Making sense of the Digital Services Act

How to define platforms' systemic risks to democracy

by Dr. Michele Loi

August 2023

Published by

ALGORITHM WATCH

with support by

civitates

/ CONTENTS

/ AN OUTLINE OF A RISK ASSESSMENT METHOD FOR MEASURING THE RISK POSED BY INTERNET SERVICES TO MEDIA FREEDOM AND DIVERSITY

1. INTRODUCTION

Internet services, especially dominant platforms like Instagram, Google, TikTok, Twitter and others have established themselves as practically essential consumer goods for much of the world. These and other services operate 24 hours a day, every day, and produce high amounts of utility for their users – indeed, one can hardly imagine a life without efficient internet searches (a field where users' preference for Google has been remarkably stable for decades), or losing the ability to interact with others via social media which is fundamental to many social and professional identities. Despite their benefits, these powerful internet services are arguably radically transforming our capitalist society into, potentially, its worst version yet .[1] In so doing, the platform economy gives rise to a myriad of risks which may negatively impact individuals and democracy itself.

The European Union's lawmakers are convinced that platforms can pose such risks, and have therefore enacted the Digital Services Act (DSA),[2] a law that requires so-called Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to "diligently identify, analyze and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." These risks include, but are not limited to, "the dissemination of illegal content" through platforms' and search engines' services, "any actual or foreseeable negative effects for the exercise of fundamental rights [...], on civic discourse and electoral processes, and public security; [...] 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." When such risks are identified, companies have to take "appropriate measures" to mitigate them.

Whereas the law provides a long list of possible measures to mitigate risks – e.g. "adapting the design, features or functioning of their services, including their online interfaces" – it is remarkably quiet on how VLOPs and VLOSEs should conduct a risk assessment and what legislators expect from them. The law does spell out that VLOPs and VLOSEs must account for the ways in which certain factors may influence systemic risks, including the design of recommender systems and any other relevant algorithmic systems, content moderation systems, applicable terms and conditions and their enforcement, systems for selecting and presenting advertisements, and data related practices of the provider. There is, however, no mention

1 Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization," *Journal of Information Technology* 30, no. 1 (March 2015): 75–89, https://doi.org/10.1057/jit.2015.5; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019* (Profile Books, 2019).2019

2 Article 34, Risk assessment, 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), https://eur-lex.europa.eu/eli/reg/2022/2065/oj

of a procedure that would delineate how to identify a systemic risk in practice. Also, the European Commission has not published any guidance for companies. This means that at this moment, it is very unclear how VLOPs and VLOSEs should assess risks to comply with the DSA's requirements.

This is the context for our paper. We will not attempt here to provide a holistic assessment of the goods and the ills of digital society as shaped by currently dominant internet services. Our goal is rather narrower and more specific. The purpose of this inquiry is to determine whether some elements of risk that an internet service generates for **freedom of speech** and **media pluralism** are identifiable. Our hope is that with this contribution, we provide a tangible point of departure for the discussion about what different stakeholders can and should expect from a risk assessment, and how it could be done in practice. It will also serve as a benchmark for measuring how we, as a civil society watchdog, will judge the risk assessments that VLOPs and VLOEs are conducting at this moment.[3]

This paper comprises six sections. Section 2, 3, and 4 define and briefly explain the value of democracy and media pluralism. Section 5 introduces the concept of risk and its measure. Sections 6 and 7 define a methodology for a measure of risk to freedom of speech and media pluralism, respectively, followed by a short conclusion.

In the final Appendix, Michele Loi, the architect of this framework, and the team at AlgorithmWatch, have collaboratively suggested seven study designs. These suggested studies enhance the framework by serving two main purposes: initially, they delve into more detailed causal hypotheses about the genesis of the risks outlined in the framework. Secondly, they examine the role of bias in risk creation; for example, when certain demographic groups are disproportionately impacted by systemic risks to freedom of speech and media pluralism.

2. WHAT IS VALUABLE IN A DEMOCRACY?

To assess the risks that internet services pose to freedom of speech and media pluralism, it is essential to first bridge the gap between descriptive/explanatory theories of democracy and the normative theory of democracy to understand the fundamental values and principles that are at stake. Descriptive and explanatory theories are concerned with how democracy operates in practice, while normative theories delve into the underlying values and principles that justify and guide the design of democratic institutions. By connecting these two perspectives, we can more comprehensively understand the specific elements of risk that internet services may pose for freedom of speech and media pluralism specifically within the context of democratic ideals.

In the normative inquiry, it is customary to distinguish instrumentalist and intrinsic justifications of democracy. According to instrumentalist theories, democratic systems are generally conducive to good decisions. This viewpoint aligns with epistemic views, which recognize that democracy is justified by its ability to both elicit good decisions and arrive at sound outcomes.[4] By considering this perspective, we can appreciate how the potential risks posed by internet services to freedom of speech and media pluralism intersect with the fundamental goals and justifications of democracy, including its aspiration for effective decision-making.

3 VLOPs and VLOSEs are required to hand in their first risk assessments to the European Commission until August of 2023. It must be said, though, that it is unlikely they will make these assessments publicly available. It is also unclear what the Commission will release about their contents, and when. See Paddy Leersen, "Counting the Days: What to Expect from Risk Assessments and Audits under the DSA – and When?," *DSA Observatory Blog* (blog), January 30, 2023, https://dsa-observatory.eu/2023/01/30/counting-the-days-what-to-expect-from-risk-assessments-and-audits-under-the-dsa-and-when/.

4 Epistemic views do not have to be epistocratic. For example, philosopher David Estlund, argues for inclusive and deliberative procedures that allow for broad participation and equal consideration of citizens' preferences. Estlund critiques epistocratic views and highlights the value of democracy in uncovering and correcting errors through open discussion. See, e.g., his *Democratic Authority: A Philosophical Framework* (Princeton, N.J: Princeton Univ Pr, 2007).

According to intrinsic theories of democracy, meanwhile, democracy is valuable because it is a form of self-rule. These accounts stress the importance of democratic participation.[5] Yet, participation is a contested idea in democratic theory. Elitist theory[6] maintains that it is normal for political leaders to appeal to uninformed and overly emotional citizens. Realistically, the function of electoral participation is to avoid the disasters that would be produced by very bad leaders if those could not be substituted. Robert Dahl[7] advances the view that each citizen is the member of one or more interest groups and votes for politicians favoring the interests of those groups. By contrast, proponents of participatory democracy argue that it is not enough for citizens to merely exercise their voting rights during elections but that they should have continuous opportunities for engagement and influence in public affairs.

We will focus now on two pre-conditions for democracy: the first is respect for freedom of speech, the second is media pluralism. Regardless of how democracy is defined or justified, these two elements play a fundamental role in ensuring the functioning and health of democratic systems.

