initiatives to make technical expertise more easily accessible.

## 7. Open Problems

Based on this review of values in recommender systems from a variety of perspectives, we propose the following list of open problems. Each of these is both consequential and challenging, and would benefit greatly from future work.

**Process for defining values and resolving trade-offs.** One of the overarching challenges in this area is that there is no widely-agreed process for deciding which values are prioritized and tracked, how they are measured, and how trade-offs are adjudicated. There are trade-offs or conflicts both between multiple

people or stakeholders, and between different values. For example, while it is clear that users should have substantial control over the recommendations they receive, it is less clear what to do when this appears to harm oneself or others. The question of what is measured, and on which subgroups, has similarly profound implications for outcomes. In an operational sense, “values” can only come from the normative debates that take place in society (Gabriel, 2020). Methods such as multi-stakeholder metric construction (Lee et al., 2019; Stray, 2020) and user juries (Fan and Zhang, 2020) may provide a way forward, but are not well developed.

**Better measurements.** Many values are not easy to operationalize. Collections of AI-relevant metrics such as the IEEE 7010 standard (Schiff et al., 2020) provide useful compendiums, but measures developed in social science and policy may not apply directly to particular recommender contexts such as news recommendation or targeted advertising, and will need to be refined.

**Controls that people want to use.** The controls offered to users so far have been remarkably sparse and slow to develop, despite the obvious practical, policy, and ethical advantages of better control. In part this may be a problem of under-investment, but there is also a fundamental unsolved problem with recommender controls: most are never used by more than a few percent of users (Jin et al., 2017). Better control designs might improve this situation, such as immediate feedback on how changing a setting changes which items are recommended (Schnabel et al., 2020). Still, it is likely that most users will never adjust default settings, so a better strategy may be to attempt to design controls that elicit user feedback in the normal course of use, e.g. voting systems or a “respect” button (Stroud et al., 2017). It will be important to differentiate between giving users a feeling of agency and giving them effective controls (Loecherbach et al., 2021; Vaccaro et al., 2018) as both are necessary.

**Non-behavioral feedback.** At present essentially all feedback signals to recommender systems are behavioral, e.g. engagement. There are a variety of well-known problems with inferring preferences from behavior, including self-control or addiction issues, uninformed users, and the contextual meaning of choices (Anderson, 2001; Thorburn et al., 2022a). In particular, many of the values discussed above involve outcomes that are unlikely to be identifiable from on-platform behavior alone, e.g. well-being. The most promising solution is simply to ask users about their experience by repurposing survey instruments that have been developed in the social science and policy communities. Only a small fraction of users can be surveyed, but the resulting data stream can be used to build and continuously validate models that predict survey responses, which can then be used as algorithmic objectives. This is already done in industry (Goodrow, 2021; Stray, 2020) but there is essentially no public research on this emerging technique.

**Long-term outcomes.** Most recommender algorithms today are myopic in the sense of optimizing only for immediate responses. Longer-term outcomes are typically managed by product teams who monitor richer feedback and make algorithmic adjustments. While today these teams are typically optimizing for purchases, subscriptions or user retention, recommender systems could also be managed on values-relevant outcomes. However, human management will always optimize against aggregated outcomes, while algorithmic optimization can be personalized and therefore has the potential to better serve subgroups and individuals. Emerging reinforcement learning methods may make this possible

(Afsar et al., 2021; Mladenov et al., 2019) but this may also require cheap and accurate individual-level proxy metrics for the outcomes of interest.

**Causal inference of human outcomes.** Recommenders may have significant effects on people and society, yet determining what those effects are remains extremely challenging due to sampling and confounding effects. Individual case studies can be instructive but are difficult to generalize to platform scale, while even large observational studies struggle to answer questions like “does social media cause depression or do depressed people use social media more?” In many cases it will not be possible to say what a platform’s effect on an individual has been because there is no counterfactual, but it may be possible to measure group-level effects through on-platform experiments. The long term effects of design decisions – and the resulting outcomes for users interacting with the system – are particularly difficult to study because so many other things are happening in a user’s life.

**Industry-academic research collaborations.** While a variety of external algorithmic auditing methods have been developed (Sandvig et al., 2014) many of the most important questions can only be answered in an ecologically valid setting using private platform data or on-platform experiments (Griffioen et al., 2020). Unfortunately, it is not easy for external researchers and platforms to work together due to concerns around access, privacy, security, research integrity, intellectual property, competitiveness, and cost (OECD, 2019, chap. 4; Shapiro et al., 2021; Singh, 2020a). External collaborations would benefit from the development of technical methods to enhance the privacy and security of shared data, and from legal or policy approaches that set the terms of engagement so as to protect all relevant interests.

**Interdisciplinary Policy-making.** There is a substantial gap between the policy community and the technical community (including between scholars of both). Not only is there an enormous amount of specialized knowledge unknown to each side, but they use very different terms to understand and describe the problems of recommenders. There are probably also fundamental differences in values and politics between disciplines that will have to be resolved, which is complicated by the fact that recommender policy touches fundamental rights such as privacy and freedom of expression. It is clear that effective policy-making must be collaborative and interdisciplinary, even if it is not yet clear how to achieve this.

## 8. Conclusion

Recommender systems are a profound technology that will continue to touch many aspects of individual and social life. At their best, they serve important interests of multiple parties. Consumers might want accurate information, good music, or useful products, producers might depend on recommenders to help them find their audience or customers, and platforms need to capture some of this value to continue operating. Yet recommenders can also cause a wide variety of harmful effects, or they may miss opportunities to do good. Building human values into recommender systems raises complex and consequential challenges, including philosophical, regulatory, political, sociological, and technical issues.

This paper contributes to this interdisciplinary conversation and guides further research in several ways. We have identified some values that are relevant to recommender systems, and discussed the main issues surrounding each. We have described current industrial practice including a sketch of modern

recommender design, and an illustrative process for shifting a value in a production system. We discussed the challenges of measuring adherence to a value, including the difficulty of “operationalizing” or translating a value into metrics, and the types of data sources that might supply useful information on values. We feel it is too early to attempt a general theory of the value-sensitive design of recommenders, because the field is still emerging. Instead, we articulated an extensive menu of design techniques that have been applied or could likely be applied to production systems. Finally, we surveyed developing approaches to recommender regulation, identifying a substantial gap in knowledge and orientation between the technology and policy communities.

While the intersection of content personalization with individual and societal flourishing is a huge and varied area, we hope that this synthesis provides a shared language, useful starting points, and essential research directions for building human values into recommender systems.

## Appendix A: Table of Values

This list was culled from a wide variety of sources and refined through multi-stakeholder consultation as described above. It includes perspectives from various traditions and cultures, but is not intended to be a comprehensive list of the values that might be important to consider in recommender systems. Nor do we attempt to prioritize or rank these values here. Rather, this table should be viewed as one list of broadly important themes.

For each value, we provide possible interpretations in the context of recommender systems, some indicators that might be used to assess whether a system enacts or supports that value, and example designs that are relevant to that value.

	Value	Interpretation in the context of recommenders	Example Indicator	Example design
1	Usefulness	<p>“The purpose of recommenders is often summarized as ‘help the users find relevant items’” (Jannach and Adomavicius, 2016)</p> <p>A recommender should provide a useful service to the user.</p>	<p>Long term user retention.</p> <p>User satisfaction surveys.</p>	<p>Show people more of what they rate highly.</p>
2	Liberty	<p>A platform should respect human dignity and protect human rights and freedoms (Fjeld et al., 2020, p. 61); recommenders should operationalize respect for autonomy (Varshney, 2020).</p> <p>A platform should not stop users from exploring certain types of information. Users should be able to pursue their own good in their own way.</p>	<p>Similar mechanisms as Freedom of Expression.</p>	<p>Protections against arbitrary removal of content.</p> <p>Eliminating threats of violence designed to compel individuals to form particular opinions, in violation of article 19 in Universal Declaration of Human Rights.</p>
3	Freedom of Expression	<p>“Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds” (United Nations, 1966 art .19)</p> <p>“The exercise of freedom of opinion, expression and information ... is a vital factor in the strengthening of peace and international understanding” (UNESCO, 1978).</p> <p>Platforms should not stop users from expressing their thoughts and opinions, or unduly suppress distribution of user posts.</p>	<p>“Conducting an ex-ante evaluation attempts to predict the future relationship between [human] rights and an on-going or proposed business activity.” (Latonero and Agarwal, 2021)</p>	<p>Transparency around content which is suppressed.</p> <p>Appeals processes for removed content.</p>

4	Control	<p>Control “includes the system’s ability to allow users to revise their preferences, to customize received recommendations, and to request a new set of recommendations” (Pu and Chen, 2010, p. 18)</p> <p>“Ability to determine the nature, sequence and/or consequences of technical and operational settings, behavior, specific events, and/or experiences.” (IEEE, 2021, p. 17)</p> <p>The platform should give users ways to control the content selected for them and it should give users ways to control how the content they create is shared.</p>	<p>“I feel in control of my news feed.” (Vaccaro et al., 2018)</p> <p>Locus of control Scale. Ex: “Other people usually control my life” (Sapp and Harrod, 1993)</p>	<p>Recommender controls (Harambam et al., 2019; He et al., 2016; Schnabel et al., 2020)</p>
5	Agency, Autonomy, Efficacy	<p>“The sense of agency can be analyzed as a compound of more basic experiences, including the experience of intentional causation, the sense of initiation and the sense of control” (Pacherie, 2007)</p> <p>The platform should provide the capacity for intentional action and help users achieve their goals (Varshney, 2020)</p> <p>The platform should not manipulate users for the benefit of other stakeholders.</p>	<p>General Self-Efficacy Scale (GSE) Ex: “It is easy for me to stick to my aims and accomplish my goals” (Chen et al., 2001)</p> <p>Autonomy Scale (AS): Measures components of autonomy including family loyalty autonomy, value autonomy, emotional autonomy, and behavioral autonomy (Anderson et al., 1994)</p>	<p>Recommender controls (Harambam et al., 2019; He et al., 2016; Schnabel et al., 2020)</p>
6	Privacy	<p>“The protection of personal data is of fundamental importance to a person’s enjoyment of his or her right to respect for private and family life” (Council of Europe, 2020)</p> <p>Users should be allowed to determine if and how their personal data is collected, processed and disseminated (Fjeld et al., 2020, p. 21; IEEE, 2021, p. 71)</p>	<p>GDPR Data Protection Impact Assessment (European Commission, 2016)</p>	<p>Controls to limit data use.</p> <p>Federated recommendation algorithms (Qi et al., 2020; C. Wu et al., 2021).</p>
7	Safety, Security	<p>The principle of ‘safety’ requires that an AI system be reliable and that ‘the system will do what it is supposed to do without harming living beings or [its] environment.’” (Fjeld et al., 2020, p. 38)</p> <p>“[System] use should not contribute to increasing stress, anxiety, or a sense of being harassed by one’s digital environment.” (Université de Montréal, 2018)</p>	<p>Sense that most people can be trusted (World Values Survey, 2020)</p>	<p>Hate speech moderation</p> <p>Parental controls</p> <p>Option to report harmful content.</p>
8	Self-expression, authenticity	<p>“Self-expression is a notion that is closely associated with a horde of positive concepts, such as freedom, creativity, style, courage, self-assurance, and even healing and spirituality” (Kim and Ko, 2007)</p> <p>The platform should empower users to express their identity (including personality, attributes, behavior) and to decide how it is presented to others. (Kerlin, 2020, p. 19).</p>	<p>Correlates with (but is not identical to) quantity and quality of user content creation</p>	<p>Easy-to-use, attractive content creation/modification tools, e.g. SnapChat filters, TikTok tools.</p> <p>Content moderation policies that allow self-expression (e.g. don’t take down art because it has nudity)</p>

