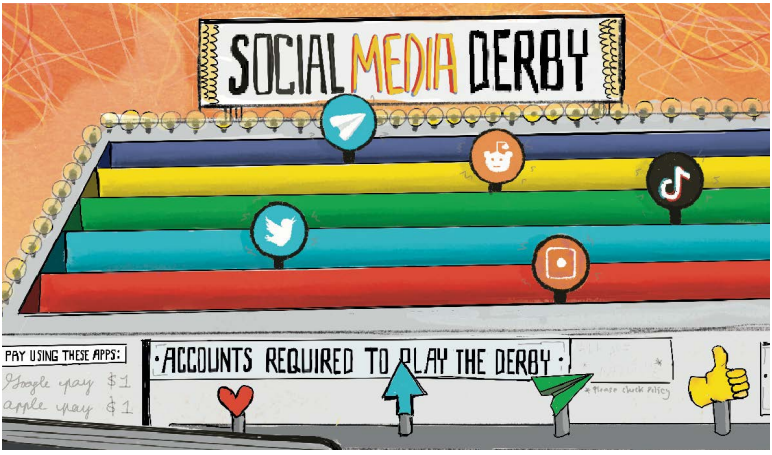
OPTIMIZING FOR WHAT?
ALGORITHMIC AMPLIFICATION AND SOCIETY



What's in an Algorithm? Empowering Users Through Nutrition Labels for Social Media Recommender Systems

By Luca Belli and Marlena Wisniak

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In April 2023, the Knight Institute hosted a symposium, "Optimizing for What? Algorithmic Amplification and Society," to explore how law and policy regulates or should regulate the algorithms that power social media platforms' recommendation systems and the challenges and opportunities that arise from these systems. The symposium was organized in partnership with the Institute's 2022-2023 Visiting Senior Research Scientist Arvind Narayanan and took place at Columbia University.

The essays in this series were originally presented and discussed at this event. Written by leading scholars, nonprofit leaders, technologists, and professionals from a wide range of disciplines, including computer science, psychology, law, social science, philosophy and other fields, these essays offer more precise explanations of how algorithms are designed, deployed, and evaluated and propose interventions that would mitigate some of the harms caused by amplification. This series also engaged with normative questions about algorithmic recommenders: What should they optimize for? How do we design systems to promote healthier public discourse? Who makes these decisions?

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, executive director; Katy Glenn Bass, research director; Arvind Narayanan, visiting senior research scientist; Alex Abdo, litigation director; and Larry Siems, chief of staff. The essay series was edited by Glenn Bass and Narayanan with additional support from Lorraine Kenny, communications director; Victoria Tang, legal research fellow; Mia Speier, research coordinator; Kushal Dev, research fellow; Avian Muñoz, intern; and Stephen Dai, intern.

The full series is available at knightcolumbia.org/research/

OVERVIEW OF RECOMMENDER SYSTEMS AND TRANSPARENCY

Enabling algorithmic transparency for users and affected communities

Social media platforms can decide what content remains online and what is removed, a process known as content moderation. The binary take down/leave up model can be extended with the option to downrank or demote problematic content or, conversely, to amplify content with recommender systems.[1] Downranking can limit the restrictions on users' freedom of expression by reducing the exposure to their content, rather than removing it. Other content can be amplified to users, otherwise known as algorithmic amplification. In both cases, questions remain around the transparency of how these systems operate, what type of content is shown or hidden, and why. This is especially important given the dominant role and resulting influence

1 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act) [DSA, 2022], Recital 55: "Restriction of visibility may consist in demotion in ranking or in recommender systems, as well as in limiting accessibility by one or more recipients of the service or blocking the user from an online community without the user being aware ('shadow banning')."

that recommender systems have on social media platforms, as demonstrated by Leerssen (2023, p. 46). As a result, the regulatory, policy, and business conversations on platform governance and strict content moderation are shifting towards content governance with a focus on curation. As Leerssen (2023, p. 45) argued, "a rare point of consensus for both scholars and policymakers appears to be the need for greater transparency" in recommender systems. It is critical to better understand, measure, and document algorithmic amplification. Understanding algorithmic amplification (see Thorburn et al. 2022) is necessary to explore the processes and tools that empower users to proactively shape and assess the content they engage with.

Narayanan (2023a) stated that, while recommender systems are understood by computer scientists, the same is not necessarily true for the broader public, especially marginalized groups who are most at risk of harm. Understanding recommender systems is not enough to prevent harm and promote content that enables users to exercise their human rights. Users and affected communities need to understand how these systems operate on a given platform; they need to know why and how the content they see is recommended to them so that they can assess its impact.

