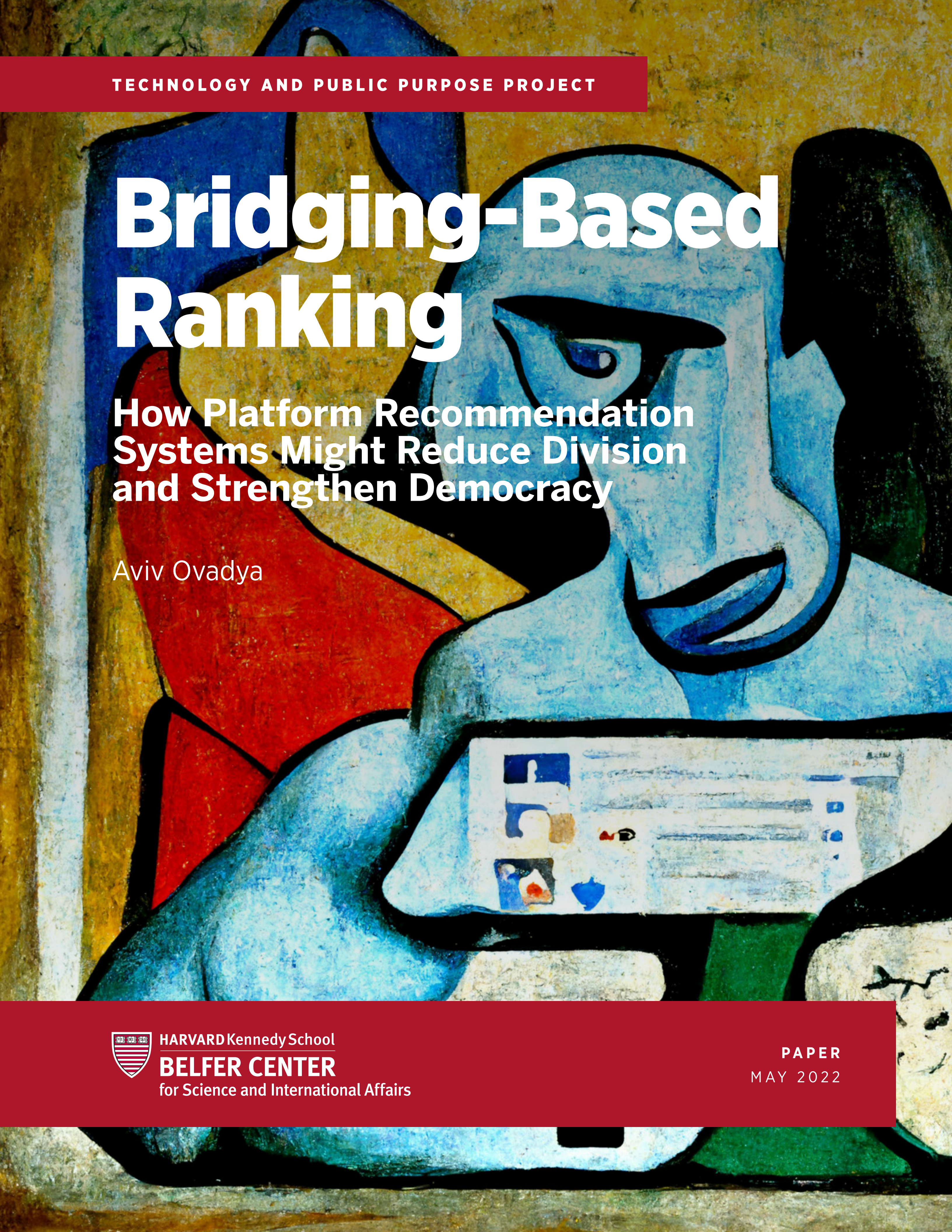


TECHNOLOGY AND PUBLIC PURPOSE PROJECT

Bridging-Based Ranking

How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy

Aviv Ovadya

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
MAY 2022

Technology and Public Purpose Project
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

belfercenter.org/project/technology-and-public-purpose

TECHNOLOGY AND PUBLIC PURPOSE PROJECT

Bridging-Based Ranking

How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy

Aviv Ovadya

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
MAY 2022

About the Technology and Public Purpose Project (TAPP)

The arc of innovative progress has reached an inflection point. It is our responsibility to ensure it bends towards public good.

Technological change has brought immeasurable benefits to billions through improved health, productivity, and convenience. Yet as recent events have shown, unless we actively manage their risks to society, new technologies may also bring unforeseen destructive consequences.

Making technological change positive for all is the critical challenge of our time. We ourselves — not only the logic of discovery and market forces — must manage it. To create a future where technology serves humanity as a whole and where public purpose drives innovation, we need a new approach.

Founded by Belfer Center Director, MIT Innovation Fellow, and former U.S. Secretary of Defense Ash Carter, the TAPP Project works to ensure that emerging technologies are developed and managed in ways that serve the overall public good.

TAPP Project Principles:

1. Technology's advance is inevitable, and it often brings with it much progress for some. Yet, progress for all is not guaranteed. We have an obligation to foresee the dilemmas presented by emerging technology and to generate solutions to them.

2. There is no silver bullet; effective solutions to technology-induced public dilemmas require a mix of government regulation and tech-sector self-governance. The right mix can only result from strong and trusted linkages between the tech sector and government.

3. Ensuring a future where public purpose drives innovation requires the next generation of tech leaders to act; we must train and inspire them to implement sustainable solutions and carry the torch.

About the Author

Aviv Ovadya is an internationally recognized expert on the societal implications of emerging technology, with a focus on internet platforms and artificial intelligence. His underlying goal, as aptly described in the New York Times, is to "incentivize...systemic changes" — to help create an information ecosystem that facilitates understanding and trust. Aviv writes, builds, advises, and conducts research as a Technology and Public Purpose fellow in Harvard's Belfer Center, as the founder of the Thoughtful Technology Project, as a non-resident fellow at the Alliance for Securing Democracy, and through his consulting work.

Previously he was the founding Chief Technologist at the Center for Social Media Responsibility at the University of Michigan, a Knight News Innovation Fellow at Columbia University, and a TED CIVIC Resident. He consults for organizations as varied as the Partnership on AI, Snopes, and Meedan, and presents his work around the world, to technology platforms, executives, and governments.

His current work navigates the interactions between many problems, tools, and domains — these include misinformation, synthetic media (e.g. deepfakes, language models), content moderation, recommendation engines, metrics, contextualization, platform design, governance, media literacy, alignment, and the operationalization of tech ethics.

