International Journal of Community Well-Being (2020) 3:443–463
https://doi.org/10.1007/s42413-020-00086-3

PERSPECTIVE ARTICLE

Aligning AI Optimization to Community Well-Being

Jonathan Stray[1]

Received: 7 February 2020 / Accepted: 9 October 2020 / Published online: 4 November 2020

Abstract

This paper investigates incorporating community well-being metrics into the objectives of optimization algorithms and the teams that build them. It documents two cases where a large platform appears to have modified their system to this end. Facebook incorporated "well-being" metrics in 2017, while YouTube began integrating "user satisfaction" metrics around 2015. Metrics tied to community well-being outcomes could also be used in many other systems, such as a news recommendation system that tries to increase exposure to diverse views, or a product recommendation system that opstimizes for the carbon footprint of purchased products. Generalizing from these examples and incorporating insights from participatory design and AI governance leads to a proposed process for integrating community well-being into commercial AI systems: identify and involve the affected community, choose a useful metric, use this metric as a managerial performance measure and/or an algorithmic objective, and evaluate and adapt to outcomes. Important open questions include the best approach to community participation and the uncertain business effects of this process.

Keywords Artificial intelligence · AI ethics · Community well-being · Optimization · Corporate social responsibility

Introduction

This paper is an extended analysis of a simple idea: large-scale commercial optimizing systems may be able to manage harmful side effects on communities by monitoring established well-being metrics. It sketches a theory that ties together quantitative measures of well-being, contemporary metrics-driven management practice, the objective function of optimization algorithms, participatory and multi-stakeholder governance of algorithmic systems, and the protection or promotion of community well-being. Detailed analyses of recent efforts by Facebook and YouTube are used to

✉ Jonathan Stray
jonathan@partnershiponai.org

[1] Partnership on AI, San Francisco, CA, USA

illustrate the challenges and unknowns of this approach, which generalizes to a variety of different types of artificial intelligence (AI) systems. The core contribution of this article is a proposed process for the use of community well-being metrics within commercial AI systems.

Well-being encompasses "people's living conditions and quality of life today (current well-being), as well as the resources that will help to sustain people's well-being over time (natural, economic, human and social capital)" (OECD 2019b, p. 2). Community well-being attempts to evaluate well-being at the level of a community defined "in geographic terms, such as a neighborhood or town … or in social terms, such as a group of people sharing common chat rooms on the Internet, a national professional association or a labor union" (Phillips and Pittman 2015, p. 3). The measurement of well-being is now a well-established field with a long history, and is increasingly used in policy-making (Exton and Shinwell 2018).

Large AI systems can have both positive and harmful side effects on communities, through effects on employment and inequality (Korinek and Stiglitz 2017), privacy and safety (OECD 2019a), addictive behavior (Andreassen 2015), fairness and discrimination (Barocas et al. 2018), human rights (Donahoe and Metzger 2019), polarization, extremism, and conflict (Ledwich and Zaitsev 2020; Stoica and Chaintreau 2019), and potentially many other areas (Kulynych et al. 2020). Importantly, AI systems can affect non-users too, as with environmental externalities.

Most AI is built around optimization "in which the aim is to find the best state according to an objective function" (Russell and Norvig 2010, p. 121) where an *objective function* is some method for quantitatively evaluating the desirability of an outcome (Dantzig 1982). Standard management practice also increasingly involves the maximization of quantitative metrics (Parmenter 2020), which can be considered an optimization process. This paper is concerned with optimizing systems composed of people and algorithms which affect communities, where the choice of objective might have significant societal influence. Examples include systems used to allocate resources or assign work, choose what news people see, recommend products to buy, or implement government policy. Many of these systems would be considered AI, but perhaps the phrase "autonomous and intelligent systems" (Schiff et al. 2020, p. 1) which appears in certain standards efforts would be better, because an automated system does not have to be very smart to cause harm. Rather, the unifying feature is optimization – both the cause of many problems and an opportunity for a response.

The central idea of this paper is to incorporate community well-being metrics into the optimization process at both the managerial and technical level. This is a sociotechnical approach to systems design (Baxter and Sommerville 2011) that considers the role of both people and technology. There are many technical interventions that could be undertaken aside from the modification of an algorithmic objective function; for example, a social media product team could choose to show a simple chronological list of posts rather than using algorithmic content personalization. However, if product managers are evaluated on community well-being outcomes, they may choose to make such a change based on the expected effects on users. The integration of the managerial and the technical in an optimization framework can motivate many possible product design changes.

Background

This paper responds most directly to recent calls for research into well-being and AI. It proposes specific "improvements to product design" (Schiff et al. 2019, p. 3) and it is interdisciplinary, systems-based, and community-oriented (Musikanski et al. 2020). It draws on and contributes to the emerging field of recommender alignment, the practice of building algorithms for content ranking and personalization that enact human values (Stray et al. 2020).

The goal of the process proposed in this paper is the governance of large-scale commercial algorithmic systems. Rahwan (2018) calls this *society-in-the-loop* control, defined as "embedding the values of society, as a whole, in the algorithmic governance of societal outcomes" (p. 3). In this sense community participation is a key element of the proposed framework, and this paper draws on approaches as diverse as participatory design (Simonsen and Robertson 2012) and corporate stakeholder engagement (Manetti 2011).

Community Well-Being

At the individual level well-being is usually studied as an experiential state, and there is now a wealth of research on the definition and reliable measurement of subjective well-being (Diener et al. 2018). Although well-being is a rich, multidimensional construct, even single questions can reveal substantial information, such as *overall, how satisfied are you with life as a whole these days?* answered on a 0–10 scale. This well-studied measure has several advantages: it correlates with how people make major life decisions, gives a similarly reliable result across cultures, and is by itself informative enough to be used in quantitative evaluations of policy choices (O'Donnell et al. 2014).

Community well-being "embraces a wide range of economic, social, environmental, political, cultural dimensions, and can be thought of as how well functions of community are governed and operating" (Sung and Phillips 2018, p. 64). In practice, community well-being is assessed using a variety of metrics across many domains. Often both subjective and objective indicators are needed to get a full picture (Musikanski et al. 2019). A survey of local and national well-being indicator frameworks in use in the United Kingdom gives an overview of the substance and range of such metrics (Bagnall et al. 2017). Community well-being frameworks can originate from consideration of geographic communities, or communities of interest (Phillips and Pittman 2015) which may be particularly relevant to online platforms.