This paper is based on the premise that meaningful transparency is needed to empower users and affected communities to enhance their engagement with social media platforms. This requires centering the users and affected groups as the target audience and providing information about recommender systems to them in an understandable way. Yet consumer-facing products, such as social media platforms, often require less technical information (Partnership on AI, 2021). This paper proposes a set of metrics to measure algorithmic applications for recommender systems that we call nutrition labels, which are user-friendly notices that can operationalize transparency goals at scale. Nutrition labels are surely no silver bullet; many challenges related to business and profit maximization, implementation, operationalization, and digital literacy can severely limit the effectiveness and impact of such labels. Nonetheless, we believe that labels can be a helpful tool in the broader algorithmic transparency toolkit and are, thus, worth exploring further. Civil society and affected communities are key stakeholders in the design and content of the labels and are critical to ensuring that meaningful transparency is attained.

Compliance with emerging legal obligations

The U.N. Guiding Principles (UNGPs) on Business and Human Rights (U.N. Human Rights Council, 2011), unanimously endorsed in 2011, are the leading international instrument outlining the responsibilities of private sector actors to respect human rights. Human rights due diligence, which consists of assessing and addressing the adverse human rights risks of a company's products and services, is a core part of the framework. While non-binding, the framework has been applied across the tech sector, leading to transparency and reporting mechanisms on issues as diverse as content moderation, government requests, and appeals processes. The UNGPs call for adequate internal and external reporting of products and services. In the case of social media platforms and algorithmic content curation, this can be achieved in part through nutrition labels for recommender systems.

As of November 2022, reporting platforms operating in the European Union (EU) are subject to strict legal requirements related to transparency under the Digital Services Act (DSA). The DSA, today's most comprehensive, legally binding instrument for digital platforms, defines recommender systems as:

> A fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service or prioritize that information, including as a result of a search initiated by the recipient of the service or otherwise determining the relative order or prominence of the information displayed (DSA, 2022, Article 3[s]).

Regarding amplification, the DSA states:

> [Recommender systems] also play an important role in the amplification of certain messages, the viral dissemination of information, and the stimulation of online behaviour. Consequently, online platforms should consistently ensure that **recipients of their service are appropriately informed about how recommender systems impact the way information is displayed** and can influence how information is presented to them. They should clearly **present the parameters for**

> **such recommender systems in an easily comprehensible manner** to ensure that the recipients of the service understand how information is prioritised for them.
>
> Those parameters should **include at least the most important criteria in determining the information suggested** to the recipient of the service **and the reasons** for their respective importance, including where information is prioritised based on profiling and their online behaviour. (DSA, 2022, Recital 70).

The devil is always in the details; determining what constitutes "appropriate information" and "comprehensible manner," what parameters should be presented, and which criteria and reasons should be included therein are open questions. We propose an approach analogous to the food and beverage industry by creating nutrition labels (based on metadata), which could help meet these requirements. Deciding on what parameters should be disclosed in the labels, however, remains a challenge; we urge policymakers and industry representatives to include external stakeholders, such as academics, civil society, and affected groups, in the decision-making process.

Very large online platforms and search engines have additional obligations. Among other obligations, they are required to conduct risk assessments of their services, which explicitly includes "the design of their recommender systems and any other relevant algorithmic systems" (DSA, 2022, Article 34(2)a). As a corollary to this obligation, the identified risks must be addressed through mitigation measures including adapting recommender systems to "mitigate the negative effects of personalised recommendations and correct the criteria used in their recommendations" (DSA, 2022, Article 35, Recital 88). Our proposal for nutrition labels, as presented below, could contribute to platforms' due diligence responsibilities under the DSA. Indeed, the recommender systems of very large online platforms, as defined by the DSA, have been listed as high-risk in the latest version adopted by the European Parliament and, thus, require heightened reporting and due diligence

obligations.[2] Nutrition labels could also respond to future obligations under the upcoming EU Artificial Intelligence Act.

Other non-binding norms and standards include the ones developed by the National Institute of Standards and Technology (NIST, 2023) and the Organization for Economic Co-operation and Development (OECD, n. d.).

Scope of the paper

This paper centers on recommender systems for user-generated content (UGC).[3] Although the focus is on content recommendations and not specific types of content, it is helpful to propose an overarching content taxonomy consisting of three categories: illegal content, content that violates platforms' terms of service, and non-violative content. First, illegal content includes, among others, terrorist content, child sexual abuse material, and content that violates laws such as copyright, anti-discrimination, and international human rights law (e.g., incitement to violence or war propaganda). Second, content that violates policies varies from platform to platform but typically includes hate speech, harassment, and misinformation. Lastly, all content includes anything that does not violate the law or platforms' terms of service and is allowed on the platform. This ranges from content that can be considered newsworthy or in the public interest, counterspeech, and satire, to potentially problematic or toxic content that is nonetheless allowed.