Aviv's work has been covered regularly, including by the BBC, NPR, Yomiuri Shimbun, the Economist, and the New York Times and his writing has been published by the Washington Post, HBR, MIT Technology Review, and Bloomberg.

He can be found on the web at aviv.me, through his newsletter, and on Twitter as @metaviv.

Table of Contents

Executive Summary

The problem

Algorithmic ranking and recommendation systems determine what kinds of behaviors are rewarded by digital platforms like Facebook, YouTube, and TikTok by choosing what content to show to users. Because these platforms dominate our attention economy, and because attention can be transformed into money and power, *platform recommendations therefore provide a reward structure for society at large*.

Platforms currently reward divisive behavior with attention due to the interactions between *engagement-based ranking* and human psychology. This helps determine the kinds of politicians, journalists, entertainers, and others who can succeed in their respective social arenas, resulting in significant impacts on the quality of our decision-making, our capacity to cooperate, the likelihood of violent conflict, and the robustness of democracy.

The opportunity

We can potentially mitigate this 'centrifugal' force toward division by deploying ranking systems that do the opposite—that provide a countervailing 'centripetal' or bridging force.

Bridging-based ranking **rewards behavior that bridges divides.** For example, imagine if Facebook rewarded content that led to positive interactions across diverse audiences, including around divisive topics.[1] How might that change what people, posts, pages, and groups are successful?

This report explores the potential of bridging and discusses some of the most common objections, addressing questions around legitimacy and practicality. It contrasts bridging with some of the most discussed approaches for reforming ranking: reverse-chronological feeds, 'middleware', and 'choose your own ranking system'. (Unfortunately, without introducing bridging, all of these proposed reforms still reward those who seek to divide.) Finally, this report explores early examples where bridging systems are already being tried with some success.[2]

1 What 'positive interactions' means will depend on the product and bridging implementation. It may be explicit based on user reactions and comments or implicit based on user behavior.

2 While we do not go into the technical details here, a collaboration will be publishing on this shortly—and the example of Twitter Birdwatch includes the complete code for their limited experimental implementation.

Summary of next steps

We can and should rapidly build capacity to develop, evaluate, and deploy bridging-based ranking systems.

- [ ] Governments, platforms, funders, and researchers must direct resources towards this goal.
- [ ] We specifically call for platforms to measure the extent to which their products divide people (bridging metrics), and to include both bridging metrics and bridging-based ranking into their product roadmaps and quarterly goals.
- [ ] To address legitimacy and platform power concerns, we suggest putting the ultimate question of 'what recommendation systems should reward' to the impacted populations through platform democracy. We further argue that the default should not be divisive engagement-based ranking, or chronological feeds (that reward those who post the most), but bridging-based ranking (as it actively mitigates divisive tendencies).
- [ ] We must involve interdisciplinary scholars and practitioners to ensure that what we create is truly beneficial for the public and democracy.

Bridging-based ranking alone is not a silver bullet—we need other reforms to address the many challenges of platform-enabled connectivity. But bridging would help address one of the most significant risks—that of being pushed past a "division threshold" beyond which democracy can no longer function.

Introduction

Ranking and recommendation systems, also known as *recommender systems* and *recommendation engines*, are one of the primary ways that we navigate the deluge of information from products like YouTube, Facebook, Amazon, and TikTok. In this deluge, recommendation systems help allocate our finite time and attention across the zettabytes of data (trillions of gigabytes) now produced each year (Statista, 2021). They can be thought of as automated content curators, with their most obvious role being to choose a small set of items from a much larger set and show that small set to a particular person.

The key question that ranking answers: *What is rewarded?*[3]

However, ubiquitous recommendation systems have far more significant societal impacts than simply choosing among a set of pre-existing items to display. Recommendation systems direct attention—and attention is a currency that can be converted into money, power, and status. Much of the focus of the current conversation around recommendation systems is around "algorithmic amplification" and the ways that they influence what content is *consumed*. However, even more important may be that recommendation systems can have *enormous impacts on what content is produced.*[4]

3 This author finds the question of *"What is rewarded?"* as particularly crucial, as he sees it as one of the fundamental 'compass questions' around the impacts of social media and related technologies.

4 The work of Lischka et al. (2021) and Tennenholtz et al. (2019) among others explore the game-theoretic dynamics of ranking and content production in much more detail.

At their core, recommendation systems decide what kinds of *behavior* a product will reward—and their impacts are the result of a very messy human world. In other words, recommendation systems *provide a reward structure for society*.

The purpose of this report is to outline some of the challenges with current approaches for improving this 'reward structure' and to argue for an alternative—*bridging-based ranking*—which explicitly seeks to counteract the divisive properties of current ranking systems.[5]

Key Terms

Algorithm: Any well-defined process or procedure. Most used to describe fully automated systems, but also applicable to procedures that involve people, such as voting.