9	Well-being	<p>Well-being is a complex and multidimensional construct that encompasses many of the values we describe here and there is no universally accepted definition (Dodge et al., 2012).</p> <p>Well-being may be measured by both objective and subjective factors including satisfaction with life, emotions and affect, psychological well-being, social relationships, and meaning (Linton et al., 2016)</p> <p>On the platforms, users should see content that leads them to experience contentment, joy and pleasure, both ephemerally and over the long term. Platforms should help users feel satisfied with their lives.</p>	<p>Satisfaction with Life Scale (SWLS). Ex: “The conditions of my life are excellent” (Diener et al., 1985)</p> <p>Positive Affect and Negative Affect Scales (PANAS). Ex: “Indicate the extent to which you have felt interested, distressed, excited, upset, etc over the past week” (Watson et al., 1988)</p> <p>Scale of Psychological Wellbeing (SPWB). Assesses psychological functioning across domains such as autonomy, self-acceptance, purpose in life, and positive relationships (Ryff and Keyes, 1995)</p>	<p>Encourage active vs. passive use of social media (Verduyn et al., 2017)</p> <p>Usage controls, e.g. screen time limits</p>
10	Inspiration, Awe	<p>“In times of uncertainty, others are sought for guidance, inspiration or motivation, to seek ideas, goals or possibilities, which can influence ambitions, choices, and achievements.” (Kerlin, 2020, p. 11)</p> <p>Platforms should show users content that will inspire, motivate, or guide them.</p>	<p>Inspiration scale: Measures the experience of being inspired, feeling inspired and motivation to do something with that inspiration (Thrash and Elliot, 2003)</p>	<p>Uprank content identified by Inspiration classifier (Ignat et al., 2021)</p>
11	Mental Health	<p>Platforms should help users protect and improve their mental health. Platforms should not encourage unhealthy types or amounts of use.</p> <p>“[A] state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community” (World Health Organization, 2004)</p>	<p>Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) (Tennant et al., 2007)</p> <p>Positive Mental Health Scale (PMH) Ex: “I feel that I am actually well equipped to deal with life and it’s difficulties” (Lukat et al., 2016)</p>	<p>Features that help users manage their screen time, e.g. “take a break” notifications, screen time limits.</p> <p>Option to report harassment or bullying.</p>
12	Physical Health	<p>Platforms should help users protect and improve their physical health, such as by providing accurate health information and encouraging behaviors that contribute to physical health.</p>	<p>WHO Wellbeing Index (WHO-5) Ex: “I have felt active and vigorous” and “I woke up feeling fresh and rested” (Topp et al., 2015)</p>	<p>Help users develop healthy habits (exercise, diet, etc.)</p> <p>Clear platform policies on removing health misinformation, ads for fake cures, etc.</p>
13	Self-actualization, Personal Growth	<p>Platforms should enable users to reach their full potential.</p>	<p>“Feeling that the things one does are worthwhile” (UK Office of National</p>	<p>Educational recommender systems (Dascalu et al., 2016)</p>

		<p>“Personal growth is a continuous improvement in life in order to find purpose and meaning” (Kerlin, 2020, p. 22)</p> <p>“[R]ather than have people choose the easiest option, we wish to have them develop a strong sense of determination of having selected the right path” (Knijnenburg et al., 2016)</p>	<p>Statistics, 2019)</p> <p>European Social Survey (ESS). Ex: “To what extent [do you] learn new things in life?” (Norwegian Centre for Research Data, 2012)</p>	
14	Recognition, Acknowledgment	<p>Platforms should provide ways for other people to recognize a user’s contributions or worth.</p> <p>“Whilst some esteem needs can be met by having self-acceptance, self-worth and self-value, validation from others is also important” (Kerlin, 2020, p. 28)</p>	<p>Social Inclusion Scale (SIS). Ex: ‘I have felt accepted by my friends’ (Wilson and Secker, 2015)</p>	<p>“Celebrate” reaction button (as LinkedIn has)</p>
15	Knowledge, Informativeness	<p>Users should see items that keep them informed about topics they care about or might care about.</p> <p>“[P]ersonalised news recommendations allow the media not only to help users find relevant information, but also to inform them better and more effectively.” (Helberger, 2019, p. 995)</p> <p>“Curiosity drives individuals to seek stimulation, information or new experiences, serving a purpose to increase knowledge and build skills.” (Kerlin, 2020, p. 18)</p>	<p>News knowledge quizzes (Allcott et al., 2020; Angelucci and Prat, 2021)</p> <p>Curiosity and exploration inventory (CEI). Ex: “I would describe myself as someone who actively seeks as much information as I can in a new situation” (Kashdan et al., 2004)</p>	<p>News recommender systems (Karimi et al., 2018)</p> <p>Trending topics</p>
16	Connection	<p>“Individuals are driven to interact and seek social closeness ... The benefits of social connections are far-reaching.” (Kerlin, 2020, p. 17)</p>	<p>Perceived Social Support scale (PSS). Ex: “I have friends with whom I can share my joys and sorrows” and “I can talk about my problems with my friends” (Zimet et al., 1988)</p> <p>Online Social Support Scale (OSSS). Ex: “Online, I belong to groups of people with similar interests” and “Online, people make me feel like I belong” (Nick et al., 2018)</p>	<p>Recommendations for people and groups with similar interests (Carullo et al., 2014)</p>
17	Civic Engagement	<p>“Working to make a difference in the civic life of one's community and developing the combination of knowledge, skills, values and motivation to make that difference.” (Ehrlich, 2000, p. vi)</p> <p>“Creating a public forum and optimal conditions for engagement” (Helberger, 2019, p. 1005)</p>	<p>Voter turnout (OECD, 2020)</p> <p>Attendance of peaceful demonstrations (World Values Survey, 2020)</p> <p>Approximate total hours a month active in voluntary organizations (World Values Survey, 2020)</p>	<p>Encourage people to vote (Bond et al., 2012)</p>

			Donations to a charity in a month (National Opinion Research Center, 2013)	
18	Community, Belonging	<p>“[T]he feeling that people matter to one another and to the group, and a shared faith that members’ needs will be met through their commitment to be together.” (McMillan and Chavis, 1986)</p> <p>“The concept of members sharing mutual concern and/or love for one another” (Kerlin, 2020, p. 15) .</p>	<p>Sense of belonging to a neighborhood (UK Office of National Statistics, 2019)</p> <p>The General Sense of Belongingness Scale (Malone et al., 2012). Ex: “I have a place at the table with others”</p>	<p>Recommend community-oriented groups</p> <p>User affordances for donating to charitable causes.</p>
19	Accessibility, Inclusiveness	Inclusiveness in design means that diverse people and perspectives should be involved in the design of AI systems. Inclusiveness in impact calls for a just distribution of AI benefits and harms.(Fjeld et al., 2020, pp. 51–53)	Assessing the experience of different user demographics through surveys and other methods, such as visually impaired or hard of hearing users (Berke et al., 2018; Morris et al., 2016)	<p>Implement Web Content and Accessibility Guidelines (WCAG)</p> <p>Variety of formats for cognitive impairments (e.g. screen readers, transcripts, captions)</p>
20	Tolerance, Constructive Discourse	<p>Tolerance “creates the opportunity for a wide range of political groups to express their ideas and to participate in public life” (Sullivan and Transue, 1999)</p> <p>Polarization divides society into “us” and ”them” camps and contributes to the erosion of democracy (McCoy and Somer, 2019)</p>	Polarization measures such as issue-position or affective polarization (Boxell et al., 2017; Iyengar et al., 2018; Yarchi et al., 2021)	Increase ideological diversity, select for civil conversations, optimize polarization measures (Stray, 2021a)
21	Duty	The notion of duty and obligation is defined in contrast to self-interest (Hursthouse and Pettigrove, 2018).	Personal control and responsibility scale (Meca, 2012)	User affordances for donating to charitable causes.
22	Care, Compassion, Empathy	<p>“The ethic of care emphasizes the importance of context, interdependence, relationships, and responsibilities to concrete others.” (Koggel and Orme, 2010) and is “based on the recognition that self and others are interconnected” (Skoe, 2014).</p> <p>These notions also appear in Confucian and Buddhist traditions (Vallor, 2016, p. 132) and in Ubuntu values where “a person is ‘a person through others’” (Ndiweni and Sibanda, 2020).</p>	Empathy quotient (Baron-Cohen and Wheelwright, 2004). Ex: “I find it easy to put myself in somebody else’s shoes” and “Seeing people cry doesn’t really upset me “	Best practices in suicide prevention (Robinson et al., 2016)
23	Fairness, Equality, Equity	<p>“A morally justifiable distribution of benefits, goods, harms, and risks” (Vallor et al., 2020).</p> <p>In multi stakeholder fairness, different stakeholders have different objectives and needs</p>	Fairness metrics in classification and recommendation (Beutel et al., 2019a; Raj et al., 2020)	Add fairness measures to the recommender training or serving objectives (Beutel et al., 2019a; Mehrotra et al.,

		from the system (Abdollahpouri et al., 2020).		2018)
24	Diversity	<p>“[D]iversity refers to the idea that in a democratic society, informed citizens construct their worldview from a diverse set of sources which helps them to make balanced and well-considered decisions.” (Helberger et al., 2018, p. 192).</p> <p>“‘Cultural diversity’ refers to the manifold ways in which the cultures of groups and societies find expression.” (UNESCO, 2005 art. 4)</p>	<p>Slate and feed diversity metrics of various kinds (Kunaver and Požrl, 2017)</p> <p>Perceived diversity and user satisfaction (Kunaver and Požrl, 2017)</p>	Methods to increase various types of diversity, including topical and source diversity, novelty, coverage, etc. (Castells et al., 2015; Kunaver and Požrl, 2017)
25	Accountability	<p>Accountability is “a component of the state of being responsible, alongside being answerable and being attributable” (Mattingly-Jordan et al., 2019)</p> <p>AI principles documents include “verifiability and replicability,” “impact assessments,” “environmental responsibility,” “evaluation and auditing requirements.” (Fjeld et al., 2020, p. 28)</p>	Ranking Digital Rights Corporate Accountability Index (Ranking Digital Rights, 2020)	<p>Explanations of decisions according to clearly defined principles.</p> <p>Processes for users to bring up problems and resolve them in a timely manner.</p>
26	Transparency & Explainability	<p>Transparency means that information provided about a system is a) meaningful, b) useful, c) accessible, d) comprehensive, and e) truthful (IEEE, 2021, p. 71)</p> <p>“Transparency is instrumental to uphold intrinsic values of human autonomy and justice” (Canca, 2020, p. 20)</p> <p>Desirable properties of transparency metrics (Spagnuolo et al., 2016)</p> <p>“Explainability implies presenting technical information in a form understandable to, and meeting the needs of, a specific user category” (IEEE, 2021, p. 49)</p>	System cards which provide stakeholders with an overview of different components of an ML system (Adkins et al., 2022)	<p>Explanations for recommendations, answering questions such as “why am I seeing this?” or “Why has this been removed?”</p> <p>Publicly disclose and discuss changes to moderation or ranking (Stray, 2021b).</p>
27	Accuracy (Factuality)	<p>Accuracy is “an AI system’s ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models” (Weiser, 2019)</p> <p>AI systems should make “accurate information readily available” (Google, 2021) and should not spread “untrustworthy information.” (Université de Montréal, 2018)</p>	Credibility metrics including prevalence of misinformation items among recommendations, and user engagement on this material.	<p>Misinformation classifiers (Bozarth and Budak, 2020)</p> <p>Prompts to consider accuracy before sharing (Pennycook and Rand, 2021).</p>
28	Tradition, History	“Heritage is a concept to which most people would assign a positive value. The preservation of material culture ... and intangible culture ... are generally regarded as a shared common good by which everyone benefits.” (Silverman and Ruggles, 2007)	Percentage of locally produced content.	Promotion of cultural events (Chaopreecha, 2019)

		<p>Practices, symbols, ideas, and beliefs that are developed by groups and represent their shared experience and fate. They often take the form of religious rites, beliefs, and norms of behavior (Schwartz, 2012, p. 6).</p> <p>Rituals or conventional norms are one of the five central virtues of Confucianism (Csikszentmihalyi, 2020)</p>		
29	Environmental Sustainability	<p>“Respect for environment and natural habitat, efficiency, maintainability, operability, supportability, reliability, durability, resilience, forgiveness, robustness, redundancy, reusability, reconfigurability, simplicity, economy, renewability” (IEEE, 2021, p. 70)</p> <p>“Develop and scale up carefully assessed technologies, infrastructure and actions that reduce climate change and its associated risks.” (UNESCO, 2017)</p>	Product level carbon footprints (Meinrenken et al., 2012)	<p>Recommender systems can potentially direct users toward climate-friendly options (Rolnick et al., 2019, p. 25)</p> <p>Reporting on the energy used by ML models (Henderson et al., 2020)</p>
30	Progress	The International Covenant on Civil and Political rights includes the right of people to “freely pursue their economic, social and cultural development” (United Nations, 1966) as expanded in the Declaration on Social Progress and Development (United Nations, 1969).	UN Human Development Index (Stanton, 2007)	To the best of our knowledge, none
31	Labor	Meaningful work is both significant and positive in valence (Rosso et al., 2010). Job insecurity, low employability, and unemployment are all detrimental to well-being, beyond the effects of income loss (Green, 2011)	Work and Meaning Inventory (Steger et al., 2012)	Ensure users see recommendations for jobs that they would want.