Although this paper does not aim to assess the human rights impacts of amplification, it is still important to note that recommender systems have significant human rights implications. These include, among others: freedom of expression; freedom of assembly and association; freedom of

---

2 AI systems intended to be used by social media platforms that have been designated as very large online platforms within the meaning of Article 33 of Regulation EU 2022/2065, in their recommender systems to recommend to the recipient of the service user-generated content available on the platform. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1). Amendment 740, Proposal for a regulation, Annex III – paragraph 1 – point 8 – point a, b. See also Amendment 73, Proposal for a regulation, Recital 40 b.

3 While we recognize the importance of recommender systems for advertising and acknowledge the relevance of DSA provisions focusing on the transparency thereof, we have chosen to exclude advertising from the scope of this paper.

information; right to privacy; right to life, liberty, and security; economic, social, and cultural rights; and prohibition of discrimination.[4]

PRIOR WORK ON METADATA FOR AI

IN RECENT YEARS, a variety of new metrics have been developed in the field of Responsible AI. Starting with classification problems, many of these metrics were proposed to address the disparate outcomes of different models. Many of these definitions are mutually exclusive. For example, it is possible to mathematically prove that a classification system cannot simultaneously satisfy two popular metrics, such as equalized odds and statistical parity (e.g., Garg et al. (2020)). Each metric has an embedded value judgment (Selbst et al. 2019), and choosing one over the other is a subjective, political decision (Narayanan, 2018). A key challenge for practitioners today is the lack of shared guidelines and understanding to navigate the ever-growing portfolio of available metrics. As a side note, many practitioners equate Responsible AI work with balancing as many metrics as possible, thus failing to account for the actual issues.

Efforts to provide metadata for data sets or models were also proposed, such as the seminal work of Gebru et al. (2021) and Mitchell et al. (2019). One valuable proposed addition is to explicitly state the intended use of both the data sets and the models. A data set collected or a model developed for a specific task or goal might not be appropriate for others. Communicating the intended purpose of an AI system can help prevent future misuse or function creep, which would have significant implications for human rights. Function creep can inadvertently lead to harm because data sets or models could be used by practitioners who are less familiar with the problem domain, or who have less understanding of the human rights and societal impacts of their application. Furthermore, function creep can also enable intentional

4 The European Commission's Code of Practice on Disinformation recognizes this and requires platforms to "[d]ilute the visibility of disinformation by improving the findability of trustworthy content" and to "invest in technological means to prioritize relevant, authentic, and authoritative information" (DSA, 2022, Recital 84).

misuse or abuse of AI systems, for example when governments or powerful actors use AI systems for surveillance, censorship, discrimination, or other human rights violations.

Yang et al. (2018) proposed "a nutritional label for rankings," building an interpretability tool called Ranking Fact. Focusing on fairness and diversity, they created multiple widgets (such as Recipe, Ingredients, Stability, Fairness, and Diversity) to convey information on the ranking algorithm. Similar to our proposal, the authors developed a system that imitated the standardized labels in the food industry to inform users about the "ingredients and production processes." They considered three areas: rankings of computer science departments, criminal risk assessments, and credit and loan ranking.

These previous approaches by scholars have had a significant influence on algorithmic transparency and moved the needle toward fairness and accountability in machine learning. Our proposal for nutrition labels for social media recommender systems builds on their important work while aiming to address some of their limitations. First, such definitions assume, by design, the full knowledge of the demographic characteristics of each user to compare performance across groups. This is not always possible; even when it is possible, it raises questions about the ethical use of data. Once a piece of data is collected, even if it is to ensure model quality or lack of bias, or a model is built, it is hard to ensure that it will never be misused in the future. Furthermore, the demographic categories commonly used today are often based on U.S. cultural and historical norms. Other categories are typically based on international human rights law, corresponding to the core international human rights treaties (specifically protecting the rights of women, racialized persons, migrants and refugees, children and the elderly,

and disabled persons, among other groups)[5]. As such, demographic categories are social and legal constructs.[6] Although international human rights law has a global scope, it is important to note that local contexts and norms should be considered when identifying marginalized and vulnerable groups.

Second, while prior work on metadata has been critical for some issues, it has mostly focused on classification problems and omits broader societal and human rights considerations besides non-discrimination. In this paper, we argue that similar documentation processes and artifacts are needed for recommender systems, especially those that curate UGC on social media platforms. We also develop this by endorsing the documentation and external communication of metadata that are not related to risks of bias and discrimination only but also risks to other rights, such as freedoms of expression, information, assembly and association, privacy, as well as democracy and rule of law more broadly.

NUTRITION LABELS FOR RECOMMENDER SYSTEMS

AS MENTIONED ABOVE, our proposal expands on previous efforts, without only focusing on fairness and contextualizing it for recommender systems. We propose to measure algorithmic amplification for a recommender system where users are both producers and consumers of content (e.g., social media).