5 Bruce Ackerman and James S. Fishkin, "Deliberation Day," *Journal of Political Philosophy* 10, no. 2 (2002): 129–52; James Fishkin, *When the People Speak: Deliberative Democracy and Public Consultation* (Oxford University Press, 2009).

6 Schumpeter, in his seminal work „Capitalism, Socialism, and Democracy" (1942 / 1965), challenged the traditional "classical" conception of democracy as a system where the people rule. Instead, he proposed an "elitist" model of democracy, where political leaders are competitively selected by citizens through periodic elections. For Schumpeter, democracy is not about the realization of the common good, or a reflection of the people's will, but rather a method for selecting political leadership. Schumpeter's theory has been influential, but also contentious, in the field of political science, provoking debates about the role of citizen participation and the nature of representation in democratic governance.

7 In „A Preface to Democratic Theory" (1959 / 2006), Dahl expounds on the complex nature of democracy, providing a critical analysis of traditional democratic theories, including majority rule and the idea of popular sovereignty. Dahl notably challenged the notion of an "unrestrained majority," proposing instead a vision of democracy that safeguards minorities and individual rights, while ensuring broad participation and political equality.

3. WHAT IS FREEDOM OF SPEECH AND WHY IS IT VALUABLE?

One key aspect of democracy in liberal systems is the defense of different freedoms (which may be regarded as precondition for democracy to exist, in a substantive sense). Freedom of speech, in particular, is a liberty whose protection is regarded inherent to democracy. Art. 11 of the EU Charter of Fundamental Rights[8] says that:

1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

2. The freedom and pluralism of the media shall be respected.

Clearly, the key functions of democracy such as taking (good) decisions, evaluating leaders (in order to confirm or remove them), and achieving compromises, become harder when citizens lack freedom of expression. Without freedom of speech, some important information and ideas will go unheard or remain underground, including opinions that are critical of those in power. Clearly, this is not the only right that is important for democracy, but we offer an analysis of risk based on freedom of expression as an example of how to build an empirical indicator for at least one salient aspect or component of this risk. If the aim was to summarize the overall risk with a single score - which isn't necessarily our goal - then a comprehensive measure of risk for democracy should encapsulate various aspects, along with a method to balance these aspects against each other. By contrast, we only include two dimensions in this inquiry: freedom of speech and media pluralism.

In the section on freedom of speech we focus on moderation with regard to content deletion, rather than on *recommendation and ranking* which leads to

8 "Article 11 - Freedom of Expression and Information," European Union Agency for Fundamental Rights, April 25, 2015, https://fra.europa.eu/en/eu-charter/article/11-freedom-expression-and-information.

the (unequal) visibility of content. Arguably, in the communication environment of the internet, search and rank functions can limit the *degree* to which content is accessible, meaning that even content that is technically accessible online can be very hard to find when it is not rated highly by search and recommendation engines. We focus on content recommendation in our analysis of the risk to media pluralism, which can also be considered an aspect of risk to democracy.

4. WHAT IS MEDIA PLURALISM AND WHY IS IT VALUABLE?

Following a customary distinction,[9] by media pluralism we either mean a significant plurality of publicly available opinions and analyses (content pluralism) or a significant plurality of media outlets (source pluralism). The two forms of pluralism are clearly not independent, since independent media sources can have incentives to differentiate their contents, for example, by seeking an identification with different parts of society or with groups of people with different worldviews. At the same time, the two are not reducible to one another: source pluralism does not guarantee content pluralism and content pluralism does not imply source pluralism.

Media pluralism can also be considered a precondition for democracy. Indeed, media pluralism is also standardly considered an aspect of freedom of speech, but for the sake of simplicity, here we consider them as separate, because this allows us to focus on two distinct types of risk: the risks connected with the elimination of content (under the heading of threats against freedom of speech) and the risks connected with algorithmic recommendations (under the heading of threats against media pluralism).

Content pluralism is valuable for more than one reason. First, it is valuable in utilitarian terms, because it fulfills the preferences of many readers. Empirically, we observe that at least citizens of democracies seek access to diversity of opinion, including opinions that differ strongly from their own (which does not contradict preferences for personalization, to a given degree).[10] Second, media pluralism can be conceived as a requirement for democracy, in particular for participatory and epistemic accounts of democracy that stress the importance of reasonable, and reasonably well-informed, political deliberation.

Source pluralism is also valuable. It is valuable, first, as a means to favor content pluralism. Second, it is valuable for oppositional social groups to identify opinion leaders, and not just information, who are independent from the dominant parties in power, either in government or in large corporations, particularly in the case of small, independent media. This can be stressed as important from the point of view of an elitist theory of democracy, which values opposition forces in the struggle for power, not necessarily from the point of view of a deliberation-based theory. Third, source pluralism is a form of market competition, so it is also a means to economic efficiency.

5. THE CONCEPT OF INTERNET EVENT RISK

When measuring risk from an internet service, we need to distinguish the following elements:

1. The *Event,* A, which is the event causing the risk.

2. The *Consequence,* C, that is produced by the event.

9 For example, "media pluralism can either mean a plurality of voices, of analyses, of expressed opinions and issues (internal pluralism), or a plurality of media outlets, of types of media (print, radio, TV or digital) and coexistence of private owned media and public service media (external pluralism)", in Reporters Sans Frontieres, "Contribution to the EU Public Consultation on Media Pluralism and Democracy," July 2016, 1.

10 Peter M. Dahlgren, "A Critical Review of Filter Bubbles and a Comparison with Selective Exposure," *Nordicom Review* 42, no. 1 (January 1, 2021): 15–33, https://doi.org/10.2478/nor-2021-0002.

3. The probability of the event A, *p,* how likely it is that the risk will materialize.

4. The (conditional) probability of the consequence C, *p',* how likely is it that something negative will materialize.[11]

A *The risk event*	**C** *The (negative) consequence*
p *The probability of the risk event*	p' *The (conditional) probability of a harmful consequence*

Let us provide an example for something that is not an internet event. Say that I want to evaluate the risk connected to forgetting my backpack somewhere. The event, A, is forgetting the backpack. The negative consequence, C, is that my backpack is stolen. The probability of the event, p is the probability that I forget my backpack. The probability p' is the (conditional) probability that, if I forget the backpack, it gets stolen. If I know the probability *p* and the (conditional) probability *p'* I can measure the risk as an expected value, which is provided by p*p'*v(C), where v(C) is the value of the stolen items. For example, suppose that we consider the backpack I use to carry my computer around. The value v(C) is then the value of the computer (assuming, for simplicity, that the backpack as such has no value). This value may decrease over time the computer gets older or even increase over time as I store more important files into my computer, if I do not do any backup.