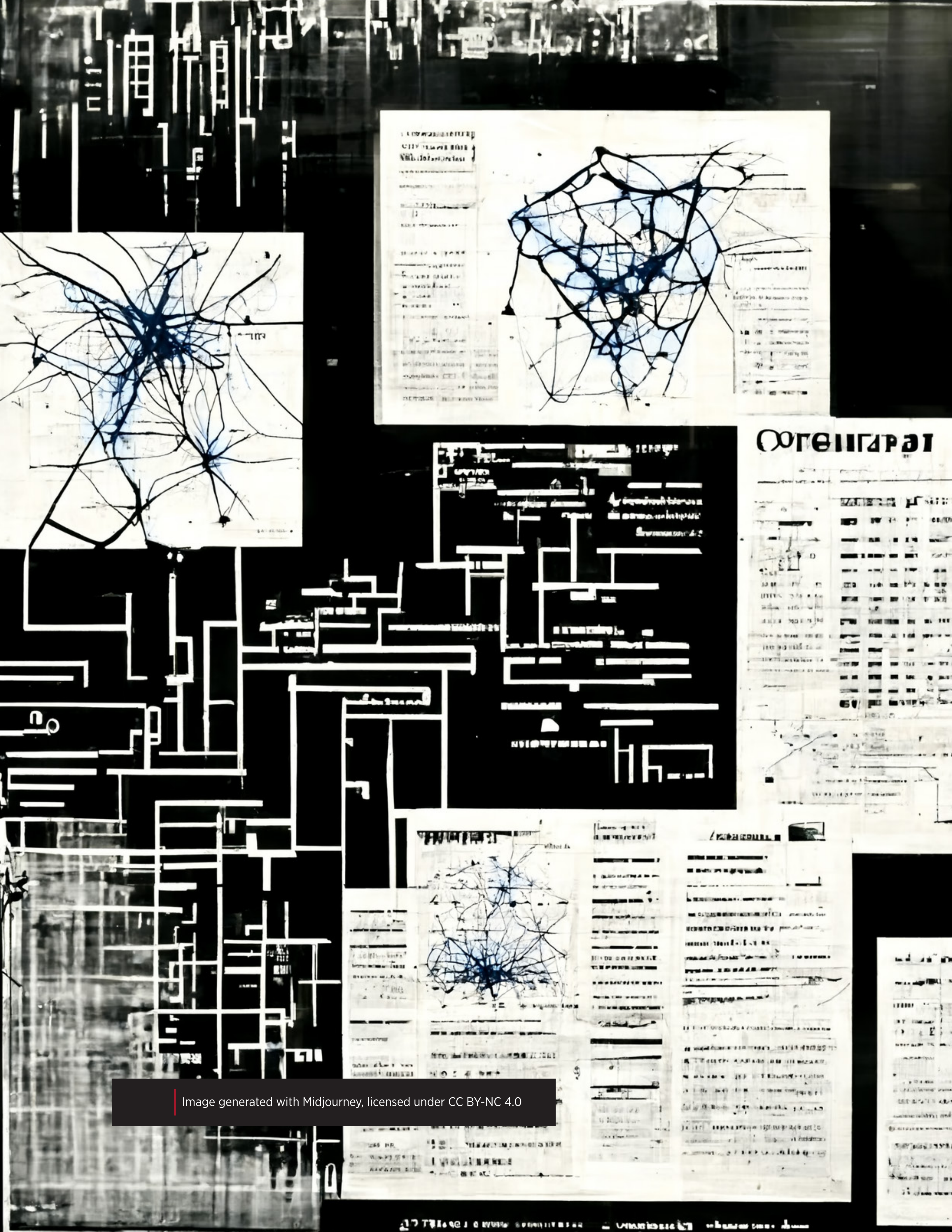
Ranking system: An algorithm used to choose and rank a set of items from a much larger set for a particular context (e.g., for a particular user in a particular product); ranking systems may or may not take explicit user-provided input such as a search term.[6]

Recommendation system: A personalized ranking system that does not take explicit input, e.g., Facebook Feed, TikTok's 'For You Page', YouTube's next video recommendations. (Also known are recommender systems or recommendation engines.)[7]

5 Given the limited access to platform data, there is some academic disagreement on the extent to which recommendation systems currently foment division, and there is significant nuance on exactly what is meant by division (Eckles, 2021; Haidt & Bail, 2021). This report assumes that different choices for ranking systems do have impacts for division and bridge-building, as described informally below.

6 Note that under this definition, democratic elections are a form of ranking system. Different election systems use different algorithms for determining the winner(s).

7 In this paper we use "ranking" and "recommendation" somewhat interchangeably as the content applies to the more general ranking, but recommendation appears the more popular term in policy and is used beyond its original definition. Computer science texts may use a much more constrained definition of a "recommender/recommendation system/problem", focused on predicting what items a user will rate highly (Aggarwal, 2016). However, this definition does not encompass many real-world uses of recommendation systems where the system objectives can vary dramatically depending on their context and impact.



Background and Context

A dominant approach: “engagement-based ranking”[8]

Recommendation systems for the prominent user-generated content platforms commonly use what has become referred to as “engagement-based ranking” (Mac, 2021). They aim to predict which content will lead a user to engage the most—for example by interacting with the content or spending more time using the product. This content is ranked higher and is the most likely to be shown to the user. The idea is that this will lead to more time using the company’s product, and thus ultimately more time viewing ads. Non-ads-based products like Netflix also use recommendations to provide value to their subscribers by showing them content they will enjoy.

The problem with engagement-based ranking: many externalities

While engagement-based ranking may be good for business, appears to serve users[9], and is relatively straightforward to implement, it is also likely to be a rather harmful approach with many negative externalities. In particular, engagement-based ranking can increase

8 Engagement itself may have many definitions which may even be in conflict with each other. For example, there may be more clicks but less time on a site due to a change in the algorithm or product design, as discussed below.

9 In some sense, engagement-based ranking is just a form of “giving people what they want”, which can seem—at least at first glance—like a reasonable and philosophically legitimate approach to building a company: the company that is explicitly connecting people to their desires.

divisiveness, misinformation, out-group attacks, and addiction while drowning out critical information and deeper human connection.[10]

It acts as a sort of content and societal centrifuge, bringing people more and more of what they already like—going deeper instead of broader—and incentivizes content producers to fulfill those surface-level desires. This has some benefits, but can unwittingly tear apart families as loved ones fall down deep 'recommendation holes'. At a societal level, as sources of distrust are emphasized repeatedly to subpopulations, societal divisions widen and the social contract supporting a democracy can falter, leading to internal strife, the breakdown of institutions, and potentially even mass violence.

If we were all 'hyper-rational Spocks' perhaps we would only pay attention to factual things that truly mattered—and engagement-based ranking would be great at rewarding the best content. However, humans have good evolutionary reasons to pay attention when, for example, someone says that our families are in danger from a malevolent enemy—and so messy humans can end up very engaged when we see sensationalism and divisiveness, regardless of its truthfulness.

The harmful feedback loop of sensationalism and engagement

In summary, sensationalism and divisiveness attract attention. This drives engagement. Engagement-based ranking systems reward this with more attention. The resulting feedback loop provides a strong incentive for anyone desiring money, status, or power to produce sensationalist and divisive content. This includes politicians and journalists. Even worse, non-sensationalist content is less likely to be seen, creating a very strong

10 For the sake of brevity, this and the following paragraphs present a simplified account of a very complex set of questions with limited data for answering them within the potentially necessary timeframe to act. A systematic review of the global literature suggests that while 'digital media' in general does e.g., increase political participation and information consumption, it also increases polarization, misinformation, populism, etc.; and some of this is context dependent (Lorenz-Spreen et al., 2021). Meta has referred to evidence suggesting that in some countries, access to the internet has led to less polarization which may suggest that recommendations are less divisive (Boxell et al., 2020), but the limitations on such a study invite some skepticism and suggest that other methodologies may be more appropriate. There is concrete evidence that e.g. moral emotional language (Brady & Bavel, 2021) and out-group attacks (Steve Rathje et al., 2021) increase engagement. Regardless of the specifics, there are significant psychological causal mechanisms that support these claims, to the extent that scholars studying civil war believe that such technologies have played a significant role in the current wave of democratic decline and polarization (Walter, 2022).

disincentive to produce grounded, nuanced, and fact-based content. All of this leads rapidly to a race to the bottom.

