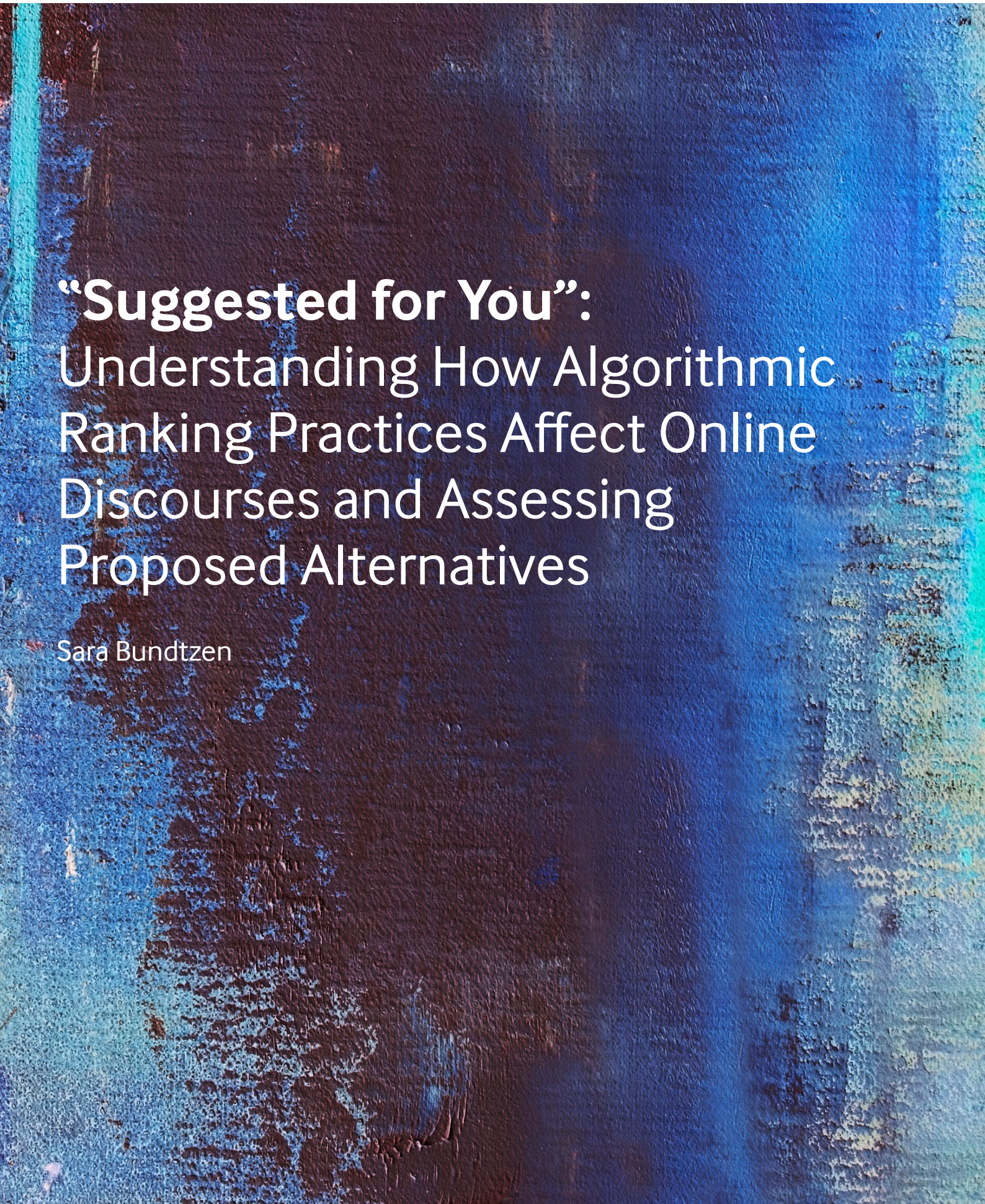
ISD Institute for Strategic Dialogue

gefördert durch:
Auswärtiges Amt



“Suggested for You”: Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives

Sara Bundtzen

About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of disinformation, hate speech, extremist and terrorist content online. It is comprised of representatives of relevant ministries and regulatory bodies from liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space. We thank the German Federal Foreign Office for their support for this project.

About this Paper

As part of the Digital Policy Lab, ISD organised two Working Group meetings on the topic of algorithmic ranking systems in July 2022. Participants represented ministries, departments, and regulatory bodies from Australia, Canada, Germany, Ireland, the Netherlands, New Zealand, the United Kingdom, and the United States. Participation included representatives from academia, civil society groups and former platform employees. The Working Group focused on the impact of algorithmic ranking on online discourse and user behaviour, as well as potential avenues toward platform accountability. This Policy Paper builds and expands on those discussions. The views expressed in this paper do not necessarily reflect the views of participants or any governments involved in this project.

About the Author

Sara Bundtzen is an Analyst at ISD, where she studies the spread of information manipulation by state and non-state political actors in multilingual online environments. As part of the Digital Policy Lab (DPL), Sara informs ISD's advisory work and analyses proposed pathways toward countering disinformation, influence campaigns, hate speech, and extremist content.

Editorial responsibility:
Huberta von Voss, Executive Director ISD Germany

Acknowledgments

We would like to give special thanks to participants from civil society and academia for providing valuable contributions to the Working Group: Sahar Massachi (Integrity Institute), Jenny Brennan (Ada Lovelace Institute), Professor Barak Richman (Duke University), Dr. Anna-Katharina Meßmer (Stiftung Neue Verantwortung), Oliver Marsh (The Data Skills Consultancy), and Marie-Therese Sekwenz (TU Delft). A special thanks also to members of the team at ISD, especially Henry Tuck and Helena Schwertheim, for their feedback and revisions.

ISD | Institute for Strategic Dialogue

www.isdgermany.org

Table of Contents

Executive Summary

Most social media platforms today host much more content than users could realistically consume. With the volume of content increasing and user attention spans remaining fixed, platforms moved on from reverse-chronological feeds to algorithmic ranking to show users the "most interesting" rather than the most recent content – with "most interesting" usually being content that has the highest predicted value to a company, for example, increasing the average time that users spend on a platform.

Algorithmic ranking systems, also known as recommender systems, make automated decisions about which pieces of content to prioritise or demote on feeds or in search results, who to connect with, who or what pages to follow – ultimately shaping the online experience of billions of users. With growing information velocity, reach and accessibility, social media platforms became integral to everyday communication and news consumption. At the same time, misleading, hateful, conspiratorial, or extremist views often find their way into the public debate long before factual or nuanced information.

This paper unravels the use of algorithmic ranking and its role in shaping online discourses and behaviour. It looks at the infamous "engagement problem" and other algorithmic phenomena, and how they may exacerbate the spread of harmful or "borderline" content, while reinforcing biases and discrimination. Acknowledging a knowledge gap and lack of clear evidence when it comes to the aggregate effects of algorithmic ranking, this paper examines current methodological and epistemic challenges of third-party algorithmic auditing, both within the research community and emerging regulatory frameworks. The paper highlights the need to develop common quality standards for independent auditing capabilities.

Looking ahead, this paper examines the benefits and drawbacks of existing proposals around advancing user agency, middleware, "positive nudges", "quality-focused" and "bridging-based" ranking systems.

Moving on from more individualistic approaches such as increasing user choice, alternative interventions in the field of ranking algorithms aim to recalibrate the incentives and intent of the metrics companies use for testing and evaluating how well ranking algorithms work. Throughout, this paper considers industry practices as well as regulatory (or co-regulatory) initiatives proposed by liberal democratic governments.

Glossary

Algorithms in computer science commonly refer to a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation.[1] An algorithm can be a simple if/then statement. A set of algorithms can be a sequence of more complex mathematical models, including machine learning algorithms, neural networks and deep learning algorithms.

Application Programming Interfaces (APIs) are software intermediaries that allow two applications to communicate with each other. APIs thereby allow researchers to access certain data from online platforms via data requests. As an intermediary, APIs provide an additional layer of security by logging, managing and controlling the volume and frequency of requests.

Content moderation practices are governance mechanisms that structure participation on a platform and enforce rules. Content moderation teams, either employed by a platform or outsourced to third parties, flag, review, demote and remove content that has violated a platform's Terms of Service. To keep up with the scale of content, platforms or moderation service providers increasingly rely on the use of automated content-moderation systems, as well as automated tools to assist content moderation teams.

Content recommendation uses algorithmic ranking to prioritise and display posts, pages, groups, or profiles to users. Examples of algorithmic ranking include YouTube's Shorts (short-form videos), TikTok's 'For You' page, Instagram's Explore page, or Facebook's News Feed.

Engagement rates are user interactions with of a piece of content. Interactions may inlcude liking, reacting to, commenting on or sharing a post, viewing a photo or video, or clicking on a link. Contrary, impressions are the number of times a piece of content was displayed on the user interface (for example, the feed), no matter if the user actually has seen (reach) or directly interacted (engagement) with the content.

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. ML algorithms use historical data as input to predict new output values. Supervised ML requires data scientists to train the algorithm with both labelled inputs and desired outputs.[2]

Harmful content and behaviours may refer to a broad spectrum of online activities that can have a negative impact on democratic and societal discourses online. Such activities can include hateful content, incitement to violence against a particular group, conspiratorial content, and false, misleading or manipulated content. In some instances, the risk of harm may be intrinsic to the content itself. In other instances, the risk of harm may be caused by aggregate patterns of behaviour rather than the nature of the content itself. In both cases, risks can be exacerbated by amplification. Depending on the legislative context, different forms of harmful content or behaviours may or may not be illegal. Depending on the platform, harmful content or behaviours may or may not be covered by a company's Community Guidelines, standards or rules. Some companies refer to **borderline content** when referring to content that comes close to infringing on the Community Guidelines.

Introduction

Since the mid-2000s, major tech companies such as Facebook (Meta), Google or Netflix have been developing and employing ranking algorithms to show content to users on their respective interfaces. While the growing volume of content required some form of ordering, most ranking algorithms today are designed to present users with the content deemed to maximise company metrics such as increased engagement, the time users spend on a platform or the number of views of a video.