What is amplified and why are questions that remain without sufficient attention or practical answers. While recommender systems, especially those

---

5 International Convention on the Elimination of All Forms of Racial Discrimination (1965); International Covenant on Civil and Political Rights (1966); International Covenant on Economic, Social and Cultural Rights (1966); Convention on the Elimination of All Forms of Discrimination against Women (1979); Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (1984); Convention on the Rights of the Child (1989); International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (1990); International Convention for the Protection of All Persons from Enforced Disappearance (2006); Convention on the Rights of Persons with Disabilities (2006). See United Nations Human Rights Office of the High Commissioner of Human Rights, The Core International Human Rights Instruments and Their Monitoring Bodies.

6 The concepts of race and racial belonging are perceived differently in different countries and cultures.

deployed in industry settings, are trained to maximize engagement, who or what kind of content gets amplified remains unanswered.

Like Kelley (2009), the Partnership on AI (2021), Food and Drug Administration (FDA) (2021), Apple (n. d.), and Google (n. d.), we take inspiration from the Nutrition Facts Label in the U.S. and food labeling rules in Europe, which provide standardized information for consumers to compare different items and understand the food they consume. A standardized list of nutrients (e.g., fat, cholesterol, sodium, total carbohydrate, protein) is presented in these labels, together with the total nutrient value and percentage of daily value.

Similarly, we propose that a recommender system should report the equivalent values for amplification. This requires defining both the nutrient groups and the unit of measurement (including where the zero lies); in other words, what would be the equivalent of sodium or protein for a recommender system? Similarly, the unit of measurement must be defined for recommender systems. In the case of food nutrients, the use of the gram (the international unit for weight) is the obvious choice. For recommender systems, there is no obvious choice. What does it mean to be 'one more unit amplified'? This section seeks to address this issue. Like measuring nutrients, there is no singular way to measure algorithmic amplification, as amplification is a complex phenomenon understood by multiple metrics, with each metric representing a different angle or value judgment. Unlike food, however, where such labels are static and can be printed on the container, recommender systems are dynamic objects that respond to and influence users in a reciprocal feedback loop, with an impact that is deeply dependent on the local political, cultural, and social context. Imagine the challenge of trying to fully picture the complex three-dimensional shape of a moving object by looking at multiple two-dimensional shadows projected from different angles; we do not envision a static system, but rather a collection of dashboards or time series.

Limitations

It is important to acknowledge some significant differences between our case and the general case of nutrition. First, we recognize that the science behind nutrition labels for food is much more consolidated and agreed upon than for

recommender systems. Indeed, an overwhelming majority of experts agree on what a balanced diet should comprise. Second, there are authoritative bodies that are responsible for creating and enforcing labels based on the latest science, and, if needed, open investigations;[7] recommender systems lack both an authoritative, legitimate enforcing body and a consolidated science.

Recommender systems for user-generated content networks

This paper centers on UGC networks, which are content networks in which users follow each other. In such systems, each user is both a consumer and a producer of content; some platforms also allow users to reshare content created by others to their network. Furthermore, we assume that some of the content might be algorithmically delivered (i.e., not via a chain of reshares) by users who do not explicitly follow the author. Using the terminology of Narayanan (2023b), we assume that our system can propagate information via subscriptions, networks, and algorithms.

Before attempting to define algorithmic amplification, we must understand what happens when a user posts a piece of online content. The content is delivered via a mixture of hard-coded rules and algorithmic systems to different users, some but not all of whom follow the author. Would we expect all the users who follow the author to see their content? It depends.

There are multiple relevant factors here; some are specific to each reader and some are shared by all readers. In the context of the reader, each one spends a variable yet finite amount of time on the platform, remaining engaged only as far as they are interested and available. Where content is shown in the reader's feed and how it is prioritized is also vital; since the time spent on the platform is finite, it is mostly a zero-sum game. If the reader only sees the first n pieces of content, then everything that is ranked in position n+1 onwards will be ignored. Notably, this is why many companies have moved to algorithmic feeds. Since algorithmic feeds increase the time users spend on the platform, they not only increase the amount of content

7 Note that in some cases there are disparities in how science is interpreted. For example, in the U.S., it is required to list sodium, while in the EU, salt is presented instead.

and advertisements seen but also collect more data for subsequent targeted advertising. Yet, even if time spent on the platform might increase, it still is finite and does not affect the underlying line of thinking. Finally, there are also endemic factors to be considered; for example, large and exceptional events, such as sports games or natural disasters, might dominate various feeds and non-relevant content might be ranked lower. At the same time, do readers who do not follow a particular author see their content? If so, how many posts or pieces of content will they see? Is such a reach organic or algorithmically inflated? Once again, it depends. If the platform allows for content resharing (e.g., the network model), the content might naturally spread beyond the author's followers. This type of spread is user-generated because the users choose to reshare it. As argued above, there is no guarantee that the reshared content will be seen, as it depends on how it will be ranked in each reader's feed. However, at each iteration, users might still decide to reshare the content themselves, creating a cascading effect that makes the content go viral organically.