The type of events that we consider here are events *produced* by internet services in their interactions with users. Internet services generate events that produce consequences that have a certain value or disvalue. We need to proceed by the following steps:

First step: identifying a type of events A that are produced by an internet service. This is an event like "forgetting a backpack", which may or may not be harmful, depending on its consequences. Our focus here will be *primarily* on those events that are automated by algorithms in concert with user/data interactions.

Second step: explaining what the relevant consequences, C, are and why the consequence of events of type A have a relevant and negative impact on individuals and/or society. In the individual example, the generally negative consequence C of forgetting a backpack is that the backpack is stolen. In our measures, the consequence C must be something with a generally negative impact on democratic society. The negative impact for freedom of speech and media pluralism that we should capture in our model is a negative impact for society as a whole.[12] We indicate the (dis)value of the consequence C as v(C).

Third step: assess the probability, *p*, that events of type A will be produced. This is comparable to the probability of forgetting a backpack. It is the probability that the relevant type of event (that may or not may be harmful) occurs.

Fourth step: assess the probability of *p'*, that a consequence with a generally negative impact C will occur, given the occurrence of the risk event A. This is comparable to the probability that the backpack holding the computer, which was forgotten, gets stolen. In our account of risk for democracy, *p'* is the probability that, when a risk event A occurs, the (impersonally) bad consequences for democratic society will occur.

11 Terje Aven and Shital Thekdi, *Risk Science: An Introduction*, 1st edition (Routledge, 2021).

12 Insofar as individual wrongs occur (e.g., violations of freedom of speech of particular individuals), they are counted as contributing to the impersonal badness of a state of affairs. The aggregate of individual wrongs can be understood as a degree of systemic risk for democracy, i.e. as something that is bad for society in general.

6. ONLINE FREEDOM OF SPEECH: CASE STUDIES AND PROPOSED MEASURE

Let us illustrate the approach with one case study and a proposed approach to measure risk to democracy in algorithmic content moderation by platforms.

In the literature, content moderation is recognized as a risk to democracy. Loi and Dehaye[13] point out three emblematic cases in which content moderation has generated risks to democracy:

- Towards the end of September, 2017, MEP Marietje Schaake uploaded a series of videos on YouTube concerning the debate in the Parliament on the new law on European trade for goods that are used for torture and in carrying out the death penalty. According to the MEP, YouTube removed one of her videos with a recording of the opinion of the European Commissioner for Trade, Cecilia Malmström. YouTube's reasons for removal were that the video was "flagged for review" by other users and that YouTube determined that YouTube Community Guidelines were violated. MEP Schaake filed a "video appeal", where she had to argue in one sentence why the video needed to stay up. After she publicized the incident through Twitter, Google reached out to one of her parliamentary assistants to smooth it out and reverse the decision. The video was back online after four hours (Loi and Dehaye 2017, p. 158)

- "Napalm Girl" is widely regarded as the most iconic documentary photograph of the Vietnam war. It shows a naked 9-year-old Phan Thj Kim Phúc running away from a Napalm attack. Norwegian author Tom Egeland, working for the newspaper Afterposten, included this picture in the context of a display of seven photographs that changed war history. Facebook promptly removed the picture, since it shows Kim Phúc's naked genitals, which violates Facebook's Community Guidelines. Subsequently, the editor of Aftenposten wrote an open letter to Facebook that circulated widely among media outlets and on the blogosphere, criticizing the company's actions. Erna Solberg, the Conservative prime minister of Norway, took to Facebook itself to voice similar criticism. Facebook then reversed its previous decision which is evidence of the difficulty of automatically filtering content of nude children, the day after the publication of the open letter (Loi and Dehaye 2017, p. 160).

- Facebook image censorship guidelines, leaked in 2012, revealed that moderators were instructed to remove any images of breastfeeding in which nipples were visible. Facebook's nipple policy could be charged with intentionally or inadvertently supporting corporate interests threatened by breastfeeding (e.g. the powder milk industry), in so far as it limits the users' exposure to pictures of women breastfeeding. This may have an influence on women's choices with respect to whether to breastfeed in public, or indeed breastfeed at all, and on their partners' motivation to support them. Second, it sends all kinds of messages about gender roles, insofar as the depiction of men's nipples is permitted but not women's. Thus, the combination of Facebook software (for signaling content) and moderation rules is an institution of social cooperation with the power to influence the conceptions of what is good, appropriate, dignified [...] (Loi and Dehaye 2017, p. 161-162).

It is therefore reasonable to assume that the moderation of content from dominant platforms affects the nature of what is debated in society. Due to the pervasiveness of certain platform services and their unique role as citizen *fora* for expressing opinion, protecting freedom of expression from the undue interference of algorithmic decisions is, nowadays, arguably as important as protecting the *legal* right of freedom of expression, which is customarily defined as a right against the interference of *public authority*

13 Michele Loi and Paul-Olivier Dehaye, "If Data Is the New Oil, When Is The Extraction of Value From Data Unjust," *Philosophy and Public Issues* 7, no. 2 (2017): 138–78.

(not against the interference of *algorithmic regulation*). Our research question concerns the identification of a measure of risk that is indicative of the degree to which content removals such as the ones figuring in the above examples may, individually and cumulatively, generate a risk for democracy.

In what follows, we will provide an outline of a methodology for providing a quantitative measure of this risk. We then briefly comment on the *assessment* of the risk. The difference between measuring risk and assessing it is that risk assessment is entirely normative, as it concerns deciding an acceptable quantity of risk (and probability of harm). While the definition of a risk measure, as we shall see, is also value laden, risk acceptability (as in risk assessment) is impossible to discuss without considering the feasible options for reducing risk. Since both risks and risk measures generate consequences for society, judgements of risk acceptability presuppose a holistic view of political priorities and the interests that can be affected by those measures.

Coherently with the conceptual framework introduced in section 5, we will focus on risk measures rather than risk assessment here. We approach the issue by showing how, for risk to democracy, we can identify the relevant type of event *A*, its consequences, *C*, the probability of the event, *p*, and the probability of the consequences, *p'*.

THE RISK EVENT, A

We propose to consider the primary risk event for freedom of speech, *A*, as the algorithmic flagging of content. This is the event in which an algorithm determines that content posted on a platform has a high risk of violating that platform's "community guidelines".[14] Guidelines may prohibit adult explicit content, offensive content, hate speech, or dangerously misleading messages about matters of high public importance ("fake news"). How the algorithm makes the determination whether to flag a piece of content may be guided by machine learning techniques which recognize and evaluate that content's verbal or pictorial features, or informed by users reactions such as reporting the content as a violation or an unusually high degree of sharing of the content.[15] A combination of both methods is also possible. This event of algorithmic flagging may have negative consequences or not, depending on what happens next.