## Acknowledgements

Thank you to Chloe Bakalar, Daphne Keller, Joelle Pineau, and Andrea Wong for helpful feedback.

## References

- Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., Pizzato, L., 2020. Multistakeholder recommendation: Survey and research directions 30, 127–158. <https://doi.org/10.1007/s11257-019-09256-1>
- Abdollahpouri, H., Burke, R., 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness.
- Adiwardana, D., Luong, M.-T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q.V., 2020. Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977 [cs, stat].
- Adkins, D., Alsallakh, B., Cheema, A., Kokhlikyan, N., McReynolds, E., Mishra, P., Procope, C., Sawruk, J., Wang, E., Zvyagin, P., 2022. Method Cards for Prescriptive Machine-Learning Transparency. Presented at the CAIN 2022 - 1st International Conference on AI Engineering.
- Afsar, M.M., Crump, T., Far, B., 2021. Reinforcement learning based recommender systems: A survey. arXiv:2101.06286 [cs].
- Allcott, H., Braghieri, L., Eichmeyer, S., Gentzkow, M., 2020. The welfare effects of social media. *American Economic Review* 110, 629–676. <https://doi.org/10.1257/aer.20190658>
- Allcott, H., Gentzkow, M., Song, L., 2021. Digital Addiction (No. 28936), NBER Working Papers. NBER.
- Al-Zaman, Md.S., 2021. Prevalence and source analysis of COVID-19 misinformation in 138 countries. *IFLA Journal* 03400352211041135. <https://doi.org/10.1177/03400352211041135>
- Anderson, E., 2001. Symposium on Amartya Sen’s philosophy: 2 Unstrapping the straitjacket of ‘preference’: A comment on Amartya Sen’s contributions to philosophy and economics. *Economics & Philosophy* 17, 21–38. <https://doi.org/10.1017/S0266267101000128>
- Anderson, R.A., Worthington, L., Anderson, W.T., Jennings, G., 1994. The development of an autonomy scale. *Contemp Fam Ther* 16, 329–345. <https://doi.org/10.1007/BF02196884>
- Andrus, M., Spitzer, E., Brown, J., Xiang, A., 2021. “What We Can’t Measure, We Can’t Understand”: Challenges to Demographic Data Procurement in the Pursuit of Fairness. arXiv:2011.02282 [cs].
- Angelucci, C., Prat, A., 2021. Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News (SSRN Scholarly Paper No. ID 3593002). Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.3593002>
- Aral, S., 2016. The Future of Weak Ties. *American Journal of Sociology* 121, 1931–1939. <https://doi.org/10.1086/686293>
- Arendt, H., 1972. *Crises of the Republic: Lying in Politics, Civil Disobedience on Violence, Thoughts on Politics, and Revolution*. Houghton Mifflin Harcourt.
- Arsht, A., Etcovitch, D., 2018. The Human Cost of Online Content Moderation. *Harvard Journal of Law & Technology*.
- Asimovic, N., Nagler, J., Bonneau, R., Tucker, J.A., 2021. Testing the effects of Facebook usage in an ethnically polarized setting. *Proc Natl Acad Sci USA* 118, e2022819118. <https://doi.org/10.1073/pnas.2022819118>
- Aust, F., Diedenhofen, B., Ullrich, S., Musch, J., 2013. Seriousness checks are useful to improve data validity in online research. *Behav Res* 45, 527–535. <https://doi.org/10.3758/s13428-012-0265-2>
- Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Mann, M., Lee, J., Volfovsky, A., Merhout, F., 2018. Exposure to opposing views on social media can increase political polarization 115, 9216–9221. <https://doi.org/10.1073/pnas.1804840115>
- Balog, K., Radlinski, F., Arakelyan, S., 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. ACM, New York, NY, USA, pp. 265–274. <https://doi.org/10.1145/3331184.3331211>
- Bandy, J., 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. arXiv:2102.04256 [cs].
- Bandy, J., Diakopoulos, N., 2021. More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proc. ACM Hum.-Comput. Interact.* 5, 1–28. <https://doi.org/10.1145/3449152>

- Baron-Cohen, S., Wheelwright, S., 2004. The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *J Autism Dev Disord* 34, 163–175. <https://doi.org/10.1023/B:JADD.0000022607.19833.00>
- Bartal, A., Pliskin, N., Tsur, O., 2020. Local/Global contagion of viral/non-viral information: Analysis of contagion spread in online social networks. *PLoS ONE* 15, e0230811. <https://doi.org/10.1371/journal.pone.0230811>
- Bauer, C., Novotny, A., 2017. A consolidated view of context for intelligent systems. *AIS* 9, 377–393. <https://doi.org/10.3233/AIS-170445>
- Baugut, P., Neumann, K., 2020. Online propaganda use during Islamist radicalization. *Information, Communication & Society* 23, 1570–1592. <https://doi.org/10.1080/1369118X.2019.1594333>
- Baum, S.D., 2020. Social choice ethics in artificial intelligence 35, 165–176. <https://doi.org/10.1007/s00146-017-0760-1>
- Beauchamp, T.L., Childress, J.F., 2019. *Principles of Biomedical Ethics*. Oxford University Press.
- Bekalu, M.A., McCloud, R.F., Viswanath, K., 2019. Association of Social Media Use With Social Well-Being, Positive Mental Health, and Self-Rated Health: Disentangling Routine Use From Emotional Connection to Use. *Health Educ Behav* 46, 69S–80S. <https://doi.org/10.1177/1090198119863768>
- Ben-Porat, O., Goren, G., Rosenberg, I., Tennenholtz, M., 2019. From Recommendation Systems to Facility Location Games, in: *Proceedings of the AAAI Conference on Artificial Intelligence*. Presented at the Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1772–1779. <https://doi.org/10.1609/aaai.v33i01.33011772>
- Ben-Porat, O., Tennenholtz, M., 2018. A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers, in: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Berke, L., Kafle, S., Huenerfauth, M., 2018. Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Presented at the CHI '18: CHI Conference on Human Factors in Computing Systems, ACM, Montreal QC Canada, pp. 1–12. <https://doi.org/10.1145/3173574.3173665>
- Bernheim, B.D., 2016. The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics 7, 12–68. <https://doi.org/10.1017/bca.2016.5>
- Bernstein, A., de Vreese, C., Helberger, N., Schulz, W., Zweig, K., Baden, C., Beam, M.A., Hauer, M.P., Heitz, L., Jürgens, P., Katzenbach, C., Kille, B., Klimkiewicz, B., Loosen, W., Moeller, J., Radanovic, G., Shani, G., Tintarev, N., Tolmeijer, S., van Atteveldt, W., Vrijenhoek, S., Zueger, T., 2020. Diversity in News Recommendations.
- Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C., 2019a. Fairness in recommendation ranking through pairwise comparisons 2212–2220. <https://doi.org/10.1145/3292500.3330745>
- Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., Chi, E.H., 2019b. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Presented at the AIES '19: AAAI/ACM Conference on AI, Ethics, and Society, ACM, Honolulu HI USA, pp. 453–459. <https://doi.org/10.1145/3306618.3314234>
- Beutel, A., Covington, P., Jain, S., Xu, C., Li, J., Gatto, V., Chi, E.H., 2018. Latent Cross: Making Use of Context in Recurrent Recommender Systems, in: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18*. Association for Computing Machinery, New York, NY, USA, pp. 46–54. <https://doi.org/10.1145/3159652.3159727>
- Biega, A.J., Gummadi, K.P., Weikum, G., 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. Presented at the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018. <https://doi.org/10.1145/3209978.3210063>
- Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization 489, 295–298. <https://doi.org/10.1038/nature11421>
- Borisov, A., Markov, I., de Rijke, M., Serdyukov, P., 2016. A context-aware time model for web search, in: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 205–214.
- Borning, A., Muller, M., 2012. Next Steps for Value Sensitive Design, in: *CHI '12*. p. 10.
- Boutilier, C., 2013. Computational Decision Support Regret-Based Models for Optimization and Preference Elicitation, in: *Comparative Decision Making*. Oxford University Press, New York. <https://doi.org/10.1093/acprof:oso/9780199856800.003.0041>

- Boutillier, C., Zemel, R.S., Marlin, B., 2003. Active collaborative filtering, in: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI-03). Presented at the 19th Conference on Uncertainty in Artificial Intelligence, Acapulco, Mexico, pp. 98–106.
- Boxell, L., Gentzkow, M., Shapiro, J., 2017. Is the Internet Causing Political Polarization? Evidence from Demographics. <https://doi.org/10.3386/w23258>
- Boxell, L., Gentzkow, Matthew, Shapiro, J.M., 2020. Cross-Country Trends in Affective Polarization.
- Bozarth, L., Budak, C., 2020. Toward a Better Performance Evaluation Framework for Fake News Classification 14, 60–71.
- Bozarth, L., Saraf, A., Budak, C., 2020. Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 U.S. presidential nominees 48–59. <https://doi.org/10.2139/ssrn.3340173>
- Brady, W.J., Van Bavel, J.J., 2021. Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis (preprint). Open Science Framework. <https://doi.org/10.31219/osf.io/s4w2x>
- Brennen, J.S., Simon, F.M., Howard, P.N., Nielsen, R.K., 2020. Types, Sources, and Claims of COVID-19 Misinformation. Reuters Institute for Politics, University of Oxford.
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O’Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T.K., Dyer, L., Khan, S., Bengio, Y., Anderljung, M., 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.
- Bruns, A., 2019. It’s Not the Technology, Stupid: How the ‘Echo Chamber’ and ‘Filter Bubble’ Metaphors Have Failed Us.
- BSR, 2018. Human Rights Impact Assessment: Facebook in Myanmar.
- Burke, R., 2017. Multisided Fairness for Recommendation. arXiv:1707.00093 [cs].
- Camerer, C.F., Fehr, E., 2004. Measuring social norms and preferences using experimental games: A guide for social scientists, in: Foundations of Human Sociality – Experimental and Ethnographic Evidence from 15 Small-Scale Societies. p. h.
- Camerer, C.F., Loewenstein, G., Rabin, M. (Eds.), 2004. Advances in Behavioral Economics. Princeton University Press.
- Cañamares, R., Castells, P., Moffat, A., 2020. Offline evaluation options for recommender systems. *Inf Retrieval J* 23, 387–410. <https://doi.org/10.1007/s10791-020-09371-3>
- Canca, C., 2020. Operationalizing AI ethics principles. *Commun. ACM* 63, 18–21. <https://doi.org/10.1145/3430368>
- Carroll, M., Hadfield-Menell, D., Dragan, A., Russell, S., 2021. Estimating and Penalizing Preference Shift in Recommender Systems. *RecSys ’21: Fifteenth ACM Conference on Recommender Systems*. <https://doi.org/10.1145/3460231.3478849>
- Carterette, B., Kanoulas, E., Yilmaz, E., 2012. Incorporating variability in user behavior into systems based evaluation, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 135–144.
- Carullo, G., Castiglione, A., De Santis, A., 2014. Friendship Recommendations in Online Social Networks, in: 2014 International Conference on Intelligent Networking and Collaborative Systems. Presented at the 2014 International Conference on Intelligent Networking and Collaborative Systems (INCoS), IEEE, Salerno, pp. 42–48. <https://doi.org/10.1109/INCoS.2014.114>
- Castells, P., Hurley, N.J., Vargas, S., 2015. Novelty and diversity in recommender systems. Springer US, pp. 881–918. [https://doi.org/10.1007/978-1-4899-7637-6\\_26](https://doi.org/10.1007/978-1-4899-7637-6_26)
- Chan, C., Sounderajah, V., Daniels, E., Acharya, A., Clarke, J., Yalamanchili, S., Normahani, P., Markar, S., Ashrafian, H., Darzi, A., 2021. The Reliability and Quality of YouTube Videos as a Source of Public Health Information Regarding COVID-19 Vaccination: Cross-sectional Study. *JMIR Public Health and Surveillance* 7, e29942. <https://doi.org/10.2196/29942>
- Chaopreecha, J., 2019. Revitalization of Tradition through Social Media: A Case of the Vegetarian Festival in Phuket, Thailand 35.
- Chen, G., Gully, S.M., Eden, D., 2001. Validation of a New General Self-Efficacy Scale. *Organizational Research Methods* 4, 62–83. <https://doi.org/10.1177/109442810141004>
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X., 2020. Bias and Debias in Recommender System: A Survey and Future Directions. arXiv:2010.03240 [cs].