As an example community well-being framework, the OECD Better Life Index (Durand 2015) aims to measure "both current material conditions and quality of life" (p. 1) across countries through the metrics shown in Table 1. This framework includes the life satisfaction measure above, as well as statistical indicators around health, education, employment, etc. in conjunction with subjective indicators such as whether one feels safe walking alone at night.

Technologists and scholars have begun to appreciate the significance of well-being measures in the design and operation of AI systems (Musikanski et al. 2020). The *IEEE 7010 Recommended Practice Standard for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* collects pre-existing measures from sources such as the OECD Better Life Index, the UN Sustainable Development Indicators, the

Table 1 Indicators from the OECD Better Life Index (Durand 2015). Each of these has a specific statistical definition and has been collected across OECD countries since 2011

Domain	Indicators
Housing	Dwellings without basic facilities
	Housing expenditure
	Rooms per person
Income	Household net adjusted disposable income
	Household net wealth
Jobs	Labor market insecurity
	Employment rate
	Long term unemployment rate
Community	Quality of support network
Education	Educational attainment
	Student skills
	Years in education
Environment	Air pollution
	Water quality
Civic engagement	Stakeholder engagement for developing regulations
	Voter turnout
Health	Life expectancy
	Self-reported health
Life Satisfaction	Life satisfaction
Safety	Feeling safe walking alone at night
	Homicide rate
Work-life balance	Employees working very long hours
	Time devoted to leisure and personal care

Human Development Index, the World Health Organization, the World Values Survey, Freedom House, and others (Schiff et al. 2020). From the point of view of a technologist who is concerned about the societal effects of their work, established well-being metrics have the advantage of representing extensive deliberation by domain experts.

Optimization

Optimization is used extensively in AI to guide training and learning. A problem to be solved is expressed as a scalar function – a method to calculate a single number that expresses the desirability of any given hypothetical solution. Solving the problem means finding a solution that maximizes this function. The encapsulation of concerns into a single function was a major conceptual advance that enabled the creation of generic optimization algorithms (Dantzig 1982). Conceptually, any problem that has some set of best solutions can be expressed as optimization with a single objective

function, though practical problem-solving often involves the optimization of multiple sub-goals.

A supervised machine learning algorithm that attempts to identify objects from images would usually be trained through a loss function that penalizes incorrect answers. A reinforcement learning approach to playing a video game might use the game score as a reward function. There are also value functions, cost functions, fitness functions, energy functions and more, all of which operate on similar principles (Russell and Norvig 2010). For simplicity, in this paper I refer to all of the scalar functions used to drive AI behavior as *objective functions*.

In this paper I refer to an *optimizing system* as if there were one optimizer and one objective. In practice such systems, especially those at platform scale, may include dozens or hundreds of optimizing components (numerous trained sub-models, for example). There isn't one objective function that can be altered, but many. Nonetheless, there are usually a few high-level goals concerned with the system's main outputs. This is the case at Groupon with many interacting models and a master objective function that aligns to company goals (Delgado et al. 2019).

Quantitative metrics analogous to objective functions are also used in corporate management. Modern management practice includes concepts such as *key performance indicators* (Parmenter 2020) and *objectives and key results* (Doerr 2017), both of which involve quantitative indicators of progress. Economic theory frequently models the corporation as a profit optimizer (e.g. Samuelson and Marks 2014). More sophisticated descriptions try to account for the creation of various types of long-term value, such as the *balanced scorecard* (Kaplan 2009) and *sustainability accounting* (Richardson 2013), both of which describe various non-financial metrics that are intended to be optimized.

Case Studies of Platform Interventions

This section presents two examples where large technology companies seem to have optimized for well-being, or a similar concept. These cases have been reconstructed through documentary evidence such as public posts, previously published interviews, financial reports, and research articles by employees.

Facebook's Well-Being Optimization

In late 2017 and early 2018, Facebook made a number of changes to their product explicitly designed to promote well-being. Facebook researchers Ginsberg and Burke (2017) wrote in a public post in December 2017:

> **What Do Academics Say? Is Social Media Good or Bad for Well-Being?** According to the research, it really comes down to *how* you use the technology. For example, on social media, you can passively scroll through posts, much like watching TV, or actively interact with friends — messaging and commenting on each other's posts. Just like in person, interacting with people you care about can be beneficial, while simply watching others from the sidelines may make you feel worse. (para. 7).

This post cites a number of peer-reviewed studies on the well-being effects of social media, some of which were collaborations between Facebook researchers and academics. Ginsberg and Burke (2017) cite Verduyn et al.'s (2017) review paper on the effects of social media on well-being, which has an obvious resonance with Facebook's framing:

> passively using social network sites provokes social comparisons and envy, which have negative downstream consequences for subjective well-being. In contrast, when active usage of social network sites predicts subjective well-being, it seems to do so by creating social capital and stimulating feelings of social connectedness. (Verduyn et al. 2017, p. 274)

A close reading of posts around this time shows that Facebook developed a well-being proxy metric. A January 2018 post by Facebook's Chief Executive Officer notes that "research shows that strengthening our relationships improves our well-being and happiness" (Zuckerberg 2018, para. 2) and mentions *well-being* twice more, then switches to the phrase "meaningful social interactions:"

> I'm changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions. (Zuckerberg 2018, para. 7)

Relevance is a term of art in recommender systems, referring to user preferences as expressed through item clicks or ratings, and is increasingly understood as a simplistic objective (Jannach and Adomavicius 2016). The algorithmic change away from relevance was described by the head of the News Feed product:

> Today we use signals like how many people react to, comment on or share posts to determine how high they appear in News Feed. With this update, we will also prioritize posts that spark conversations and meaningful interactions between people. To do this, we will predict which posts you might want to interact with your friends about, and show these posts higher in feed (Mosseri 2018, para. 3).

Facebook created a well-being metric and assigned it as a goal to a product team, which incorporated it into an existing algorithmic objective function. This objective function was augmented by creating a model that uses existing data such as past user behavior and post content to predict whether a user will have a *meaningful social interaction* if shown any particular post. There is little public documentation of how *meaningful social interactions* are measured. The most detailed description is from the transcript of a call where Facebook reported earnings to investors, which explains that *meaningful social interactions* are measured through user surveys:

> So the thing that we're going to be measuring is basically, the number of interactions that people have on the platform and off because of what they're seeing that they report to us as meaningful…the way that we've done this for

> years is we've had a panel, a survey, of thousands of people who basically we asked, what's the most meaningful content that they had seen in the platform or they have seen off the platform. (Facebook 2018, p. 13)

The resulting system is reconstructed in Fig. 1.