THE CONSEQUENCE, C

When algorithmic systems are deployed to recognize guideline violations, flagged content can then be eliminated directly without any human review or submitted to a human moderator to take the ultimate decision. Alternatively, flagged content may be algorithmically demoted, i.e., prevented to diffuse widely.[16] In every case, the risk event, A – the fact that content has been *flagged* – is a necessary but not a sufficient condition for a negative consequence. We will consider a negative consequence to be the fact that content is deleted when the content in question should not have been deleted (a false positive).

This begs the question what is meant by "should". It could be (a) content that should not have been deleted because it did not actually violate community guidelines; or (b) it could be content that actually violates community guidelines but where this still undermines the right to freedom of expression (so here the community guidelines themselves would be the problem and pose a risk to democratic debate). Our emphasis here is on (b), for two reasons. First, (a) is anyway extremely difficult to assess. Guidelines are subject to interpretation and there is no ultimate standard for the correct interpretation other than the past behavior of the moderators (which may, in turn, be systematically biased or, indeed, incompatible with freedom of expression properly understood). Second, relying on the guidelines would prevent the possibility of criticizing the company's moderation

14 These guidelines are not defined by a community but by the platform; they are guidelines the users ("the community") of the platform have to obey under penalty of having their posts or profiles suspended or deleted.

15 Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

16 Zuckerberg.

policy, other than pointing to incoherence in its application. This is a limited critique because, even when the application of guidelines is coherent and fair (e.g., all content judged disturbing by any user is removed, without any ideological, religious, or political bias), it can nonetheless have negative repercussions on freedom of speech.

We will not at this stage specify the criteria for "content that should not be deleted", as a wide range of specifications are compatible with the broad approach we outline here, and different interpretations of "content that should not be deleted" will lead to different measures of the risk in question. We illustrate one meaningful option:[17] defining a negative consequence for freedom of speech to be any deletion of content that is neither criminal nor pornographic, as defined by a clear legal standard.[18] If the elimination of content is strictly required by the law, then we stipulate that it does not count as a distinct contribution of the algorithmically flagged event to the restriction of freedom of speech.[19]

The (dis)value of the consequence v(C) can also be assessed with some conventional measure and will depend on the context of the case. The value v(C) corresponds mathematically to the disvalue for society of the fact that *a unit* of content that is a false-positive is eliminated. Clearly, it is inappropriate to measure the value of the expression of legitimate content in commercial terms. The disvalue is more *symbolic* in that it consists in the disvalue of freedom of expression being arbitrarily limited. This disvalue will depend on the type of normative explanation we assume for freedom of expression.

From a utilitarian standpoint, the value of freedom of expression is a function of *both* the importance for society of the *platform* in which the content was expressed *and* of the importance for society of the voice whose content is eliminated. In this case, legitimate content eliminated from Twitter may have a greater disvalue than legitimate content eliminated from a gaming platform, and legitimate content from an influential activist or politician may have greater disvalue than legitimate content from a less influential voice. This approach is in line with a utilitarian account of the value of democracy and freedom of expression within it. A thorough utilitarian view however must also consider that the voice of marginalized groups is particularly important, precisely because it reduces the uniformity of the value judgments in society – if, as argued among others by the Utilitarian philosopher John Stuart Mill, dissenting voices[20] are especially valuable for the pursuit of the truth in the long run.[21]

17 By this we do not mean to suggest that this is the only valid or the best option.

18 A complication here is that empirically determining what is pornography also involves value judgements that can be contested, and a country may lack established legal standards for evaluating what counts as pornography. Moreover, countries with equivalent liberal-democratic credentials (e.g., USA and Germany) may offer different degrees of protection for certain highly contested contents (e.g., Nazi ideology) on account of their specific history and culture.

19 This is a stipulation, not an argument, and it is a reasonable stipulation when we deal with legal systems that realize a reasonable protection of freedom of speech. If we wanted, for example, to gauge the risk to freedom of speech due to algorithms in a society characterized by illiberal legal standards, we could instead count as contributions to risk to democracy by platforms precisely those cases in which the platform eliminates content *because* it is illegal – given that the law of the country itself is, in this case, a threat to freedom of speech and democracy. In that context, we would be interested in the danger posed by platforms in that they provide illiberal governments with effective means to implement their illiberal policies in the online world. We are assuming that this is not the type of risk we are trying to measure in this framework when we used the proposed definition. The framework can only do the work required in the context of an illiberal regime after changing the definition of "content that should not be deleted" to reflect this fact.

20 It's worth noting a distinction between the terms „marginalized" and „dissenting," as they imply different concepts in certain contexts. The term „marginalized" often refers to groups such as racial minorities, LGBTQ communities, or those traditionally disenfranchised. For instance, in the U.S., discussions around achieving greater „equity" for marginalized groups do not necessarily imply dissent from mainstream views, although there may be perceived dissent from the status quo. Conversely, „dissenting" is commonly associated with views that challenge the establishment, the „mainstream media," or even scientific consensus, especially evident in the era of COVID-19. Nonetheless, it's important to highlight that the Millian argument places an enhanced emphasis on the perspectives of marginalized groups when (and only when) they dissent from established norms or the status quo, a situation they often find themselves in.

21 John Stuart Mill, "Utilitarianism," in *Utilitarianism and Other Essays*, ed. Alan Ryan (Harmondsworth, Middlesex, England; New York, N.Y., U.S.A.: Penguin Books, 1987), 272–338; John Stuart Mill, *On Liberty* (London: Penguin Classics, 1859).

Alternatively, a deontological or dignity-based approach may attribute the same degree of disvalue to the elimination of legitimate content from every individual, since the deontological approach may treat every unjustified elimination to be equally morally bad (in that it is morally wrong) and the dignity-based approach may consider the harm to dignity of an individual, produced by the violation of freedom of expression, to be independent from the value of the freedom of that individual to others in society.[22]

THE PROBABILITY OF THE RISK EVENT, P

The risk event is a piece of content being algorithmically flagged on a platform. The probability of such events can be assessed by measuring the frequency of algorithmically flagged content, *F*, relative to all content in a platform, N_C, in a given unit of time, e.g., one year. N_C provides the *denominator* of the fraction that expresses the probability of the risk of the event. The *numerator* of the relevant fraction, F_C, is the amount of content that is flagged by every platform. To evaluate the risk posed by *all platforms* on democracy, N_C must include the entirety of all online content posted in every platform (some estimation method may be developed to compute this number) and F_C must include the content flagged by any platform. The probability *p* is therefore the ratio of flagged content F_C to total content N_C. The probability *p* understood as the frequency of risky events for democracy by *one* platform, S, can be measured as the frequency of content flagged by S, F_S, relative to all content that is to say, p_S= . This provides a measure of the weight of the algorithms of S in regulating speech, relative to the total amount of online communication. The probability of the flagging of content from a specific (high-valued) group of users, for example journalists, activists, or opposition politician, can be calculated by only considering in the numerator and in the denominator only the occurrences of content from such users, etc.