It is this combination of human psychology (what we pay attention to) and societal incentives (our desire for attention and its rewards) that leads to harm; engagement-based recommendations are just a particular way to increase the reward and thus the harm.

Key Insight

Engagement based-ranking rewards conflict creators over bridge builders.

How ranking actually works: metrics and weighting[11]

In practice, most mature recommendation systems are not purely engagement-based. They take into account many different ranking signals—and these different signals all have different *weightings* (i.e., the extent to which they impact what is actually recommended). There are also many definitions and forms of engagement depending on the particular product, and different *metrics* can be used to measure the "success" of a ranking system.

For example, in 2012 YouTube's core business metrics were refocused from maximizing the number of 'video views' to maximizing the amount of 'watch time' (Doerr, 2018). This perhaps is a subtle difference, but had significant impacts on what kinds of recommendation systems were built, and thus what sort of content was rewarded, even though both are engagement-based ranking. As a very specific example, focusing on watch time encouraged the creation of longer videos; some allege that this helped make YouTube a conspiracy and radicalization engine.

11 Thorburn et al. provide a more in-depth but still succinct overview of recommendation system design (2022).

If recommendation systems create a reward function for society, it is *metrics that define the reward function for the recommendations systems themselves*. Core metrics such as YouTube's 'watch time' and Facebook's 'meaningful social interactions' (MSI) not only influence the design of recommendation systems. They also influence the incentives and structures of the organization themselves (Mac, 2021; Ovadya, 2021a).

Key Terms

Ranking signal: Any signal which is used by a ranking system; this might be based on e.g. the item, the user, the context, how similar users interacted with this item, etc.

Ranking signal weights: Mature ranking algorithms usually incorporate many ranking signals; the weight for a signal determines the influence of that signal on the final ranking.

Engagement-based ranking: A ranking algorithm that significantly weights items likely to increase engagement (e.g. clicks, interactions) over other factors (this is a more informal term used to describe a class of algorithm).

Metric: A quantitative objective for the system.

Insufficient 'fixes': chronological feeds and middleware

There are a number of approaches that have been proposed to address the externalities of recommendation systems. Unfortunately, as we will see, most of the commonly cited approaches come up short when it comes to addressing divisiveness.

Chronological feeds are just 'recency-based ranking' systems with their own problems

The perils of engagement-based ranking have led some advocates, policymakers, and even former tech employees to want to replace recommendation systems with chronological feeds.[12] This appears to make sense at first glance. If engagement-based recommendation systems have many externalities, then it seems ideal to simply rank posts by time, completely ignoring engagement and every other factor.

However, while chronological feeds address some of the problems with engagement ranking, they also cause many problems. On one hand, time-ordered feeds may help by bringing us closer to the baseline incentives for sensationalism. On the other hand, if there existed a "divisiveness threshold"—defined as a degree of divisive sensationalism that is simply too much for a democracy to survive—it is eminently plausible that one could get there without even needing engagement-based ranking to supercharge it.[13] While it would be difficult to prove this, there are at least significant indications, e.g. in Kenya, India, and Brazil, that messaging systems such as text messaging and WhatsApp share many of the same problems as engagement-based social media—even though they use a chronological feed.

12 Technically "reverse chronological feeds" as they show the most recent content first.

13 Notably, simulations predict that there are "polarization tipping points" which would be very difficult to reverse (Macy et al., 2021)2021.

One of the reasons is that *chronological feeds are a simple form of recommendation system, ranked by recency*. We think of them as "*recency-based ranking*."[14] Since recommendation systems determine "What is rewarded?" and recency is rewarded, this means that chronological feeds will primarily reward those who post the most. This is something that engagement-based ranking systems can help mitigate, by not showing too many things from the same person. It's not clear that this would lead to better outcomes—and there is evidence that at least in some cases, Facebook switching to a chronological feed made things worse.

Chronological feeds also only apply to particular kinds of products (e.g., feeds where one has explicitly chosen accounts to follow), and do not provide an answer on how recommendations outside of that context should operate for products like YouTube's video recommendations, Facebook Groups, Twitter follows, or TikTok shorts. They also provide no answers for what new users are recommended before they follow anything (the 'cold start problem')—and if nothing is recommended at all, this provides significant advantages to incumbent platforms.

Key Insight

Chronological feeds reward those who post the most.

Where chronological feeds are applicable, they are the most minimal possible "fix"—and that may not be good enough. While they avoid one of the harmful feedback loops caused by engagement-based recommendations, a movement to chronological feeds doesn't even attempt to mitigate and overcome the underlying problems.

"Choose your own ranking systems" are not quite good enough

Another popular approach to addressing the harms of recommendation systems is enabling individual users to choose their own recommendation systems. This could be implemented by allowing users to choose amongst

14 Or recency-based ranking.

a set of options for their recommendations, including e.g., chronological, engagement-based, and so on. Relatedly, platforms could be required to let third parties do ranking on their behalf—the "middleware" recommendation approach advocated by Fukuyama et al (2021).[15]

This is the kind of approach that this author would *personally* be excited about—it would be wonderful to have full control over everything one sees. However, just because the author (perhaps selfishly) wants complete control over their environment does not mean that giving everyone that control is sufficient to support a stable democracy.

As the proponents of middleware recommendations themselves state:

> Empowering each individual to tailor their algorithms might encourage a further splitting of the American polity, allowing groups to more easily find voices that echo their own views, sources that confirm their factual beliefs, and political leaders that amplify their own fears.

Unfortunately, this is not the worst of it. As discussed earlier, recommendation systems provide a reward system for society. While such proposals could be effective at helping information *consumers* curate their own information sources, they will likely do little to address the problems for information *producers*. Proponents of "choose your own ranking system" should consider if such a system would significantly alter the current problematic incentives for journalists and politicians.[16]

Fukuyama et al. go on to argue that these risks are "outweighed by the dangers of concentrated platform power," but as we have seen over the past few years in the United States, the degree of divisiveness fomented by platforms is at least as large a risk to democratic functioning as that of concentrated

15 Other aspects of the middleware proposal may have fewer downsides, e.g., middleware as a supplement for content labeling.

16 The one significant caveat here is if there are very strict bounds on the properties of allowable recommendation systems, e.g. if the only options are essentially variants on the bridging-based ranking described later.

platform power.[17] Moreover, platform democracy makes this a false dichotomy; it is possible to both devolve platform power and reduce societal division as is described below (Ovadya, 2021b).