While there is no public account of the effects of the incorporation of the *meaningful social interactions* prediction model on the *meaningful social interactions* metric as measured by Facebook through user surveys, Facebook has reported reduced engagement on at least one product, suggesting that the *meaningful social interactions* objective was weighted strongly enough to cause significant changes in which items are presented to users:

> video is just a passive experience. To shift that balance, I said that we were going to focus on videos that encourage meaningful social interactions. And in Q4, we updated our video recommendations and made other quality changes to reflect these values. We estimate these updates decreased time spent on Facebook by roughly 5% in the fourth quarter. To put that another way: we made changes that reduced time spent on Facebook by an estimated 50 million hours every day to make sure that people's time is well spent. (Facebook 2018, p. 2)...

YouTube's User Satisfaction Metrics

John Doerr's *Measure What Matters* (2017) documents YouTube's multi-year effort to reach one billion hours of daily user watch time through interviews with Susan Wojcicki, Chief Executive Officer and Cristos Goodrow, Vice President of Engineering at YouTube (Doerr 2017, pp. 154–172). Goodrow describes the inception of

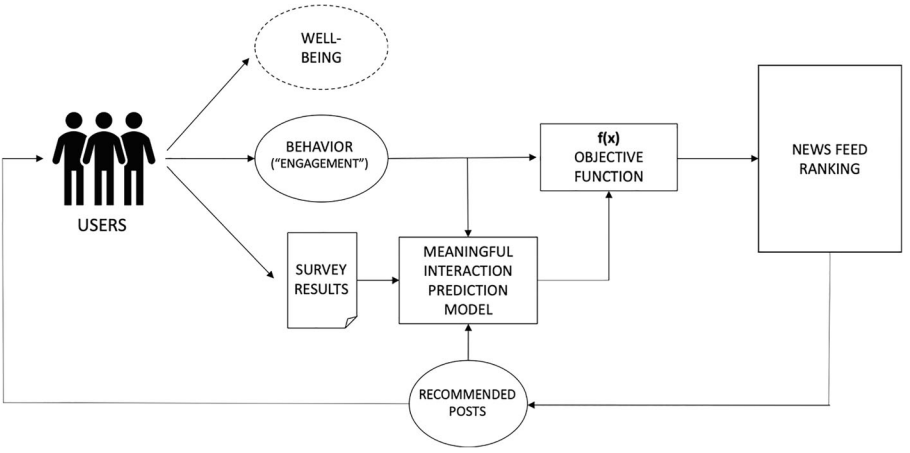

Fig. 1 A reconstruction of Facebook's use of *meaningful social interactions* circa 2018. Well-being effects are unobserved because they happen outside of user interactions with Facebook

YouTube's recommendation system in 2011, and how he advocated to optimize for watch time instead of video views as:

> On a dedicated team named Sibyl, Jim McFadden was building a system for selecting "watch next" recommendations, aka related videos or "suggestions." It had tremendous potential to boost our overall views. But were views what we really wanted to boost?...
>
> I sent a provocative email to my boss and the YouTube leadership team. Subject line: "Watch time, and only watch time." It was a call to rethink how we measured success: "All other things being equal, our goal is to increase [video] watch time."...
>
> Our job was to keep people engaged and hanging out with us. By definition, viewers are happier watching seven minutes of a ten-minute video (or even two minutes of a ten-minute video) than all of a one-minute video. And when they're happier, we are, too. (Goodrow quoted in Doerr 2017, p. 162)...

Goodrow's retelling includes user *happiness* and *satisfaction* as goals along with the more business-oriented *engagement*. For the purposes of this paper, I assume user *happiness* and *satisfaction* are analogous to well-being, but unlike the Facebook case, YouTube's public statements have not mentioned well-being. In accordance with the unified treatment of managerial and technical optimization proposed in this paper, Goodrow confirms that a team-level metric drove engineering decisions:

> Reaching one billion hours was a game of inches; our engineers were hunting for changes that might yield as little as 0.2 percent more watch time. In 2016 alone, they would find around 150 of those tiny advances. We'd need nearly all of them to reach our objective. (Goodrow quoted in Doerr 2017, p. 169)

Yet watch time was not the only objective, and YouTube incorporated other changes to improve the quality of the product and the effects on users:

> In fact, we'd commit to some watch-time-negative decisions for the benefit of our users. For example, we made it a policy to stop recommending trashy, tabloid-style videos—like "World's Worst Parents," where the thumbnail showed a baby in a pot on the stove. Three weeks in, the move proved negative for watch time by half a percent. We stood by our decision because it was better for the viewer experience, cut down on click bait, and reflected our principle of growing responsibly. Three months in, watch time in this group had bounced back and actually increased. Once the gruesome stuff became less accessible, people sought out more satisfying content. (Goodrow quoted in Doerr 2017, p. 164)

This was the beginning of a move away from strict maximization of time spent. Starting in 2015 YouTube began to incorporate user satisfaction metrics (Doerr 2017, p. 170). As in the Facebook case, these are derived from surveys:

> we learned that just because a user might be watching content longer does not mean that they are having a positive experience. So we introduced surveys to ask users if they were satisfied with particular recommendations. With this direct feedback, we started fine-tuning and improving these systems based on this high-fidelity notion of satisfaction. (Google 2019, p. 21)

These *user satisfaction* survey results were incorporated directly into the objectives of the YouTube recommendation system, as discussed in a recent YouTube technical paper:

> we first group our multiple objectives into two categories: 1) engagement objectives, such as user clicks, and degree of engagement with recommended videos; 2) satisfaction objectives, such as user liking a video on YouTube, and leaving a rating on the recommendation. (Zhao et al. 2019, p. 43)

Analysis of Facebook and YouTube Cases

The Facebook and YouTube cases are significant because they are examples of a major platform operator explicitly saying that they have decided to monitor and optimize for a well-being proxy, operationalized at both the management and algorithmic levels.

Facebook has provided a public justification for its *meaningful social interaction* metric in terms of prior research which suggests that active use of social media improves well-being while passive use decreases it. While this is far from a holistic measure of well-being, let alone community well-being, at least it connects to previous work in a clear way. Public statements from YouTube have not mentioned well-being, instead focusing on "responsibility" (Wojcicki 2019, para. 2) and *user satisfaction* as assessed through surveys.