At the same time, an emerging body of – mostly qualitative – research indicates negative effects of algorithmic ranking practices on user experience and online discourse.[3] In response, many governments are debating how to regulate algorithmic recommender systems to ensure safer and more transparent online environments.

Prominently, in late 2021, Frances Haugen, a former Data Engineer and Product Manager at Facebook, leaked internal company documents that seemed to indicate how deliberate business decisions to increase user interactions led to algorithms prioritising and amplifying negative, toxic or hateful content. In mid-2022, algorithmic ranking received wider public interest again, with macro-influencers such as Kim Kardashian and Kylie Jenner joining a backlash against Instagram's algorithm changes to their feeds. In particular, users complained about being swamped with videos from accounts they did not follow. A Change.org petition "Make Instagram Instagram again" quickly gained more than 300,000 signatures, calling for the return of reverse-chronological timelines, an algorithm that favours photos, and more consideration for content creators.[4]

While those demands would not necessarily resolve the issues raised, the outcry of the user and creator community highlights a fundamental concern of policy and scholarly debates: the power imbalances behind the decision-making and design processes of algorithmic ranking.

In summer 2022, Meta CEO Mark Zuckerberg noted in an earnings call that the "main transformations" of the company's business model would be "that social feeds are going from being driven primarily by the people and accounts you follow to increasingly also being driven by AI recommending content that you'll find interesting from across Facebook or Instagram, even if you don't follow those creators". More so, Meta CFO David Wehner explained that the company's "explicit strategy was to get Reels [a short-form video feature] in front of more users".[5] As it contends with competition from rivals such as Chinese-owned TikTok, Meta is reacting to and encouraging a trend towards short-format vertical video. However, rather than users organically or genuinely favouring Reels as a type of content format and platform functionality, it is company metrics incentivising content creators to post more videos in order to be recommended in the ranking feed (which is not to deny that some users do favour videos over photos). With the impact of algorithms gaining more prominence, observers noticed a phenomenon among content creators of "openly and aggressively going for virality" when posting short-form video content.[6] Of course, such dynamics also arise when borderline or harmful content goes "viral", reaching a high number of users, especially as short-form videos are introduced on top of existing content moderation designed for other formats or product features.[7]

While creators claim to "game the algorithms", there remains very little transparency and scrutiny of algorithmic ranking practices. More quantitative evidence of the (unintended) effects of those algorithms not only on individuals, but also broader societal views is necessary to enable sound regulatory accountability and due diligence obligations. Despite (or indeed because of) limited availability of evidence on the societal effects of algorithmic ranking, this paper sets out to explain current ranking practices, scrutinise methods for auditing algorithms, and reflect on potential interventions.

Section 1: Ranking algorithms

1.1 Ranking systems basics

The algorithmically curated News Feed introduced by Facebook sparked a fundamental shift from reverse-chronological feeds, which had been a common default in the early days of social media[8] – acknowledging that for platforms such as YouTube or TikTok, chronological feeds have not been an obvious "default" ranking. Against this backdrop, this section primarily considers the ranking practices of the News Feed and its core metrics to understand the state-of-the-art of ranking practices as well as the algorithmic phenomena that can shape people's information consumption and behaviour online.

In 2018, Mark Zuckerberg publicly announced an overhaul of the News Feed to help users have what the company calls "meaningful social interactions" (MSI).[9] According to Meta, the News Feed algorithm would prioritise posts from friends and family over public content, which would include "posts that inspire discussion in the comments" and posts that users might "want to share and react to".[10] Before looking at the effects of introducing MSI in algorithmic ranking, this section explains the key components of ranking processes.

Most algorithmic ranking systems today use "predicted engagement" to score and rank content items on the platform interface – in other words, ranking maximises whatever engagement goals (metrics) the company has set, usually at the level of each user. As visualised in *Figure 1*, the lifecycle of any type of content builds on inventory, features, statistical machine learning (ML) models, predictions, and final scores that are in return all rooted in company metrics.

Figure 1: Top Line Company Metrics, Goals, Expectations. Graphic by Integrity Institute.

Ranking algorithms first gather a query inventory of all applicable content on a platform that a user is eligible to see and engage with. **Algorithms use computed features (signals) that comprise millions of historical data points about the exposure, characteristics, and distribution of user behaviour and content.**

For example:

- Has the user liked similar content previously?
- What is the relationship between the viewer and the producer of the content? (e.g., Is the content a post from the user's friend, or a public post?)
- What is the type of the content? (e.g., Is the post a video, text, image or a combination?)

Large statistical ML models use these historical features to predict answers to various probability questions such as, "If a user has seen this content before, what is the probability that the user would like it, comment on it, share it, or watch it?" For example, the more a user previously interacted with a post's author, the more likely they would be predicted to engage with a post of this author. If a user tends to engage more with videos, ML models would predict a higher probability that a user will like, comment, share, or watch that type of format. Ranking algorithms also predict answers to questions such as "Will the user follow a page or friend another user involved in the content item?" With the optimisation toward MSI, content items are weighed according to what the company deems "meaningful conversations". Thereby, the News Feed algorithm uses different weights for different predicted actions, e.g., a "like" would be weighted one point, a "reaction or reshare without text" five points, and "significant comments or reshares" 30 points.[11]

Finally, ML models rank each of the items available to a user. All scores are combined to sort and generate a final ranking list that near-instantaneously generates each users' Feed. Final ranking of the News Feed also runs a "contextual pass" to ensure "content type diversity" and a mix of content formats.[12] **While ranking teams and managers may decide to maximise for specific metrics, the company leadership decides about the top line metrics to evaluate how well the ranking algorithm works. Top line metrics can include daily active users, overall number of posts, and average time spent on the platform.** For example, YouTube's core metrics maximise toward views (number of times that a video was viewed), watch time (number of minutes that users watched videos) and engagement (comments, likes, dislikes, shares, etc.).[13] While this section focuses on the News Feed, it is important to remember that social media platforms use algorithmic ranking for recommending many different types of content, especially audio-visual content, and channels such as suggesting accounts to follow or autocompleting search functions.

1.2 The engagement problem, superusers and other algorithmic phenomena

A key concern of engagement-based ranking is the "natural engagement pattern" of users. In simple words, users tend to engage (such as liking, sharing, commenting) more with content that nears the "cut-off point" of what is allowed on a platform. Meta defines this type of content as "borderline to the Community Standards", which is not prohibited by Community Standards but comes close to the limits of these policies, for example, borderline bullying and harassment, hateful speech, violence and incitement, or content that is misleading or sensationalised.[14]

Figure 2 visualises the resulting engagement problem (x-axis = allowed vs. prohibited content, y-axis = level of engagement). In 2018, Mark Zuckerberg wrote, "our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average – even when they tell us afterwards they don't like the content". In a similar concession, when Guillaume Chaslot, a former YouTube engineer, alerted colleagues that the video platform recommended conspiratorial content from Alex Jones' Infowars, a colleague responded, "People are clicking on it. What are we going to do?"[15]

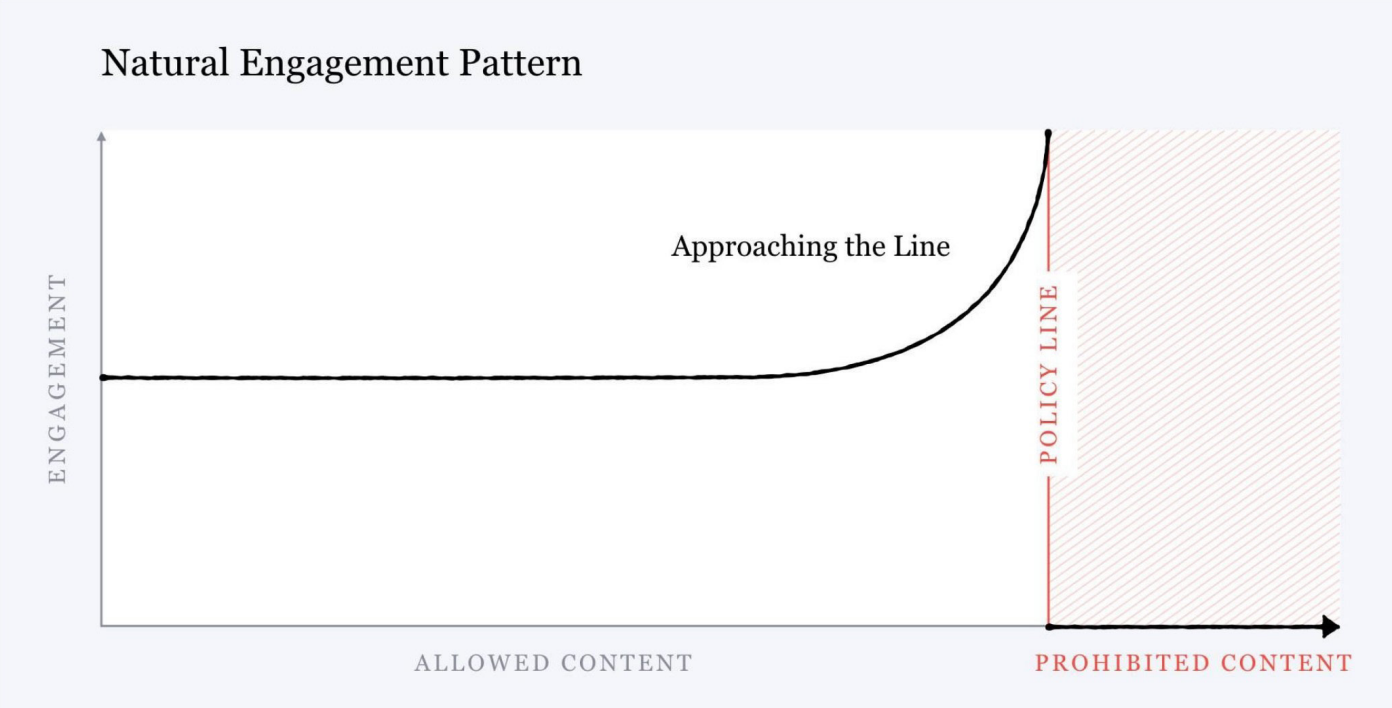


Figure 2: Natural Engagement Pattern. Graphic by Integrity Institute.