This doesn't mean that we can't have *some* individual choice over what our recommendation systems do. But it does mean that all of our choices must have as a baseline, mechanisms to address harmful incentives around production. You can sell any kind of car you want, as long as you first ensure it is unlikely to harm people during normal operation.[18]

17 Some would argue that at least in the United States, cable news and radio have had more of an impact on divisiveness than technologies like social media. This may be true to some extent, but three other things are also likely true: (a) social media changes the incentives on mainstream media providers, creating a pull to the extremes, (b) that any meaningful addition to divisiveness—or opportunity to reduce it—should be addressed if one is close to the threshold of democratic dysfunction (c) it may be easier to improve recommendation system and the incentives they provide than it is to directly improve the US cable news and radio ecosystem. Of course, it is also likely that some political systems are more likely to drive divisiveness than others, but fixing political systems outside of the platform context is beyond the scope of this work.

18 There are incidentally a number of other challenges with the middleware approach e.g. involving privacy, though there are partial solutions (Keller, 2021).

Bridging-Based Ranking

What would it look like to create a sort of opposite to an engagement-based ranking system? Something that explicitly aligns incentives such that sensationalism and divisiveness are no longer favorable—thus leveling the initial playing field so that bridge-building and nuance have a fighting chance?

We can call such a system a **bridging recommender**—*a recommendation system that rewards content that helps bridge divides*. It would use **bridging-based ranking** to reward content that leads to positive interactions across diverse audiences, even when the topic may be divisive.

For example, imagine two potential articles that Twitter's feed might show someone about immigration. One appears likely to increase divisions across opposite sides, another is more likely to decrease divisions. Engagement-based ranking would not try to take this into account, it would simply factor in how likely one is to engage or stay on the app—which is likely higher for a divisive article that leaves users ranting and doom-scrolling. Bridging-based ranking would instead reward the article that helps the opposing sides understand each other, and that bridges the divide.

Key Terms

Bridging model: A representation of the 'divisions and bridges' in a population.

Bridging metrics: Numeric measures of the extent of division, and of the impact of recommendations on that.

Bridging-based ranking: A ranking algorithm that rewards behavior that bridges divides.

For example, by rewarding content that leads to positive interactions across diverse audiences—including around divisive topics.[19] [20]

Bridging recommendation system: A recommendation system using bridging-based ranking.

19 This involves the system having some model of societal divides and then recommending content that is likely to decrease them; in contrast to engagement-based ranking, which has a model of what items lead to interactions and recommends content that is likely to increase those.

20 Note that this is not specific to recommendation systems, it is also applicable to search engines and any other system where ranking is necessary (including e.g. contextualization engines).

Another way to think about this is that engagement-based systems serve as a form of "centrifugal ranking," dividing users and encouraging the "othering" of different groups. In contrast, bridging-based ranking is a kind of "centripetal ranking"— helping re-integrate groups and foster trust.

If we can do this well, the implications are significant. Bridging-based ranking would reward the ideas and policies that can help bridge divides in our everyday lives, beyond just online platforms. Moreover, it can help change incentives on politicians, many of whom have been forced to play to the whims of the engagement-based algorithms (Horwitz, 2021). It might even make journalism more profitable and better able to realign with its professional ethos. Bridging could reduce the likelihood that fact-based reporting would be buried under a sea of cheap (but highly engaging) sensationalism and would reward empathic and bridge-building journalistic practices (Ripley, 2020). Finally, it may dramatically reduce the need for moderation. If division and hate are no longer rewarded with attention, there will likely be much less of it.

This provides a *highly simplified* illustration of the differences between traditional engagement-based ranking and bridging-based ranking. In this scenario, there are just two distinct parties, green and blue, where we assume everyone in each party reacts fairly similarly (and so engagement-based ranking using standard algorithms will automatically infer party affiliation). Alice, Bob, and Igor are green party members, Oscar and Wendy are blue party members.

Engagement-Based vs. Bridging-Based Ranking

(in a simplified two party society)

In Example 1, we see on the left how Alice, Bob, Oscar, and Wendy reacted (or did not react) to posts A, B, C, and D. The color of the post represents the party affiliation of the original poster. On the right, we can see the resulting ranking for the posts for Igor (also green party) under bridging-based vs. engagement-based ranking (posts toward the top are ranked higher).

With bridging, the post ranked the highest does not originate from a green party member but is the only post actively liked by both parties. The lowest-ranked post is the one that is clearly the most divisive.

Example 2 is similar but shows how additional kinds of interactions data might be used to determine ranking. The exact ranking outcomes in this slightly more complicated case will depend on algorithm and weighting details; here we give one potential set of outcomes.

This diagram is excerpted from a forthcoming technical working paper and was created in collaboration with Luke Thorburn. (Ovadya & Thorburn, 2022)

Illustrative examples

While there are no perfect examples of publicly disclosed bridging-based ranking 'in the wild' that the author is aware of, there are some exciting bright spots to call out. For example, there are reasons to believe that Facebook has already begun experiments along these lines, determining what comments receive a positive reaction across a diverse audience. These comments are then ranked higher, potentially decreasing the need for moderation.

Twitter has also done some experimentation with something analogous to bridging-based ranking in its 'Birdwatch' system, also in a content-agnostic way—suggesting that bridging may even help identify misinformation. Birdwatch lets users provide context (e.g., fact check) labels on tweets that "[take] into account […] whether people who rated it seem to come from different perspectives." It seems to work: "In surveys of people on Twitter in the US, the majority of respondents found notes designated 'currently rated helpful' by Birdwatch contributors to be "somewhat" or "extremely" helpful — this includes people from across the political spectrum" (Twitter, n.d.). In a peer-reviewed paper on Birdwatch, Allen et al. support this, writing that: "there is a strong, positive relationship between the percent of co-partisan ratings and the overall helpfulness rating of the note"—and that the ratings of sufficient Birdwatch users match fairly closely to that of professional fact-checkers (2021).

There are also promising examples outside of the social media platform context. Polis, for example, has been used around the world, most notably in Taiwan, in order to surface perspectives that bridge divides for policy-making (Small et al., 2021). Tools like Remesh have been used by the United Nations to similarly surface perspectives across divides in war-torn Libya (Brown, 2021).

It is important to note that all of these examples share a crucial property: their goal is not to push content to people "from the other side" in order to change them. There is evidence that suggests this can actually *increase* polarization (Bail et al., 2018). Instead, the goal of these systems is to surface content that is received positively across diverse audiences. This can build mutual understanding and trust, similar to what a human facilitator might do in some kinds of depolarization workshops. This creates a space that can then enable deeper conversations and deliberation, moving beyond vitriolic reactivity.

Challenges to deployment

The underlying goal we can aim for is to ensure that ranking supports the bridging of divides, or more generally 'rewards content that 'facilitates understanding, trust, and wise decision-making.' While making bridging-based ranking a reality at scale may not be easy, there are clear steps forward.