Explicit user surveys are an improvement on YouTube's previous identification of watch time with user happiness. Researchers report a negative correlation between TV watching and well-being that suggests there is something like an addiction mechanism involved: "individuals with incomplete control over, and foresight into, their own behavior watch more TV than they consider optimal for themselves and their well-being is lower than what could be achieved" (Frey et al. 2007, p. 283). Similar effects have been observed in social media use where addicted users "have typically attempted to cut down on social networking without success" (Andreassen 2015, p. 176). Google now publicly recognizes that maximizing watch time does not optimize for "positive" outcomes (Google 2019, p. 21).

A more systematic conception of well-being would articulate what aspects of well-being matter to YouTube and why *user satisfaction* is a good proxy. Of course, well-being outcomes depend enormously on who a user is and what they watch. A user might learn valuable and fulfilling skills from how-to videos, become more politically engaged, consume worthwhile art, or they might be radicalized into violence (Ledwich and Zaitsev 2020).

Another issue is that both companies are optimizing for individual outcomes: well-being but not necessarily community well-being. Community well-being "is more than

an aggregate of individuals' satisfaction" (Sung and Phillips 2018, p. 65) and cannot be assessed simply by adding up the well-being of all individuals in the community. This is analogous to the classic problem of aggregating utilities in welfare economics (Foster and Sen 1997, p. 16). Conversely, optimizing for each person individually will not necessarily promote community well-being due to problems of externalities, collective action, and conflicting preferences (Baum 2020; Milano et al. 2019b). Attention to aggregates may also miss local problems, such as negative effects in a particular city or for a particular subgroup, or run into Simpsons' paradox issues where the sign of the effect depends on the granularity of the groups studied (Kievit et al. 2013). For all these reasons, clarity on the definition of community or communities matters greatly.

Perhaps the biggest weakness of these cases is that there is no record of consultation with the putative beneficiaries of these algorithmic changes, and no public evaluation of the results. Hopefully algorithmic interventions of this magnitude were informed by user research or some sort of consultative process, but none was reported. Presumably *meaningful social interactions* and *user satisfaction* were increased, but there has been no disclosure of how much. Absent also is any report of effects on any other components of well-being, such as feelings of social connectedness or life satisfaction, or even objective indicators like employment status. It's similarly unclear how these changes affected not just individual well-being but community well-being for different communities; there may even have been negative effects on certain types of users. Information about outcomes is especially important because the link between Facebook's *meaningful interactions* and well-being is theoretical, deduced from previous research into active and passive social media use, while YouTube has said their *user satisfaction* surveys are included in a "responsibility" metric (Bergen 2019, para. 10) and that they aim for "positive" experiences (Google 2019, p. 21) without providing any further explanation of their goals or results. Determining the actual effect of these large-scale interventions is itself a significant social science research effort, and if Facebook or YouTube have these answers, they have not been shared. This is algorithmic management, but not yet the algorithmic governance that the *society-in-the-loop* model envisions (Rahwan 2018).

The reported business outcomes are also instructive, as both the Facebook and YouTube changes resulted in at least temporary reductions in engagement metrics. Facebook reports that the incorporation of a *meaningful social interactions* metric into their video product caused a 5% reduction in time spent, which was considered significant enough to be discussed with investors (Facebook 2018) but the longer-term effects are unclear. YouTube described changes that reduced watch time but also reports that watch time recovered over a time span of months as users changed their behavior. This demonstrates both that major corporations are willing to accept reductions in engagement to pursue social ends, and that the long-term business effects of incorporating well-being metrics are not necessarily negative.

Generalization to Other Domains

The Facebook and YouTube cases suggest the possibility of a general method for managing the well-being outcomes of commercial optimizing systems, which is the core contribution of this article. This section begins by arguing that some type of

metric-driven community well-being optimization is not only useful but likely necessary for any AI system with broad social impacts, because individual user control will not be sufficient. It then shows how this general method could apply to diverse domains by working through potential applications to news recommendation and online shopping. These hypothetical applications demonstrate the generality of a metrics-driven approach and illuminate further possibilities and challenges that shape the recommendations in this paper.

User Control is not Sufficient for Community Well-Being

This article recommends participatory processes to involve users and other stakeholders in metric-driven optimization for community well-being. A potential alternative is to provide increased user control directly, so that people can choose what is best for themselves. Many authors have pointed to the central role of user agency in the ethics of AI systems (Floridi and Cowls 2019) and in the important context of content ranking Paraschakis (2017) has proposed "controls [that] enable users to adjust the recommender system to their individual moral standards" (p. 6). However, increasing user agency will not by itself solve the problem of ensuring good outcomes at the community level because many users will not customize the systems they use, and because individually good choices do not necessarily produce socially good outcomes.

Any set of controls must necessarily be few enough to be humanly manageable. This restricts the number of dimensions that can be controlled and will make it difficult to express nuanced conceptions of well-being. Natural language interfaces e.g. Yu et al. (2019) may allow the expression of more complicated concepts. Nonetheless users will probably leave most parameters at default settings, which means that the defaults must promote well-being.

Even if all users in fact succeeded in directing an AI system to do exactly as desired this would not necessarily result in the best community outcomes. As Ostrom (2000) has articulated, individual action does not succeed in producing social goods without the concurrent evolution of social norms. These challenges of collective action have been explored in the context of AI systems from the perspective of social choice theory (Baum 2020) and multi-stakeholder recommendation systems (Milano et al. 2019a). Further, existing societal inequalities can constrain users' ability to exploit algorithmically provided choices (Robertson and Salehi 2020), for example due to a lack of information or the cost burden of choosing the "best" option.

User control is essential, perhaps even necessary for community well-being, but it is not sufficient. Collective algorithmic governance is needed for much the same reasons societal governance is needed, and appropriate well-being metrics are useful in algorithmic governance just as they are in public policy.

Diverse News Recommendations

News recommenders are the algorithms that choose, order, and present journalism content to users. The potential application of community well-being metrics to these systems illustrates the challenges around defining a community and choosing metrics. News recommendation algorithms can have societal consequences (Helberger 2019) but it is not clear how to manage such algorithms for community well-being. To begin

with, there is no single community that consumes news, but many overlapping communities organized around different geographic regions and different topics (Reader and Hatcher 2011, p. 3). Each of these communities may have different concerns at any given moment. Incorporating social network analysis or country-specific data can improve the performance of recommender systems as measured by traditional relevance metrics (Chen et al. 2018; Roitero et al. 2020) but the question of how a recommender system impacts pre-existing communities, e.g. a city, has not been explored. Conversely, existing community well-being indicators have not been designed to capture the consequences of news recommender systems.