- Chen, L., Pu, P., 2004. Survey of Preference Elicitation Methods 23.
- Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., Chi, E.H., 2019. Top-k off-policy correction for a reinforce recommender system 456–464. <https://doi.org/10.1145/3289600.3290999>
- Chen, Y.-Y., Chen, T., Hsu, W.H., Liao, H.-Y.M., Chang, S.-F., 2014. Predicting Viewer Affective Comments Based on Image Content in Social Media, in: Proceedings of International Conference on Multimedia Retrieval. Presented at the ICMR '14: International Conference on Multimedia Retrieval, ACM, Glasgow United Kingdom, pp. 233–240. <https://doi.org/10.1145/2578726.2578756>
- Chien, S., Jain, P., Krichene, W., Rendle, S., Song, S., Thakurta, A., Zhang, L., 2021. Private Alternating Least Squares: Practical Private Matrix Completion with Tighter Rates, in: Proceedings of the 38th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 1877–1887.
- Chowdhury, R., 2021. Sharing learnings about our image cropping algorithm [WWW Document]. Twitter. URL [https://blog.twitter.com/engineering/en\\_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm](https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm) (accessed 10.14.21).
- Christakopoulou, K., Radlinski, F., Hofmann, K., 2016. Towards Conversational Recommender Systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 815–824. <https://doi.org/10.1145/2939672.2939746>
- Cobbe, J., Singh, J., 2019. Regulating Recommending: Motivations, Considerations, and Principles 10. <https://doi.org/10.2139/ssrn.3371830>
- Council of Europe, 2020. Guide on Article 8 of the European Convention on Human Rights.
- Courtois, C., Slechten, L., Coenen, L., 2018. Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics* 35, 2006–2015. <https://doi.org/10.1016/j.tele.2018.07.004>
- Cowlshaw, T., Burlington, T., Man, D., Fiala, J., Barrington, R., Wright, G., 2018. Personalizing the Public: Personalising Linear Radio at a Public Service Broadcaster. British Broadcasting Corporation.
- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B., 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model User-Adap Inter* 18, 455. <https://doi.org/10.1007/s11257-008-9051-3>
- Csikszentmihalyi, M., 2020. Confucius, in: Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., Hormozdiari, F., Houlisby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T.F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., Sculley, D., 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning.
- Das, A., Mathieu, C., Ricketts, D., 2009. Maximizing profit using recommender systems.
- Das, A.S., Datar, M., Garg, A., Rajaram, S., 2007. Google news personalization: scalable online collaborative filtering. ACM Press, New York, New York, USA, p. 271. <https://doi.org/10.1145/1242572.1242610>
- Dascalu, M.-I., Bodea, C.-N., Mihailescu, M.N., Tanase, E.A., Ordoñez de Pablos, P., 2016. Educational recommender systems and their application in lifelong learning. *Behaviour & Information Technology* 35, 290–297. <https://doi.org/10.1080/0144929X.2015.1128977>
- Deldjoo, Y., Schedl, M., Cremonesi, P., Pasi, G., 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 1–38. <https://doi.org/10.1145/3407190>
- Delgado, J., Lind, S., Radecke, C., Konijeti, S., 2019. Simple objectives work better.
- Díaz, Á., Hecht-Felella, L., 2021. Double Standards in Social Media Content Moderation. Brennan Center for Justice.
- Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B., 2020. Evaluating Stochastic Rankings with Expected Exposure, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20. Association for Computing Machinery, New York, NY, USA, pp. 275–284. <https://doi.org/10.1145/3340531.3411962>
- Diener, E., 2000. Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist* 55, 34–43. <https://doi.org/10.1037/0003-066X.55.1.34>
- Diener, E., Emmons, R.A., Larsen, R.J., Griffin, S., 1985. The Satisfaction With Life Scale. *Journal of Personality Assessment* 49, 71–75. [https://doi.org/10.1207/s15327752jpa4901\\_13](https://doi.org/10.1207/s15327752jpa4901_13)

- Dodge, R., Daly, A.P., Huyton, J., Sanders, L.D., 2012. The challenge of defining wellbeing. *International Journal of Wellbeing* 2.
- Donahoe, E., Metzger, M.M., 2019. Artificial intelligence and human rights 30, 115–126. <https://doi.org/10.1353/jod.2019.0029>
- Douek, E., 2021. Governing Online Speech: From “Posts-as-Trumps” to Proportionality and Probability 121.
- Douek, E., 2020. The Limits of International Law in Content Moderation. *SSRN Journal*. <https://doi.org/10.2139/ssrn.3709566>
- Drev, M., Delak, B., 2021. Conceptual Model of Privacy by Design. *Journal of Computer Information Systems* 1–8. <https://doi.org/10.1080/08874417.2021.1939197>
- Dulac-Arnold, G., Evans, R., van Hasselt, H., Sunehag, P., Lillicrap, T., Hunt, J., Mann, T., Weber, T., Degris, T., Coppin, B., 2015. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*.
- Dunn, K., Singh, S.P., 2014. Pluralistic conditioning: social tolerance and effective democracy. *Democratization* 21, 1–28. <https://doi.org/10.1080/13510347.2012.697056>
- Dwork, C., Roth, A., 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 211–407. <https://doi.org/10.1561/04000000042>
- Ehrlich, T., 2000. Civic responsibility and higher education, American Council on Education/Oryx Press series on higher education. Oryx Press, Westport, Conn.
- Ekstrand, M., Das, A., Burke, R., Diaz, F., 2021. Fairness in Recommender Systems, in: Ricci, F., Roach, L., Shapira, B. (Eds.), *Recommender Systems Handbook*. Springer-Verlag.
- Ekstrand, M.D., Das, A., Burke, R., Diaz, F., 2021. Fairness and Discrimination in Information Access Systems. *arXiv:2105.05779 [cs]*.
- Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S., 2018. 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*. Presented at the Conference on Fairness, Accountability and Transparency, PMLR, pp. 172–186.
- Ellison, N.B., Steinfield, C., Lampe, C., 2007. The Benefits of Facebook “Friends:” Social Capital and College Students’ Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12, 1143–1168. <https://doi.org/10.1111/j.1083-6101.2007.00367.x>
- Eskildsen, J.K., Kristensen, K., 2011. The gender bias of the Net Promoter Score, in: *2011 IEEE International Conference on Quality and Reliability*. Presented at the 2011 IEEE International Conference on Quality and Reliability, pp. 254–258. <https://doi.org/10.1109/ICQR.2011.6031720>
- European Commission, 2020. Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act).
- European Commission, 2016. General Data Protection Regulation.
- Evans, C., Kasirzadeh, A., 2021. User Tampering in Reinforcement Learning Recommender Systems. *arXiv:2109.04083 [cs]*.
- Exton, C., Shinwell, M., 2018. Policy use of well-being metrics: Describing Countries’ Experiences 33. <https://doi.org/10.1787/d98eb8ed-en>
- Facebook, 2021a. Facebook’s Corporate Human Rights Policy.
- Facebook, 2021b. Our Approach to Ranking [WWW Document]. Facebook Transparency Center. URL <https://transparency.fb.com/features/ranking-and-content/> (accessed 10.28.21).
- Facebook, 2019. People, Publishers, the Community. Facebook Newsroom. URL <https://about.fb.com/news/2019/04/people-publishers-the-community/> (accessed 10.8.21).
- Faddoul, M., Chaslot, G., Farid, H., 2020. A Longitudinal Analysis of YouTube’s Promotion of Conspiracy Videos 1–8.
- Faltings, B., Jurca, R., Pu, P., Tran, B.D., 2014. Incentives to Counter Bias in Human Computation, in: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. pp. 59–66.
- Fan, J., Zhang, A.X., 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance 1–14. <https://doi.org/10.1145/3313831.3376293>
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S., 2015. Certifying and removing disparate impact, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 259–268.
- Fernandez, M., Bellogin, A., 2020. Recommender Systems and Misinformation: The Problem or the Solution?, in: *OHARS Workshop, 14th ACM Conference on Recommender Systems*. p. 9.
- Fields, B., Jones, R., Cowlshaw, T., 2018. The case for public service recommender algorithms, in: *FATREC*

- Workshop at RecSys 2018. Vancouver.
- 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. Berkman Klein Center for Internet & Society, Cambridge, MA.
- Fletcher, R., Kalogeropoulos, A., Nielsen, R.K., 2021. More diverse, more politically varied: How social media, search engines and aggregators shape news repertoires in the United Kingdom. *New Media & Society* 14614448211027392. <https://doi.org/10.1177/14614448211027392>
- Fletcher, R., Nielsen, R.K., 2018. Are people incidentally exposed to news on social media? A comparative analysis 20, 2450–2468. <https://doi.org/10.1177/1461444817724170>
- Freedman, R., Shah, R., Dragan, A., 2020. Choice Set Misspecification in Reward Inference, in: ICJCAI-PRICAI 2020 Workshop on Artificial Intelligence Safety.
- Frey, B.S., Benesch, C., Stutzer, A., 2007. Does Watching TV Make Us Happy? 28, 283–313.
- Friedman, B., Hendry, D.G., Borning, A., 2017. A survey of value sensitive design methods 11, 63–125. <https://doi.org/10.1561/11000000015>
- Friedman, B., Kahn, P.H., Borning, A., 2002. Value Sensitive Design: Theory and Methods 8.
- Frison, E., Eggermont, S., 2020. Toward an Integrated and Differential Approach to the Relationships Between Loneliness, Different Types of Facebook Use, and Adolescents' Depressed Mood. *Communication Research* 47, 701–728. <https://doi.org/10.1177/0093650215617506>
- Furnham, A., 1986. Response bias, social desirability and dissimulation. *Personality and Individual Differences* 7, 385–400. [https://doi.org/10.1016/0191-8869\(86\)90014-0](https://doi.org/10.1016/0191-8869(86)90014-0)
- Gabriel, I., 2020. Artificial Intelligence, Values, and Alignment. *Minds & Machines* 30, 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gabriel, I., Ghazavi, V., 2021. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety.
- Gaertner, W., 2009. *A Primer in Social Choice Theory: Revised Edition*. OUP Oxford.
- Gauci, J., Conti, E., Liang, Y., Virochsiri, K., He, Y., Kaden, Z., Narayanan, V., Ye, X., Chen, Z., Fujimoto, S., 2019. Horizon: Facebook's Open Source Applied Reinforcement Learning Platform. arXiv:1811.00260 [cs, stat].
- Geyik, S.C., Ambler, S., Kenthapadi, K., 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Presented at the KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Anchorage AK USA, pp. 2221–2231. <https://doi.org/10.1145/3292500.3330691>
- Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., Dollé, S., 2018. Offline A/B Testing for Recommender Systems, in: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Presented at the WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining, ACM, Marina Del Rey CA USA, pp. 198–206. <https://doi.org/10.1145/3159652.3159687>
- Global Internet Forum to Counter Terrorism, 2021. Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes.
- Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., Sculley, D., 2017. Google Vizier: A Service for Black-Box Optimization, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*. Association for Computing Machinery, New York, NY, USA, pp. 1487–1495. <https://doi.org/10.1145/3097983.3098043>
- Goodrow, C., 2021. On YouTube's recommendation system [WWW Document]. YouTube blog. URL <https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/> (accessed 11.19.21).
- Google, 2021. AI At Google: Our Principles [WWW Document]. URL <https://ai.google/principles/> (accessed 11.2.21).
- Google, 2020. Recommendation Systems Overview [WWW Document]. Google Developers. URL <https://developers.google.com/machine-learning/recommendation/overview/types> (accessed 1.5.22).
- Gordon, M.L., Zhou, K., Bernstein, M.S., 2021. The Disagreement Deconvolution : Bringing Machine Learning Performance Metrics In Line With Reality.
- Graham, C., Laffan, K., Pinto, S., 2018. Well-being in metrics and policy. *Science* 362 (6412), 287–288.
- Green, F., 2011. Unpacking the misery multiplier: How employability modifies the impacts of unemployment and job insecurity on life satisfaction and mental health 30, 265–276. <https://doi.org/10.1016/j.jhealeco.2010.12.005>
- Gretzel, U., Fesenmaier, D.R., 2006. Persuasion in Recommender Systems. *International Journal of Electronic Commerce* 11, 81–100.
- Griffioen, N., van Rooij, M., Lichtwarck-Aschoff, A., Granic, I., 2020. Toward improved methods in social media