Implementation challenges

One of the biggest challenges to developing alternatives to engagement-based ranking systems is that many of the platforms and regions where the harms of recommendations are greatest are the same regions where software has limited capacity to understand the local languages. If bridging-based ranking required content analysis, that would limit its impact and potentially make it less effective for many formats such as audio and video. However, while challenging to implement, there are ways to adopt bridging-based ranking that are content-neutral. For example, like engagement-based ranking, it can rely primarily on modeling the patterns of interactions and sharing over time—just with a different purpose.

There are currently a number of different approaches to ranking that are being explored to develop recommendation systems that e.g. aim to reduce polarization (Stray, 2021), and some exciting directions for bridging are discussed in 'Existing Implementations' below and a forthcoming technical working paper.[21]

21 While this report is not meant to be a deep technical exploration, the author and collaborators aim to share a technical working paper within the next few months that explore more existing implementation approaches and the open research questions needed to advance the field (Ovadya & Thorburn, 2022).

Platform incentives

Of course, there is still the question of why a platform might even implement this in the first place. Even barring a government force, there are a number of self-interested reasons why bridging-based ranking might be of interest to platforms:

1. *Brand impact*: Being seen as a place for bridge-building is likely better for a brand than being seen as a place for outrage—which may also make the platform more valuable to advertisers concerned about their own brands.

2. *Regulator pressure relief*: If regulators don't see platforms as being as threatening to their constituents, they may be less likely to push for harsh fines or potentially counterproductive mitigations.

3. *Potential engagement benefits*: Depending on the specific context, there is a possibility that some forms of engagement may actually go down in the short run with bridging-based ranking—but it is also likely that engagement could go up in the long run as the experience of using a product becomes more positive, connecting, and meaningful.

4. *Decrease moderation costs*: If divisive attacks and sensationalist misinformation are rewarded, there will be a lot more of them—and therefore more content that will need to be reported and moderated. With companies like Meta paying over $500 million USD a year for content moderation from just one partner, decreases in moderation costs could significantly impact the bottom line.

Of course, these reasons alone may not be enough to get a company that already has something that is "working" to change—especially if managers with compensation tied to business metrics believe that there is any risk of a negative impact in the short term. Platform intrapreneurs, external pressure, and foundation-supported external pilots can all help incentivize and de-risk bridging enough to get to critical mass.

Addressing potential objections

However, while it may be heartening that recommendation systems *could* be made supportive of a healthy democracy, does that mean that platforms *should* deploy such systems?

System neutrality is a mirage

One of the often-implicit assumptions of both chronological feeds and "choose your own ranking system" as solutions is that there can be a sort of "neutrality" to recommendation systems. The advocates of such systems might be thinking that if everyone is just seeing the most recent things, or choosing whatever they want to see, there is no "hand on the scale". Unfortunately, there is no avoiding that hand on the scale. Any decision relating to recommendation systems—including one of those two 'fixes'—is not neutral and has significant societal implications. *Even requiring chronological feeds is a form of wide-scale social engineering*—it will impact what kinds of content is made and what kinds of people succeed (and those impacts may not be positive).

Arguably the 'most neutral' option would be *non-divisive recommendations*, which given the psychological and social factors mentioned earlier would still require some bridging (in order to counteract the divisive impacts that even appear to be present with chronological feeds).

Key Insight

Any technology that impacts society is implicitly a form of "social engineering."

There are no 'true defaults' and no neutrality.

The 'most neutral' approach likely still requires bridging.

The ‘philosophical legitimacy’ of recommendations

It is worth taking what may appear to be a short digression here to explore philosophy and legitimacy. The idea that a recommendation system can be neutral, or that giving an option to choose one’s own recommendation system can be neutral is based on a set of implicit philosophical assumptions of legitimacy, in addition to practical considerations.

Engagement-based ranking is considered by some to be legitimate because it “gives people what they want”—though most people do not “want to want” addiction or the erosion of democracy (among other potential downsides). Chronological feeds can be seen as a nod to the “legitimacy of simplicity.” Middleware recommendations are aligned with more libertarian and individualistic values. However, as we have discussed, none of these may be sufficient to sustain democracy.

The ideal way to provide deeper legitimacy for a particular recommendation approach would be through the end-game of the philosophy of political legitimacy: democratic decision-making (Peter, 2017)

‘Platform Democracy’ as an approach to address legitimacy challenges

Unfortunately, electoral democracies have their own legitimacy challenges, as recommendation systems can be used—and have been used—to entrench partisan power. The Facebook Papers include quotes such as “[the Public Policy team] commonly veto launches which have significant negative impacts on politically sensitive actors” (Edgerton et al., 2021). We must therefore also be cautious about giving political actors too much influence over recommendations.

If we do not fully trust incumbent politicians to govern the use of money in politics, the same should apply to the governance of attention—both directly enable the entrenchment of power. Platforms also cannot be fully trusted given their incentives to protect their political benefactors, for attention hoarding, and for profit. Traditional multi-stakeholder bodies can be co-opted by platforms or governments and may not have democratic legitimacy.

This seems to leave us with no alternative for legitimacy, but newer forms of democracy may provide a solution, as outlined by this author in the working paper "Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure"[22] (Ovadya, 2021b):

> Many purported fixes to platform problems involve giving individual users more options; e.g. if you want to see less sensationalist and divisive content, a platform might let you tweak your personal recommendations. But this individual agency "solution" does not solve the collective problems that the sensationalism and divisiveness might cause for a community, nation, or the planet—and it could even make those problems worse.

It is outside of the scope of this report to explain the mechanics in too much detail, so we just provide a brief summary of one of the most promising approaches to platform democracy: the representative deliberative process (e.g., citizens' assembly style processes). This democratic process may provide a robust approach to address such incentives, as they let stakeholders and experts present their perspectives to a representative sample of the impacted populations—i.e. to people who lack such strong vested interests (OECD, 2020). In other words, the ultimate choice of ranking approach can in some cases be given to the collective population being impacted to ensure that such systems are not co-opted by those in power. Such representative deliberative bodies can be convened by platforms, NGOs, governments[23], or ideally a combination (as they already have been for significant issues around the world from nuclear policy by South Korea, to climate change in France). This is not just idealism; several platforms are now seriously exploring the applicability of these such democratic processes and running limited pilots to better understand the operational mechanics of such processes.

Of course, given that these are democratic processes, it is very possible that the members of such a deliberative body will end up supporting alternative

22 The research that led to the report on platform democracy was also highly motivated by the challenge of legitimate decision-making for ranking and recommendations. Liscow et al. make a related proposal focused on the related behavioral economics applied to public policy (2022).