Engagement-based ranking is biased to prioritise content nearing that borderline, as content predicted to be engaging may be more likely to be harmful. **The problem becomes the gravitation towards borderline and harmful content.** ML models score user behaviour that maximises engagement highly, but some user behaviour that maximises engagement violates Community Guidelines. The focus on achieving metrics can therefore result in algorithmic ranking optimising for engagement, and thereby for more toxic, sensationalist, and borderline content.

This has real-world consequences. For example, in a leaked company note from April 2019, European political parties stated that Facebook's ranking algorithms forced them to use "far more negative content than before" because engagement on positive and policy posts had fallen dramatically given that MSI "systematically" rewarded "provocative, low-quality content".[16] The company metrics rewarded users or groups that post divisive, shocking or misleading content. In turn, they risked disincentivising producers of content, such as political parties or politicians, to post more nuanced and fact-based information.

Engagement-based ranking further risks creating unintended power imbalances between the users themselves. In the summer of 2020, researchers investigated the phenomenon of superuser-supremacy on Facebook, a class of users that produce more likes, shares, reactions, comments, and posts than 99 percent of users in the United States. Researchers analysed 52 million users, looking at 500 US-run pages with the highest average engagement as well as the highest-interaction posts from more than 41,000 of the highest-membership US public groups. Researchers found that the top one percent of accounts were responsible for 35 percent of all observed interactions; the top three percent were responsible for 52 percent. Many users rarely, if ever, interact with public groups or pages. **They found that these hyper-influential users were also the most abusive, skewing the publicly available inventory towards borderline content.** Among a randomly selected sample of 30,000 users, focusing on the 219 accounts with at least 25 public comments, 68 percent spread misinformation, reposted in spammy ways, published comments that were racist or sexist or anti-Semitic or anti-gay, or incited violence.[17]

As ranking algorithms reward "meaningful interactions" such as comments highly, superusers would have a disproportionate influence over how ranking weighs what could be interesting to other users. In this context, Sahar Massachi from the Integrity Institute further explains a stakeholder trap effect: if a platform rewards the "wrong" users for long enough, those users can become very powerful, trapping platforms and their algorithms.[18]

The superuser-supremacy problem points toward another risk of ranking algorithms that is the lack of equity of all users. Research shows that ML algorithms are often biased and risk disproportionately impacting already marginalised communities in negative ways. For example, researchers examined racial bias in hate speech and abusive language detection datasets, all of which use data collected from Twitter. Research showed evidence of systematic racial bias as classifiers trained on them tend to predict that tweets written in African-American English are abusive at substantially higher rates.[19] In 2018, Dr. Safiya Umoja Noble famously looked at online search results, showing how algorithms perpetuate societal biases, notably discrimination against people of colour, specifically women of colour.[20] **Ranking algorithms today rely on a massive inventory of content, which includes racism, sexism, and other intersecting forms of discrimination, but lack sufficient human oversight capabilities.** Ranking teams of (mostly) US-based social media platforms often do not have the necessary skills to estimate the algorithmic impact on content and user behaviour in widely different geographic, linguistic or legal contexts[21], further reinforcing user inequality and discrimination.

Importantly, statistical biases and the engagement problem are not necessarily new phenomena; they strengthen discriminatory behaviour and cognitive human biases. Business incentives exacerbated those cognitive phenomena in the past, notably in the traditional mass media, in particular tabloid newspapers, using sensationalist headlines or framings. Aviv Ovadya from the Belfer Center at Harvard Kennedy School points toward evolutionary reasons – "messy humans can end up very engaged when we see sensationalism and divisiveness, regardless of its truthfulness"[22]. In short, the above outlined phenomena are not so much just algorithmic, but human experiments.

In conclusion, algorithmic ranking practices are about power and "who controls what we see and how we see it"[23] – and currently, company metrics that evaluate the "success" of these ranking practices reward predominantly higher engagement rates, which in turn thrive on borderline content and human biases. This section highlighted some of the emerging societal risks of established ranking practices.

Section 2: Algorithm audits

Platform employees responsible for building and testing ranking algorithms are not necessarily incentivised to uncover risks of societal divisions and biases the way external researchers would be. And yet, given the limited access to proprietary data for researchers, these platform employees are often the only people who can truly study them. Consequently, ranking algorithms and their actual effects are not entirely understood by external researchers and organisations, and by extension policymakers and the public.

Third-party auditors with no contractual relationship to the audit target should independently assess how ranking practices work, and how to mitigate any potential biases or harms. However, without clear methods and standards of auditing practices, such audits will face difficulties of conducting sound investigations, verifying findings and offering clear evidence.

This section considers methods for conducting third-party audits of ranking algorithms, acknowledging the inherent limitations of current approaches. The section further elaborates on the need for transparent, clear and standardised auditing practices to enable platform accountability and evidence-based policymaking in privacy-preserving ways.

2.1 Algorithmic auditing methods

Audits are a central feature for public policy interventions that seek an evidence base and compliance with due diligence obligations. As AI-based technology infiltrates more aspects of society – for example, algorithms support decision-making about who is hired, laid off, or granted a loan – such systems are increasingly the subject of debates over ethics assessments in a range of policy areas.[24] Auditing such algorithms can serve monitoring over time, investigating a specific harm or risk, or ensuring compliance with regulatory obligations. For example, in the case of algorithms that assign scores to humans, audits have focused on discrimination against certain groups based on potential biases inherent in ML models and labelled data used to train them. An audit of commercial automated facial analysis algorithms and datasets by researchers Joy Buolamwini and Timnit Gebru found that the system did not recognise darker-skinned people as well as white people.[25]

With regard to algorithmic ranking on social media, researchers proposed a range of research methods to assess how systems influence online discourse and user behaviour, each with their own advantages and drawbacks. Importantly, the range of research methods mostly derives from a lack of meaningful, privacy-preserving, and consistent access to internal platform data and research experiments. The Ada Lovelace Institute produced an overview of methods for inspecting algorithmic ranking systems, outlining both purpose and drawbacks of each method.[26]

Audit method	Description	Purpose	Challenges
Code audit	Auditors have direct access to the codebase of the underlying the system, or 'pseudocode', i.e., plain-English descriptions of what the code does.	Understanding intentions of algorithms; in the case of machine learning, useful for understanding objectives are being optimised.	Codebases can be huge – individual engineers in large companies rarely understand how all parts of the platform operate. Hard to see effects through code. Concerns about IP and security.
User survey	Auditors conduct a survey and/ or perform user interviews, to gather descriptive data of user experience on the platform.	Gathering information about user experience on a platform – to paint a rough picture of the types of problematic behaviour that could then be further investigated.	Vulnerable to common social science concerns with surveys – pressure to answer in a particular way, unreliable human memory and difficulty to attribute causation to findings.
Scraping audit	Auditors collect data directly from a platform, typically by writing code to automatically click or scroll through a webpage to collect data of interest (for instance, text that users post).	Understanding content as presented on the platform – particularly making descriptive statements (e.g., 'this proportion of search results contained this term') or comparing results for different groups or terms.	Requires the development of a custom tool for each social media platform, which can be brittle as small (legitimate) changes to a website's layout can break the program.
API-Audit	Auditors access data through a programmatic interface provided by the platform that allows them to write computer programs to send and receive information to/ from a platform, e.g., an API might allow a user to send a keyword and get back the number of	matches. Easier programmatic access to data than a scraping audit – allowing easier automation of collection for descriptive statements or comparative work.	Publicly available APIs may not provide a regulator with the data they need. With information-gathering powers, they could compel a platform to provide access to further APIs or even a custom API, but this may require
Sock-puppet audit	Auditors use computer programs to impersonate users on the platform. The data generated by the platform in response to the programmed users is recorded and analysed.	Understanding what a particular user profile, or set of user profiles, may experience on a platform.	Sock puppets are only impersonating users – they are not real users and so are at best a proxy for individual user activity and experience. Yet, sock-puppet audits have been useful for research purposes that seek to gain a basic understanding of the algorithmic ranking system.
Crowd-sourced audit	A crowd-sourced audit uses real users who collect information from the platform while they are using it – either by manually reporting experiences or through automated means like a browser extension.	Observing what content users are experiencing on a platform, and whether different profiles of users are experiencing different content.	Requires custom data-collection approach for each media platform being audited, often relying on web-scraping techniques; so far only demonstrated on desktop not mobile devices so may skew results or overlook mobile experiences.

Table by Ada Lovelace Institute.

Policy Initiative: The EU Digital Services Act (DSA)

The Digital Services Act (DSA), which has been signed and published in the Official Journal of the European Union on 27 October 2022[27], will require that providers of very large online platforms (VLOPs) and of very large online search engines (VLOSEs)[I] comply with a range of obligations related to transparency and due diligence. Thereby, the regulation introduces several layers of auditing: mandatory internal risk assessments conducted by the platforms (first party audits), mandatory external audits by independent contractors to assess compliance with obligations (including the risk assessments), and third-party assessments conducted by vetted researchers as part of data access provisions.

Mandatory risk assessments (Article 34) will require platforms to identify, analyse and assess any systemic risks stemming from the design or functioning of their service and its related systems, including algorithmic systems. This includes systemic risks stemming from; illegal content; any actual or foreseeable negative effects for the exercise of fundamental rights; on civic discourse and electoral processes, and public security; or in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being. Importantly, platforms must specifically consider the design of their algorithmic ranking systems. Where the algorithmic amplification of information contributes to systemic risks, providers will need to reflect this in their risk assessments, and ultimately their risk mitigation measures (Article 35).