- research. *Technology, Mind, and Behavior* 1. <https://doi.org/10.1037/tmb0000005>
- Guess, A., Lyons, B., Nyhan, B., Reifler, J., 2018. Avoiding the Echo Chamber about Echo Chambers: Why selective exposure to like-minded political news is less prevalent than you think.
- Gupta, V., Ouyang, Y., 2020. Rise of the Machines: Removing the {Human-in-the-Loop}, in: 2020 USENIX Conference on Operational Machine Learning. Presented at the 2020 USENIX Conference on Operational Machine Learning.
- Hadfield-Menell, D., Hadfield, G.K., 2019. Incomplete contracting and AI alignment 417–422. <https://doi.org/10.1145/3306618.3314250>
- Halevy, A., Ferrer, C.C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., Stoyanov, V., 2020. Preserving Integrity in Online Social Networks.
- Harambam, J., Bountouridis, D., Makhortykh, M., van Hoboken, J., 2019. 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. Presented at the RecSys '19: Thirteenth ACM Conference on Recommender Systems, ACM, Copenhagen Denmark, pp. 69–77. <https://doi.org/10.1145/3298689.3347014>
- Harambam, J., Helberger, N., Van Hoboken, J., 2018. Democratizing algorithmic news recommenders: How to materialize voice in a technologically saturated media ecosystem 376. <https://doi.org/10.1098/rsta.2018.0088>
- Hase, P., Bansal, M., 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? arXiv:2005.01831 [cs].
- Hassan, A., White, R.W., 2013. Personalized models of search satisfaction, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. pp. 2009–2018.
- He, C., Parra, D., Verbert, K., 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56, 9–27.
- Helberger, N., 2019. On the Democratic Role of News Recommenders 7, 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 21, 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 [WWW Document]. URL <https://blogs.lse.ac.uk/mediase/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/> (accessed 10.25.20).
- Helberger, N., von Drunen, M., Vrijenhoek, S., Möller, J., 2021. Regulation of news recommenders in the Digital Services Act: empowering David against the Very Large Online Goliath. *Internet Policy Review*.
- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J., 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research* 21.
- Hicks, S., Tinkler, L., Allin, P., 2013. Measuring Subjective Well-Being and its Potential Role in Policy: Perspectives from the UK Office for National Statistics. *Soc Indic Res* 114, 73–86. <https://doi.org/10.1007/s11205-013-0384-x>
- Hochschild, J.L., Einstein, K.L., 2015. *Do Facts Matter?: Information and Misinformation in American Politics*. University of Oklahoma Press.
- Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., Aral, S., 2020. The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify, in: Proceedings of the 21st ACM Conference on Economics and Computation, EC '20. Association for Computing Machinery, New York, NY, USA, pp. 75–76. <https://doi.org/10.1145/3391403.3399532>
- Hosseiniardi, H., Ghasemian, A., Clauset, A., Rothschild, D.M., Mobius, M., Watts, D.J., 2020. Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube.
- Hsu, S., Vaccaro, K., Yue, Y., Rickman, A., Karahalios, K., 2020. Awareness, Navigation, and Use of Feed Control Settings Online, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20. Association for Computing Machinery, New York, NY, USA, pp. 1–13. <https://doi.org/10.1145/3313831.3376583>
- Hursthouse, R., Pettigrove, G., 2018. Virtue Ethics, in: Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- Huszár, F., Ktena, S.I., O'Brien, C., Belli, L., Schlaikjer, A., Hardt, M., 2021. Algorithmic Amplification of Politics on Twitter.
- Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., Cheng, H.-T., Chandra, T., Boutilier, C., 2019. SlateQ:

- A Tractable Decomposition for Reinforcement Learning with Recommendation Sets, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Presented at the Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}, International Joint Conferences on Artificial Intelligence Organization, Macao, China, pp. 2592–2599. <https://doi.org/10.24963/ijcai.2019/360>
- IEEE, 2021. 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design. IEEE, 2020. IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being. IEEE. <https://doi.org/10.1109/IEEESTD.2020.9084219>
- Ignat, O., Boureau, Y.-L., Yu, J.A., Halevy, A., 2021. Detecting Inspiring Content on Social Media. arXiv:2109.02734 [cs].
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J., 2018. The Origins and Consequences of Affective Polarization in the United States 1–35. <https://doi.org/10.1146/annurev-polisci-051117-073034>
- Jackman, S., 2008. Measurement, in: Box-Steffensmeier, J.M., Brady, H.E., Collier, D. (Eds.), *The Oxford Handbook of Political Methodology*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199286546.003.0006>
- Jacobs, A.Z., Wallach, H., 2021. Measurement and Fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Presented at the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, pp. 375–385. <https://doi.org/10.1145/3442188.3445901>
- Jain, P., Thakkar, O.D., Thakurta, A., 2018. Differentially Private Matrix Completion Revisited, in: Proceedings of the 35th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 2215–2224.
- Jannach, D., Adomavicius, G., 2017. Price and Profit Awareness in Recommender Systems. Como, Italy.
- Jannach, D., Adomavicius, G., 2016. Recommendations with a purpose 7–10. <https://doi.org/10.1145/2959100.2959186>
- Jannach, D., Jugovac, M., 2019. Measuring the business value of recommender systems 10, 1–22. <https://doi.org/10.1145/3370082>
- Jannach, D., Manzoor, A., Cai, W., Chen, L., 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 1–36. <https://doi.org/10.1145/3453154>
- Jannach, D., Naveed, S., Jugovac, M., 2017. User Control in Recommender Systems: Overview and Interaction Challenges, in: Bridge, D., Stuckenschmidt, H. (Eds.), *E-Commerce and Web Technologies, Lecture Notes in Business Information Processing*. Springer International Publishing, Cham, pp. 21–33. [https://doi.org/10.1007/978-3-319-53676-7\\_2](https://doi.org/10.1007/978-3-319-53676-7_2)
- Jawaheer, G., Szomszor, M., Kostkova, P., 2010. Comparison of implicit and explicit feedback from an online music recommendation service, in: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec '10. Association for Computing Machinery, New York, NY, USA, pp. 47–51. <https://doi.org/10.1145/1869446.1869453>
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 78, 15169–15211. <https://doi.org/10.1007/s11042-018-6894-4>
- Jeunen, O., Goethals, B., 2021. Top-K Contextual Bandits with Equity of Exposure, in: Fifteenth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, pp. 310–320.
- Jiang, R., Chiappa, S., Lattimore, T., György, A., Kohli, P., 2019. Degenerate Feedback Loops in Recommender Systems. <https://doi.org/10.1145/3306618.3314288>
- Jin, Y., Cardoso, B., Verbert, K., 2017. How Do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations? Presented at the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, Como, Italy., p. 8.
- Kalimeris, D., Bhagat, S., Kalyanaraman, S., Weinsberg, U., 2021. Preference Amplification in Recommender Systems, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Presented at the KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Virtual Event Singapore, pp. 805–815. <https://doi.org/10.1145/3447548.3467298>
- Kantrowitz, A., 2021. Facebook Removed The News Feed Algorithm In An Experiment. Then It Gave Up. *Big Technology*. URL <https://bigtechnology.substack.com/p/facebook-removed-the-news-feed-algorithm> (accessed 4.2.22).
- Karimi, M., Jannach, D., Jugovac, M., 2018. News recommender systems – Survey and roads ahead 54, 1203–1227. <https://doi.org/10.1016/j.ipm.2018.04.008>
- Kashdan, T.B., Rose, P., Fincham, F.D., 2004. Curiosity and Exploration: Facilitating Positive Subjective

- Experiences and Personal Growth Opportunities. *Journal of Personality Assessment* 82, 291–305.  
[https://doi.org/10.1207/s15327752jpa8203\\_05](https://doi.org/10.1207/s15327752jpa8203_05)
- Keller, D., 2021. Amplification and Its Discontents. Knight First Amendment Institute at Columbia University.
- Keller, D., 2018. Internet platforms: Observations on speech, danger, and money 5–8.
- Kerlin, L., 2020. Human values: understanding psychological needs in a digital age.
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D., 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv:2005.04790 [cs].
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R., 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in: Proceedings of the 35th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 2668–2677.
- Kim, H., Ko, D., 2007. Culture & Self-Expression 1 Culture and Self-Expression, in: *Frontiers of Social Psychology: The Self*. pp. 325–342.
- Klonick, K., 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131.
- Knijnenburg, B.P., Sivakumar, S., Wilkinson, D., 2016. Recommender systems for self-actualization 11–14.  
<https://doi.org/10.1145/2959100.2959189>
- Koggel, C., Orme, J., 2010. Care Ethics: New Theories and Applications. *Ethics and Social Welfare* 4, 109–114.  
<https://doi.org/10.1080/17496535.2010.484255>
- Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 [cs, stat].
- Krafft, T.D., Gamer, M., Zweig, K.A., 2019. What did you see? A study to measure personalization in Google’s search engine. *EPJ Data Sci.* 8, 38. <https://doi.org/10.1140/epjds/s13688-019-0217-5>
- Krueger, A.B., Stone, A.A., 2014. Progress in measuring subjective well-being. *Science* 346, 42–43.  
<https://doi.org/10.1126/science.1256392>
- Krueger, D.S., Maharaj, T., Leike, J., 2020. Hidden incentives for auto-induced distributional shift.
- Kuhlman, C., Gerych, W., Rundensteiner, E., 2021. Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’21).  
<https://doi.org/10.1145/3461702.3462588>
- Kunaver, M., Požrl, T., 2017. Diversity in recommender systems – A survey 123, 154–162.  
<https://doi.org/10.1016/j.knosys.2017.02.009>
- Lada, A., Wang, M., Yan, T., 2021. How machine learning powers Facebook’s News Feed ranking algorithm. Engineering at Meta. URL <https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/> (accessed 12.13.21).
- Lambrecht, A., Tucker, C., 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65, 2966–2981.  
<https://doi.org/10.1287/mnsc.2018.3093>
- Latonero, M., Agarwal, A., 2021. Human Rights Impact Assessments for AI: Learning from Facebook’s Failure in Myanmar. Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University.
- Law, E., Ahn, L. von, 2011. Human Computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 1–121. <https://doi.org/10.2200/S00371ED1V01Y201107AIM013>
- Le, H., Maragh, R., Ekdale, B., High, A., Havens, T., Shafiq, Z., 2019. Measuring Political Personalization of Google News Search, in: The World Wide Web Conference. Presented at the WWW ’19: The Web Conference, ACM, San Francisco CA USA, pp. 2957–2963. <https://doi.org/10.1145/3308558.3313682>
- Ledwich, M., Zaitsev, A., 2019. Algorithmic Extremism: Examining YouTube’s Rabbit Hole of Radicalization. arXiv:1912.11211 [cs].
- 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 3.  
<https://doi.org/10.1145/3359283>
- Leibowicz, C., Moon Sehat, C., Stephan, A., Stray, J., 2021. If We Want Platforms to Think Beyond Engagement, We Have to Know What We Want Instead [WWW Document]. URL <https://medium.com/partnership-on-ai/if-we-want-platforms-to-think-beyond-engagement-we-have-to-know-what-we-want-instead-a8cfbfb6688> (accessed 11.17.21).
- Leonhardt, J., Anand, A., Khosla, M., 2018. User Fairness in Recommender Systems, in: Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18. Presented at the Companion of the The