22 See for example the operationalization of deontological moral judgments in the *expected choiceworthiness* framework by MacAskill and Ord. In William MacAskill and Toby Ord, "Why Maximize Expected Choice-Worthiness?1," *Noûs* 54, no. 2 (2020): 327–53, https://doi.org/10.1111/nous.12264.

THE PROBABILITY OF THE HARMFUL CONSEQUENCES, P'

The probability *p'* is a conditional probability. It tells us how probable it is that a flagging event will produce negative consequences, given that the flagging event occurred. In our example, we are interested in how often platforms delete content that is neither illegal nor pornographic, assuming that the content was already flagged as a potential violation of community guidelines. In this case, *p'* is the ratio of eliminated content (that is neither illegal nor pornographic), E_C, expressed *as a proportion of the flagged content* F_C. Notice that E_C only concerns content eliminated *which was also flagged.* Eliminated content that was not among the flagged content is not considered in E_C. By assumption, the elimination of this content is independent from the automated flagging – the risk cause we consider. Thus, we can estimate p' as . The probability of harmful consequences produced by the events of a single platform p_S, can be estimated as , the frequency with which a specific service S eliminates the content after it has flagged it as potentially dangerous.

The total risk produced by a given service, S, can be measured as the product $R_S = p_S * p_S' * v(C)$, i.e. the expected (dis)value from eliminating legitimate content due to automated flagging.

RISK DEFINITIONS VS. RISK PRESCRIPTIONS

This measure of democracy risk is not a *prescriptive* measure. Platforms may have good reasons to moderate content that is neither criminal nor pornographic: this is a legitimate interest of companies insofar as the spaces they create are voluntarily joined by individuals. There is no general implication that all platforms should always allow all non-criminal and non-pornographic content generated by any user, no matter how unfit to the context. Moreover, a metric of value may attach very little (dis)value to the elimination of some legal and non-pornographic content. So, the elimination of speech on a gaming

platform may be associated with very little or even of the platform is not highly relevant for democracy.[23]

The purpose of non-prescriptive measures is to underscore the potential dangers to democracy of content deletion driven by automatic flagging, even when content deletions are justified. The more frequently that content which is neither illegal nor explicit gets removed due to automatic flagging, the more diminished the oversight of democratic institutions over the limits of free speech becomes.

Of course, completely stopping content moderation would open the floodgates to a variety of harmful practices that would significantly impact the quality of discourse and potentially harm users. So, while the risk associated with the diminishing oversight of democratic institutions over free speech could be eliminated if platforms ceased all content deletion, this cannot be a solution from a democratic perspective. Allowing unrestricted speech on platforms can lead to a proliferation of hate speech and cyberbullying. These offensive communications have the potential to marginalize, intimidate, or silence certain groups or individuals, contradicting the principle of free and equal participation in discourse, which is a pillar of democracy.

The issue of misinformation and fake news also becomes rampant without moderation. These false narratives can distort public perceptions, fuel societal divisions, and undermine trust in democratic institutions. During critical times, like elections or public health crises, the spread of misinformation can have particularly dire consequences. Privacy violations, another serious concern, could multiply without regulation. Individuals' sensitive information could be exposed without their consent, leading to potential harassment, identity theft, or other forms of exploitation. Without moderation, online platforms could also become a hotbed for illegal activities. Unrestricted posting might allow for the sharing of illicit content, promotion of violence, or selling of prohibited items, posing a serious threat to societal safety and order. Finally, the absence of content deletion could lead to the manipulation and abuse of the platform. This includes the propagation of fear, harassment, and public opinion manipulation, particularly during political campaigns or public crises.

Therefore, while ceasing content deletion might eliminate one risk, it introduces several others. Striking a balance is key - platforms must protect freedom of speech and democratic participation, while also mitigating the dangers associated with harmful content. This reinforces the need for a normative debate, engaging experts and democratic representatives, to guide the development of acceptable and effective moderation strategies.[24]

It's crucial to understand that highlighting potential drawbacks of reducing a risk measure to zero is not a critique against the measure itself. Rather, it

23 The evolution of gaming platforms into public spaces has become an increasingly noteworthy phenomenon. As these platforms transform, they host a wide array of interactions, and they become sites where social and cultural norms can be productively challenged. However, they can also become spaces where harmful behaviors may emerge. Given this transformation, gaming platforms are increasingly relevant to discussions about democracy as public spaces. This change in perspective becomes particularly salient as the nature of gaming evolves from being product-centric to interaction-focused, entering a domain commonly referred to as the ‚metaverse.' The metaverse, as it is commonly understood, is a virtual-reality space where users can interact with a computer-generated environment and other users. It's an expansive, immersive digital space where many aspects of social and economic life can occur, similar to those in the physical world. Thus, as gaming platforms increasingly resemble these virtual public squares, their impact on, and relevance to, democratic discourse and behavior cannot be overlooked.

24 Reducing this risk can be achieved in two ways: either platforms could adopt a more lenient approach towards potentially troublesome content, or governments could enact stricter laws rendering such content illegal on these platforms. For instance, legislative bodies could reasonably classify new types of hate speech directed at groups as illegal to safeguard individuals from online harms.

However, this introduces another layer of complexity: should online platforms uphold freedom of speech by allowing even those expressions considered illegal when the legal constraints on speech are excessively restrictive? In theory, platforms can endanger freedom of speech and democracy by enforcing the mandates of an oppressive legal system. To incorporate this aspect within our framework, we would need to reevaluate the definition of „content that should not be deleted". However, undertaking such a revision falls outside the scope of our current discussion.

emphasizes that the goal of regulation should not always be to eliminate the risk entirely. Balancing risk, rather than absolute elimination, often results in a more nuanced and effective approach. This concept acknowledges that even when a certain risk could theoretically be reduced to zero, doing so may not always be the most beneficial or desirable outcome from a broader perspective.

7. MEDIA PLURALISM: CASE STUDY AND PROPOSED MEASURE

In this section we apply the framework of risk measure presented in section 3 to source pluralism and content pluralism.

THE RISK EVENT, A, FOR SOURCE PLURALISM

In the context of source pluralism, a "risk event" is defined as an event that could potentially lead to the communication dominance of a specific media source, impacting the diversity and independence of news, views, and information. This is similar to how one can conceive a risk event in a competitive market: one that might lead to the economic dominance of a specific actor or entity.

The concept of a risk event for source pluralism revolves around the idea of an "invitation for a communication transaction." It's important to note that this does not necessarily entail an economic transaction, but rather a communication-based interaction. It covers a broader spectrum of sources than economic agents, including non-profit entities. This reflects the idea that dominance in the media landscape is not always tied to wealth but can also be attributed to the volume and reach of the content produced.