Risk assessments and risk mitigation will both be subject to annual independent audits. The European Commission may adopt delegated acts for more specific rules, for example, as regard to auditing methodologies and reporting templates (Article 37). Contributing scientific and technical expertise to the Commission's exclusive supervisory and enforcement role, the European Centre for Algorithmic Transparency, managed in close cooperation with the Directorate-General for Communications Networks, Content and Technology and the Joint Research Centre, will provide technical support such as "technical tests on algorithmic systems" and scientific research such as "practical methodologies towards transparent and accountable algorithmic approaches".[28]

For the purpose of compliance, the national regulatory bodies of EU Member States or the Commission will be able to require access to or reporting of specific data, including information on the "design, logic, functioning and/or testing of algorithmic systems". Upon request of regulatory bodies, platforms will also be required to provide access to vetted researchers for the purpose of "detection, identification and understanding of systemic risks" and the "assessment of the adequacy, efficiency and impacts of the risk mitigation measures" (Article 40). Mathias Vermeulen notes how vetted researchers could be considered quasi-auditors: they can assess emerging risks that may not have been covered by a platform's internal risk assessment report, and assess whether mitigation measures have been effective in practice.[29]

Voluntary Codes of Conduct will support the implementation of the DSA, including through commitments to take specific risk mitigation measures, as well as regular reporting frameworks on any measures taken and their outcomes (Article 45). In June 2022, the Commission published the strengthened Code of Practice on Disinformation (CoPD)[30], which requires relevant signatories, including Google (YouTube), Meta (Facebook, Instagram), Microsoft (LinkedIn), TikTok, and Twitter, to minimise the risks of "viral propagation of disinformation" by adopting "safe design practices". Importantly, companies commit to invest in research efforts on the spread of harmful disinformation online and related safe design practices, including the development of an "independent, third-party body that can vet researchers and research proposals". Vermeulen notes that, while the Commission will consider adherence to the Code's commitments when assessing compliance with the obligations of the DSA, the ultimate purpose of the data access regime under the Code is different as access can be granted for any research that deals with "disinformation"[31], thereby not necessarily being related to the assessment of systemic risks and risk mitigation measures.[32]

[I] VLOPs and VLOSEs cover platforms and search engines that reach a number of average monthly active users in the European Union equal to or higher than 45 million.

2.2 Methodological and epistemic limitations

In view of the varying levels of data access, exter-nal researchers develop and employ different metho-dological approaches to collecting and analysing platform data. Despite initial research findings, the field continues to face barriers to conducting systematic, longitudinal, large-scale data analyses that offer a fuller understanding of the consequences of algorithmic ranking on online discourses. For example, platforms may deliberately restrict access to data (due to user privacy, trade secrets, innovation or other concerns), or they may have other technological features that inadvertently create barriers (e.g., video and audio content cannot easily be searched or analysed in the same manner as text content). Access to some online spaces, for example, private and/or encrypted messages, or closed channels or groups, may also be restricted by practical and/or ethical considerations. In other cases, theoretically accessible, public content cannot be sufficiently analysed due to lack of disclosures, API or other data access infrastructure.[33] In brief, a knowledge gap exists when it comes to the full scale and complexity of ML algorithms that use millions of signals to make predictions. Indeed, ranking algorithms are difficult to make sense of, and impossible for external researchers to assess without access to the methodology and experiments conducted by the company's ranking teams.

The dependence on transparency and data access reinforces the epistemic limitations in the field of studying the impact of social media on society. From a data availability perspective, researchers have been focusing on specific platforms that provide more comprehensive API access, notably Twitter. However, a too narrow scope ultimately over- or under-represents certain demographic or geographic groups. For example, young audiences in particular use TikTok[34], while its API remains severely limited and the platform reveals very little useful information about its algorithms.[35] At the same time, some audiences consume most of their information through traditional means such as television, while others increasingly use messaging apps such as Telegram or WhatsApp for their information consumption.[36]

Researchers also lack common terminology related to the risks and mitigation measures that should be analysed and assessed. For example, in the context of the German federal election in 2021, the Sustainable Computing Lab sought to conduct an external risk assessment of the systemic risks to the right to free and fair elections, focusing on Twitter and Facebook. Researchers noted difficulties of clearly identifying the boundaries of the categories for identifying electoral rights violations and disinformation. Their findings suggested that categories such as "systemic electoral risk" need to be more clearly delineated and easier to reproduce and compare in future research projects. More crucially, policymakers and researchers often use terminology related to ranking systems such as "algorithmic amplification" or "demotion" without a clear and reasonable understanding, risking to set false ideas and expectations about the auditing and mitigation measures.

In conclusion, while the study of social media has been a rapidly growing research field, limited access to proprietary data sustains the field's "unknown unknowns" and contributes to varying standards of research design – leading to unreliable outcomes, sometimes limited in scope and analytic value. As mandatory rules proposed by legislative frameworks anticipate that platforms, regulators, as well as third parties conduct algorithmic auditing, the development of transparent research design, agreeing on clear terminology, and employing robust methodology will be essential to probe these systems and establish consistent as well as effective third-party auditing capabilities.

Quality standards and transparency of algorithmic audits

A range of relevant stakeholders, including civil society, academia, policymakers, regulatory bodies and the platforms themselves, should convene open consultations to develop common standards and working definitions for streamlining the quality and transparency of third-party algorithm audits. Importantly, algorithmic audits will need to develop a shared epistemic understanding of research methods and terminology.

Such consultations should include cross-disciplinary expertise, and be demographically, regionally and linguistically diverse to ensure some level of legitimacy. In particular, algorithmic auditing capabilities require more clarification on the types of societal risks and mitigation measures, including changes to the company metrics subject to third-party auditing. It will be important to clarify working definitions to operationalise frequently used concepts such "algorithmic amplification" or "demotion" into testable hypotheses.[37] **More so, it is important to acknowledge the widely different baselines depending on the features of a platform (such as type of post) when it comes to assessing "algorithmic amplification" (e.g., content is "amplified" in comparison to what baseline?).**

From a platform perspective, internal risk assessments will both require but also benefit from engagement with external researchers to better anticipate what the ranking teams should look for when experimenting with algorithm changes to prevent ML modelling causing shifts toward harmful content. Third-party auditors, on the other hand, will require that platforms disclose their internal decision-making processes when it comes to methodology and experimentation. Making ranking methodology more transparent could include disclosures about the underlying intentions of ranking. For example, platforms could disclose summary statistics available in the company's decision-making processes. Disclosures of information about internal platform experiments, such as randomised trials that assign users different algorithmic ranking, could further inform research on the (unintended) consequences of ranking. Platforms could disclose lists of experiments with hypotheses, dates and decisions made.[38] **Such level of meaningful transparency could help researchers understand the effects of changes to ranking metrics as well as the cause-and-effect relationships.**

As anticipated by the emerging EU regulatory framework, and given the sensitivities of algorithm data, an independent intermediary body, acting as a facilitator between regulatory bodies, researchers and platforms, could help developing and overseeing guidance on vetting processes, data privacy and transparency of research documentation.

Independent algorithmic auditing capabilities could also be prioritised in transatlantic and transpacific policy circles, for example, in the remit of meetings of the EU-US Trade and Technology Council (TTC) or the Christchurch Call Community (such as through the recently launched Initiative on Algorithmic Outcomes[39]). Specifically, the EU-US TTC could advance coordination of different approaches of upcoming AI regulations and frameworks, including the White House's Blueprint for an AI Bill of Rights[40], and how those address algorithmic ranking used by social media platforms. **Any intergovernmental consultation on algorithmic ranking should be informed by diverse expertise and perspectives (including geographic, gender, and linguistic diversity), the non-governmental and research community, as well as company ranking teams to enable informed and verifiable third-party auditing capabilities.**

This section explores benefits and drawbacks of proposed interventions that aim to rethink user experience and algorithmic ranking practices of social media platforms. It considers a variety of approaches – from those focusing on individual user experience to those reconsidering the inherent metrics of ranking practices – to examine pathways toward more democratic, transparent and safer online environments.

Section 3: Potential interventions and alternatives

3.1 'Choose your own ranking system'

Increasing user agency, especially through a middleware market, focuses on enabling more individual choice and competitive avenues to tackle some of the negative effects associated with platform monopoly. **Scholarly discussions long debated potential opportunities for users to opt in or out of ranking practices, and how to make users a more active part of ranking systems.**

Feasibility of a decentralised middleware market

In 2020, a Working Group on Platform Scale at Stanford University[II] proposed introducing middleware to tackle centralised platform power. They describe middleware as "software and services that would add an editorial layer between the dominant internet platforms and internet users". **The intention of middleware is to dilute the concentrated power of the tech companies to control information flows on their platforms, and reduce the impact of algorithmic amplification.** With middleware or related proposals such as Mike Masnick's "Protocols, not platforms" (which advocates for the internet of many different decentralised protocols rather proprietary platform)[41], power should move away from centralised platforms and towards individual users and other providers.

This idea is not new. For example, Gobo.social has been allowing users to connect up to three accounts from different social media platforms to a single page, showing a combined feed – and giving users more control over their feeds.[42] Another example is Twitter's Developer Toolbox, which includes 11 expression, safety and measurement tools that users can choose from to add new functionalities to the platform. For example, Block Party, an anti-harassment tool, lets users protect their mentions from trolls with block lists, automatic muting, and community support.[43] All developers and their tools are vetted based on Twitter's quality and safety standards, while tools control their own monetisation structures and pricing.[44] Google's Perspective API can also act as middleware technology, using ML to identify "toxicity" comments to improve the moderation of conversations. It can provide scores for severe toxicity, insults, profanity, attacks on identity, threats, and sexually explicit content.[45]

In short, middleware proposes to give individual users more choice over what they consume by offering a variety of providers. Daphne Keller from Stanford's Cyber Policy Center, notes that "users might choose a racial-justice–oriented lens on YouTube from a Black Lives Matter–affiliated group, layer a fact-checking service from a trusted news provider on top of Google News, or browse a G-rated version of Facebook from Disney".[46]

Introducing a middleware market for algorithmic ranking, however, poses new challenges.