- Web Conference 2018, ACM Press, Lyon, France, pp. 101–102. <https://doi.org/10.1145/3184558.3186949>
- Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y., 2021. User-oriented Fairness in Recommendation, in: Proceedings of the Web Conference 2021. Association for Computing Machinery, New York, NY, USA, pp. 624–632.
- Linton, M.-J., Dieppe, P., Medina-Lara, A., 2016. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. *BMJ Open* 6, e010641. <https://doi.org/10.1136/bmjopen-2015-010641>
- Loeberbach, F., Moeller, J., Trilling, D., van Atteveldt, W., 2020. The Unified Framework of Media Diversity: A Systematic Literature Review 8, 605–642. <https://doi.org/10.1080/21670811.2020.1764374>
- Loeberbach, F., Welbers, K., Moeller, J., Trilling, D., Van Atteveldt, W., 2021. Is this a click towards diversity? Explaining when and why news users make diverse choices, in: 13th ACM Web Science Conference 2021. Presented at the WebSci '21: WebSci '21 13th ACM Web Science Conference 2021, ACM, Virtual Event United Kingdom, pp. 282–290. <https://doi.org/10.1145/3447535.3462506>
- Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., Larson, H.J., 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav* 5, 337–348. <https://doi.org/10.1038/s41562-021-01056-1>
- Louca, R., Bhattacharya, M., Hu, D., Hong, L., 2019. Joint Optimization of Profit and Relevance for Recommendation Systems in E-commerce 2440, 4–7.
- Lukat, J., Margraf, J., Lutz, R., van der Veld, W.M., Becker, E.S., 2016. Psychometric properties of the Positive Mental Health Scale (PMH-scale). *BMC Psychology* 4, 8. <https://doi.org/10.1186/s40359-016-0111-x>
- Malone, G.P., Pillow, D.R., Osman, A., 2012. The General Belongingness Scale (GBS): Assessing achieved belongingness. *Personality and Individual Differences* 52, 311–316. <https://doi.org/10.1016/j.paid.2011.10.027>
- Manheim, D., Garrabrant, S., 2018. Categorizing Variants of Goodhart’s Law 1–10.
- Martinez, J.E., Paluck, E.L., 2020. Quantifying shared and idiosyncratic judgments of racism in social discourse.
- Mattingly-Jordan, S., Donaldson, B., Gray, P., Ingram, L.M., 2019. *Ethically Aligned Design First Edition Glossary*.
- Mattis, N., Masur, P., Möller, J., van Atteveldt, W., 2021. A theoretical framework for facilitating diverse news consumption through recommender design. *SocArXiv*. <https://doi.org/10.31235/osf.io/wvxf5>
- Mccaffery, C., 2020. An algorithm for empowering public service news. *Polis*. URL <https://blogs.lse.ac.uk/polis/2020/09/28/this-swedish-radio-algorithm-gets-reporters-out-in-society/> (accessed 10.8.21).
- McCoy, J., Somer, M., 2019. Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies. *The ANNALS of the American Academy of Political and Social Science* 681, 234–271. <https://doi.org/10.1177/0002716218818782>
- McMillan, D.W., Chavis, D.M., 1986. Sense of community: A definition and theory. *J. Community Psychol.* 14, 6–23. [https://doi.org/10.1002/1520-6629\(198601\)14:1<6::AID-JCOP2290140103>3.0.CO;2-I](https://doi.org/10.1002/1520-6629(198601)14:1<6::AID-JCOP2290140103>3.0.CO;2-I)
- Meca, A., 2012. *Personal Control and Responsibility Measure: A Psychometric Evaluation (Master of Science Psychology)*. Florida International University. <https://doi.org/10.25148/etd.FI12072703>
- Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., Yilmaz, E., 2017. Auditing search engines for differential satisfaction across demographics, in: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 626–633.
- Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., Diaz, F., 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. ACM, New York, NY, USA, pp. 2243–2251. <https://doi.org/10.1145/3269206.3272027>
- Mehrotra, R., Xue, N., Lalmas, M., 2020. Bandit based Optimization of Multiple Objectives on a Music Streaming Platform, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20. Association for Computing Machinery, New York, NY, USA, pp. 3224–3233. <https://doi.org/10.1145/3394486.3403374>
- Meinrenken, C.J., Kaufman, S.M., Ramesh, S., Lackner, K.S., 2012. Fast Carbon Footprinting for Large Product Portfolios 16, 669–679. <https://doi.org/10.1111/j.1530-9290.2012.00463.x>
- Midgley, C., Thai, S., Lockwood, P., Kovacheff, C., Page-Gould, E., 2021. When every day is a high school reunion: Social media comparisons and self-esteem. *Journal of Personality and Social Psychology* 121, 285–307. <https://doi.org/10.1037/pspi0000336>
- Milano, S., Taddeo, M., Floridi, L., 2020. Recommender systems and their ethical challenges. *AI & Soc* 35, 957–967. <https://doi.org/10.1007/s00146-020-00950-y>
- Milli, S., Belli, L., Hardt, M., 2020. From Optimizing Engagement to Measuring Value 1–14.

- Mladenov, M., Creager, E., Ben-Porat, O., Swersky, K., Zemel, R., Boutilier, C., 2020. Optimizing Long-term Social Welfare in Recommender Systems: A Constrained Matching Approach, in: Proceedings of the 37th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 6987–6998.
- Mladenov, M., Meshi, O., Ooi, J., Schuurmans, D., Boutilier, C., 2019. Advantage Amplification in Slowly Evolving Latent-State Environments 2019-Augus, 3165–3172. <https://doi.org/10.24963/ijcai.2019/439>
- Mohajan, H.K., 2017. Two Criteria for Good Measurements in Research: Validity and Reliability. ASHU- ES 17, 59–82. <https://doi.org/10.26458/1746>
- 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 21, 959–977. <https://doi.org/10.1080/1369118X.2018.1444076>
- Mooijman, M., Hoover, J., Lin, Y., Ji, H., Dehghani, M., 2018. Moralization in social networks and the emergence of violence during protests. *Nat Hum Behav* 2, 389–396. <https://doi.org/10.1038/s41562-018-0353-0>
- Morris, M.R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J.P., Kane, S.K., 2016. “With most of it being pictures now, I rarely use it”: Understanding Twitter’s Evolving Accessibility to Blind Users, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Presented at the CHI’16: CHI Conference on Human Factors in Computing Systems, ACM, San Jose California USA, pp. 5506–5516. <https://doi.org/10.1145/2858036.2858116>
- Mozilla, 2021. How Artificial Intelligence Fuels Online Disinformation: Relevant Legislation [WWW Document]. Mozilla Foundation. URL <https://foundation.mozilla.org/en/campaigns/trained-for-deception-how-artificial-intelligence-fuels-online-disinformation/relevant-legislation/> (accessed 11.19.21).
- Mullainathan, S., Obermeyer, Z., 2021. On the Inequity of Predicting A While Hoping for B. *AEA Papers and Proceedings* 111, 37–42. <https://doi.org/10.1257/pandp.20211078>
- Munger, K., Phillips, J., 2020. Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics* 194016122096476. <https://doi.org/10.1177/1940161220964767>
- Munn, L., 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday*. <https://doi.org/10.5210/fm.v24i6.10108>
- Nabi, R.L., Prestin, A., So, J., 2013. Facebook Friends with (Health) Benefits? Exploring Social Network Site Use and Perceptions of Social Support, Stress, and Well-Being. *Cyberpsychology, Behavior, and Social Networking* 16, 721–727. <https://doi.org/10.1089/cyber.2012.0521>
- Näher, A.-F., Krumpal, I., 2012. Asking sensitive questions: the impact of forgiving wording and question context on social desirability bias. *Qual Quant* 46, 1601–1616. <https://doi.org/10.1007/s11135-011-9469-2>
- Narayanan, A., Huey, J., Felten, E.W., 2016. A Precautionary Approach to Big Data Privacy, in: Gutwirth, S., Leenes, R., De Hert, P. (Eds.), *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection, Law, Governance and Technology Series*. Springer Netherlands, Dordrecht, pp. 357–385. [https://doi.org/10.1007/978-94-017-7376-8\\_13](https://doi.org/10.1007/978-94-017-7376-8_13)
- Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F., 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv:1802.00682 [cs]*.
- National Opinion Research Center, 2013. *General Social Survey Cumulative File 1972-2012*.
- Ndiweni, E., Sibanda, W., 2020. CSR governance framework of South Africa, pre, during and post-apartheid: a manifestation of ubuntu values? *International Journal of Business Governance and Ethics* 14, 363–383. <https://doi.org/10.1504/IJBGE.2020.110820>
- 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 90, 298–307. <https://doi.org/10.1016/j.chb.2018.07.043>
- Nick, E.A., Cole, D.A., Cho, S.-J., Smith, D.K., Carter, T.G., Zerkowicz, R., 2018. The Online Social Support Scale: Measure Development and Validation. *Psychol Assess* 30, 1127–1143. <https://doi.org/10.1037/pas0000558>
- Noorshams, N., Verma, S., Hofleitner, A., 2020. TIES: Temporal Interaction Embeddings for Enhancing Social Media Integrity at Facebook 3128–3135. <https://doi.org/10.1145/3394486.3403364>
- Norwegian Centre for Research Data, 2012. *ESS Round 6: European Social Survey Round 6 Data (2012)*.
- Nunes, I., Jannach, D., 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model User-Adap Inter* 27, 393–444. <https://doi.org/10.1007/s11257-017-9195-0>
- O’Donnell, G., Deaton, A., Durand, M., Halpern, D., Layard, R., 2014. *Wellbeing and Policy [WWW Document]*.

