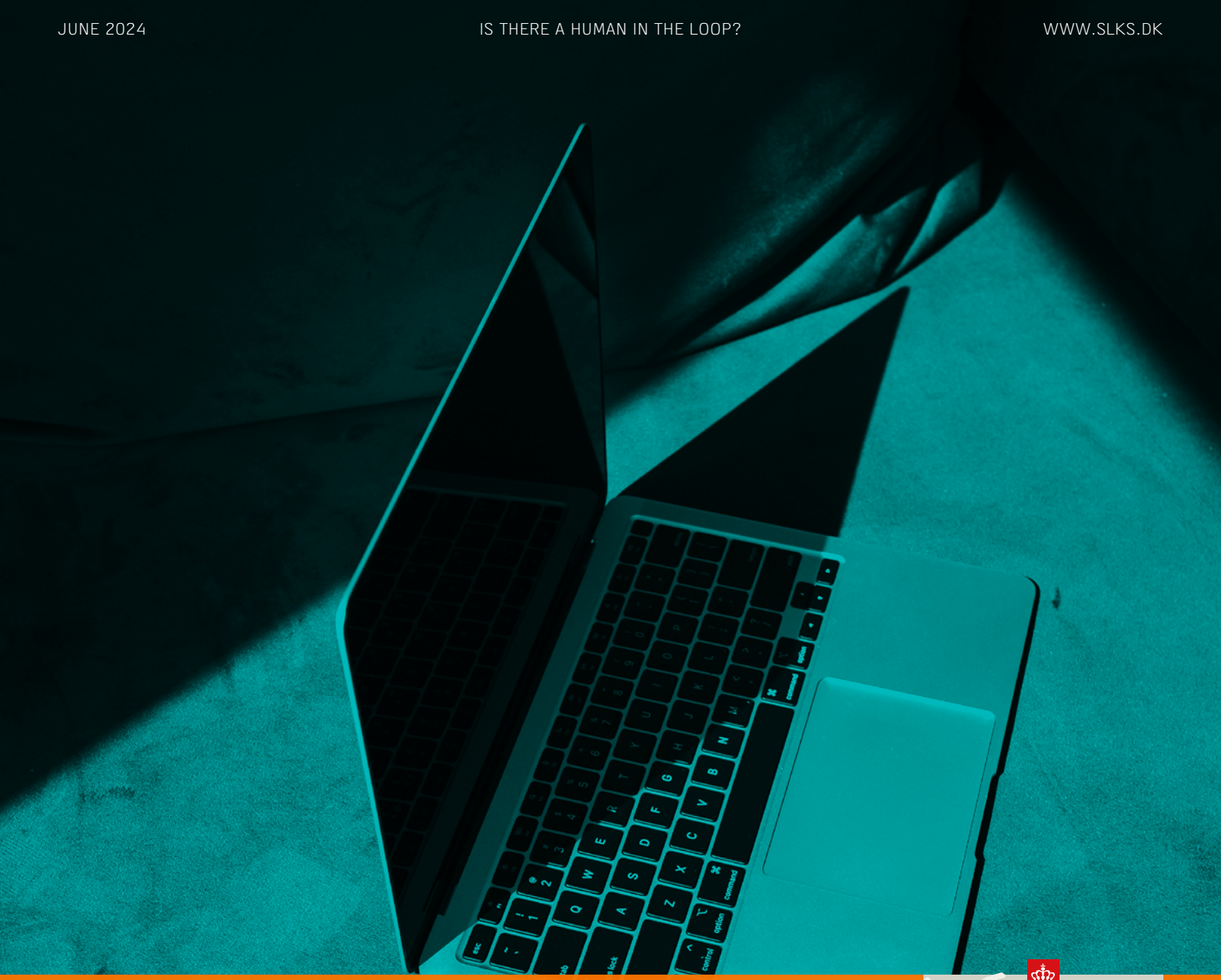


CENTER FOR
SOCIAL MEDIA
TECH AND DEMOCRACY

KULTUR
MINISTERIET
AGENCY FOR CULTURE AND PALACES

IS THERE A HUMAN IN THE LOOP?

A COMPARATIVE ANALYSIS OF VERY LARGE SOCIAL MEDIA PLATFORMS' CONTENT MODERATORS

CONTENT

INTRODUCTION

The public debate is increasingly taking place on social media. Enormous amounts of content are shared on the platforms every day, including content that can be harmful especially to children and young people or content that can be used to distort and mislead the public debate through the spread of misinformation and disinformation. It can ultimately affect elections and undermine trust in democratic institutions, with potentially far-reaching consequences for our society and democracy. It is therefore important that content on the platforms is moderated.