- For one, allowing individual users to effectively choose their own information environments could encourage user groups to more easily consume content that echoes their own views, beliefs, and fears – exacerbating "in and out group" thinking. Users who may choose a "positivity" middleware could opt out of any negative news. Simply switching on such "positivity modes" may decrease a users' exposure to important issues, and a range of perspectives. For example, social justice activists could end up with a smaller audience, if middleware helped users shield themselves from "negative" topics.[47] Other users may simply want to see and engage with borderline and harmful content.

- A diverse group of actors, including political parties, extremist groups or hostile states, could exploit a middleware market as way to promote their own content to a subscribed audience.

- Some users who may be particularly vulnerable to harmful content may not be sufficiently informed or educated to consider the potential risks and negative implications of middleware.

II The Program on Democracy and the Internet convened a working group in January 2020 to consider the scale, scope, and power exhibited by the digital platforms, study the potential harms they cause, and, if appropriate, recommend remedial policies. The group included an interdisciplinary group of scholars, namely Francis Fukuyama, Barak Richman, Ashish Goel, Roberta R. Katz, A. Douglas Melamed, and Marietje Schaake.

- Such concerns aside, middleware options face uncertainty around data privacy. Providers of alternative ranking systems would need to access both publicly available and platform proprietary content. For example, to what extent would privately-shared content from a user's friends be used in middleware technology to create a sufficiently large inventory for the middleware to function? While middleware that labels content can be designed as a browser extension (which has its own data privacy concerns), alternative ranking algorithms would require a more integrated data-sharing framework to be established between platforms and middleware providers.[48] In 2021, the OECD released a report on the role that data portability and interoperability measures can play in promoting competition both within and among digital platforms. The report discusses standards that enable real-time data sharing across services (e.g., cross-posting social media content on multiple platforms), and those that enable the combination of functionalities (e.g., having a single account log-in across multiple different online services).[49] Essentially, providers of platforms and middleware options would need to grapple with the division of liabilities and mandated responsibilities of due diligence and accountability.

- Finally, a sustainable business model for middleware providers would be required to induce an adequate supply and competition. Providers of middleware could use paid subscriptions, or platforms and middleware providers could agree on revenue-sharing[50] – **which still leaves the question of whether alternative ranking providers would actually use different metrics and reformulate the goals of their algorithms.**

Active user choice at the design level

Introducing greater user agency (also referred to as user choice or empowerment) at the design level remains a popular approach to platform regulation. In the UK, the Behavioural Insights Team (BIT) partnered with the Centre for Data Ethics and Innovation (CDEI) and Doteveryone to explore how to create "active" choices for users. The underlying notion of the project was that designing online services that enable people to use them "in line with their preferences" would be an important part of creating a "positive technology landscape". Notably, the research defined active choices as "choices that reflect users' wishes without obstruction, and are based on an understanding of the likely consequences".[51]

BIT conducted experiments using prototypes that demonstrate what active choice could look like in three online contexts – smartphone operating systems, web browsers and social media. BIT ran three online experiments, each with approximately 2,000 participants, to test how the alternative interfaces performed against the controls.

The social media experiment considered common types of preferences such as organising feed content (chronologically or algorithmically), filtering of untrustworthy sources, and privacy settings. The research used three trial designs for testing the ability of users to make informed choices about their settings. The social media experiment, as visualised in *Figure 3*, included:

- A slider mode, allowing manual customisation along a spectrum;
- A private mode, bundling choices together into a simple binary;
- Responsive toggles that combined choices by topic but were not bundled into a single toggle.

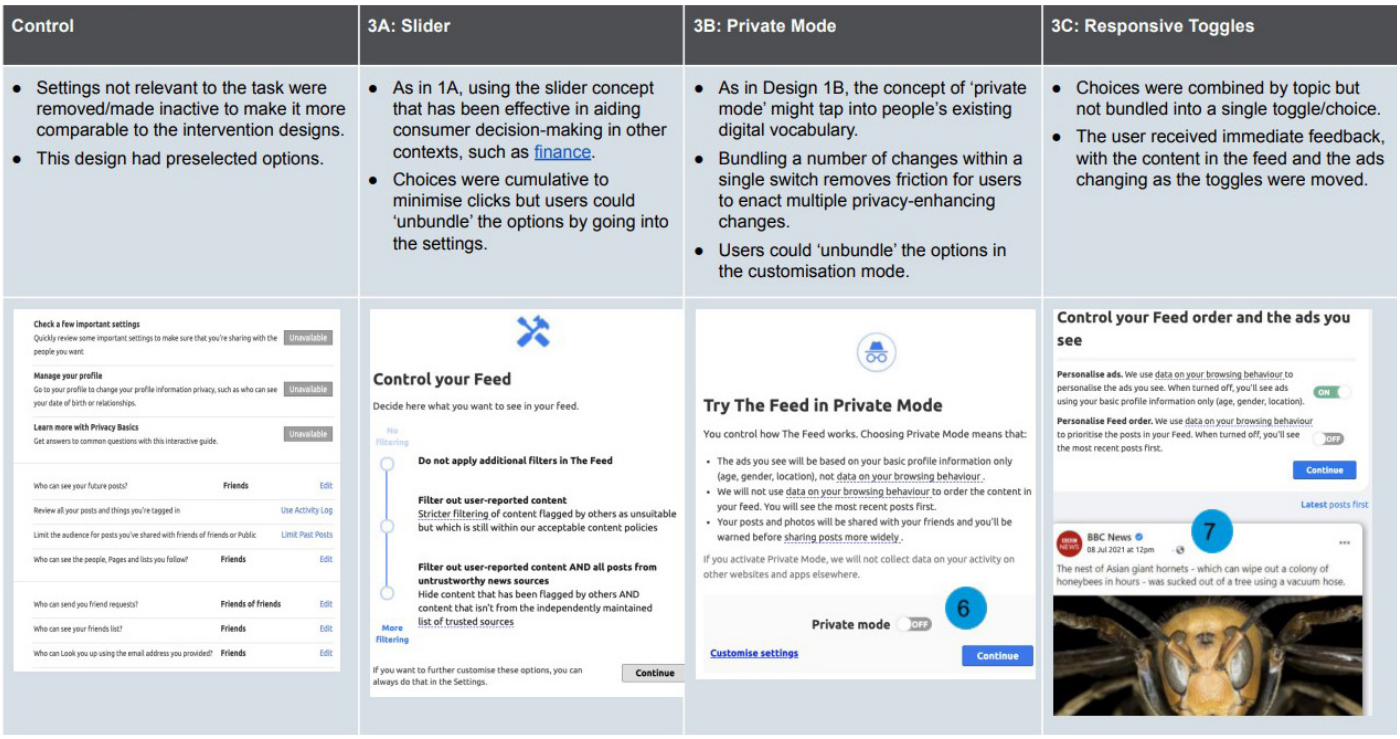


Figure 3: Design of the social media experiment. Graphic by the Behavioural Insights Team (BIT).

The findings of the experiments offer some learnings for proposed interventions such as introducing middleware options. To begin with, self-reported metrics appeared to be not a good indicator of the ability of people to make choices in line with their preferences. For example, self-reported feelings of control improved even though task accuracy (i.e., could participants adjust settings to match the preferences of a fictional persona) and understanding of consequences (i.e., whether participants could correctly indicate the implications of their choices) did not. Secondly, the design needs to be carefully tailored to users' levels of knowledge. Feedback suggested that the designs used for the social media prototype were too complex. Clear labelling of options, transparency, and testing would be crucial so that simplification and bundles do not disadvantage specific user groups, and so that users are able to fully align choices with their preferences. Lastly, performance of the trusted third-party prototype, where participants could delegate choices to a third-party organisation (which could be a middleware provider), appeared to be driven by familiarity and association of that organisation with the digital world. For example, users may not pay much attention to what actually represents bundled choices recommended by a well-recognised digital technology company, and choose it even if it does not match their preferences. In short, there would be no "one size fits all" solutions for enabling user agency, as performance and outcomes will likely depend on the individual context, user knowledge and familiarity.

In conclusion, increased user agency and a decentralised middleware market could give more individual choice and power to users. This approach could generate greater competition among providers, and therefore dilute the impact of a small number of particularly powerful companies. Yet, middleware providers would need to grapple with recurring challenges of data privacy, content moderation, and sustainable business models. Moreover, without additional risk mitigation measures, more individual choice over the content environment could create a fragmentation of harmful or borderline content and behaviour, further divide user groups, and enable ways for those wishing to disseminate negative or divisive products to selected audiences.

Policy Initiative: Australia's Safety by Design Framework

In June 2018, eSafety, Australia's independent regulator and educator for online safety, stated their intention to develop a Safety by Design (SbD) Framework. The Framework is a broad and iterative program that aims to guide organisations as they seek to embed the rights of users and user safety into the design and functionality of their products and services.

The accompanying SbD Principles are underpinned by human rights and have been developed from information collected through eSafety's research and reporting schemes, outreach programs, industry and key stakeholder consultations. The second phase of the SbD initiative focuses on developing guidance to assist industry partners in operationalising the Principles and elevating online safety practices and interventions. These guidance materials are used by social media platforms, as well as broader online sectors, such as gaming, dating and banking.

The framework provides principles on user empowerment and autonomy. Thereby, online services should provide "technical measures and tools" that allow users to manage their own safety, and that are set to the most secure privacy and safety levels by default. Services should "leverage the use of technical features to mitigate against risks and harms", referring to technical tools that can filter content—not necessarily for removal but, for example, to place it behind age-gates or interstitial warning pages. Services should evaluate design and features to ensure that risk factors for all users, particularly for those with distinct characteristics and capabilities, have been mitigated before products or features are released to the public.

As technologies and online environments evolve, the Safety by Design approach, alongside Privacy by Design and Security by Design approaches, aims to ensure harms can be mitigated before they occur.