One well-developed concern with news recommenders is exposure diversity, meaning the range of sources, topics, and viewpoints that each person is algorithmically presented (Bernstein et al. 2020). Taking political theory as a starting point Helberger et al. (2018) identify liberal, deliberative, and radical approaches to the design of diverse news recommenders. Consider the problem of designing a national news recommender that supports a deliberative view of diversity, one in which:

> exposure to diverse viewpoints is considered valuable because it helps citizens develop more informed opinions and less polarized, more tolerant attitudes towards those with whom they disagree … it is conceivable to design metrics that would focus, for example, on user engagement with opposing political views, cross-ideological references in public debates or social media connections between people who represent different ideological positions. (Helberger et al. 2018, p. 195)

Diversity metrics could be constructed from algorithmic methods to estimate the ideological position of users or posts (Budak et al. 2016; Garimella and Weber 2017). These give a measure of distance between any two items, which could then be used to define the diversity of a set of recommended items according to various standard formulas such as the average distance between any pair (Kunaver and Požrl 2017).

Such a metric would capture the output of the system, not its effects on users. Facebook and YouTube use user surveys to tie algorithmic changes to human outcomes. It may be possible to establish a causal connection from news diversity metrics to existing well-being metrics such as voter turnout, and Facebook has already demonstrated a substantial effect on voter turnout by presenting users with personalized messages (Bond et al. 2012). It would be better to direct the optimization process towards more closely related outcomes like polarization or tolerance that are not included in current well-being frameworks. Directly measuring these outcomes is crucial because exposure to diverse opinions can actually increase polarization (Bail et al. 2018). Polarization and tolerance outcomes are also explicitly relational, and thus indicate aspects of community well-being not captured in individual-level metrics.

Low Carbon Shopping

Large-scale product recommender systems have profound influence over what is purchased. One reason for this is that it is not possible to navigate millions of possible products without them. Rolnick et al. (2019) have proposed using these systems to

direct consumers to lower-carbon alternatives. This possibility highlights two problems that may arise in the course of modifying AI objective functions: obtaining the data needed to evaluate a metric and understanding the business impacts of such a change.

Climate change is a key issue for many communities (Fazey et al. 2018) and carbon emissions appear in a number community well-being frameworks (Bagnall et al. 2017). Carbon emissions from recommended products are also a key example of AI system side effects on non-users. From a technical point of view, carbon footprint can be incorporated using multi-stakeholder recommendation algorithms that explicitly consider the effect on parties other than the user (Abdollahpouri et al. 2020).

This is possible only if the carbon footprint of each product is available. There are now established methods to estimate product carbon footprints (BSI 2011; ISO 2018) but there are no product carbon footprint (PCF) databases comprehensive enough to cover the millions of different products sold by a large online retailer. However, it may be possible to use machine learning methods to estimate the PCF values of an entire product portfolio starting from a comparatively small database of examples (Meinrenken et al. 2012). Robust, scalable product carbon footprint estimation could be a key enabling technology for low-carbon commerce and, ultimately, long-term community well-being.

A commercial operator will want to know the business effects before any such system is implemented, and it is tempting to evaluate the potential revenue effect of incorporating a carbon term into the objective function by testing against historical purchase data. Such back-testing will show that optimizing for anything other than profit must drive the system away from a profit maximum, but offline estimates will not give the full story because both consumer and producer behavior may change if carbon footprint starts to affect product ranking. Users might appreciate being informed of low-carbon alternatives and buy more from that retailer or pay a premium for lower carbon items, while producers will have an incentive to sell lower carbon products. The case of organic food demonstrates the existence of such market dynamics, as it is 22–35% more profitable globally than conventional alternatives even though it is typically more expensive to produce (Crowder and Reganold 2015).

Recommendations

The incorporation of community well-being metrics into both managerial and algorithmic optimization is a very general method for managing the effects of commercial optimizing systems, yet good management is only part of good governance. This section synthesizes the analysis and discussion above with previous work on algorithmic governance, participatory design, best use of metrics, and corporate stakeholder engagement to make recommendations for fostering community well-being in AI systems in ways that are both effective and accepted as legitimate. It also identifies gaps and unknowns where future research would be valuable.

Identifying and Involving Communities

An attempt to optimize for community well-being is an attempt to benefit a particular group of people, who need to have a say in what is done on their behalf. In some cases

it would be reasonable to say that every user of the system (potentially billions of people) is a member of the community, but that would preclude the management of local outcomes such as a system's effects on the residents of a particular city, or on people of a certain age, or workers in particular professions. Non-users can be affected as well, as in environmental externalities or a navigation system that routes cars to a formerly quiet street. Each view of community is a choice about who counts, and this choice should be made explicit before any intervention begins.

Once a community is identified, there are many approaches to try to integrate its members into the process of selecting and using metrics. Participatory design is an orientation and a set of practices that attempts to actively involve all stakeholders in a system design process (Simonsen and Robertson 2012). It is a promising framework for algorithmic governance. The WeBuildAI method (Lee et al. 2019) demonstrates what participatory design of metrics might look like. Researchers worked with a food-delivery non-profit to design an algorithm to match donated food with volunteer drivers and local food distribution charities. Stakeholders from each of these groups worked with researchers to build quantitative models of their preferred trade-offs between factors such as driver travel time, time since last donation, neighborhood poverty level, etc. At run time this system ranks the possible matches for each donation according to the models representing the preferences of each stakeholder, with the final result chosen through a ranked-choice voting rule. Future work could investigate participatory metric design in the context of a large commercial platform.

There are both instrumental and political goals when attempting to integrate communities into the selection and use of metrics. Without engaging the community, it is not possible to know which aspects of well-being matter most to them and how serious these issues are, and therefore how to make tradeoffs. Engagement is also necessary for credibility. When choosing community indicators, "most communities consider input by its residents and others to be vital; it builds support for the use of indicators as well as help vest those most impacted by subsequent actions in decision-making processes" (Sung and Phillips 2018, p. 73). In the context of commercial systems it will also be important to draw on the experience of corporate stakeholder engagement efforts such as those found in sustainability reporting (GSSB 2016; Manetti 2011).