23 And audited by third parties.

approaches. That is the ideal of deliberative democracy—if such a body was to be convened around ranking for a platform, the role of this report would simply be to provide a strong foundation for its perspective—so that such a democratic body would make the best possible decision given the needs and values of its members.

Could bridging actually be bad for democracy?

There are also critiques that embedding bridging into our ranking systems is anti-democratic or anti-pluralistic (Clyde, 2022)[24]. The motivation of this work is of course to support democracy and pluralism, but it is plausible that bridging, implemented in a too heavy-handed way, might stifle the democratic process or hamper necessary activism and advocacy. As with engagement-based rating, the exact choices made are likely to have significant impacts on what kind of content is successful. One of the necessary areas of research includes work around distinguishing constructive conflict from destructive conflict at platform scale—something which is plausibly evident from user interactions and which can inform bridging system design. By clearly laying out metrics and goals for such systems, ideally through un-co-opted democratic processes, we can aim to ensure that bridging based-ranking—and ranking systems in general—act as a force in support of a vibrant and evolving democracy.

24 This reference refers to an earlier version of this proposal, so not all of it may remain applicable.

Next Steps and Action Items

Given the vast sea of information that we navigate every day, it is not possible to avoid the impacts of recommendation systems, both on consumers and producers. It would also be folly to give up on their potential when the recommendation system can be harnessed to nurture an attention economy that is not tilted to reward those who aim to divide us. We must at least level the playing field so that bridge-building has a fighting chance against divisiveness—and neither chronological feeds nor a free for all "choose your own ranking system" are likely to be sufficient.

There are two core technologies that can contribute to bridging:

- **Bridging Metrics** for measuring the extent that a recommendation system rewards divisive action vs. bridging action.
- **Bridging Methods** for weighting and ranking content in ways that support bridging action (as opposed to weighting primarily on engagement).

Resourcing: research, development, evaluation, deployment, and community

We need investment by both the public and private sectors in order to rapidly develop, deploy, and scale viable versions of these technologies for bridging-based ranking.

Governments, **platforms**, **funders**, and **researchers** should direct resources towards:

- [ ] Studying (and likely expanding) existing early-stage potentially bridging systems such as Twitter's Birdwatch and Facebook's 'diverse positive interactions comment ranking.'
- [ ] Accelerating research into approaches for viable implementation appropriate for different contexts (platform types, cultures, languages, etc.).
- [ ] Developing open-source implementations that can act as models and which can be directly adopted by startups.
- [ ] Evaluating efficacy and unintended consequences through both quantitative and qualitative methods.
- [ ] Carefully deploying these systems at global scale.[25]
- [ ] Sustaining a strong interdisciplinary community of practitioners and scholars working on refining recommendation systems in the service of democracy.[26]

25 As described below, this would ideally be the result of approval by representative deliberative bodies—but waiting to improve systems that were not created democratically embeds a status-quo bias, and there is no reason that engagement-based ranking is a more legitimate 'initial condition' than bridging-based ranking.

26 There is a tremendous need for interdisciplinary work around bridging systems, this should most certainly not be left entirely to computer scientists and user researchers; as just one example, theories of conflict management, conflict resolution, and conflict transformation will likely all prove valuable (Miall, 2004).

Platform action items: OKRs, roadmaps, and metrics

Platforms should implement bridging metrics and ranking as a default (and use platform democracy to navigate potential disagreements).

Concretely:

- [ ] Each of the action items in the section above should be tied to specific objective key results (OKRs) and put on concrete roadmaps with quarterly goals for every product org involving recommendation and for relevant cross-functional teams/orgs such as integrity, trust and safety, civic health, etc.
- [ ] Bridging metrics should be developed for each product surface across a platform's ecosystem of products. They should be shown alongside engagement and growth metrics, including in regular reports and dashboards for leadership and shareholders.
- [ ] Platforms of sufficient size and influence should regularly convene representative deliberative bodies (ala 'platform democracy') to navigate potential decisions around how to weigh different ranking factors; these should be convened through neutral 3rd parties, compliant with the appropriate norms/standards, and the relevant democratic governments and civil society organizations should be included as co-conveners and/or stakeholders.

Policymaker action items: funding, metrics, and platform democracy

As discussed earlier, we must be careful about embedding counterproductive 'false fixes' into law. We must also be cautious about giving either platforms or political actors too much power over recommendations.

Concretely:

- [ ] Governments should help fund the crucial R&D work to accelerate the development and deployment of bridging-based ranking systems.
- [ ] If governments are planning to encode ranking system requirements into law, then those requirements *should* include bridging based-ranking and *should not* include chronological feeds or middleware (unless they can be shown to not reward division).[27]
- [ ] Governments should require that platforms develop rigorous metrics on the extent to which they reward division, ideally with third-party auditing, report those measures publicly, and weight bridging enough to counteract any rewards for division.

In addition, governments, intergovernmental organizations, and non-governmental organizations can help ensure that those impacted by recommendation systems can have a voice in determining what such systems reward.

Concretely:

- [ ] Governments should participate in, convene, and/or help fund representative deliberative bodies on a regular basis to evaluate recommendation system policy. One particular focus should be the opportunity for well-designed recommendation systems to support democracy.[28]
- [ ] In lieu of government or platform convened deliberations, intergovernmental and non-governmental organizations should initiate such deliberations and invite governments to participate as stakeholders.

27 For example, middleware regulation might require that all recommendation system options highly weight bridging-based ranking, with complete flexibility beyond that requirement.

28 Ideally, such processes should be institutionalized within regulators, platforms, platform consortia, and/or intergovernmental organizations.

Conclusion

Transitioning to ranking systems more aligned with democracy is a logical progression for a new technology, and even specifically for a new means of societal communication. We have developed other institutions and practices, such as the professionalization of journalism and the development of evidence-gathering practices, in order to mitigate sensationalism and support democracy in the past. We can do it again in this new algorithmic arena.

Bridging-based ranking is just one kind of "bridging system" we may need to invest in to ensure that a deeply connected world can be kept above the "divisiveness threshold" necessary for democracy. It is also perhaps only one of several advances we must make to address the challenges of ranking and recommending the zettabytes of content produced every year. For example, we may also need some of our ranking systems to be able to identify and promote content that helps satisfy the critical information needs of the public over that which is bridge-building but frivolous (Friedland et al., 2012). There is much to do, and there may be little time to do it before we reach a "divisiveness threshold"—a point of no return.

At the end of the day, we become what we reward. Let us reward effective bridge-building over engaging divisiveness. Democracy depends on it.