Still, companies' adoption of the principles remains voluntary. While no enforcement mechanisms are in place to compel compliance with the principles, they are intended to serve as a foundational self-regulatory layer underpinning the enforceable co-regulatory and regulatory requirements of Australia's Online Safety Act 2021. The Act provides for industry bodies or associations to develop mandatory codes that will be registered by the eSafety Commissioner, if they meet "appropriate community safeguards" (otherwise, eSafety can determine a "standard").[52] It further sets out Basic Online Safety Expectations for online service providers to be proactive in how they protect people from abusive conduct and harmful content online. eSafety has the power to require providers to report on how they are meeting any or all of those Expectations.[53]

3.2 Quality-focused ranking algorithms

In recent years, experts proposed that companies should rethink the metrics used in algorithmic ranking to change the types of content and behaviour these systems are rewarding. For example, Renee DiResta from the Stanford Internet Observatory noted, "Platforms can decide to allow Pizzagate[III] content to exist on their site while simultaneously deciding not to algorithmically amplify or proactively proffer it to users".[54] Rethinking metrics does not aim to completely eradicate the existence of harmful content, but make the algorithms more favourable to less divisive and toxic discourses. **The notion of "quality-focused" ranking is not new and has not necessarily proven to be effective when employed by companies, but it is worth exploring further to have an informed debate about successful enforcement, including through regulatory frameworks.**

Industry: Facebook's 'Remove, reduce, and inform' strategy

Since 2016, Facebook's News Feed has deployed what the company calls a 'remove, reduce, and inform' strategy to manage misleading or harmful content. This means that besides removing content that violates Community Standards, the News Feed demotes borderline "low-quality" content such as engagement bait or web pages with little substance and disruptive ads. Doing so, the News Feed algorithm uses signals such as whether a domain's Facebook traffic is highly disproportionate to their place in the web graph[55], i.e., pages and hyperlinks on the Internet may be viewed as nodes and arcs in a directed graph. Facebook teams reviewed web pages linked to and from Facebook to identify those that contain "little substantive" content and have "a large number of disruptive, shocking or malicious ads". It then uses algorithms to identify new web pages of low quality.

According to Meta's Content Distribution Guidelines (last updated in October 2022)[56], the News Feed also demotes:

- Low-quality comments (meaning comments that do not "add meaningfully to the discourse around a post");
- Low-quality events (including events missing key details such as time, location and/or sign-up information);
- Low-quality videos (including videos predicted to be static, animated, looping, polls-only or pre-recorded);
- Pages posting unoriginal videos (including videos that are repurposed from other sources with limited added value);
- Pages predicted to be spam;
- Sensationalist health content and commercial health posts;
- Domains with limited original content (such as containing a large amount of content from other publishers);
- Fact-checked misinformation;
- Inauthentic sharing (including Pages engaging in behaviour that artificially boosts views or engagement);
- Links to domains and Pages with high "click-gap" (meaning domains that receive a disproportionate amount of their traffic directly from Facebook compared to the amount of traffic from the rest of the Internet);
- Posts from broadly untrusted news publishers;
- Posts from Pages that artificially inflate their distribution;
- Posts from users who 'hypershare' content in groups;
- Unoriginal news articles.

Additionally, Facebook intends for the News Feed to inform users with contextual information about articles using "credibility signals". For example, showing information about the Page that published the original article, providing context from external experts or third-party organisations (Wikipedia), or showing "Related articles or more from the publisher".[57] **Despite this effort, Facebook has not yet provided clear definitions for terms like "demotion", for example, how long would users need to scroll to see reduced content – making it harder to assess the success of this strategy. It also remains unclear whether low-quality content like engagement bait, spam or unoriginal videos is demoted to the same extent as sensationalist health content or fact-checked misinformation.[58]**

[III] A viral conspiracy theory known as 'Pizzagate' that claims that Hillary Clinton and democratic operatives placing orders at a pizzeria in DC, called Comet Ping Pong, were actually using code to talk about underage prostitutes.

Industry: Google's Search Quality Rating

Google's Search Quality Rating serves as an example of including "quality" scoring in algorithmic ranking practices. Still, it should be acknowledged that researchers have criticised the Search algorithms for past failures to sufficiently demote or remove harmful content, including discriminatory and misleading content in the search results. Researchers should thereby continue to hold Google's Search algorithms accountable to their quality promise.

Besides user query and relevancy of websites (in relation to the query), Google's Search algorithms aim to surface "relevant information from the most reliable sources". In a White Paper published in February 2019, Google notes that ranking algorithms are "an important tool in our fight against disinformation". Google's Search algorithms would "relegate lower quality or outright malicious results (such as disinformation or otherwise deceptive pages) to less visible positions in Search or News".[59]

The White Paper notes that "algorithms cannot determine whether a piece of content on current events is true or false, nor could they assess the intent of its creator just by reading what is on a page", but they are able to identify manipulation or deception tactics such as spammy behaviours at scale.[60] **Meanwhile, Google's Search Quality Rater Guidelines define the goals of the ranking systems, guiding third-party evaluators who provide ongoing assessments of the algorithms. According to the Guidelines, so-called "High Quality Pages" would have the following characteristics:**

- High level of Expertise, Authoritativeness, and Trustworthiness (E-A-T).
- Satisfying amount of high-quality Main Content (MC), including a descriptive or helpful title. High quality MC must be factually accurate for the topic and must be supported by expert consensus where such consensus exists.
- Satisfying website information and/or information about who is responsible for the website. If the page is primarily for shopping or includes financial transactions, then it should have satisfying customer service information.
- Positive website reputation for a website that is responsible for the MC on the page (or positive reputation of the creator of the MC, if different from that of the website).[61]

In 2017, Google announced a partnership with *The Trust Project*[IIII] to help news sites use eight "trust indicators" to distinguish between " quality journalism" and "promotional content or misinformation":[62]

1. Best Practices: What are the standards of the news outlet? Who funds it? What is the mission?
2. Expertise: Who reported this? Are we given details about the journalists, including their expertise?
3. Type of Work: Does the content have labels with clear definitions to distinguish opinion, analysis and sponsored content from news stories?
4. Citations and References: What is the source? Are we given access to the sources behind the facts and assertions?
5. Methods: Are we given context about why journalists chose to pursue a story and how they went about the process?
6. Locally Sourced: Was the reporting done on the scene, with deep knowledge about the local situation or community?
7. Diverse Voices: Does the newsroom bring in diverse perspectives across social and demographic differences?
8. Actionable Feedback: Can readers participate? Does the newsroom engage the public when setting coverage priorities?

[IIII] Sally Lehrman, an award-winning journalist, founded and leads The Trust Project, an international collaboration that she began building in 2014 to strengthen public confidence in the news through accountability and transparency. The project is funded by the Trustworthy Journalism Initiative of Craig Newmark Philanthropies and Google. Funders also included the Democracy Fund, the John S. and James L. Knight Foundation and Facebook.

Google previously stated that these E-A-T based criteria and rater data are not used directly in the ranking algorithms. Rather, they are used as feedback to help ranking teams understand if the systems are working.[63] **To reflect these criteria, a prominent signal of "authoritativeness" has been PageRank, which uses the number of links on the web, link attributes, anchor text, and the likelihood of being clicked.** In simple words, the more sources linking to a page, the more valuable the information on that page and the more likely users are to visit it. In 2006, Google redesigned the algorithms to select a few trusted sources referred to as seed pages and assess the quality of other pages based on links coming from such pages. For example, Google labels The New York Times as a seed page as it covers a wide range of topics that interest users and features many outgoing links.[64]

The Integrity Institute suggested that ranking algorithms of social media platforms such as Facebook could incorporate PageRank to track and score Pages or Accounts according to media literacy checks.[65] This way, content that would fail media literacy checks would include content from anonymous sources (which may not include those posting content anonymously for legitimate reasons, for example, whistle-blowers or human rights activists), content from sources that are systematically copying content from other creators, content from sources that use networked assets (bots), or content that violates community standard. **Importantly, the development and use of "quality" criteria as metrics, and *how* they are tested, experimented with and enforced, should be linked to transparent stakeholder consultations together with civil society and external researchers.**

3.3 Positive friction, nudges and bridging-based ranking

Part of the problem of most ranking algorithms is the ease and speed of frictionless feeds that allow users to post and share content constantly. **Positive friction aims to slow down posting and user interactions, giving users a chance to have a break and think before sharing.[66] Such intention could be incorporated into algorithmic ranking.**

Facebook whistle-blower Frances Haugen previously emphasised that many superusers would engage with borderline content most late at night. Slowing down algorithmic ranking towards the evening could help incentivise superusers to switch off earlier, so algorithms receive fewer toxic signals from such users. Specifically, Haugen advocated for "break-glass" measures that simply limit the number of times a piece of content can be shared, no matter what content, to help reduce the disproportionate influence of superuser activity. Such friction would also add another layer of decision-making, for example, requiring users to copy and paste content, if they wanted to share it.[67]

A working paper published by MIT Sloan examined how and why misinformation about COVID-19 spreads on social media, by testing a simple technical intervention, a small nudge, that could limit this spread.[68] Through experiments researchers observed that nudges to get users thinking about the accuracy of a piece of content made them more discerning when it came to sharing true or false news. **Notably, users who performed the task of rating the accuracy first were less likely to share inaccurate news, and more likely to share accurate news. Moreover, researchers note that the cumulative effects of such an intervention may be substantially larger than what is observed when only examining the tested individuals.**

Policy Initiative: The Nudging Users to Drive Good Experiences on Social Media (Social Media NUDGE) Act in the United States

In February 2022, Senators Amy Klobuchar (D-MN) and Cynthia Lummis (R-WY) introduced the Social Media NUDGE Act.[69] The bill directs the National Science Foundation (NSF) to work with the National Academy of Sciences, Engineering, and Medicine (NASEM) to "conduct ongoing studies to identify content-agnostic interventions" that the larger social media platforms could implement "to reduce the harms of algorithmic amplification and social media addiction".

The bill defines "content-agnostic interventions" as an action that alters a user's experience and "does not rely on the substance" of content. It lists potential interventions such as "screen time alerts", "labels and alerts that require a user to read or review user-generated content before sharing such content, or "prompts to users, which may help users identify manipulative and microtargeted advertisements".

The US Federal Trade Commission (FTC) would be required to initiate a rulemaking proceeding to determine which of the recommended interventions should be made mandatory.