Choosing Metrics

Aside from the well-known issues with using metrics in a management context generally (Jackson 2005) metrics pose a problem for AI systems in particular because most AI algorithms are based on strongly optimizing a narrow objective (Thomas and Uminsky 2020). Poor use of metrics can result in a damaging emphasis on short term outcomes, manipulation and gaming, and unwanted side effects (Jackson 2005; Thomas and Uminsky 2020). Even a successful metric cannot remain static, as the structure of the world it measures is constantly changing. In addition, there are many domains without a clear consensus on well-being goals, necessitating a process of normative deliberation before metrics can be chosen. The following issues should be considered in choice of metrics:

Deciding What to Measure In many cases existing well-being metrics will not be directly usable because they are too expensive to collect at scale or don't readily apply

in the company's domain. These issues drove Facebook's substitution of *meaningful social interactions* for more general measures of user well-being. Creating a custom metric is challenging because community well-being is a theoretical construct, not an observable property, and there may be misalignment between the designer's intentions and what is actually measured. For example, decreasing polarization measures may just indicate that minority voices have been effectively silenced. The particular well-being aspect of interest must first be "operationalized" and tested for reliability and validity (Jacobs and Wallach 2019).

Long-Term Outcomes If a metric is evaluated only over the short term it may lead to poor longer-term outcomes. As the YouTube case demonstrates, a video platform that tries to maximize user watch time may encourage binging behavior where users eventually regret the time they spend. While effective AI optimization requires frequent feedback, it is critical to pick shorter-term metrics that are thought to drive longer-term outcomes (Lalmas and Hong 2018).

Gaming Any measure that becomes a target will change meaning as people change their behavior, a very general problem that is sometimes known as Goodhart's law (Manheim and Garrabrant 2018). This is particularly relevant to large platforms that must defeat adversarial efforts to gain exposure for financial or political ends. While there are emerging methods to use causal inference to design metrics that resist gaming (Miller et al. 2019), a more robust solution is to continuously monitor and change the metrics in use.

Dynamism The metrics employed need to be able to change and adapt, a property that Jackson (2005) names *dynamism*. This is necessary because of gaming and other behavior change in response to metrics, but more importantly the world can and does change; at the onset of the COVID-19 pandemic many existing machine learning models stopped working (Heaven 2020). Dynamism also avoids the serious problems that can arise from over-optimization for a single objective, such as a robot which injures humans in an attempt to fetch a coffee more quickly (Russell 2019). In the context of contemporary commercial optimization, there are always humans supervising and operating the AI system, and they are free to change the objective function as needed.

Normative Uncertainty Catalogs such as IEEE 7010 (Schiff et al. 2020) provide a long list of consensus metrics but not all of them will correspond to community needs, and not all AI systems can be effectively evaluated using metrics originally designed for public policy use. In short, many systems will face a lack of consensus around what a "good" outcome would be. Appropriate values for AI systems cannot be derived from first principles but must be the result of societal deliberation (Gabriel 2020), which again underscores the necessity for participatory processes.

Evaluating Outcomes

It may be very challenging to determine the actual well-being effects of incorporating a metric into an optimization process. Facebook uses ongoing user panels to count *meaningful social interactions,* but this is a narrow facet of user well-being, let alone community well-being. They could use broader well-being instruments such as a life satisfaction survey question, but it would be difficult to assess the causal contribution of Facebook use to any changes. In other cases, such as the diverse news recommender, pre-existing well-being indicators would not apply so assessing societal impact would require the creation and validation of new community well-being metrics.

Outcome evaluation at scale is essentially corporate social science. The *IEEE 7010 Recommended Practice Standard for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* proposes what amounts to a difference-in-differences design between users and non-users before and after an algorithmic change (Schiff et al. 2020). This is a promising approach, but there do not seem to be any published examples.

Business Implications

For commercial AI systems, metrics-driven changes must also integrate legitimate business concerns such as the cost of implementation and the effects on business outcomes. Although a naïve analysis of multi-objective optimization suggests that considering anything other than revenue can only reduce revenue, this assumes everything else is equal. In reality there are complex secondary effects, such as changes in user and supplier behavior. YouTube's experience demoting clickbait videos is a documented case where doing the *responsible* thing led to a short-term decrease in the primary watch time metric, but then a long-term increase. It is difficult to predict the financial effects of incorporating well-being into optimization. Business and social objectives may turn out to be aligned, but this cannot be expected to be true as a rule. While ethical outcomes can sometimes be achieved through changes to optimization goals, there are also situations that could conceivably require avoiding features, products, or business models altogether (Barocas et al. 2020).

Case studies are one promising avenue for progress on the problem of uncertain business outcomes. If companies are already incorporating well-being metrics into their management and algorithms then documenting these cases will let others learn from their experiences, develop the field, and normalize the idea that companies should proactively manage the effects of their optimizers. This underscores the need for transparency around work that is explicitly designed to improve the lives of great numbers of people.

Conclusion

This paper has explored the integration of community well-being metrics into commercially-operated optimizing systems. Community well-being is an attractive goal because it is well-developed in public policy contexts and practically measurable. At least two large technology companies, Facebook and YouTube, have explicitly

modified their objective functions in pursuit of well-being, demonstrating the practicality of this approach.

There are still a number of weaknesses in the interventions that Facebook and YouTube have undertaken, at least in terms of what has been reported publicly. The community that these interventions are intended to serve has not been well defined; rather, these metrics and interventions are oriented towards the individual level and do not account for existing communities such as cities or discussion groups. It is not clear if or how users were engaged in selecting the *meaningful social interactions* and *user satisfaction* metrics; there is no report of the outcomes either in terms of these metrics or with respect to broader well-being metrics; and although both companies reported reduced short term engagement, the broader business effects have not been discussed.

However incomplete, the Facebook and YouTube cases suggest that the optimization of community well-being metrics may be a powerful general method for managing the societal outcomes of commercial AI systems. The same methods could be applied to many other types of systems, such as a news recommender system that incorporates measures of content diversity in an attempt to increase tolerance and reduce polarization, or an online shopping platform that uses product-level estimates of carbon footprint to steer users toward more environmentally friendly purchases. Although many scholars and critics have stressed the importance of increased user control over AI systems, no amount of user control can replace appropriate well-being metrics due to issues of collective action and the need for reasonable defaults.

An analysis of the above cases suggests that the following multi-step process may be effective:

Identify a community to define the scope of action. In online settings this may be a challenging decision.

Select a well-being metric, perhaps from existing frameworks. This stage frames the problem to be solved in concrete terms, so it may be where community involvement matters most.

Use this metric as a performance measure for the team building and operating the system.