Acknowledgments

Many thanks to the researchers and practitioners that have made this work possible, providing feedback, critiques, and sounding boards, including Luke Thorburn, Jessica Yu, Jonathan Stray, Cathy Wu, Priyanjana Bengani, Natania Antler, Colin Megill, Joaquin Quiñonero Candela, Andrew Konya, Kathy Pham, Austin Clyde, and many more.

Thanks especially to the incredible TAPP community of fellows that have made this year special, the staff and funders that have made it possible, and my RA, Yi-Ting Lien.

Thanks finally to the developers of Connected Papers, Elicit, and Semantic Scholar for supporting the exploration necessary for such interdisciplinary work.

References

Aggarwal, C. C. (2016). An Introduction to Recommender Systems. In C. C. Aggarwal (Ed.), *Recommender Systems: The Textbook* (pp. 1–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-29659-3_1

Allen, J. N. L., Martel, C., & Rand, D. (2021). *Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program.* PsyArXiv. https://doi.org/10.31234/osf.io/57e3q

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

Boxell, L., Gentzkow, M., & Shapiro, J. (2020). *Cross-Country Trends in Affective Polarization* (No. w26669; p. w26669). National Bureau of Economic Research. https://doi.org/10.3386/w26669

Brady, W. J., & Bavel, J. J. V. (2021). *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis*. OSF Preprints. https://doi.org/10.31219/osf.io/s4w2x

Brown, D. (2021, April 23). *The UN boosts peacekeeping efforts with AI mediator—The Washington Post*. https://www.washingtonpost.com/technology/2021/04/23/ai-un-peacekeeping/

Clyde, A. (2022, February 17). *Algorithmic Systems Designed to Reduce Polarization Could Hurt Democracy, Not Help It. Tech Policy Press*. https://techpolicy.press/algorithmic-systems-designed-to-reduce-polarization-could-hurt-democracy-not-help-it/

Doerr, J. (2018). *Measure what matters: How Google, Bono, and the Gates Foundation rock the world with OKRs*. Portfolio/Penguin.

Eckles, D. (2021). *Algorithmic transparency and assessing effects of algorithmic ranking.*

Edgerton, A., McLaughlin, D., & Forden, S. (2021, October 25). *Facebook Papers Reveal Internal Struggle: Profit Vs Public Safety—Bloomberg.* https://www.bloomberg.com/news/articles/2021-10-25/facebook-whistle-blower-files-reveal-internal-struggle-profit-vs-public-safety?sref=ExbtjcSG

Friedland, L., Napoli, P., Ognyanova, K., Weil, C., & III, E. J. W. (2012, July 16). *Review of the Literature Regarding Critical Information Needs of the American Public*. Federal Communications Commission. https://www.fcc.gov/news-events/blog/2012/07/25/review-literature-regarding-critical-information-needs-american-public

Fukuyama, F., Richman, B., Goel, A., Katz, R. R., Melamed, A. D., & Schaake, M. (2021). *Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy* (p. 14).

Haidt, J., & Bail, C. A. (2021). *Social Media and Political Dysfunction: A Review*. https://tinyurl.com/PoliticalDysfunctionReview

Horwitz, K. H. and J. (2021, September 15). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. *Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Keller, D. (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy*, 32(3), 168–172. https://doi.org/10.1353/jod.2021.0043

Lischka, J. A., & Garz, M. (2021). Clickbait news and algorithmic curation: A game theory framework of the relation between journalism, users, and platforms. *New Media & Society*, 14614448211027174. https://doi.org/10.1177/14614448211027174

Liscow, Z. D., & Markovits, D. (2022). *Democratizing Behavioral Economics* (SSRN Scholarly Paper ID 4012996). Social Science Research Network. https://doi.org/10.2139/ssrn.4012996

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2021). *Digital Media and Democracy: A Systematic Review of Causal and Correlational Evidence Worldwide.* SocArXiv. https://doi.org/10.31235/osf.io/p3z9v

Mac, R. (2021, October 5). Engagement ranking boost, M.S.I., and more. *The New York Times*. https://www.nytimes.com/2021/10/05/technology/engagement-ranking-boost-msi-facebook.html

Macy, M. W., Ma, M., Tabin, D. R., Gao, J., & Szymanski, B. K. (2021). Polarization and tipping points. *Proceedings of the National Academy of Sciences*, 118(50), e2102144118. https://doi.org/10.1073/pnas.2102144118

Miall, H. (2004). Conflict Transformation: A Multi-Dimensional Task. In A. Austin, M. Fischer, & N. Ropers (Eds.), *Transforming Ethnopolitical Conflict* (pp. 67–89). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-663-05642-3_4

OECD. (2020). *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD. https://doi.org/10.1787/339306da-en

Ovadya, A. (2021a, October 5). Platforms Are Still Flying Blind When It Comes to Societal Impact—Better Metrics Could Help. *OneZero*. https://onezero.medium.com/platforms-are-still-flying-blind-thats-not-ok-but-it-is-fixable-85d2c8212774

Ovadya, A. (2021b). *Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure.* https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure

Ovadya, A., & Thorburn, L. (2022). *Bridging Systems*.

Peter, F. (2017). Political Legitimacy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/legitimacy/

Ripley, A. (2020, September 29). Complicating the Narratives. *Solutions Journalism*. https://thewholestory.solutionsjournalism.org/complicating-the-narratives-b91ea06ddf63

Small, C., Bjorkegren, M., Erkkilä, T., Shaw, L., & Megill, C. (2021). Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi, 26*(2), Article 2. https://doi.org/10.6035/recerca.5516

Statista. (2021, June 1). *Total data volume worldwide 2010-2025*. https://www.statista.com/statistics/871513/worldwide-data-created/

Steve Rathje, Bavel, J. J. V., & Linden, S. van der. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26). https://doi.org/10.1073/pnas.2024292118

Stray, J. (2021). Designing Recommender Systems to Depolarize. *ArXiv*.

Tennenholtz, M., & Kurland, O. (2019). Rethinking search engines and recommendation systems: A game theoretic perspective. *Communications of the ACM, 62*(12), 66–75.

Thorburn, L. (2022, February 8). How Platform Recommenders Work. *Understanding Recommenders*. https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a

Twitter. (n.d.). Diversity of perspectives | *Birdwatch Guide*. Retrieved April 7, 2022, from https://twitter.github.io/birdwatch/diversity-of-perspectives/

Walter, B. F. (2022). *How civil wars start* (First Edition). Crown.

Technology and Public Purpose Project
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
belfercenter.org/project/technology-and-public-purpose

Printed in the United States of America