Pushing for more transparency and reporting obligations, the bill also requires platforms to publicly disclose information about their compliance, the impact of the interventions, and statistics related to required changes and content on their platforms (including the total number of views for each piece of publicly visible content posted during the month and sample randomly from the content). As violations are treated as unfair or deceptive acts or practices, the bill gives enforcement authority to the FTC.

The proposed legislation is seen as unique among various other legislation introduced to Congress that aim to regulate social media platforms in that it does not focus on specific types of content, but on *how* content spreads online.

Ellen P. Goodman from Rutgers Law School notes that while it is "good to support more research on which kinds of interventions are effective", the "content-agnostic" terminology would be unlikely to save the bill from the First Amendment challenges that confront any legislation that touches platform content moderation. Goodman argues that design interventions to demote certain content, or nudge users towards other content, probably will not be "content-agnostic" – highlighting that if interventions are differentially attached to "microtargeted ads or content deemed "manipulative", those interventions are not *per se* agnostic anymore.[70]

Industry: Twitter's Birdwatch (Community Notes)

In 2021, Twitter introduced Birdwatch in the US as a pilot of a "community-driven approach" to help address misleading tweets (initially keeping it separate from Twitter). In brief, contributors, meaning users who sign up for Birdwatch, can identify tweets they believe are misleading, write notes that provide context, and rate the quality of other contributors' notes.

Rather than majority rules or popularity, Birdwatch shows notes that are found to be "helpful" by people who tend to disagree.[71] Hence, in order to be shown on a tweet, Birdwatch notes need to be found "helpful" by contributors who have tended to disagree in their past ratings. This should increase the odds that context added to tweets is helpful to wide audiences. The algorithm technique behind determining which notes are ultimately displayed as "helpful" or "unhelpful" is matrix factorisation on a note-rater matrix. This technique, originally explored by Funk in the 2006 Netflix prize recommender system competition, seeks a latent representation of users and items to explain the affinity of certain users for certain items. For a note to achieve a high intercept term (which is the note's helpfulness score), it must be rated helpful by raters with a diversity of viewpoints (factor embeddings). This way, the algorithm aims to identify notes with broad appeal across viewpoints. Birdwatch includes the top two "explanation tags" that were given by raters to explain why they rated the note helpful or not (e.g., for helpful notes tags include "UnbiasedLanguage", "UniqueContext", "Empathetic", "GoodSources", "ImportantContext"). Birdwatch's algorithm is publicly available on GitHub.[72]

Initial feedback suggests that people valued notes for the "community's voice" (rather than that of Twitter or a central authority) and appreciated that notes provided useful context to help them better understand and evaluate a tweet (rather than focusing on labelling content as "true" or "false").[73] According to Twitter's surveys, with between 3,000 and 19,000 participants run between August 2021 and August 2022, a person who sees a Birdwatch note would be, on average, 20-40 percent less likely to agree with the substance of a potentially misleading tweet than someone who sees the tweet alone with no added context. In October 2022, Twitter announced that it is expanding the visibility of Birdwatch notes that have been rated "Helpful" by contributors to Twitter in the US.[74]

Aviv Ovadya describes Twitter's Birdwatch as bridging-based ranking and proposes to base ranking systems on rewarding content that "leads to positive interactions across diverse audiences, even when the topic may be divisive".[75] A bridging algorithm would act as "centripetal ranking" and reward posts ranked the highest in terms of positive reactions from across diverse audiences. Bridging would not be about showing opposing opinions (which can lead to further divisions), but rank higher content that has more "common ground". Ovadya notes that bridging will require more research on how scoring and ranking can use signals such as user interactions to distinguish "constructive conflict" from "destructive conflict" at platform scale. Another concern could be whether such ranking could end up with lower quality due to skewed collective ratings, as bridging is not dealing with content *per se*, but with the diversity of users engaging positively with that content. Ultimately, as with every other intervention, developing bridging-based ranking will require sufficient levels of transparency to ensure accountability of the outcomes.

In conclusion, bridging-based ranking tackles divisive content as primary concern of online discourses and user behaviour. The approach combines different elements of proposed interventions. It incorporates the discussions on the incentives behind metrics, i.e., changing metrics from rewarding engagement to rewarding "diversity of audience". At the same time, it includes the potential benefits of crowd-sourcing, fact-checking and accuracy nudges. An advantage seems to be that it does not directly concern itself with the toxicity, falsehood, or divisiveness of the content item itself, but aims to incorporate user interactions and behaviour to provide context that could be helpful to users who tend to disagree. While this approach is in the initial stages, the idea behind and the tools of bridging-based algorithms could well be explored further by the research community, platforms themselves as well as policymaking consultations.

Conclusion

The societal effects of people's information consumption and behaviour online, influenced by algorithmic ranking practices of social media platforms, demands a multi-stakeholder approach that is inclusive, transparent and democratic. Transparency about platform methodology (what do ranking algorithms incentivise?) and experiments testing their (unintended) effects on online discourses remains at the core of evidence-based policy interventions.

Algorithmic ranking is not neutral nor is it unchangeable. Contrary, company leadership decisions change the metrics of algorithms all the time. And these metrics can be adapted to help ensure they work toward safeguarding democratic discourses, not toward undermining them.

Given the methodological and epistemic limitations to external research, the "unknown unknowns" sustain a knowledge gap between tech companies on the one side, and policymakers, regulators, researchers and not least the public on the other. Governments should invest in research that aims to build an evidence base on how algorithmic ranking impacts society today. A transparent and collaborative effort to understand algorithmic systems remains central to sound policy interventions and continuous evaluation of their effectiveness.

At the same time, governments should invest in researching already existing measures that aim to tackle the risks of engagement-based ranking. **Evidence on the effects of changing ranking metrics can help to align, and integrate, proposed interventions with the desired outcome of supporting safe and more democratic online discourses.**

Making algorithmic ranking practices work for open and safe online experiences will ultimately comprise various layers of intervention, some that may prove inefficient or counter-productive, and others that could entirely change how liberal-democratic societies interact with those algorithms. Diversity of expertise and perspective will be fundamental to ensuring sufficient re-evaluation and scrutiny of social media platforms and their ranking algorithms. This way, a revitalised multi-stakeholder effort can conquer the prevailing knowledge gap and power imbalances in the online information environment.

Endnotes

1 Koshiyama, A. et al. (2021). Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998

2 Burns, E. (2022). Machine Learning. *TechTarget*. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

3 For example, a qualitative study conducted by ISD indicates how YouTube's algorithms contribute to promoting misogynistic, anti-feminist and extremist content to Australian boys and young men. The report finds that YouTube's 'Shorts' feature seems to optimise more aggressively in response to user behaviour and show more extreme videos within a relatively brief time-frame. See here: Thomas, E. & Balint, K. (2022). Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere'. *ISD*. Available at: https://www.isdglobal.org/isd-publications/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/

4 Make Instagram Instagram again (2022). *change.org*. Available at: https://www.change.org/p/make-instagram-instagram-again-saveinstagram

5 Q2 2022 Earnings (2022). Meta Investor Relations. *Meta*. Available at: https://investor.fb.com/investor-events/event-details/2022/Q2-2022-Earnings/default.aspx

6 Patel, N. (2022, September 13). Everyone knows what YouTube is — few know how it really works. *The Verge*. Available at: https://www.theverge.com/2022/9/13/23349037/mark-bergen-youtube-creators-tiktok-algorithm

7 ISD analysed election disinformation on YouTube, Meta and TikTok in the weeks leading up to 2022 US mid-term elections, demonstrating how platforms failed to prepare for short-form videos as a vector for election disinformation. See here: Martiny, M., Jones, I. & Cooper, L. (2022). Election disinformation thrives following social media platforms' shift to short-form video content. *ISD*. Available at: https://www.isdglobal.org/digital_dispatches/election-disinformation-thrives-following-social-media-platforms-shift-to-short-form-video-content/

8 Chung, A. (2019). News Feeds, Old Content: A Brief History of Algorithmically Curated Feeds on Facebook and Twitter. *Medium*. Available at: https://medium.com/@annawchung/news-feeds-old-content-a-brief-history-of-algorithmically-curated-feeds-on-facebook-and-twitter-85b5e5d8e30a

9 Mark Zuckerberg (2018, January 12). Status update. *Facebook*. Available at: https://www.facebook.com/zuck/posts/10104413015393571#

10 Mosseri, A. (2018). Bringing People Closer Together. *Meta*. Available at: https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/

11 Internal document obtained by CNN. Available at: https://s3.documentcloud.org/documents/21093256/internal-document-obtained-by-cnn.pdf

12 Lada, A., Wang, M., & Yan, T. (2021). How does news feed predict what you want to see? Facebook Newsroom. *Meta*. Available at: https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

13 YouTube (2022). Analytics and Reporting APIs. Metrics. *YouTube*. Available at: https://developers.google.com/youtube/analytics/metrics

14 Meta (2022). Content borderline to the Community Standard. *Meta Transparency Center*. Available at: https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards

15 Patel, N. (2022, September 13). Everyone knows what YouTube is — few know how it really works. *The Verge*. Available at: https://www.theverge.com/2022/9/13/23349037/mark-bergen-youtube-creators-tiktok-algorithm

16 Metz, R. (2021, October 27). Likes, anger emojis and RSVPs: the math behind Facebook's News Feed — and how it backfired. *CNN Business*. Available at: https://edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math/index.html

17 Hindman, M., Lubin, N. and Davis, T. (2022, February 10). Facebook Has a Superuser-Supremacy Problem. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/

18 Edelman, G. (2021, November 19). How Facebook Could Break Free From the Engagement Trap. *Wired*. Available at: https://www.wired.com/story/jeff-allen-interview-facebook-engagement-trap/

19 Davidson, T., Bhattacharya, D. and Weber, I. (2019) Racial Bias in Hate Speech and Abusive Language Detection Datasets. *arXiv*. Available at: https://arxiv.org/pdf/1905.12516.pdf

20 Noble, S. (2018). Algorithms of oppression. Databite No. 109. *Data & Society*. Available at: https://datasociety.net/library/safiya-umoja-noble-algorithms-of-oppression/