Finally, users might also see content from authors that they do not explicitly follow; popular or trending content is usually a candidate for these kinds of recommendations. In some cases, explaining why the content was explicitly recommended could increase trust in the system, mitigate concerns about bias, and even absolve the platform for faulty and/or inappropriate recommendations. Some common explanations for recommended content across all major social media platforms include "content similar to X," "because person X also liked this content," or "trending topic."

A note on systems

Modern recommender systems are a complex combination of different systems that interact together. In deployed systems, it is common to have a candidate generation step, a ranker step, and post-ranking rules that might satisfy business or product needs. Furthermore, the websites or apps that users interact with are not just collections of backend systems but also have product and design choices embedded therein. Since those choices influence how users interact with the system, they also influence how content is amplified. The platform performs various A/B tests at a given time. This begs the question: what stage(s) or version of the recommender system is

being focused on to measure amplification? In this paper, we focus on the end-to-end user experience of most users. We ignore all the sub-components (i.e., considering everything as a black-box inscrutable model) and focus on the main version of the algorithm that serves most users (e.g., the one that other models are measured against).

Measuring amplification with nutrition labels

Amplification of a post or piece of content is derived from the corresponding ranking in the readers' feeds. A similar definition can be derived for an author (or a group of authors) by looking at the amplification of the content they create.

Let us consider two extreme examples. In the first case, content that is consistently ranked last, regardless of how interesting or relevant it is, may seldom or never be seen by readers. Conversely, content that is always ranked highly is much more likely to be seen and/or actioned by users. The open-source parts of Twitter's ranking system offer a real-world case of the latter example, where posts created by the website's owner receive an artificial boost, regardless of users' level of interest or if users actively try to hide the posts from their feeds (Twitter, 2023).

A first attempt to define amplification could be to count the number of unique, raw impressions that the content receives, i.e., how much other users see it. As assumed above, not all ranked content will is seen in practice. However, this definition does not consider the size of a user's following. Those with a larger following are expected to receive more impressions than those with a smaller one. However, the situation might be more complex, as a second-order network effect might happen if users reshare that content themselves. In this case, the actual reach (i.e., impressions) of the content might be greater than the size of the original network. Similarly, it does not consider how often each author posts. While, in theory, it is only the quality of the content that determines its reach, people might get fatigued or bored if they look at too much content from the same author.

Suggested nutrition labels for recommender systems

TOP PRODUCERS

To circumvent this problem, Belli et al. (2022) devised the concept of normalized impressions, i.e., the impressions that each author receives per user and piece of content. This is a useful metric to compare users with different network sizes and different amounts of content produced. However, the assumption that the relationship between network size, amount of content produced, and reach is linear might not be true for real networks.

Considering the above, the first proposed nutrition label is the highest and lowest producers (by absolute number or percentage) by normalized impressions each time.

IN- AND OUT-OF-NETWORK PRODUCERS

When content is delivered to someone who is explicitly following the author, a recommendation is considered in-network (IN). Conversely, we refer to out-of-network (OoN) recommendations. The case of users resharing the content could be argued either way; even if the reader might not explicitly follow the original author, an argument could be made that someone they do follow is chosen by explicitly sharing this content with their network. However, the previous nutrition label only considers total impressions without considering the type of impression (i.e., how they were delivered); IN and OoN recommendations are treated equally.

Let us consider the following scenario: two users, A and B, have the same number of total normalized impressions, but A's impressions are 10% IN and 90% OoN, whereas B's impressions are split evenly 50/50. If we were to consider total impressions only, as, in the above example, both users would be equally amplified. We argue, however, that A should be considered more algorithmically amplified than B given the source of impressions. Another way of thinking about this is to consider IN impressions as organic (or expected) and OoN ones as an editorial choice of the platform, even if it is only implicit.

We, therefore, propose to use the ratio of IN to OoN normalized impressions as a second nutrition label.

Defining groups and content of interest

The above definitions can naturally be extended to groups of users via summary statistics (e.g., mean, average) of the amplification of users who

comprise the group. Given the scale of platforms and their number of users, it is important to reflect on whether nutrition labels should apply to all users, or only be limited to some groups of users whose content is likely to pose high human rights risks. Indeed, reporting such measures for the totality of authors is neither actionable nor effective for meaningful transparency. From a computational perspective, it would require too many resources and, from a reporting perspective, it may lead to an overload of information. Considering that dominant social media platforms tend to have hundreds of millions of users, it would be hard to sift among all of them to find relevant and critical information. For example, one could track the amplification of people who like dogs and report that as nutrition labels. However, information about this group is mostly inconsequential and unlikely to significantly impact human rights and democracy.