- URL <https://li.com/reports/the-commission-on-wellbeing-and-policy/>
- OECD, 2020. Better Life Index [WWW Document]. URL <https://www.oecdbetterlifeindex.org/> (accessed 11.11.21).
- OECD, 2019. Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies. Organisation for Economic Co-operation and Development, Paris.
- O’Neil, M., Jensen, M.J., 2020. Australian Perspectives on Misinformation. News Media Research Centre, University of Canberra.
- Ostrom, E., 2000. Collective action and the evolution of social norms 14, 137–158. <https://doi.org/10.1257/jep.14.3.137>
- Ovadya, A., 2021. Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure.
- Pacherie, E., 2007. The Sense of Control and the Sense of Agency. *Psyche* 13, 1.
- Paolini, S., Harwood, J., Rubin, M., 2010. Negative Intergroup Contact Makes Group Memberships Salient: Explaining Why Intergroup Conflict Endures. *Pers Soc Psychol Bull* 36, 1723–1738. <https://doi.org/10.1177/0146167210388667>
- Papcunová, J., Martončík, M., Fedáková, D., Kentoš, M., Bozogánová, M., Srba, I., Moro, R., Pikuliak, M., Šimko, M., Adamkovič, M., 2021. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex Intell. Syst.* <https://doi.org/10.1007/s40747-021-00561-0>
- Patja Howell, J., 2021. The Arrival of International Human Rights Law in Content Moderation. The Lawfare Podcast.
- Pennycook, G., Rand, D.G., 2021. Reducing the spread of fake news by shifting attention to accuracy: Meta-analytic evidence of replicability and generalizability (preprint). *PsyArXiv*. <https://doi.org/10.31234/osf.io/v8ruj>
- Pennycook, G., Rand, D.G., 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci USA* 116, 2521–2526. <https://doi.org/10.1073/pnas.1806781116>
- Pettigrew, T.F., Tropp, L.R., 2006. A meta-analytic test of intergroup contact theory. 90, 751–783. <https://doi.org/10.1037/0022-3514.90.5.751>
- Pinna, M., Picard, L., Goessmann, C., 2021. Cable News and COVID-19 Vaccine Compliance (SSRN Scholarly Paper No. ID 3890340). Social Science Research Network, Rochester, NY. <https://doi.org/10.2139/ssrn.3890340>
- Prior, M., 2013. Media and Political Polarization 16, 101–27. <https://doi.org/10.1146/annurev-polisci-100711-135242>
- Przybylski, A.K., Weinstein, N., 2017. A Large-Scale Test of the Goldilocks Hypothesis: Quantifying the Relations Between Digital-Screen Use and the Mental Well-Being of Adolescents. *Psychol Sci* 28, 204–215. <https://doi.org/10.1177/0956797616678438>
- Pu, P., Chen, L., 2010. A User-Centric Evaluation Framework of Recommender Systems 612, 8.
- Qi, T., Wu, F., Wu, C., Huang, Y., Xie, X., 2020. Privacy-Preserving News Recommendation Model Learning. *arXiv:2003.09592 [cs]*.
- Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R., 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54, 209–228. <https://doi.org/10.1111/j.1540-5907.2009.00427.x>
- Raj, A., Ekstrand, M.D., 2022. Measuring Fairness in Ranked Results 11.
- Raj, A., Wood, C., Montoly, A., Ekstrand, M.D., 2020. Comparing Fair Ranking Metrics. *arXiv:2009.01311 [cs]*.
- Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P., 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44.
- Ranking Digital Rights, 2020. 2020 Ranking Digital Rights Corporate Accountability Index. Ranking Digital Rights.
- Rasmussen, C.E., Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*: MIT Press.
- Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F., Meira, W., 2020a. Auditing radicalization pathways on YouTube, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT\* ’20*. Presented at the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, pp. 131–141. <https://doi.org/10.1145/3351095.3372879>
- Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F., Wagner Meira, W.M., 2020b. Auditing radicalization pathways on YouTube, in: *Conference on Fairness, Accountability, and Transparency (FAT\* ’20)*, Barcelona, pp. 131–141. <https://doi.org/10.1145/3351095.3372879>
- Robertson, S., Salehi, N., 2020. What If I Don’t Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *arXiv:2007.06718 [cs]*.
- Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., Herrman, H., 2016. Social media and suicide prevention: a systematic review: Suicide prevention and social media. *Early Intervention in*

- Psychiatry 10, 103–121. <https://doi.org/10.1111/eip.12229>
- Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E.D., Mukkavilli, S.K., Kording, K.P., Gomes, C., Ng, A.Y., Hassabis, D., Platt, J.C., Creutzig, F., Chayes, J., Bengio, Y., 2019. Tackling Climate Change with Machine Learning.
- Roose, K., 2019. The Making of a YouTube Radical.
- Rosso, B.D., Dekas, K.H., Wrzesniewski, A., 2010. On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior* 30, 91–127. <https://doi.org/10.1016/j.riob.2010.09.001>
- Rudin, C., 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154 [cs, stat].
- Russell, S., 2019. Human compatible: Artificial intelligence and the problem of control. Viking, New York.
- Rychwalska, A., Roszczyńska-Kurasińska, M., 2018. Polarization on Social Media: When Group Dynamics Leads to Societal Divides. <https://doi.org/10.24251/hicss.2018.263>
- Ryff, C.D., Keyes, C.L.M., 1995. The Structure of Psychological Well-Being Revisited. *Journal of Personality and Social Psychology* 69.
- Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C., 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Presented at the Data and Discrimination: Converting Critical Concerns into Productive Inquiry, at the 64th Annual Meeting of the International Communication Association.
- Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N.A., 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. arXiv:2111.07997 [cs].
- Sapiezynski, P., Zeng, W., Robertson, R.E., Mislove, A., Wilson, C., 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. arXiv:1901.10437 [cs].
- Sapp, S.G., Harrod, W.J., 1993. Reliability and Validity of a Brief Version of Levenson’s Locus of Control Scale. *Psychol Rep* 72, 539–550. <https://doi.org/10.2466/pr0.1993.72.2.539>
- Saveski, M., Roy, B., Roy, D., 2021. The Structure of Toxic Conversations on Twitter. *Proceedings of the Web Conference 2021* 1086–1097. <https://doi.org/10.1145/3442381.3449861>
- Schiff, D., Ayes, A., Musikanski, L., Havens, J.C., 2020. IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence.
- Schnabel, T., Amershi, S., Bennett, P.N., Bailey, P., Joachims, T., 2020. The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation 991–1000. <https://doi.org/10.1145/3397271.3401117>
- Schwartz, S.H., 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2. <https://doi.org/10.9707/2307-0919.1116>
- Sen, A., 2017. *Collective choice and social welfare*. Harvard University Press.
- Seymour, B., Getman, R., Saraf, A., Zhang, L.H., Kalendarian, E., 2015. When Advocacy Obscures Accuracy Online: Digital Pandemics of Public Health Misinformation Through an Antifluoride Case Study. *Am J Public Health* 105, 517–523. <https://doi.org/10.2105/AJPH.2014.302437>
- Shakya, H.B., Christakis, N.A., 2017. Association of Facebook Use With Compromised Well-Being: A Longitudinal Study. *Am. J. Epidemiol.* [amjepid;kww189v2. https://doi.org/10.1093/aje/kww189](https://doi.org/10.1093/aje/kww189)
- Shani, G., Heckerman, D., Brafman, R.I., 2005. An MDP-Based Recommender System. *Journal of Machine Learning Research* 6, 1265–1295.
- Shapiro, E.H., Sugarman, M., Media, F.B., Zuckerman, E., 2021. New approaches to Platform Data Research.
- Silverman, H., Ruggles, D.F., 2007. Cultural Heritage and Human Rights, in: Silverman, H., Ruggles, D.F. (Eds.), *Cultural Heritage and Human Rights*. Springer, New York, NY, pp. 3–29. [https://doi.org/10.1007/978-0-387-71313-7\\_1](https://doi.org/10.1007/978-0-387-71313-7_1)
- Simonsen, J., Robertson, T., 2012. *Routledge international handbook of participatory design*. Routledge. <https://doi.org/10.4324/9780203108543>
- Singh, A., Halpern, Y., Thain, N., Christakopoulou, K., Chi, E.H., Chen, J., Beutel, A., 2021. Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach 10.
- Singh, A., Joachims, T., 2018. Fairness of Exposure in Rankings, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’18*. Association for Computing Machinery, New York, NY, USA, pp. 2219–2228. <https://doi.org/10.1145/3219819.3220088>
- Singh, S., 2021. *Charting a Path Forward*.
- Singh, S., 2020a. *Regulating Platform Algorithms: Approaches for EU and U.S. Policymakers*.
- Singh, S., 2020b. *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*. New America Foundation.

- Singh, S., 2020c. Promoting Fairness, Accountability, and Transparency Around Algorithmic Recommendation Practices [WWW Document]. New America. URL <http://newamerica.org/oti/reports/why-am-i-seeing-this/> (accessed 11.19.21).
- Singh, S., Doty, L., 2021a. Cracking Open the Black Box. New America Foundation.
- Singh, S., Doty, L., 2021b. The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules [WWW Document]. New America. URL <http://newamerica.org/oti/reports/transparency-report-tracking-tool/> (accessed 10.9.21).
- Skoe, E.E.A., 2014. Measuring care-based moral development: The Ethic of Care Interview. *Behavioral Development Bulletin* 19, 95–104. <https://doi.org/10.1037/h0100594>
- Smith, B., 2021. How TikTok Reads Your Mind. *The New York Times*.
- Sørensen, J.K., Hutchinson, J., 2018. Algorithms and Public Service Media, in: *Public Service Media in the Networked Society*. pp. 91–106.
- Spagnuolo, D., Bartolini, C., Lenzini, G., 2016. Metrics for Transparency, in: Livraga, G., Torra, V., Aldini, A., Martinelli, F., Suri, N. (Eds.), *Data Privacy Management and Security Assurance, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 3–18. [https://doi.org/10.1007/978-3-319-47072-6\\_1](https://doi.org/10.1007/978-3-319-47072-6_1)
- Stanton, E.A., 2007. The Human Development Index: A History. PERI Working Papers 127.
- Steger, M.F., Dik, B.J., Duffy, R.D., 2012. Measuring Meaningful Work: The Work and Meaning Inventory (WAMI). *Journal of Career Assessment* 20, 322–337. <https://doi.org/10.1177/1069072711436160>
- Steinfeld, C., Ellison, N.B., Lampe, C., 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology* 29, 434–445. <https://doi.org/10.1016/j.appdev.2008.07.002>
- Stoutenborough, J.W., Robinson, S.E., Vedlitz, A., 2016. Is “fracking” a new dirty word? The influence of word choice on public views toward natural gas attitudes. *Energy Research & Social Science* 17, 52–58. <https://doi.org/10.1016/j.erss.2016.04.005>
- Strauss, M.E., Smith, G.T., 2009. Construct Validity: Advances in Theory and Methodology. *Annu Rev Clin Psychol* 5, 1–25. <https://doi.org/10.1146/annurev.clinpsy.032408.153639>
- Stray, J., 2021a. Designing Recommender Systems to Depolarize. *First Monday*.
- Stray, J., 2021b. Show me the algorithm: Transparency in recommendation systems [WWW Document]. Schwartz Reisman Institute. URL <https://srinstitute.utoronto.ca/news/recommendation-systems-transparency> (accessed 10.4.21).
- Stray, J., 2020. Aligning AI Optimization to Community Well-being. *International Journal of Community Well-Being* 3, 443–463. <https://doi.org/10.1007/s42413-020-00086-3>
- Stray, J., 2012. Who should see what when? Three principles for personalized news. *Nieman Lab*. URL <https://www.niemanlab.org/2012/07/who-should-see-what-when-three-principles-for-personalized-news/> (accessed 6.15.21).
- Stroud, N.J., Muddiman, A., Scacco, J.M., 2017. Like, recommend, or respect? Altering political behavior in news comment sections. *New Media & Society* 19, 1727–1743. <https://doi.org/10.1177/1461444816642420>
- Su, S., 2017. New Test With Related Articles. *Facebook Newsroom*. URL <https://about.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles/> (accessed 9.19.21).
- Subrahmanyam, K., Reich, S.M., Waechter, N., Espinoza, G., 2008. Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology* 29, 420–433. <https://doi.org/10.1016/j.appdev.2008.07.003>
- Sullivan, J.L., Transue, J.E., 1999. The Psychological Underpinnings of Democracy: A Selective Review of Research on Political Tolerance, Interpersonal Trust, and Social Capital. *Annu. Rev. Psychol.* 50, 625–650. <https://doi.org/10.1146/annurev.psych.50.1.625>
- Sundararajan, M., Taly, A., Yan, Q., 2017. Gradients of Counterfactuals. arXiv:1611.02639 [cs].
- Suzor, N., Van Geelen, T., Myers West, S., 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 385–400. <https://doi.org/10.1177/1748048518757142>
- Taghipour, N., Kardan, A., Ghidary, S.S., 2007. Usage-based web recommendations: a reinforcement learning approach, in: *Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys '07*. Association for Computing Machinery, New York, NY, USA, pp. 113–120. <https://doi.org/10.1145/1297231.1297250>
- Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., Stewart-Brown, S., 2007. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation.