21 Simonite, T. (2021, October 25) Facebook Is Everywhere; Its Moderation Is Nowhere Close. *Wired*. Available at: https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

22 Ovadya, A. (2022). Bridging-Based Ranking. *Belfer Center*. Available at: https://www.belfercenter.org/publication/bridging-based-ranking

23 Chung, A. (2019). News Feeds, Old Content: A Brief History of Algorithmically Curated Feeds on Facebook and Twitter. *Medium*. Available at: https://medium.com/@annawchung/news-feeds-old-content-a-brief-history-of-algorithmically-curated-feeds-on-facebook-and-twitter-85b5e5d8e30a

24 O'Neil, C. (2017) *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

25 Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of Machine Learning Research 81:1–15, *2018 Conference on Fairness, Accountability, and Transparency*. Available at: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

26 Ada Lovelace Institute (2021) Technical methods for regulatory inspection of algorithmic systems. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

27 Official Journal of the EU (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (Digital Services Act). Volume 65. 27 October 2022. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2022:277:FULL&from=EN

28 For more information on the European Centre for Algorithmic Transparency (ECAT), see the website of the European Commission available at: https://digital-strategy.ec.europa.eu/en/policies/ecat [Accessed 31 Oct. 2022]

29 Vermeulen, M. (2022). Researcher Access to Platform Data: European Developments. *Journal of Online Trust and Safety*. Vol. 1 No. 4 (2022). Available at: https://tsjournal.org/index.php/jots/article/view/84

30 European Commission (2022). 2022 Strengthened Code of Practice on Disinformation. *European Commission*. Available at: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

31 The Code considers "Disinformation" to include "misinformation", "disinformation", "information influence operations", and "foreign interference in the information space", which are defined in the European Commission's Communication on the European Democracy Action Plan, p.18. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

32 Vermeulen, M. (2022). Researcher Access to Platform Data: European Developments. Journal of Online Trust and Safety. Vol. 1 No. 4 (2022). Available at: https://tsjournal.org/index.php/jots/article/view/84

33 Guhl, J., Marsh, O. and Tuck, H. (2022). Researching the Evolving Online Ecosystem: Barriers, Methods and Future Challenges. *ISD*. Available at: https://www.isdglobal.org/wp-content/uploads/2022/07/Researching-the-Evolving-Online-Ecosystem_Main-report.pdf

34 Roumeliotis, G. (2019, November 4). Exclusive: U.S. Opens National Security Investigation into TikTok – Sources. *Reuters*. Available at: https://www.reuters.com/article/us-tiktok-cfius-exclusive-idUSKBN1XB4IL

35 O'Connor, C. (2021) Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok. Available at: https://www.isdglobal.org/wp-content/uploads/2021/08/HateScape_v5.pdf

36 Camargo, C. Q. and Simon, F. M. (2022). Mis- and disinformation studies are too big to fail: Six suggestions for the field's future. *Harvard Kennedy School Misinformation Review*. September 2022, Volume 3, Issue 5. Available at: https://misinforeview.hks.harvard.edu/article/mis-and-disinformation-studies-are-too-big-to-fail-six-suggestions-for-the-fields-future/

37 See upcoming paper: Meßmer, A.-K. and Degeling, M. (202x). Auditing Recommender Systems with Risk Cards. *Stiftung Neue Verantwortung*. Available here: https://www.stiftung-nv.de/en/subproject/approaches-analyse-and-evaluate-ai-based-recommendation-systems-internet-intermediaries

38 Dimson, T. (2022). How Recommendation Algorithms Actually Work. *Future*. Available at: https://future.com/forget-open-source-algorithms-focus-on-experiments-instead/

39 Christchurch Call (2022). *Christchurch Call Initiative on Algorithmic Outcomes*. Available at: https://www.christchurchcall.com/media-and-resources/news-and-updates/christchurch-call-initiative-on-algorithmic-outcomes/

40 White House (2022). *Blueprint for an AI Bill of Rights*. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

41 Masnick, M. (2019) Protocols, Not Platforms: A Technological Approach to Free Speech. *Knight First Amendment Institute*. Available at: https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

42 Gobo.social. MIT Media Lab's Center for Civic Media. Available at: https://www.media.mit.edu/projects/gobo/overview/

43 Block Party. Available at: https://www.blockpartyapp.com/

44 Twitter (2022). Twitter Toolbox. *Developer Platform*. Available at: https://developer.twitter.com/en/community/toolbox

45 Perspective (2022). Available at: https://developers.perspectiveapi.com/s/about-the-api

46 Keller, D. (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy*. Vol. 32. Issue 3. 168-72. Available at: https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/

47 Marsh, O. (2022). Social Media Futures: Interventions Against Online Unpleasantness. *Tony Blair Institute for Global Change*. Available at: https://institute.global/policy/social-media-futures-interventions-against-online-unpleasantness

48 Keller, D. (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy*. Vol. 32. Issue 3. 168-72. Available at: https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/

49 OECD (2021). Data portability, interoperability and digital platform competition. *OECD Competition Committee Discussion Paper*. Available at: http://oe.cd/dpic

50 Fukuyama, F. et al.. Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy. *Stanford Cyber Policy Center*. 3. Available at: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf

51 BIT, CDEI and Doteveryone (2022). Active Online Choices: Designing to Empower Users. *Centre for Data Ethics and Innovation (CDEI)*. Available at: https://www.bi.team/wp-content/uploads/2021/08/CDEI-Active-Online-Choices_Final-Report-1.pdf

52 eSafety Commissioner (2021). Development of industry codes under the Online Safety Act. Position Paper. *eSafety Commissioner*. Available at: https://www.esafety.gov.au/sites/default/files/2021-09/eSafety%20Industry%20Codes%20Position%20Paper.pdf

53 eSafety Commissioner (2021). Online Safety Act 2021. Fact sheet. *eSafety Commissioner*. Available at: https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf

54 Bush, D. and Zaheer, A. (2019) Bing's Top Search Results Contain an Alarming Amount of Disinformation. Freeman Spogli Institute for International Studies. Stanford University. Available at: https://fsi.stanford.edu/news/bing-search-disinformation

55 Meta (2019). People, Publishers, the Community. *Facebook Newsroom*. Available at: https://about.fb.com/news/2019/04/people-publishers-the-community/

56 Meta (2022). Types of content we demote. *Transparency Center*. Available at: https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/

57 Smith, J., Leavitt, A. and Jackson, G. (2018). Designing New Ways to Give Context to News Stories. *Facebook Newsroom*. Available at: https://about.fb.com/news/2018/04/inside-feed-article-context/

58 Singh, S. (2021). Facebook Releases Information on Algorithmic Content Ranking, but More Transparency Is Needed. Open Technology Institute. *New America*. Available at: https://www.newamerica.org/oti/blog/facebook-releases-information-on-algorithmic-content-ranking-but-more-transparency-is-needed/

59 Google (2019). How Google Fights Disinformation, p.12. *Google*. Available at: https://kstatic.googleusercontent.com/files/388aa7d18189665e5f5579aef18e181c2d4283fb7b-0d4691689dfd1bf92f7ac2ea6816e09c02eb-98d5501b8e5705ead65af653cdf94071c47361821e362da55b

60 Ibid, p. 4, p.12.

61 Google (2022). Quality Raters. General Guidelines, p. 23. *Google*. Available at : https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

62 Chang, J. (2017). Identifying credible content online, with help from the Trust Project. *Google Keyword*. Available at : https://blog.google/outreach-initiatives/google-news-initiative/sorting-through-information-help-trust-project/

63 Google (2019) What site owners should know about Google's August 2019 core update. Google Search Central Blog. Available at: https://developers.google.com/search/blog/2019/08/core-updates

64 Google Patents. Producing a ranking for pages using distances in a web-link graph. *Google Patents*. Available at: https://patents.google.com/patent/US9165040B1/en

65 Allen, J. (2021). The Integrity Institute's Analysis of Facebook's Widely Viewed Content Report (Q4 2021 - Q1 2022). *Integrity Institute*. Available at: https://integrityinstitute.org/widely-viewed-content-analysis-tracking-dashboard

66 Brown, S. (2019). Social media is broken. A new report offers 25 ways to fix it. *MIT Sloan School of Management*. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/social-media-broken-a-new-report-offers-25-ways-to-fix-it

67 Parliament of Australia (2022). Select Committee on Social Media and Online Safety - 03/02/2022 - Online harms that may be faced by Australians on social media and other online platforms. *Parliamentary Business*. Available at: https://www.aph.gov.au/Parliamentary_Business/Hansard/Hansard_Display?bid=committees/commrep/25635/&sid=0000

68 Walsh, D. (2020). Study: 'Accuracy nudge' could curtail COVID-19 misinformation online. *MIT Sloan School of Management*. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/study-accuracy-nudge-could-curtail-covid-19-misinformation-online

69 Congress.gov (2022). S.3608 - Social Media NUDGE Act. *Congress.gov*. Available at: https://www.congress.gov/bill/117th-congress/senate-bill/3608?s=1&r=5

70 Goodman, E.P. (2022). Assessing the NUDGE Act. *Tech Policy Press*. Available at: https://techpolicy.press/assessing-the-nudge-act/

71 Birdwatch (2022). Note ranking. Twitter, Inc.. Available at: https://twitter.github.io/communitynotes/ranking-notes/

72 Birdwatch (2022). Matrix Factorization. *Twitter, Inc.*. Available at: https://twitter.github.io/communitynotes/ranking-notes/#matrix-factorization

73 Coleman, K. (2021). Introducing Birdwatch, a community-based approach to misinformation.*Twitter, Inc.*. Available at: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

74 Birdwatch (2022). Helpful Birdwatch notes are now visible to everyone on Twitter in the US. *Twitter, Inc.*. Available at: https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us

75 Ovadya, A. (2022). Bridging-Based Ranking. *Belfer Center*. Available at: https://www.belfercenter.org/publication/bridging-based-ranking

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdgermany.org

gefördert durch:

Auswärtiges Amt