Regarding groups of interest, we believe it is important to monitor elected officials. Following Husz´ar et al. (2022), who demonstrated that the political right is more amplified on Twitter than the political left in most countries, it makes sense to consider elected officials as a group of interest. Similarly, we believe platforms should track the amplification of single accounts for high-profile politicians, such as presidents or prime ministers, as their content may cause or contribute to harm. That said, determining what type of content and accounts should be measured and scrutinized is a sensitive decision with many associated values. We argue that it is very context-dependent and impossible to give guidance in the abstract.

Furthermore, tracking and reporting how individuals are amplified can potentially have implications regarding data protection and privacy law, as they inevitably require more processing and sharing of personal data. Platforms may have to assess whether affected users have a right to object to the nutrition label created about them and, if so, what internal complaints mechanisms should exist. In any case, platforms should establish clear criteria for determining who is subjected to nutrition labels and duly inform those affected in advance. As consistent with international human rights laws and the constitutional law of democratic countries worldwide, the right to privacy of public figures can be restricted to enable public debate and respect the right to freedom of expression and information. By limiting nutrition

labels to high-visibility users, we believe that the potential infringements on privacy could be proportionate and necessary in the context of social media platforms.

Defining the target group for nutrition labels is no small feat and should be done collectively with external stakeholders, such as civil society and academics. For example, online political discourse does not only include state representatives and political candidates. From participants in the Arab Spring to Black Lives Matter, users such as activists, protesters, political dissidents, journalists, grassroots movements, and civil society organizations, especially those from marginalized and vulnerable groups, have shown that they are a critical part of the online civic space. Understanding whether some of their content has been demoted can help assess platforms' impact on users' freedoms of expression, information, assembly, and association.

Applying nutrition labels to a narrow scope of content might also be more impactful than monitoring all content available on a platform for amplification. Indeed, given the sheer volume of content that is produced every second (Twitter, 2014), there is no way to label and monitor it all in real time. However, interesting categories of content can still be found to investigate. Platforms have their terms of service that prescribe what type of content is prohibited. Furthermore, platforms might also want to limit exposure to some content that does not violate their terms of service but that, nonetheless, should not be amplified because it is seen as problematic or 'borderline,' e.g. see Twitter (n. d.[b]), Yee et al. (2022). Limiting visibility and engagement could be achieved via a combination of design choices (Twitter, n. d.[a]) or technical interventions (e.g., by downranking such content in recommender systems). Yet, we have seen repeatedly how such content can be widely shared before being actioned, thus creating harm. Focusing on such content would enable assessments of how much content was amplified, even though it was later deemed to violate the platform's terms of service.

Responsibility for nutrition labels and data sharing

As is current practice in the food industry, companies and platforms should be responsible for funding, creating, and maintaining nutrition labels. That said, external, independent, and authoritative bodies should be able to audit and verify that labels are accurate and trustworthy.

The proposed labels would require data that is not currently available via most Application Programming Interfaces (APIs). For instance, impression data is unavailable, except for some aggregate data that would not enable such audits. Nonetheless, the DSA now requires platforms to give access "to data that are necessary to monitor and assess compliance with this Regulation."[8] We believe that our proposal responds to these obligations and helps define the kind of data sharing required by platforms.

ENGAGING EXTERNAL STAKEHOLDERS IN DEVELOPING NUTRITION LABELS

THE FORMAT AND CONTENT of nutrition labels should be developed in close consultation with external stakeholders, especially marginalized and vulnerable groups. such as women and non-binary persons, racialized persons, LGBTQIA+ persons, migrants and refugees, disabled persons, children and the elderly, and those of lower socioeconomic status, among others. Engaging with users and affected communities is particularly effective in developing an understanding of the potential problems or opportunities of recommender systems. This also helps to determine what information is useful for these groups and how it can be conveyed. By centering the risks and concerns of these groups and designing mechanisms that address them, platforms will likely address the concerns of all groups.

We urge platforms to take a decolonial, trans-feminist, and anti-racist lens when developing communication and documentation strategies. This can take many forms but, at its core, it aims to ensure that the language and content center around the needs and concerns of marginalized groups around the world, especially those in the Global Majority. To best understand and address these needs, marginalized groups, including regional and local civil society and affected communities, need to meaningfully engage in the

8 DSA, 2022, Article 40(1): "Providers of very large online platforms or very large online search engines shall provide the Digital Services Coordinator of establishment or the Commission, at their reasoned request and within a reasonable period specified in that request, access to data that are necessary to monitor and assess compliance with this Regulation."

process.