Directly translate the metric into code as a modification to an algorithmic objective function or use these measured outcomes to evaluate more general design changes.

Evaluate the results, in terms of actual human outcomes, and adjust accordingly. This may require adjusting the chosen metric in response to changing conditions, or if it is found to be causing side effects of its own.

Require transparency throughout to make participation possible and to hold companies accountable to the communities who are meant to be served by this process.

Funding The author is an employee of Partnership on AI. Partnership on AI is supported by donations from companies and philanthropies, including Facebook and Google. The author did not receive funding from Facebook or Google for the creation of this article.

Data Availability N/A

Code AvailabilityN/A

Compliance with Ethical Standards

Conflict of Interest The author declares that they have no conflicts of interest.

Ethics Approval This paper does not include any studies with human participants or animals performed by the author.

References

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

Andreassen, C. S. (2015). Online social network site addiction: a comprehensive review. *Current Addiction Reports, 2*(2), 175–184. https://doi.org/10.1007/s40429-015-0056-9.

Bagnall, A., South, J., Mitchell, B., Pilkington, G., & Newton, R. (2017). *Systematic scoping review of indicators of community wellbeing in the UK*. 1–71. http://eprints.leedsbeckett.ac.uk/5238/1/community-wellbeing-indicators-scoping-review-v1-2-aug2017.pdf.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedingss of the National Academy of Sciences of the United States of America, 115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115.

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

Barocas, S., Biega, A. J., Fish, B., Niklas, J., & Stark, L. (2020). When not to design, build, or deploy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 695–695. https://doi.org/10.1145/3351095.3375691.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI and Society, 35*(1), 165–176. https://doi.org/10.1007/s00146-017-0760-1.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: from design methods to systems engineering. *Interacting with Computers, 1*, 4–17. https://doi.org/10.1016/j.intcom.2010.07.003.

Bergen, M. (2019). YouTube executives ignored warnings, Let Toxic Videos Run Rampant. *Bloomberg*. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant.

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., … Zueger, T. (2020). *Diversity in News Recommendations*. http://arxiv.org/abs/2005.09495.

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. *Nature, 489*(7415), 295–298. https://doi.org/10.1038/nature11421.

BSI. (2011). *PAS 2050:2011 Specification for the assessment of the life cycle greenhouse gas emissions of goods and services*. https://shop.bsigroup.com/en/Browse-By-Subject/Environmental-Management-and-Sustainability/PAS-2050/.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly, 80*(Specialissue1), 250–271. https://doi.org/10.1093/poq/nfw007.

Chen, R., Hua, Q., Chang, Y. S., Wang, B., Zhang, L., & Kong, X. (2018). A survey of collaborative filtering-based recommender systems: from traditional methods to hybrid methods based on social networks. *IEEE Access, 6*, 64301–64320. https://doi.org/10.1109/ACCESS.2018.2877208.

Crowder, D. W., & Reganold, J. P. (2015). Financial competitiveness of organic agriculture on a global scale. *Proceedings of the National Academy of Sciences of the United States of America, 112*(24), 7611–7616. https://doi.org/10.1073/pnas.1423674112.

Dantzig, G. B. (1982). Reminiscences about the origins of linear programming. *Operations Research Letters, 1*(2), 43–48. https://doi.org/10.1016/0167-6377(82)90043-8.

Delgado, J., Lind, S., Radecke, C., & Konijeti, S. (2019). Simple objectives work better. *Workshop on Recommendation in Multi-stakeholder Environments, RecSys*. http://ceur-ws.org/Vol-2440/paper5.pdf.

Diener, B. E., Oishi, S., & Tay, L. (2018). *Handbook of well-being*. DEF Publishers. nobascholar.com.

Doerr, J. E. (2017). *Measure what matters: How Google, Bono, and the Gates Foundation rock the world with OKRs*. Portfolio Penguin.

Donahoe, E., & Metzger, M. M. (2019). Artificial intelligence and human rights. *Journal of Democracy, 30*(2), 115–126. https://doi.org/10.1353/jod.2019.0029.

Durand, M. (2015). The OECD better life initiative: How's life? And the measurement of well-being. *Review of Income and Wealth, 61*(1), 4–17. https://doi.org/10.1111/roiw.12156.

Exton, C., & Shinwell, M. (2018). Policy use of well-being metrics: describing countries' experiences. *OECD Statistics Working Papers, 33*(94). https://doi.org/10.1787/d98eb8ed-en.

Facebook. (2018). *Facebook, Inc. (FB) Fourth Quarter and Full Year 2017 Results Conference Call*. https://investor.fb.com/.

Fazey, I., Carmen, E., Chapin, F. S., Ross, H., Rao-Williams, J., Lyon, C., Connon, I. L. C., Searle, B. A., & Knox, K. (2018). Community resilience for a 1.5 °C world. *Current Opinion in Environmental Sustainability, 31*, 30–40. https://doi.org/10.1016/j.cosust.2017.12.006.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1.

Foster, J., & Sen, A. (1997). *On economic inequality*. Clarendon Press.

Frey, B. S., Benesch, C., & Stutzer, A. (2007). Does watching TV make us happy? *Journal of Economic Psychology, 28*(3), 283–313. https://doi.org/10.1016/j.joep.2007.02.001.

Gabriel, I. (2020). *Artificial intelligence, values, and alignment*. https://arxiv.org/abs/2001.09768.

Garimella, V. R. K., & Weber, I. (2017). A long-term analysis of polarization on twitter. *Proceedings of the 11th international conference on web and social media, ICWSM 2017,* 528–531. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15592.

Ginsberg, D., & Burke, M. (2017). *Hard questions: Is spending time on social media bad for us?* Facebook. https://about.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us/.

Google. (2019). *How Google Fights Disinformation*. https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf.

GSSB. (2016). *Global Reporting Inititative 101: Foundation*. https://www.globalreporting.org/standards/gri-standards-download-center/gri-101-foundation-containing-standard-interpretation-1/.

Heaven, W. D. (2020). *Our weird behavior during the pandemic is messing with AI models*. MIT Technology Review. https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700.

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

ISO. (2018). *Greenhouse gases — Carbon footprint of products — Requirements and guidelines for quantification (ISO 14067:2018)*. International Organization for Standardization. https://www.iso.org/standard/71206.html.

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

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

Jannach, D., & Adomavicius, G. (2016). Recommendations with a purpose. *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*, 7–10. https://doi.org/10.1145/2959100.2959186.