- Health and Quality of Life Outcomes 5, 63. <https://doi.org/10.1186/1477-7525-5-63>
- The Markup, 2020. The Citizen Browser Project—Auditing the Algorithms of Disinformation.
- Thompson, N., 2018. How Facebook Wants to Improve the Quality of Your News Feed | WIRED. Wired.
- Thorburn, L., Stray, J., Bengani, P., 2022a. What Does it Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems. *Understanding Recommenders*. URL <https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157> (accessed 3.25.22).
- Thorburn, L., Stray, J., Bengani, P., 2022b. What Will “Amplification” Mean in Court? [WWW Document]. Tech Policy Press. URL <https://techpolicy.press/what-will-amplification-mean-in-court/> (accessed 6.30.22).
- Thrash, T., Elliot, A., 2003. Inspiration as a Psychological Construct. *Journal of personality and social psychology* 84, 871–89. <https://doi.org/10.1037/0022-3514.84.4.871>
- Tintarev, N., Masthoff, J., 2015. Explaining Recommendations: Design and Evaluation, in: Ricci, F., Rokach, L., Shapira, B. (Eds.), *Recommender Systems Handbook*. Springer US, Boston, MA, pp. 353–382. <https://doi.org/10.1007/978-1-4899-7637-6>
- Topp, C.W., Østergaard, S.D., Søndergaard, S., Bech, P., 2015. The WHO-5 Well-Being Index: A Systematic Review of the Literature. *Psychother Psychosom* 84, 167–176. <https://doi.org/10.1159/000376585>
- UK Office of National Statistics, 2019. Measuring national well-being: domains and measures [WWW Document]. URL <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/measuringnationalwellbeingdomainsandmeasures> (accessed 11.11.21).
- UNESCO, 2017. Declaration of Ethical Principles in relation to Climate Change [WWW Document]. URL [http://portal.unesco.org/en/ev.php-URL\\_ID=49457&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/en/ev.php-URL_ID=49457&URL_DO=DO_TOPIC&URL_SECTION=201.html) (accessed 12.2.21).
- UNESCO, 2005. Convention for the Protection and Promotion of the Diversity of Cultural Expressions [WWW Document]. URL <https://en.unesco.org/creativity/convention/texts> (accessed 12.8.21).
- UNESCO, 1978. Declaration on Fundamental Principles concerning the Contribution of the Mass Media to Strengthening Peace and International Understanding, to the Promotion of Human Rights and to Countering Racism, apartheid and incitement to war. [WWW Document]. URL [http://portal.unesco.org/en/ev.php-URL\\_ID=13176&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/en/ev.php-URL_ID=13176&URL_DO=DO_TOPIC&URL_SECTION=201.html) (accessed 12.8.21).
- United Nations, 1969. Declaration on Social Progress and Development.
- United Nations, 1966. International Covenant on Civil and Political Rights.
- Université de Montréal, 2018. Montreal Declaration for a Responsible Development of Artificial Intelligence [WWW Document]. URL <https://www.montrealdeclaration-responsibleai.com/the-declaration> (accessed 11.2.21).
- Vaccaro, K., Huang, D., Eslami, M., Sandvig, C., Hamilton, K., Karahalios, K., 2018. The Illusion of Control: Placebo Effects of Control Settings. Presented at the CHI 2018, ACM, New York, NY, USA. <https://doi.org/10.1145/3173574.3173590>
- Valenzuela, S., Park, N., Kee, K.F., 2009. Is There Social Capital in a Social Network Site?: Facebook Use and College Students’ Life Satisfaction, Trust, and Participation. *Journal of Computer-Mediated Communication* 14, 875–901. <https://doi.org/10.1111/j.1083-6101.2009.01474.x>
- Vallor, S., 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, New York. <https://doi.org/10.1093/acprof:oso/9780190498511.001.0001>
- Vallor, S., Raicu, I., Green, B., 2020. *Technology and Engineering Practice: Ethical Lenses to Look Through*. The Markkula Center for Applied Ethics at Santa Clara University.
- van Drunen, M.Z., Helberger, N., Bastian, M., 2019. Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law* 9, 220–235. <https://doi.org/10.1093/idpl/ipz011>
- van Stekelenburg, J., 2014. Going all the way: Politicizing, polarizing, and radicalizing identity offline and online 8, 540–555. <https://doi.org/10.1111/soc4.12157>
- Varshney, L.R., 2020. Respect for Human Autonomy in Recommender Systems. arXiv:2009.02603 [cs].
- Vecchione, B., Barocas, S., Levy, K., 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. arXiv:2109.06974 [cs]. <https://doi.org/10.1145/3465416.3483294>
- Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., Kross, E., 2017. Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review 11, 274–302. <https://doi.org/10.1111/sipr.12033>
- Vorm, E.S., Miller, A.D., 2018. Assessing the Value of Transparency in Recommender Systems: An End-User

Perspective 8.

- Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. 359, 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., Helberger, N., 2020. Recommenders with a mission: assessing diversity in news recommendations 1, 554–561. [https://doi.org/10.1007/978-3-030-65965-3\\_38](https://doi.org/10.1007/978-3-030-65965-3_38)
- Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR 1–52. <https://doi.org/10.2139/ssrn.3063289>
- Waern, A., 2004. User Involvement in Automatic Filtering: An Experimental Study. *User Model. User-Adapt. Interact.* 14, 201–237. <https://doi.org/10.1023/B:USER.0000028984.13876.9b>
- Wall Street Journal, 2021. Inside TikTok’s Algorithm: A WSJ Video Investigation. Wall Street Journal.
- Wang, L., Joachims, T., 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets, in: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. pp. 23–41.
- Washingtonian, 2019. What Happened After My 13-Year-Old Son Joined the Alt-Right. Washingtonian.
- Watson, D., Anna, L., Tellegen, A., 1988. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales 8.
- Weaver, P., Choi, J., Kaufman, T., 1997. Question Wording and Response Bias: Students’ Perceptions of Ethical Issues in the Hospitality and Tourism Industry. *Journal of Hospitality & Tourism Education* 9, 21–26. <https://doi.org/10.1080/10963758.1997.10685307>
- Weiser, S., 2019. Requirements of Trustworthy AI [WWW Document]. FUTURIUM - European Commission. URL <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1> (accessed 10.21.21).
- Weld, D., Bansal, G., 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM* 62, 70–79.
- White, J.B., Langer, E.J., Yariv, L., Welch, J.C., 2006. Frequent Social Comparisons and Destructive Emotions and Behaviors: The Dark Side of Social Comparisons. *J Adult Dev* 13, 36–44. <https://doi.org/10.1007/s10804-006-9005-0>
- Whittaker, J., Looney, S., Reed, A., Votta, F., 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review* 10. <https://doi.org/10.14763/2021.2.1565>
- Wilhelm, M., Ramanathan, A., Bonomo, A., Jain, S., Chi, E.H., Gillenwater, J., 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM ’18*. Association for Computing Machinery, New York, NY, USA, pp. 2165–2173. <https://doi.org/10.1145/3269206.3272018>
- Wilson, C., Secker, J., 2015. Validation of the Social Inclusion Scale with Students. *Social Inclusion* 3, 52–62. <https://doi.org/10.17645/si.v3i4.121>
- World Health Organization, 2004. Promoting Mental Health. World Health Organization, Geneva.
- World Values Survey, 2020. WVS Wave 7 [WWW Document]. URL <https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp> (accessed 11.11.21).
- Wu, C., Wu, F., Cao, Y., Huang, Y., Xie, X., 2021. FedGNN: Federated Graph Neural Network for Privacy-Preserving Recommendation. arXiv:2102.04925 [cs].
- Wu, H., Ma, C., Mitra, B., Diaz, F., Liu, X., 2021. Multi-FR: A Multi-Objective Optimization Method for Achieving Two-sided Fairness in E-commerce Recommendation. arXiv:2105.02951 [cs].
- Wu, H., Mitra, B., Ma, C., Diaz, F., Liu, X., 2022. Joint Multisided Exposure Fairness for Recommendation.
- Xian, Y., Zhao, T., Li, J., Chan, J., Kan, A., Ma, J., Dong, X.L., Faloutsos, C., Karypis, G., Muthukrishnan, S., Zhang, Y., 2021. EX3: Explainable Attribute-aware Item-set Recommendations, in: *Fifteenth ACM Conference on Recommender Systems, RecSys ’21*. Association for Computing Machinery, New York, NY, USA, pp. 484–494. <https://doi.org/10.1145/3460231.3474240>
- Yan, H., Fitzsimmons-Craft, E.E., Goodman, M., Krauss, M., Das, S., Cavazos-Rehg, P., 2019. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *Int J Eat Disord* 52, 1150–1156. <https://doi.org/10.1002/eat.23148>
- Yarchi, M., Baden, C., Kligler-Vilenchik, N., 2021. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication* 38, 98–139. <https://doi.org/10.1080/10584609.2020.1785067>
- Yi, X., Hong, L., Zhong, E., Liu, N.N., Rajan, S., 2014. Beyond clicks: Dwell time for personalization 113–120. <https://doi.org/10.1145/2645710.2645724>
- York, J., Zuckerman, E., 2019. Moderating the Public Sphere, in: Jørgensen, R.F. (Ed.), . MIT Press.
- Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., Feng, L., 2020. Deep attention based music genre classification.

- Neurocomputing 372, 84–91. <https://doi.org/10.1016/j.neucom.2019.09.054>
- Zamani, H., Schedl, M., Lamere, P., Chen, C.-W., 2019. An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation. arXiv:1810.01520 [cs].
- Zehlike, M., Yang, K., Stoyanovich, J., 2022. Fairness in Ranking, Part I: Score-based Ranking. ACM Comput. Surv. <https://doi.org/10.1145/3533379>
- Zehlike, M., Yang, K., Stoyanovich, J., 2021. Fairness in Ranking: A Survey. arXiv:2103.14000 [cs].
- Zhang, A.X., Hugh, G., Bernstein, M.S., 2020. PolicyKit: Building governance in online communities 365–378. <https://doi.org/10.1145/3379337.3415858>
- Zhang, Y., Chen, X., 2020. Explainable Recommendation: A Survey and New Perspectives 14, 1–101. <https://doi.org/10.1561/1500000066>. Yongfeng
- Zhao, X., Gu, C., Zhang, H., Yang, X., Liu, X., Tang, J., Liu, H., 2021. DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems. arXiv:1909.03602 [cs].
- Zhao, X., Zhang, W., Wang, J., 2013. Interactive collaborative filtering, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13. Association for Computing Machinery, New York, NY, USA, pp. 1411–1420. <https://doi.org/10.1145/2505515.2505690>
- Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X., Chi, E., 2019. Recommending what video to watch next. ACM, New York, NY, USA, pp. 43–51. <https://doi.org/10.1145/3298689.3346997>
- Zhu, H., Yu, B., Halfaker, A., Terveen, L., 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. Proc. ACM Hum.-Comput. Interact. 2, 1–23. <https://doi.org/10.1145/3274463>
- Ziegler, C.-N., McNee, S.M., Konstan, J.A., Lausen, G., 2005. Improving recommendation lists through topic diversification. ACM Press, New York, New York, USA, p. 22. <https://doi.org/10.1145/1060745.1060754>
- Zimet, G.D., Dahlem, N.W., Zimet, S.G., Farley, G.K., 1988. The Multidimensional Scale of Perceived Social Support. Journal of Personality Assessment 52, 30–41. [https://doi.org/10.1207/s15327752jpa5201\\_2](https://doi.org/10.1207/s15327752jpa5201_2)
- Zuiderveen Borgesius, F.J., Trilling, D., Möller, J., Bodó, B., de Vreese, C.H., Helberger, N., 2016. Should we worry about filter bubbles? 5, 1–16. <https://doi.org/10.14763/2016.1.401>