A risk event is triggered when a user query results in an invitation to engage with a specific media source. This is comparable to the definition of risk events for competitive markets, where a user query might lead to an invitation for a potential purchase. However, the key difference lies in the nature of the interaction – communication versus economic.

It's worth noting that the concept of a risk event isn't limited solely to search queries, but can also encompass various facets of algorithmic systems that affect searches, such as auto-complete functions in the search bar. The definition of a risk event can also be extended to incorporate various recommender systems that present news and information even without explicit queries from users. This could reflect additional risks to media pluralism as dominant media can afford more targeted exposure; users might tend to follow major brands, which perpetuates a lack of diversity in their feeds, and so forth.

For illustrative purposes, we will proceed with the paradigmatic example of a search query as our model for further discussion. This choice does not encompass the full complexity of the media landscape or the diversity of ways in which users engage with media sources, but provides a clear and tangible point of reference. Using this model allows us to examine the dynamics of media engagement, highlighting possible points of risk and their implications for pluralism. As we move forward, it's essential to keep in mind that this is only one of many possible approaches and that real-world applications will likely need to account for a wider range of factors and scenarios.

In this context, an invitation for a communication transaction arises when the result of a user query prominently features a media source, depicted in a neutral or positive manner, making it instantly recognizable to the user. In addition, the access to this source should be straightforward to obtain, whether it is through an online hyperlink or easily found offline.

Not all user queries that mention a media source are defined as risk events. For instance, if a media source is portrayed negatively, or if it is not immediately identifiable to the user, or if access to the source requires considerable effort, the query does not qualify as a risk event.

In sum, a risk event for source pluralism is an incident where the outcome of a user query may lead to the prominence of a single media source, thereby potentially skewing the diversity and plurality of

information consumed by the public. The challenge, therefore, is to monitor these events, mitigate their potential risks, and ensure a balanced and diverse media landscape.

THE (NEGATIVE) CONSEQUENCE, C, FOR SOURCE PLURALISM

In the context of media pluralism, a negative consequence, which we can denote as C, occurs when the diversity and representation of sources in the media landscape are compromised. This is similar to how a negative consequence for competitive markets occurs when the market's competitive nature is compromised, often due to the dominance of a small set of market actors.

To define dominant media sources, we may adopt a method akin to defining dominant market actors in an economic market. For instance, the dominant media sources could be the smallest set of media platforms that, combined, have an audience share exceeding 50%. It's important to bear in mind that this percentage is somewhat arbitrary and might need to be tailored to the specific media landscape and societal context in question.

A negative consequence for source pluralism can be defined as an event where a user query generates an invitation for a communication transaction that solely includes these dominant media sources. In other words, if the result of a user query leads to engagement with only these dominant media platforms, it can be regarded as a harmful invitation, akin to an invitation in the economic market context promoting market dominance by few actors.

Similarly, it's crucial to consider the distribution of user attention when defining a negative consequence. Search engines often return multiple results, but user attention is unevenly distributed, with a significant preference for the first few results. Studies show that a vast majority of clicks go to the first few organic search results, with the first three results often accounting for about 80% of total clicks.[25] This implies that the inclusion of non-dominant sources in lower-ranking positions doesn't significantly contribute to media pluralism.

Thus, a risk event for source pluralism is said to have a negative consequence if the top results from a user query predominantly or exclusively feature dominant media sources. This situation reinforces the dominance of these sources and undermines the diversity of the media landscape, which is a fundamental aspect of media pluralism. Therefore, to maintain a diverse and balanced media landscape, it is important to ensure that dominant media sources do not monopolize the top-ranking positions in search engine results.

THE PROBABILITY OF THE CONSEQUENCE, C

The probability of a harmful consequence for source pluralism (given a risk event) is the (conditional) probability that a negative consequence is produced, given that a risk event took place. In this case, it is the probability that an invitation for a communication transaction only includes dominant parties (given that an invitation has been made). This can be measured as the ratio of negative consequences (i.e., invitations only including dominant communication actors) over risk events (i.e., total invitations).

Next, we suggest an empirical measure of risk to *content* pluralism. This is much harder to conceptualize than a measure of risk to *source* pluralism because the concept of *content* pluralism raises deeper philosophical questions about the importance of diversity and the nature of content.

THE RISK EVENT, A, FOR CONTENT PLURALISM

In our proposed model, the definition of a risk event for content pluralism bears resemblance to

25 According to a study by Johannes Beus, 99.1% of clicks are received by the results in positions 1-10. The first three results alone comprise roughly 84% of the total clicks. See Johannes Beus, "Click Probabilities in the Google SERPs," SISTRIX, October 27, 2015, https://www.sistrix.com/blog/click-probabilities-in-the-google-serps/.

the definition for source pluralism we previously discussed. As with source pluralism, the risk event in the case of content pluralism arises from the response given by an internet service (be it a platform or a search engine) to a user query. This response results in an invitation for a communication transaction involving specific media sources. However, while the risk events are similar, the potential negative consequences for content pluralism differ significantly from those related to source pluralism, as we will explain in more detail below.

THE PROBABILITY OF THE RISK EVENT

The probability of the risk event is defined as the fraction of search queries on a platform that return invitations for communication transactions with media sources.

THE CONSEQUENCE, C, FOR CONTENT PLURALISM

Undoubtedly, defining and empirically determining the negative impact for content pluralism poses a challenge. The nature of the consequences for content pluralism differs notably from those for source pluralism. Source pluralism engages with a concept of diversity that can be relatively straightforwardly operationalized by referencing the ownership and organizational structure of media outlets. That is, different media sources are usually owned by distinct entities, maintain separate formal editorial organizations, and so forth. Conversely, defining diversity in the context of content pluralism requires a more nuanced approach.

Content pluralism is not primarily concerned with matters of ownership or organizational dependencies. Instead, it is intrinsically interested in the distinctiveness of viewpoints, ideas, and analyses. This distinctiveness can be appreciated from three different and complementary perspectives:

A) From a commercial perspective, where diversity is subjectively determined by the consumer. If a consumer is willing to invest in content that they perceive as unique or different in some respect – even if the difference is purely aesthetic – from a commercial standpoint, this represents a valuable aspect of diversity. This perspective acknowledges the consumers' appreciation for variety in media content.

B) From an informational perspective, the emphasis is on objective differences in the data conveyed. Regardless of varying expression or presentation styles, content is deemed diverse *only* if it imparts distinct fundamental information. "Information" in this context refers to all elements contributing to the enhancement of audience knowledge. This includes the depth and quality of the analysis, innovative interpretations, and other elements that objectively differentiate one piece of information from another.

C) From the perspective of performative content, where the emphasis lies on the diversity of speech acts.[26] The idea of speech acts implies that identical content can be used to achieve diverse effects, depending on its mode of presentation. For example, the same piece of content, when articulated in a certain manner, could contribute to a sociological treatise, while if presented differently, it could act as a catalyst for political action. Given that the method of expression facilitates different outcomes, this constitutes a form of diversity contributing to pluralism.