Kaplan, R. S. (2009). Conceptual foundations of the balanced scorecard. *Handbook of Management Accounting Research, 3*, 1253–1269. https://doi.org/10.1016/S1751-3243(07)03003-9.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology, 4*(August), 1–14. https://doi.org/10.3389/fpsyg.2013.00513.

Korinek, A., & Stiglitz, J. E. (2017). Artificial intelligence and its implications for income distribution and unemployment. In *National Bureau of Economic Research*. https://doi.org/10.7208/chicago/9780226613475.003.0014.

Kulynych, B., Overdorf, R., Troncoso, C., & Gürses, S. (2020). POTs: Protective Optimization Technologies. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372853.

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

Lalmas, M., & Hong, L. (2018). Tutorial on metrics of user engagement: Applications to news, search and E-commerce. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining,* 3, 781–782. https://doi.org/10.1145/3159652.3162010.

Ledwich, M., & Zaitsev, A. (2020). *Algorithmic extremism: Examining YouTube's rabbit hole of radicalization*. https://doi.org/10.5210/fm.v25i3.10419.

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

Manetti, G. (2011). The quality of stakeholder engagement in sustainability reporting: empirical evidence and critical points. *Corporate Social Responsibility and Environmental Management, 18*, 110–122. https://doi.org/10.1002/csr.255.

Manheim, D., & Garrabrant, S. (2018). *Categorizing Variants of Goodhart's Law*. 1–10. http://arxiv.org/abs/1803.04585.

Meinrenken, C. J., Kaufman, S. M., Ramesh, S., & Lackner, K. S. (2012). Fast carbon footprinting for large product portfolios. *Journal of Industrial Ecology, 16*(5), 669–679. https://doi.org/10.1111/j.1530-9290.2012.00463.x.

Milano, S., Taddeo, M., & Floridi, L. (2019a). *Ethical aspects of multi-stakeholder recommendation systems*. https://ssrn.com/abstract=3493202.

Milano, S., Taddeo, M., & Floridi, L. (2019b). *Recommender systems and their ethical challenges*. https://ssrn.com/abstract=3378581.

Miller, J., Milli, S., & Hardt, M. (2019). *Strategic classification is causal modeling in disguise*. http://arxiv.org/abs/1910.10362.

Mosseri, A. (2018). *Bringing people closer together*. Facebook. https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/.

Musikanski, L., Phillips, R., & Jean Crowder. (2019). *The happiness policy handbook: How to make happiness and well-being the purpose of your government*. Gabriola: New Society Publishers.

Musikanski, L., Rakova, B., Bradbury, J., Phillips, R., & Manson, M. (2020). Artificial intelligence and community well-being: a proposal for an emerging area of research. *International Journal of Community Well-Being, 3*, 39–55. https://doi.org/10.1007/s42413-019-00054-6.

O'Donnell, G., Deaton, A., Durand, M., Halpern, D., & Layard, R. (2014). *Wellbeing and Policy*. Legatum Institute. https://li.com/reports/the-commission-on-wellbeing-and-policy/.

OECD. (2019a). *Artificial intelligence in society*. OECD Publishing. https://doi.org/10.1787/eedfee77-en.

OECD. (2019b). *Measuring well-being and progress*. https://www.oecd.org/sdd/OECD-Better-Life-Initiative.pdf.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives, 14*(3), 137–158. https://doi.org/10.1257/jep.14.3.137.

Paraschakis, D. (2017). Towards an ethical recommendation framework. *Proceedings of the International Conference on Research Challenges in Information Science 2017*, 211–220. https://doi.org/10.1109/RCIS.2017.7956539.

Parmenter, D. (2020). *Key performance indicators: Developing, implementing, and using winning KPIs* (4th ed.). Wiley.

Phillips, R., & Pittman, R. H. (Eds.). (2015). *An introduction to community development*. Routledge.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8.

Reader, B., & Hatcher, J. A. (Eds.). (2011). *Foundations of community journalism*. SAGE Publications Inc.

Richardson, J. (2013). Accounting for sustainability. In A. Henriques & J. Richardson (Eds.), *The triple bottom line - Does it all add up?* (pp. 34–44). Routledge.

Robertson, S., & Salehi, N. (2020). What if I don't like any of the choices? The limits of preference elicitation for participatory algorithm design. *Participatory Approaches to Machine Learning Workshop, ICML 2020*. http://arxiv.org/abs/2007.06718.

Roitero, K., Carterette, B., Mehrotra, R., & Lalmas, M. (2020). *Leveraging behavioral heterogeneity across markets for cross-market training of recommender systems*. 694–702. https://doi.org/10.1145/3366424.3384362.

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., … Bengio, Y. (2019). *Tackling climate change with machine learning*. http://arxiv.org/abs/1906.05433.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach 3rd edition*. Prentice Hall.

Samuelson, W. F., & Marks, S. G. (2014). *Managerial economics* (8th ed.). Wiley.

Schiff, D., Murahwi, Z., Musikanski, L., & Havens, J. C. (2019). A New Paradigm for Autonomous and Intelligent Systems Development : Why Well-being Measurement Matters. *Workshop on Designing Digital Wellbeing, CHI 2019*. https://digitalwellbeingworkshop.files.wordpress.com/2019/04/02-wellbeing_measurement_schiff_murahwi_musikanski_havens.pdf.

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

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

Stoica, A.A., & Chaintreau, A. (2019). Hegemony in social media and the effect of recommendations. *The Web Conference 2019*, 2, 575–580. https://doi.org/10.1145/3308560.3317589.

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

Sung, H., & Phillips, R. G. (2018). Indicators and community well-being: exploring a relational framework. *International Journal of Community Well-Being, 1*(1), 63–79. https://doi.org/10.1007/s42413-018-0006-0.

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

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review, 11*(1), 274–302. https://doi.org/10.1111/sipr.12033.

Wojcicki, S. (2019). *Preserving openness through responsibility*. Inside YouTube Blog. https://blog.youtube/inside-youtube/preserving-openness-through-responsibility.

Yu, T., Shen, Y., & Jin, H. (2019). A visual dialog augmented interactive recommender system. *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, 157–165. https://doi.org/10.1145/3292500.3330991.

Zhao, Z., Chi, E., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., & Yi, X. (2019). Recommending what video to watch next: a multitask ranking system. *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*, 43–51. https://doi.org/10.1145/3298689.3346997.

Zuckerberg, M. (2018). *No Title*. https://www.facebook.com/zuck/posts/10104413015393571.