Speaking with diverse groups can enable a deeper understanding of their various needs and concerns; it also makes clear that there will never be a one-size-fits-all option. Users and affected communities of diverse backgrounds have highly relevant knowledge, observations, expertise, and lived experiences. These are all necessary inputs to design and develop disclosure tools, such as nutrition labels, for recommender systems. Failure to adequately involve external stakeholders, including the public and academics, in the process would hinder the effective and trustworthy development of metrics.

When mapping the potential stakeholders that should be consulted for the development of nutrition labels or other forms of transparency measures, it is important to consider who is directly, indirectly, and potentially impacted by recommender systems and who has expertise in amplification and content curation. As recommended by the European Center for Not-for-Profit Law (2023), the consultation process should, in any case, be trustworthy, whereby "[p]otential participants can see that the process is designed to be inclusive, open, fair and respectful and is delivered with integrity and competence. Where there are limitations or barriers to delivery or impact, the organisation is open and honest about these."

Most importantly, any transparency measure that targets a large audience, such as nutrition labels, must be broadly understandable and accessible. Marginalized and vulnerable groups are disproportionately at risk of harm by recommender systems, yet they face unique and additional barriers to engaging in standard-setting processes. Centering nutrition labels around the needs and concerns of those most at risk will ensure that the needs of all users are met. Nutrition labels should, therefore, be tailored to the specific needs of these groups, focusing on content, level of disclosure, and the most appropriate format (Partnership on AI, 2021).

CONCLUSION

THERE ARE RIGHTFUL CONCERNS around the operation of algorithmic amplification, particularly regarding the harms they cause when optimized for engagement. Yet, to effectively address the adverse impacts of amplification on human rights and the civic space, a nuanced and commonly agreed definition of amplification is required. Amplification is often treated as an innate property of an algorithmic system, hard-coded by the developers to reflect their values and objectives (usually profit-driven). While human input certainly influences what these models are optimized for, the reality is much more complex, as these socio-technical systems respond to users' behavior themselves. We focus on UGC networks, where users follow each other and function as both consumers and producers of content. As such, we consider content to be recommended, or amplified, when it is ranked higher in a reader's feed. Our interpretation is aligned with the EU DSA, today's leading and most comprehensive regulation on digital platforms. Moreover, a corollary to defining amplification is to measure it. In this paper, we argued that algorithmic amplification cannot be measured in one dimension only; rather, it is a complex phenomenon that can be better understood through multiple metrics.

In this paper, we proposed the introduction of nutrition labels for recommender systems to enable meaningful transparency for social media platforms and empower users to engage with platforms in a way that mitigates harm and promotes their rights. While we are aware that the main platforms do not offer any options for ranking algorithms at this time, there are encouraging signals from the protocol-based decentralized platforms that might offer meaningful choices to users. Nutrition labels and agreed-upon metrics would be useful to understand how these systems operate and to ensure that their use protects and promotes human rights.

We proposed two key metrics that would be valuable to include in the nutrition labels. First, platforms should report the highest and lowest (by number of normalized impressions) producers in a given time frame. Second, they should report the highest and lowest producers by the ratio of IN to OoN impressions. Furthermore, we offered a framework to narrow the scope of groups and accounts of interest, centering on those which may have the highest risks for human rights and democracy.

Determining which metrics should be assessed and reported, and how to design nutrition labels in a way that is useful and accessible to all users, requires meaningful engagement with external stakeholders, such as civil society and academics. This process should center marginalized groups and those living in the Global Majority by elevating their needs and concerns. Our paper, therefore, aimed to spark an inclusive conversation on how to measure and report amplification with meaningful participation from civil society, academia, policymakers, international organizations, and the private sector.

While our research focused on recommender systems for social media timelines, these metrics and transparency are not necessary only for today's dominant platforms. They are also critical for alternative and emerging models, such as the Fediverse, web3 technologies, decentralized platforms, large language models, and generative AI. We encourage further research and multi-stakeholder collaboration to explore how nutrition labels could enable recommender system transparency in all types of AI systems.

REFERENCES

Apple. n. d. *Privacy - Labels.* . Accessed on June 19, 2023.

Luca Belli, Kyra Yee, Uthaipon Tantipongpipat, Aaron Gonzales, Kristian Lum, and Moritz Hardt. 2022. County-level algorithmic audit of racial bias in Twitter's home timeline. *arXiv preprint arXiv:2211.08667*.

DSA. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union L* 227 (10), 1–102. https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Accessed on April 17, 2023.

European Center for Not-for-Profit Law. 2023. *Framework for Meaningful Engagement*. https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai Accessed on April 17, 2023.