In the biggest election year in history, where more than half of the world's population will vote, it is crucial to keep a watch on the integrity of free and democratic elections. Access to reliable information and a strong response against the spread of misinformation and disinformation are important prerequisites in this regard.

It is a large and important task that, among others, has been placed in the hands of big tech. A task which, if misused or overlooked, can put freedom of expression under pressure and weaken democracy.

It is crucial for the democratic control of big tech and their content moderation that the public and the individual user are given an insight into big tech's engine room in a transparent and uniform way. But there is a lack of transparency with who moderates the content that users encounter online and how.

The biggest tech platforms are subject to a number of obligations to publish information about their content moderation. They must, among other things, report on their moderation practices and their content moderation teams, including the qualifications and language expertise of moderators, which may impact platforms' ability to moderate content across languages and cultural contexts in the EU.

These reports form the starting point for this comparative analysis of big tech's content moderation, which raises the question: Is there a human in the loop?

ABOUT THE STUDY

In this analysis, the Danish Agency for Palaces and Culture's Center for Social Media, Tech and Democracy (CSTD) examines eight major social media platforms' transparency reports. The study takes a closer look at parts of the platforms' content moderation and compares their teams of moderators and their language skills across the EU.

The eight major social media platforms selected are Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, X (Twitter) and YouTube.

The eight social media platforms have been selected because they are among the 23 large platforms and search engines that, due to their size, are required to meet particularly strict transparency requirements in the Digital Services Act (DSA). In addition, it is especially on these platforms that citizens participate in the digital public conversation, i.e. debate, express their views, share content and search for knowledge. The platforms therefore play a central role in digital democratic processes.

Thus, large market platforms, pornographic video sharing platforms, as well as large search engines and knowledge platforms are excluded from this study.

The study is based on a comparative reading of the eight selected platforms' transparency reports, and the reported values are compared across the platforms. The reporting periods in the selected transparency reports vary from three to six months and run from the third quarter of 2023 to the first quarter of 2024. The reported data were collected on May 27th 2024.

The study is based on the platforms' own reports. It therefore only provides insights to the platforms' self-reported approach and resources for moderating content in the EU.

SUMMARY

SEVERAL PLATFORMS HAVE NO OR FEW DANISH-SPEAKING MODERATORS

- LinkedIn and X have no Danish-speaking moderators. Facebook, Instagram and Pinterest have less than 10.
- The number of moderators is rarely proportional to the number of users.

GREAT VARIATION IN CONTENT MODERATORS ACROSS THE EU

- Moderators at LinkedIn and X cover approximately one third of the EU's 24 official languages.
- Snapchat and Pinterest cover approximately half of the EU's 24 official languages.
- Only Facebook and Instagram have moderators who can cover all of the EU's 24 official languages.
- The most widespread first languages in Northern and Eastern Europe are not as well covered by the platforms' content moderators compared to widespread first languages in Western and Southern Europe.

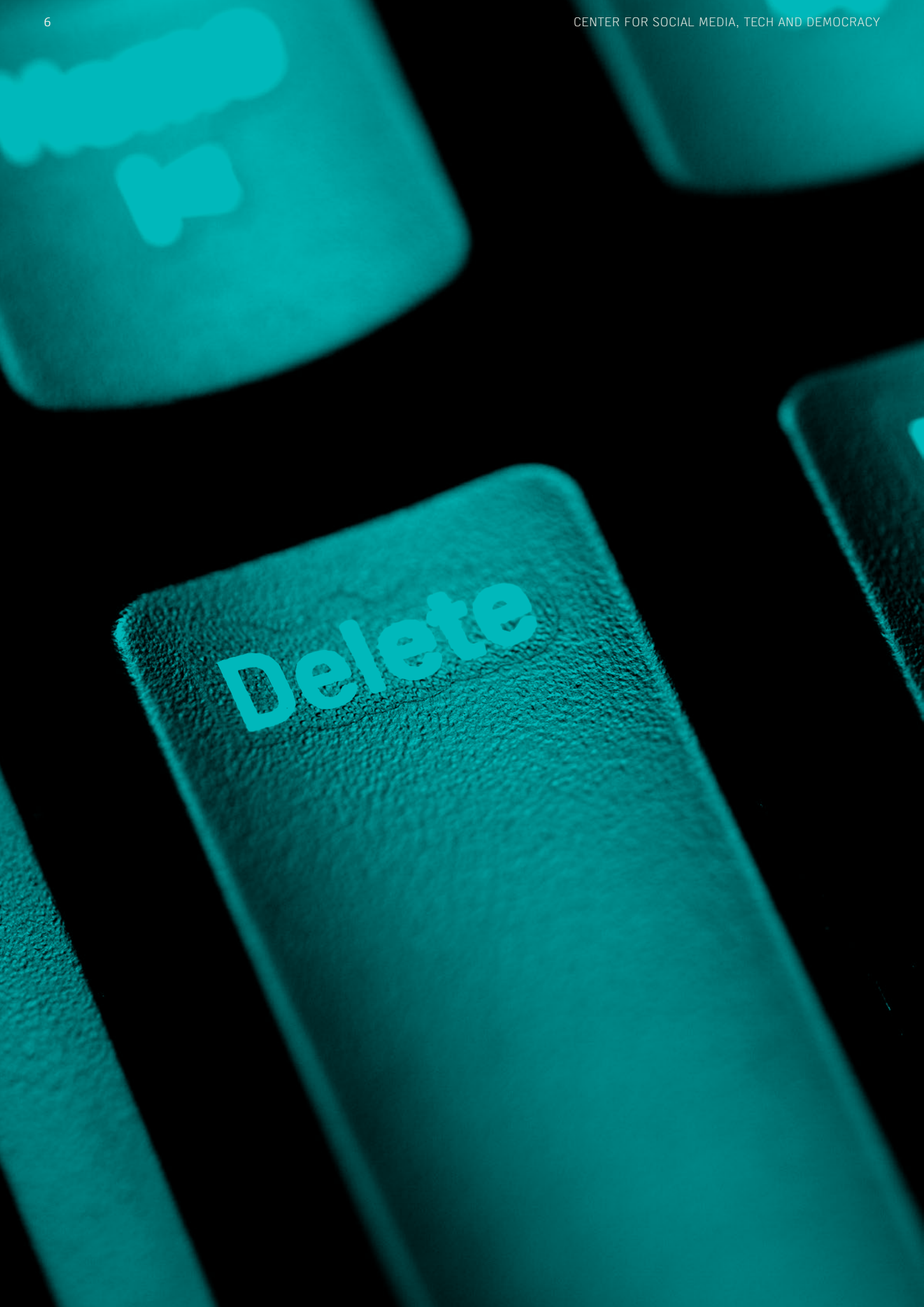
DIFFERENT REPORTING METHODS COMPLICATE COMPARISON

- The platforms do not report how many unique moderators they have associated in a comparable way. Nor do they report how many moderators are counted multiple times because they speak several languages.
- The platforms do not report how large a proportion of the content in the respective languages that is moderated.
- The platforms' reporting periods vary, which makes comparison difficult.

RECOMMENDATIONS FOR INCREASED TRANSPARENCY AND STANDARDIZATION IN BIG TECH'S TRANSPARENCY REPORTS

To ensure that the transparency reports of big tech provide better insights into their content moderation in the future, and that the reports can be compared across the board, Center for Social Media, Tech and Democracy recommends:

- Minimum requirements for how many moderators big tech have in the countries in which they are present.
- Common and comparable methods for counting moderators, so that the number can be better compared across platforms.
- Clear and precise descriptions of how big tech moderate content.


Delete

1

CONTENT MODERATION - HOW AND WHY

"Every day, we remove millions of violating pieces of content and accounts on Facebook. In most cases, this happens automatically, with technology to detect, restrict, and remove content and accounts that may go against our Community Standards, Advertising Standards, and Commerce Policies. In other cases, our technology selects content for human review. Our review teams review a blend of user reports and content surfaced by our technology [...]."

Facebooks transparency report, 2024

When platforms limit content or sanction accounts that share content, it is called content moderation. When the platforms find content that needs to be moderated, it can be done in several ways. Research in the field typically mention:

- **Hard content moderation**, such as removing content or blocking accounts that have shared content.
- **Soft content moderation**, such as labelling misleading content with a warning or limiting the content's reach and distribution[1].

Content moderation is a complex system. The platforms use both automated tools and moderators as well as help from users who report content that they believe exceeds applicable guidelines[2]. Automated tools are responsible for the vast majority of the moderation of the large amounts of content that is shared daily on the platforms. Content moderators assist particularly with questions of doubt, and help train the algorithms with their decisions.

BOX 1 **THE EUROPEAN COMMISSION OPENS FORMAL PROCEEDINGS AGAINST X FOR INSUFFICIENT CONTENT MODERATION ON THE PLATFORM**

In December 2023 the European Commission opened formal proceedings against X under the Digital Services Act in areas linked to eg content moderation. The Commision suspects X for not having sufficient mechanisms including alocated ressources to moderate content and thus not being able to sufficiently reduce the risk of illegal and 'misleading or deceptive content, spreading on the platform[3].

After Elon Musk's takeover of X in 2022, a number of media could report that a large part of the platform's associated moderators were fired[4]. At the same time, X launched that the platform's own users would in future be responsible for fact-checking each other via the "Community notes" function, where users can write notes that are placed at the bottom of a post, for example that the content is fake. Which notes appear at the bottom of a post is decided by users voting on the notes. It has since been met with criticism that content moderation must be managed by the platform's users and not professional content moderators. The Danish fact-check media TjekDet was, among other things, able to document how the vast majority of notes that mark content as misleadingor deceptive contentare never displayed on the platform.

BOX 2 **NEW REQUIREMENTS FOR BIG TECH' TRANSPARENCY UNDER THE DSA**

The recently adopted EU Regulation, Digital Services Act (DSA), has imposed a number of obligations on digital services in the EU to, among other things, promote transparency and accountability[5]. Special obligations apply to online platforms and online search engines that have more than 45 million active users in the EU per month. The obligations include that the platforms publish reports (transparency reports) twice a year on e.g. content moderation of the services.

The transparency reports must contain information about how the platforms moderate content, e.g.

- information on what measures their content moderation practices consist of,
- the number of orders they have received from i.a. administrative authorities,
- information about how much content has been removed,
- the accuracy and error rate of their automated content moderation systems, and
- the qualifications and linguistic expertise of the content moderation team.

The new regulationto promote transparency and accountability also include certain obligations to notify users why their content has been removed or why access to an account has been restricted, and users must have the opportunity to challenge these and similar decisions.

In addition, the Commission has launched a transparency database which collects and publishes reasons for the restriction or removal of content or users to enable verification of content moderation decisions made by online services. The DSA Transparency Database is publicly available and can be found **her**.[6]

WHAT CONTENT SHOULD THE PLATFORMS MODERATE?

The EU Digital Services Act (DSA) aims to ensure that citizens are not exposed to illegal content on online platforms and search engines. The rules therefore oblige the platforms to limit the spread of illegal content, and to remove the content as soon as they become aware that the content exists.

Illegal content includes any content that does not comply with EU law or the legislation of a Member State. This can be anything from content with sexual abuse material against children, the sale of illegal or counterfeit products or the illegal sharing of private images.

In addition, the platforms must assess and act to minimize "systemic risks". Examples of systemic risks could be if the services are used to disseminate or amplify misleading or deceptive content. In addition to risk assessments in relation to combating e.g. the spread of disinformation, the platforms are also obliged to implement risk-limiting measures – for example in the form of content moderation.

Big tech also have their own rules and community guidelines, for content on their platforms.

The EU's stricter requirements for the transparency of big tech under the Digital Services Act, includes requirements for big tech to publish so called transparency reports. This report examines parts of these transparency reports, and examples of the requirements for these reports are described in Box 2.

2

SEVERAL PLATFORMS HAVE NO OR FEW DANISH-SPEAKING MODERATORS

In their transparency reports, the major social media platforms state how many of the official languages of the EU their moderators cover. The reports also provide an insight into how many of the platforms' moderators speak Danish.

LINKEDIN AND X HAVE NO DANISH-SPEAKING MODERATORS

Of the eight major social media platforms included in this study, two platforms, LinkedIn and X, have no Danish-speaking moderators, see Table 1. Pinterest has one moderator who speaks Danish, and TikTok, with its 27 Danish-speaking moderators, has the most moderators with Danish language skills.

In other words, there is a big difference in the number of moderators at the eight platforms that can moderate Danish content.

It's worth noting that some platforms count one moderator who can speak multiple languages per each language the moderator speaks. Therefore, the numbers don't necessarily reflect how many unique moderators each platform has, but they do provide insight into how many different languages the platforms' content moderators can and cannot cover.

Table 1
Overview of content moderators with Danish language skills on major social media platforms

Platform	Number of moderators with Danish language skills
LinkedIn	0
X (Twitter)	0
Pinterest	1
Facebook*	6
Instagram*	6
Snapchat	15
YouTube	18
TikTok	27

Note.: *Facebook and Instagram, both owned by Meta, report the number of moderators combined for both platforms.
Source: Tech giants' transparency reports, released Apr. 2024.

THE NUMBER OF MODERATORS IS RARELY PROPORTIONAL TO THE SIZE OF THE PLATFORM

The assessment of whether the tech giants' number of moderators in a given language is sufficient must be seen in relation to how many users the platform has in the country in question. There are big differences between platforms when comparing the number of users to the number of moderators.

Of the six platforms with Danish-speaking moderators, YouTube is the most common platform in Denmark. The platform has a number of moderators corresponding to one moderator for every 290,000 users of the service in Denmark, see Table 2. In comparison, TikTok has a number of moderators corresponding to one moderator for approximately 52,000 active users.

While Instagram and Facebook report a number of moderators corresponding to one moderator per approximately 630,000 and 730,000 active users on the platform, it should be noted that Meta only reports an average number of moderators across both platforms. Therefore, in practice, each of the two platforms will likely have fewer moderators to moderate content.

Table 2
Number of monthly active users on major social media platforms in Denmark per moderator with Danish language skills

Platform	Number of moderators with Danish language skills	Monthly active users in Denmark**	Number of monthly active users in Denmark per moderator with Danish language skills
LinkedIn	0	1.400.000	No moderators
X (Twitter)	0	750.744	No moderators
Pinterest	1	1.200.000	1.200.000
Facebook*	6	4.400.000	733.333
Instagram*	6	3.800.000	633.333
Snapchat	15	2.677.066	178.471
YouTube	18	5.200.000	288.889
TikTok	27	1.400.000	51.852

Note.: *Facebook and Instagram, both owned by Meta, report the number of moderators combined for both platforms, **This includes the platforms' own counts of users or accounts that have logged in with an account on the platform. The actual number of citizens in Denmark using the platforms may vary.
Source: Calculations by the Center for Social Media, Tech and Democracy based on data reported in tech giants' transparency reports released Apr. 2024.

The number of active users on a platform does not necessarily determine how much and what kind of content is shared on the platform, but comparing this number to the number of moderators does indicate that the platforms' ability to moderate Danish-language content isn't prioritized equally, and that the number of moderators is rarely proportional to the number of users.

Platforms measure and report figures for average monthly active users differently, and several transparency reports lack sufficient information about how these figures are measured. For example, if a user uses multiple accounts on a platform, they may risk being counted multiple times. Therefore, this comparison between platforms must take into account the different calculation methods between platforms and the uncertainty that this entails.

This analysis reports on the platforms' account of how many users that have on average been logged in with an account, and have been active on the platform in a given period (e.g. within the last month). To ensure the best possible basis for comparison, this analysis reports on the platforms' account of how many users that have on average been logged in with an account, and have been active on the platform in a given period (e.g. within the last month).

The DSA suggests that users who are not logged in or do not have an account but simply visit the platform should be included in active monthly users of the platform, as this type of user can make up a significant part of the user group. Therefore, it is to be expected that the actual number of users using the platforms may vary from the tech giants' own reported figures.

*The platforms Pinterest, Instagram, Facebook and Snapchat list their active monthly users, as users or accounts that are logged in and active on the platform. LinkedIn and YouTube do not calculate a total figure for active monthly users, but calculate 1) active users who are logged in to the platform and 2) number of visits on their platforms from users who are not logged in. Figures for active users who have logged into the platforms are reported here. Although X calculates active monthly users as both active users who are logged in combined with active users who are not logged in, only figures for active users who are logged in on the platform are reported here, in order to ensure the best possible basis for comparison between the platforms.

3

WIDE VARIATION IN CONTENT MODERATORS ACROSS THE EU

The EU is an area of great linguistic diversity and Danish is just one of the 24 official EU languages that tech giants are required to report on.

MODERATORS AT LINKEDIN AND X ONLY COVER AROUND A THIRD OF THE EU'S 24 OFFICIAL LANGUAGES

Only Facebook and Instagram have moderators who can cover all 24 official EU languages, see Table 3. TikTok and YouTube have moderators on their platforms who can speak 22 of the 24 official EU languages.

In comparison, Snapchat and Pinterest cover about half and LinkedIn and X only about a third of the EU's 24 official languages with their moderators' language skills.

Table 3
Number of moderators associated with large social media who have language skills within the official EU languages

Language	Facebook	Instagram	TikTok	YouTube	Snapchat	Pinterest	LinkedIn	X (Twitter)
Bulgarian	20	20	43	19	0	0	0	0
Danish	6	6	27	18	15	1	0	0
English	98	98	2334	17507	1148	365	1152	1570
Estonian	3	3	7	7	0	0	0	0
Finnish	15	15	34	24	6	0	0	0
French	211	211	650	439	228	15	31	58
Greek	23	23	65	45	0	0	0	0
Irish	39	39	0	0	0	1	0	0
Italian	164	164	439	229	14	3	23	1
Croatian	19	19	17	34	0	0	0	0
Latvian	2	2	10	7	0	1	0	0
Lithuanian	6	6	9	14	0	0	0	0
Maltese	1	1	0	0	0	0	0	0
Dutch	52	52	162	132	20	2	8	1
Polish	66	66	201	353	4	0	10	0
Portuguese	45	45	172	370	18	10	34	25
Romanian	35	35	136	93	5	0	1	0
Slovak	12	12	38	11	0	0	0	0
Slovenian	7	7	39	7	0	0	0	0
Spanish	147	147	515	675	72	20	40	25
Swedish	42	42	111	37	24	1	0	0
Czech	18	18	57	73	0	0	0	0
German	223	223	837	352	73	17	23	61
Hungarian	24	24	47	52	0	0	0	0
Number of languages covered by moderators' language skills**	**24**	**24**	**22**	**22**	**12**	**11**	**9**	**7**

Note.: *Facebook and Instagram, both owned by Meta, report the number of moderators for both platforms combined, **If a service has not reported any moderators who speak a specific language, it is assumed that the service has 0.
Source: Tech giants' transparency reports, released Apr. 2024

NORDIC LANGUAGES IN THE EU ARE EQUALLY COVERED BY CONTENT MODERATORS' LANGUAGE SKILLS

The Nordic languages*** are covered fairly equally across the platforms, with between 2-3 platforms having no moderators who speak these languages, see Table 4. The number of active users on the platforms per moderator who speak the country's most common first language is relatively similar for the three languages. The biggest differences are seen on Instagram and Facebook, which underprioritize Danish compared to the other Nordic languages.

Overall, the three Nordic languages are the least covered by moderators on X, LinkedIn, and Pinterest compared to the other platforms.

*** The reports on which this study is based only provide figures for EU countries. In the following, the term "Nordic languages" covers Danish, Swedish and Finnish, but not Norwegian and Icelandic

Table 4
Comparison between Denmark, Sweden and Finland: Number of moderators and monthly active users on large social media platforms in each country per moderator speaking the most common first language

	Denmark	Sweden	Finland
Platform	Number of users per moderator with Danish language skills (and total number of moderators)	Number of users per moderator with Swedish language skills (and total number of moderators)	Number of users per moderator with Finnish language skills (and total number of moderators)
X (Twitter)	No moderators	No moderators	No moderators
LinkedIn	No moderators	No moderators	No moderators
Pinterest	1.200.000 (1 moderator)	1.900.000 (1)	No moderators
YouTube	288.889 (18)	281.081 (37)	254.166 (24)
Instagram*	733.333 (6)	171.429 (42)	213.333 (15)
Facebook*	733.333 (6)	171.429 (42)	213.333 (15)
Snapchat	178.471 (15)	182.355 (24)	285.491 (6)
TikTok	51.851 (27)	29.729 (111)	47.058 (34)

Note: *Facebook and Instagram, both owned by Meta, report the number of moderators combined for both platforms
Source: Calculations by CSTD based on tech giants' transparency reports published Apr. 2024.

THE MOST WIDELY SPOKEN FIRST LANGUAGES IN NORTHERN AND EASTERN EUROPE ARE LESS COVERED BY PLATFORM CONTENT MODERATORS THAN IN WESTERN AND SOUTHERN EUROPE

Not all countries across the EU are equally covered by content moderators' language skills.

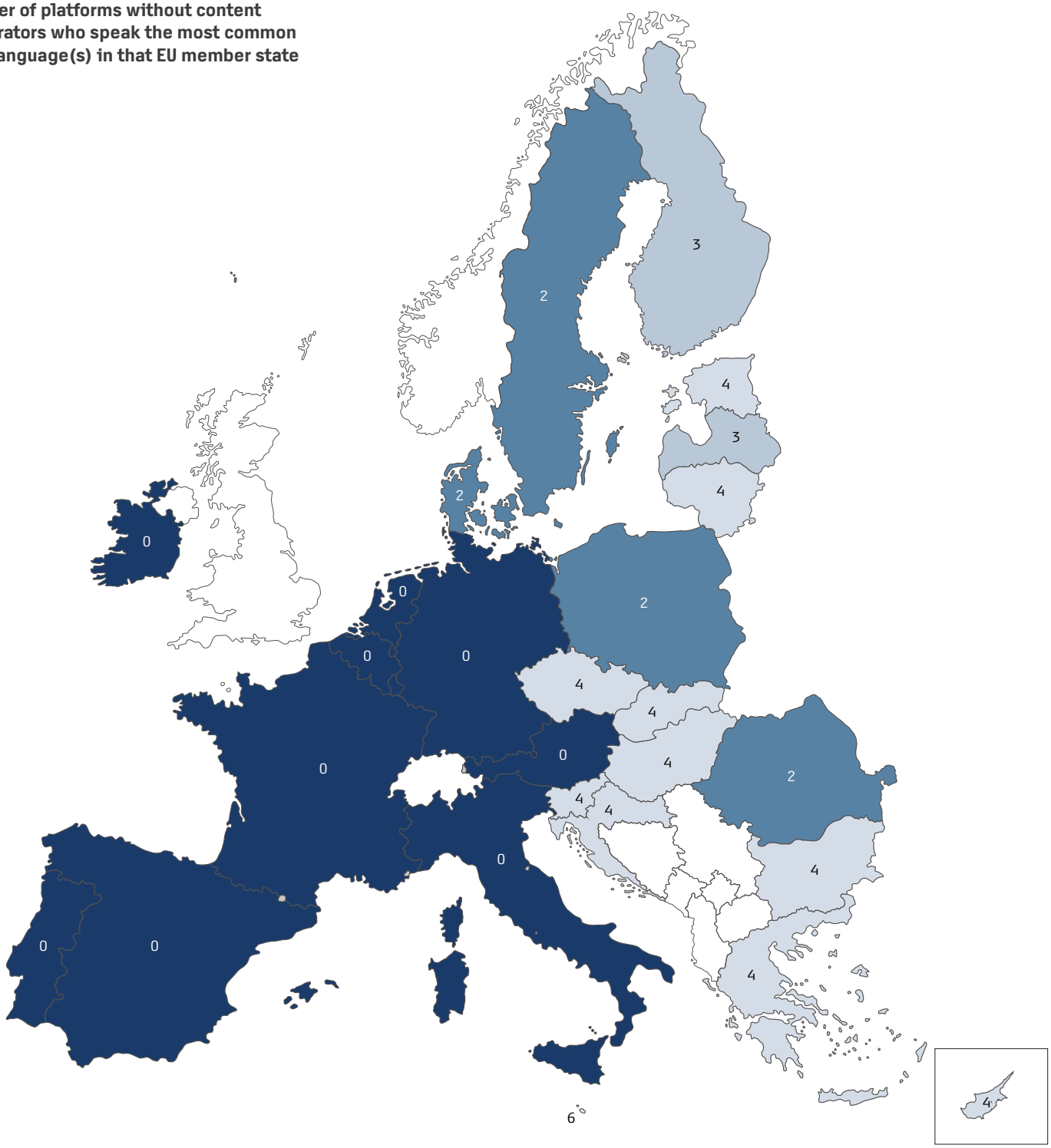
All eight platforms employ moderators who speak the most common first languages[7] of all Western and Southern European countries in the EU (Belgium, France, Netherlands, Italy, Ireland, Spain, Portugal and Austria), with the exception of Cyprus, Greece and Malta, see Figure 1.

In comparison, the most widely spoken first languages in the Northern and Eastern European countries of the EU are particularly poorly covered by the platforms' content moderators' language skills, with at least two platforms not having moderators who speak the most widely spoken languages of these countries.

For Estonia and Lithuania, as well as the Czech Republic, Hungary, Slovenia and Slovakia, four out of eight platforms have no moderators who speak the most common first language.

Cyprus, Greece and Malta, where the most common first languages are Greek and Maltese, are also poorly covered by the language skills of the platforms' moderators, with only four and six platforms respectively having moderators who speak the language

Figure 1
Number of platforms without content moderators who speak the most common first language(s) in that EU member state



Note: In Luxembourg, the most common first language is Luxembourgish. As this is not one of the 24 official languages, the next most common first languages are French and German.
Source: Tech giants' transparency reports released Apr. 2024

Number of platforms, where zero content moderators speak the country's most commond first language(s)

0 2 3 4 6


Ctrl

4

DIFFERENT REPORTING METHODS COMPLICATE COMPARISON

Although the transparency reports provide insights into which languages the platforms' moderators speak, reviewing the reports shows that big tech often measure and report these numbers differently. In practice, this complicated the comparison of the platforms' approach to content moderation in the different languages:

It is not clear how many unique moderators each platform has associated, and how many moderators are counted several times because they speak several languages.

It is unclear how many hours a moderator must dedicate to moderating content in a particular language in order to count as a moderator. For example, YouTube includes moderators who have viewed at least 10 videos on the platform during the reporting period of six months, while other platforms simply list moderators who can speak a given language.

The platforms count and calculate active monthly users differently. Where some platforms distinguish between active users who are logged in or not logged in to the service, other services calculate one overall figure. This makes it difficult to compare how widespread the services are in a particular country, and thus whether there are sufficient moderators in a given language area.

The platforms do not list in their reports how much content is shared or moderated in different languages on the platform. This makes it difficult to compare the platforms' actual moderation of content in different languages.

The platforms report over different periods, which vary between two to six months.

The lack of standardization for how big tech must count and calculate figures in their transparency reports makes it difficult to compare the different platforms and reduces the real transparency in the platforms' work with content moderation[8].

Consequently, this study, which is based on big techs reported number of moderators, cannot assess the platforms' actual prioritization of the work of moderating content. The study can shed light on whether they have moderators and thus prerequisites for moderating content in a certain language area.

First of all, the study shows how many language areas are not covered by moderators from the eight selected platforms.

The study also points to an overall need for big tech to provide increased insight into their content moderation in a transparent and standardized way in the future, so that the basis for future studies is uniform.

In December 2023, the European Commission published a draft of a template for future transparency reports, which should ensure comparison of the reports in the future[9]. This will also include a higher degree of detail for the reporting, e.g. more details about the number of moderators who have the necessary language skills to moderate in different languages, uniform reporting periods and statements of which content is removed.

In addition to this, Center for Social Media, Tech and Democracy proposes a number of recommendations for increased transparency in big tech's work with content moderation, cf. Box 3:

It is not decisive whether a given platform has five or ten Danish-speaking moderators, if it is unclear how much time these moderators spend on moderation work, and whether they are also listed as moderators for other languages. Stricter requirements must be placed on the reporting, which is already part of the commission's work. Although the transparency reports are a step in the right direction, they do not yet provide the necessary insight into the moderation practices of big tech so that citizens, companies and authorities can have an informed conversation about big tech' impact on society and the democratic conversation online.

BOX 3

RECOMMENDATIONS FOR INCREASED TRANSPARENCY AND STANDARDIZATION IN BIG TECH' TRANSPARENCY REPORTS

To ensure that the transparency reports enable authorities, actors and interested citizens to understand and compare big techs content moderation, more specific requirements for the reports is necessary. Therefore, the Center for Social Media, Tech and Democracy recommends the following requirements for the reports:

- The minimum requirement is that all official EU languages are sufficiently covered by content moderators' language skills
- Clear guidelines for how big tech must list moderators, including clear indication of time (for example, number of annual work units) spent on moderation work in different languages and what requirements are placed on the moderators' language skills.
- The reports should contain statements of what proportion of the content posted within the period is reported, flagged and removed in different languages.
- There should be clear and precise descriptions of how the services moderate content
- All previous versions of big tech' transparency reports must be available to make developments in the field visible.
- All official EU languages must be covered by content moderators' language skills.

REFERENCES

1 Sharevski et al. 2022: “Misinformation Warnings: Twitter’s Soft Moderation Effects on COVID-19 Vaccine Belief Echoes”

2 Drolsbach and Pröllochsouek 2023: “Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database”

3 The European Commission 2023: ”Press release: Commission opens formal proceedings against X under the Digital Services Act”. Udgivet d. 18. december 2023

4 Tjekdet 2024: ”Moderatorerne er fyret: Nu skal brugerne kontrollere brugerne på X”

5 The European Parliament and the Council’s regulation on a Single Market For Digital Services, The Digital Services Act (DSA)

6 The European Commission’s DSA Transparency Database

7 The European Commission (2012): “Europeans and their Languages - Europeans and their Languages : Report”

8 Global Witness 2023: “How Big Tech platforms are neglecting their non-English language users”

9 The European Commission 2023: “Digital Services Act – transparency reports (detailed rules and templates)”

ABOUT CENTER FOR SOCIAL MEDIA, TECH AND DEMOCRACY

Center for Social Media, Tech and Democracy is established as a part of the Media Agreement 2023-2026. The center is placed in the Danish Agency for Palaces and Culture, which already provides professional advice to the Ministry of Culture on the media and tech area, and which deals with digital media, e.g. through the secretariat of the Media Council for Children and Young People.

The center's tasks include, among other things, to contribute knowledge about users' mental well-being, about the importance and consequences of big tech for Danish media, and about the impact, which the spread of misinformation and disinformation on digital platforms has on the democratic conversation.

Read more about the center and its other publications **her**.