These dimensions are highlighted to underscore the inherent challenge in establishing a universal criterion that encapsulates what makes content diverse in a morally and politically significant way. This complexity, however, does not preclude the potential to gauge risk, provided one can hypothesize a robust diversity criterion. The element of significance has a normative aspect to it; for instance, it would denote importance in a democratic context, if we were to consider content pluralism crucial for democratic deliberation.

26 In „How to Do Things with Words" (1962), Austin proposes the theory of speech acts, suggesting that when we use language, we're not merely stating facts but also performing actions. These actions fall into three categories: ‚locutionary acts' (the act of saying something meaningful), illocutionary acts' (the action performed by saying something, such as commanding, questioning, or promising), and ‚perlocutionary acts' (the effect or outcome that results from the act of speaking, such as persuading, deterring, or inspiring).

Assuming the existence of a criterion for *significant diversity*, we can conceive a negative impact on content pluralism as any outcome that does not enable such diversity. To achieve this, we need to first elaborate on what it entails for a platform's query response to provide ample diversity of content. This calls for a metric that can accurately measure diversity levels, defining sufficient diversity whenever this metric is satisfied, and deeming diversity inadequate whenever it falls short. This notion of inadequate diversity, then, furnishes our understanding of a (negative) consequence.

Next, we need to identify a unit of analysis. In the case of a platform that logs all user-received content, it becomes meaningful to assess whether the array of media sources presented to a user over a significant time span (e.g., the two months surrounding a pivotal democratic event such as a referendum or election) exhibits insufficient content diversity. If aggregating results presented to the same user is unattainable, one must assess if each individual query result independently offers inadequate diversity. For instance, we could evaluate whether a query concerning a referendum vote leads to a result that provides links to media sources that present balanced arguments for both sides, or whether it skews towards sources advocating for a single side or demonstrating heavy bias. Transforming this into an empirical measure involves numerous value judgments, which may restrict the potential measures to narrow case studies. Nonetheless, these case studies could offer valuable insights to society.

However, we must qualify this discussion by noting that the concept of "balance" is a normative one, and its interpretation will differ based on each specific context. In situations where the epistemic quality of differing opinions is significant, 'balance' does not necessarily denote equal space for every opinion. For instance, a search for the "truth about climate change" should not be obliged to present 'balanced' arguments when one side may largely be supported by scientific consensus and the other by disinformation. The idea of balance in such contexts should be rooted in presenting reliable, fact-checked information that reflects the prevailing scientific understanding.

Also, we must acknowledge that the definition of balance might change depending on the political context, the scientific context, or the user's information needs. In a political context, balance might mean presenting diverse viewpoints within the boundaries of democratic discourse, while in a scientific context, it might involve giving prominence to views supported by robust empirical evidence. Therefore, the balance in this discussion is not a call for false equivalence, but an invitation to examine the complexity and diversity of perspectives within the appropriate boundaries of truth and credibility.

THE PROBABILITY OF THE RISK EVENT, P

The probability of the risk event is defined in a similar way as the probability of a risk event for source pluralism. It is the fraction of search queries that return invitations for communication transactions with media sources.

THE PROBABILITY OF THE CONSEQUENCE, C

The probability that a single event results in harmful consequences to media *content* pluralism can be measured as the fraction of risk events referring to content *lacking significant diversity*. The probability of harmful consequences, given a risk event, for a single service provider, S, is the fraction of invitations for potential media transactions that are produced by that service provider, S, that are lacking significant diversity.

8. CONCLUSION

In summary, it is feasible to construct a notion of risk to democracy that can be empirically quantified through reasonably straightforward observations. However, the true challenge lies in resolving the normative question of determining which observations hold significance and why.

Indeed, these definitions aren't simple, containing assumptions that could be contested. Moreover, these definitions incorporate evaluative judgments like moral ones, which could lead to disagreements.

For instance, people might dispute what constitutes a false positive in content moderation scenarios.

Clearly, this is only feasible in practice if the relevant data can be accessed. This methodology points to specific concepts of probability and explains how those probabilities can be measured by counting the frequency of certain types of events. It may prove impossible (or simply, unreasonable) to collect *all the data* that are relevant for the frequencies in question. For example, it may be impossible to count all the units of content posted by users across all platforms (which is required for measuring risk to freedom of speech, quantifying the denominator) and to observe each and every post that has been flagged by an algorithm as a potential violation of community guidelines. Yet, reasonably accurate estimations could be made – for example, by building representative samples, or by limiting the scope of research to only the most important algorithms by the most important platforms.

With this approach, we do not intend to be naïve about the normative and technical challenges that measuring risk would pose. And yet, we want to show that the concept of risk to democracy is not so ineffable that we should simply give up any attempt to deliver an empirical quantification. Clearly, such measures will be imprecise, and the methods and definitions suggested here are perfectible. But this is something that the current endeavor has in common with every risk model, as it unavoidably requires simplification and abstraction of reality in some respects.

9. REFERENCES:

Ackerman, Bruce, and James S. Fishkin. "Deliberation Day." *Journal of Political Philosophy* 10, no. 2 (2002): 129–52.

Aven, Terje, and Shital Thekdi. *Risk Science: An Introduction*. 1st edition. Routledge, 2021.

Beus, Johannes. "Click Probabilities in the Google SERPs." SISTRIX, October 27, 2015. https://www.sistrix.com/blog/click-probabilities-in-the-google-serps/.

Dahl, Robert A. *A Preface to Democratic Theory, Expanded Edition*. Chicago, IL: University of Chicago Press, 2006. https://press.uchicago.edu/ucp/books/book/chicago/P/bo4149959.html.

Dahlgren, Peter M. "A Critical Review of Filter Bubbles and a Comparison with Selective Exposure." *Nordicom Review* 42, no. 1 (January 1, 2021): 15–33. https://doi.org/10.2478/nor-2021-0002.

Estlund, David M. *Democratic Authority: A Philosophical Framework*. Princeton, N.J: Princeton Univ Pr, 2007.

European Union Agency for Fundamental Rights. "Article 11 - Freedom of Expression and Information," April 25, 2015. https://fra.europa.eu/en/eu-charter/article/11-freedom-expression-and-information.

Fishkin, James. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, 2009. https://www.google.com/

Leersen, Paddy. "Counting the Days: What to Expect from Risk Assessments and Audits under the DSA – and When?" *DSA Observatory Blog* (blog), January 30, 2023. https://dsa-observatory.eu/2023/01/30/counting-the-days-what-to-expect-from-risk-assessments-and-audits-under-the-dsa-and-when/.

Loi, Michele, and Paul-Olivier Dehaye. "If Data Is the New Oil, When Is The Extraction of Value From Data Unjust." *Philosophy and Public Issues* 7, no. 2 (2017): 138–78.

MacAskill, William, and Toby Ord. "Why Maximize Expected Choice-Worthiness?1." *Noûs* 54, no. 2 (2020): 327–53. https://doi.org/10.1111/nous.12264.

Mill, John Stuart. *On Liberty*. London: Penguin Classics, 1859.

———. "Utilitarianism." In *Utilitarianism and Other Essays*, edited by Alan Ryan, 272–338. Harmondsworth, Middlesex, England; New York, N.Y., U.S.A.: Penguin Books, 1987.

Reporters Sans Frontieres. "Contribution to the EU Public Consultation on Media Pluralism and Democracy," July 2016.

Schumpeter, Joseph A. *Capitalism, Socialism and Democracy. (Tenth Impression.).* Pp. xiv. 431. Unwin University Books: London, 1965.

Zuboff, Shoshana. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30, no. 1 (March 2015): 75–89. https://doi.org/10.1057/jit.2015.5.

———. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019. Profile Books, 2019.

Zuckerberg, Mark. "A Blueprint for Content Governance and Enforcement." Facebook, November 15, 2018. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

10. APPENDIX: SEVEN STUDY DESIGNS: ON RISKS TO FREEDOM OF INFORMATION AND MEDIA PLURALISM

1. STUDY DESIGN: ON THE RISK OF ARBITRARY POWER IN THE EXERCISE OF CONTENT RANKING

HYPOTHESIS:

— We assume that platforms label particular media sources as "authoritative" in order to prioritize these sources in algorithmic rankings (particularly in searches pertaining to sensitive subjects like coronavirus)

— The determination (for the purposes of algorithmic ranking) of which media source are "authoritative" likely favors established media outlets to the exclusion of smaller players (e.g. independent journalists)

- Platforms may also deem media outlets as "authoritative" that do not adhere to journalistic standards and thus needlessly elevate risks like the spread of misinformation

2. STUDY DESIGN: ON THE RISK OF ALGORITHMIC INCENTIVES INFLUENCING MEDIA PRODUCTION AND PROMINENCE

HYPOTHESIS:

— We assume that algorithmic ranking systems contain a logic that can be exploited by media players most willing and capable of "gaming" recommender systems in order to receive greater prominence in rankings — whether by focussing on certain topics, using search terms favored by the algorithm, embedding video content, etc.

— This may lead to a range of undesirable effects:

  - Reinforcing dominant media content (as well-resourced media producers are more capable of gaming the system)
  - Preferencing entertainment over journalistic content
  - Preferencing newness over quality
  - Misrepresentation of content (low-quality content pretending to contain news; misinformation packaged as journalism)

3. STUDY DESIGN: ON THE RISK OF PLATFORMS BEING "PAID TO PLAY"

HYPOTHESIS:

— We ask whether media outlets are recommended proportionate to their ad budgets, thereby entrenching the market dominance of established players in search rankings

  - Do publishers that enter into agreements with platforms receive higher rankings in recommender system (e.g., content of Google Partners)?

4. STUDY DESIGN: ON THE RISK OF REINFORCING MARGINALIZATION AND INFORMATION GAPS

HYPOTHESIS:

— We assume that platforms used automated filters to flag and potentially block or shadow-ban content containing particular words (e.g., "sex," "black,") or "gay"

— This automated filtering may systematically censor against media outlets that regularly publish valuable content on sensitive topics (e.g., human trafficking, colonialism, LGBTQ rights) using blacklisted words

— This may have the effect of further marginalizing already marginalized voices in public discourse

5. STUDY DESIGN: VALIDITY IN CONTENT MODERATION

HYPOTHESIS:

— Algorithms used for content moderation produce a score based on predictive features derived from machine learning or based on user-reported warnings

— These scores can be interpreted as a measure of the risk that the content violates an important community guideline or is even illegal

— This measure of risk is very imperfect, it may lack validity altogether (for example, censoring words like "gay" to mitigate hate speech also censors valuable discourse around LGBTQ subjects)

EXAMPLE OF METHODOLOGY:

One could check whether the scores are calibrated, for example by asking

— whether there is proportionality between the score and the probability that the content is removed by platform moderators

— whether there is proportionality between the score and the judgment of expert moderators of how clearly it is a violation of the platform community guidelines

— where there is proportionality between the score and the perception of the content as morally or politically harmful by independent reviewers

TYPES OF DATA NEEDED:

a) scores used to flag content for removal or attention by moderators

AND

b) decision thresholds (desirable, not necessary)
c) expert moderators' content removal decisions (historical, real life data)

d) expert moderators' evaluations (in experimental conditions, to be set up through a collaboration)

OR

e) flagged and non-flagged content and independent evaluators' judgements

6. STUDY DESIGN: BIAS IN CONTENT MODERATION

HYPOTHESIS:

In the application of algorithmic outputs by humans, there will be systematic differences in decisions, given the same outputs, depending on the group to which the person posting content belongs.

EXAMPLE OF METHODOLOGY:

As above, but we verify if calibration within groups is achieved when scores are used to make decisions, or whether deployment bias (the bias when algorithmic outputs are interpreted or used by humans) exist. That is to say, given a score used to trigger deletion or attention by moderators, we check that the score is conducive to similar results (e.g., probability of deletion by a human moderator) conditional on the content being generated by members of different groups. For groups, we refer to typical protected group features, in particular, gender, political orientation, religious orientation, and trade union membership.

TYPE OF DATA NEEDED:

a) scores used to flag content for removal or attention by moderatos

b) decision thresholds

c) group membership features of participants

7. STUDY DESIGN: ARBITRARY POWER IN CONTENT MODERATION

HYPOTHESIS:

We assume that, for an efficient and fair system, *procedural fairness* needs to be balanced with post-hoc adjustment based on outcomes. However, the (unavoidable) ad hoc adjustment increases the risk of the exercise of *arbitrary power*.

EXAMPLE OF METHODOLOGY:

— We require the first live instantiation of code of a tool used for content moderation.

— We ask whether the current code has been changed in significant ways

— We require access to documentation that explains why the code has been changed

— We evaluate how often the changes have been made in response to specific outcomes regarded as problematic

TYPE OF DATA NEEDED:

— First instantiation of code for live use

— Subsequent versions of code

— Records about the reasons leading to changes in the code

/ IMPRINT

Making sense of the Digital Services Act
How to define platforms' systemic risks to democracy

by Dr. Michele Loi

August 2023

Publisher:
AW AlgorithmWatch gGmbH
Linienstr. 13
10178 Berlin

Contact:
info@algorithmwatch.org

Website:
https://algorithmwatch.org/en/

A project by

ALGORITHM WATCH

supported by

civitates

Layout / Illustration:
Beate Autering