FDA. 2021. *Nutrition Facts Labels for AI/ML Transparency and Trust*. https://www.fda. gov/media/153321/download. Accessed on April 17, 2023.

Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*. 3662–3666. https://doi.org/10.1109/BigData50022.2020.9378025

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daum´e Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

Google. n. d. *Get More Information About Your Apps in Google Play*. https://blog. google/products/google-play/data-safety/. Accessed on June 19, 2023.

Human Rights Council. 2011. *Guiding Principles on Business and Human Rights*. https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Accessed on April 17, 2023.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2025334119.

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–12.

Paddy Leerssen. 2023. *Seeing What Others Are Seeing*. Ph. D. Dissertation. University of Amsterdam.

Luke Thorburn, Priyanjana Bengani, and Jonathan Stray. 2022. *What Will "Amplification" Mean in Court?* https://techpolicy.press/what-will-amplification-mean-in-court/. Accessed on April 17, 2023.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.

Arvind Narayanan. 2018. 21 *Fairness Definitions and Their Politics*. https://www. youtube.com/embed/jIXIuYdnyyk. Accessed on April 17, 2023.

Arvind Narayanan. 2023(a). *TikTok's Secret Sauce*. https://knightcolumbia.org/ blog/tiktoks-secret-sauce. Accessed on April 17, 2023.

Arvind Narayanan. 2023(b). *Understanding Social Media Recommendation Algorithms*. https://knightcolumbia.org/content /understanding-social-media-recommendation-algorithms. Accessed on April 17, 2023.

NIST. 2023. AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework

OECD. n. d. *The OECD Artificial Intelligence Policy Observatory*. https://oecd. ai/en/ai-principles. Accessed on June 19, 2023.

Partnership on AI. 2021. *About ML Reference Document*. https:// partnershiponai.org/paper/about-ml-reference-document/10/#Section-3-4-2. Accessed on April 17, 2023.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) (FAT* '19). 59–68.

Twitter. n. d.(a). *Notices on Twitter and What They Mean*. https://help.twitter.com/en/rules-and-policies/notices-on-twitter. Accessed on April 17, 2023.

Twitter. n. d(b). *Our approach to Recommendations*. https://help.twitter.com/en/rules-and-policies/recommendations. Accessed on April 17, 2023.

Twitter. 2014. *The 2014 Year On Twitter*. https://blog.twitter.com/official/en_us/ a/2014/the-2014-yearontwitter.html. Accessed on April 17, 2023.

Twitter. 2023. *Remove Stats Collection Code Measuring how Often Tweets From Specific User Groups are Served*. https://github.com/twitter/the-algorithm/commit/ec83d01dcaebf369444d75ed04b3625a0a6645eb9. Accessed on April 17, 2023.

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data* (Houston, TX, USA) (SIGMOD '18). 1773–1776.

Kyra Yee, Alice Schoenauer Sebag, Olivia Redfield, Emily Sheng, Matthias Eck, and Luca Belli. 2022. A keyword based approach to understanding the overpenalization of marginalized groups by English marginal abuse models on Twitter. *arXiv preprint arXiv:2210.06351*.

About the Authors

Luca Belli is the founder and CEO of Sator Labs, a University of California, Berkeley Tech Policy fellow, and a NIST AI visiting fellow. Previously, he was the co-founder and research lead for Twitter's Machine Learning, Ethics, Transparency, and Accountability team where he guided industry leading approaches for responsible machine learning practices and product changes. Before that, he operated as a data science and machine learning engineer at Conversant and WolframAlpha. His research interests lie at the intersection of feedback loops, algorithmic amplification (with a special eye on politics), and algorithmic audits. He holds a Ph.D. in mathematics from the University of Rome Tor Vergata.

Marlena Wisniak leads policy and advocacy on digital rights at the European Center for Not-for-Profit Law, with a focus on algorithmic systems and digital platforms in security and counterterrorism. Until recently, she oversaw content governance on Twitter's legal team and had previously led the civil society and academic portfolios at the Partnership on AI. As a lawyer, Marlena has over a decade of experience working globally across sectors at the intersection of technology and human rights, with expertise in business and human rights. Admitted to the Swiss Bar, Marlena graduated from the University of Geneva and Stanford Law School.

Cite as: Luca Belli & Marlena Wisniak, *What's in an Algorithm? Empowering Users Through Nutrition Labels for Social Media Recommender Systems*, 23-06 Knight First Amend. Inst. (Aug. 22, 2023), https://knightcolumbia.org/content/whats-in-an-algorithm-empowering-users-through-nutrition-labels-for-social-media-recommender-systems [https://perma.cc/ZBU3-TTWY].

© 2023, Luca Belli & Marlena Wisniak.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Cover illustration: Emilie Flamme


KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY