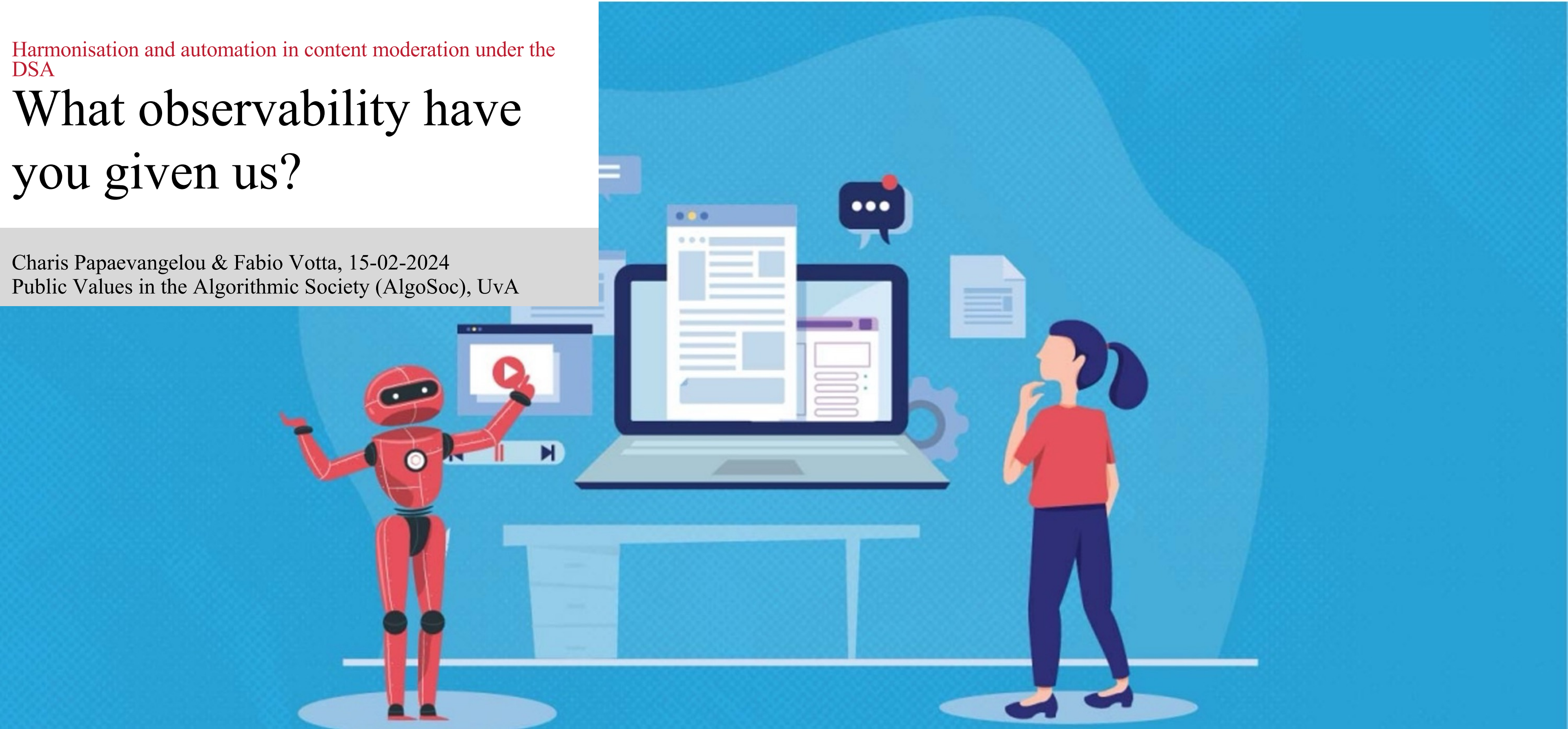
University of Amsterdam
Institute for Information Law

Harmonisation and automation in content moderation under the DSA

What observability have you given us?

Charis Papaevangelou & Fabio Votta, 15-02-2024
Public Values in the Algorithmic Society (AlgoSoc), UvA



Introduction

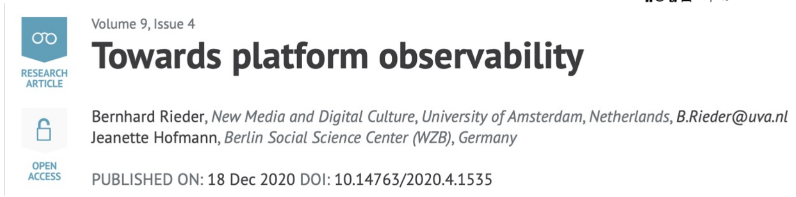
- 🤔 X/Twitter handling all content moderation in one day **manually**
- **Research background**
- **Research questions**
  - RQ1: What are the differences in the decisions, and the implementation thereof, between platforms and across the EU?
  - RQ2: How does automation in moderation vary between platforms and across the EU?

Conceptual framework

- **Platform observability (Rieder & Hofmann, 2020)**
  - Three principles: (i) observation in relation to the public interest; (ii) continuous/dynamic observation; (iii) reinforced capacity for analytical observation

- **Automation**
  - Algorithmic content moderation (Gorwa et al., 2020)
  - Language in AI-backed CoMo (Nicholas & Bhatia, 2023)

- **DSA**
  - 'Regulating for observability' & 'observability as part of regulation' (Rieder & Hofmann, 2020)
  - Also: legal harmonisation & avoidance of regulatory fragmentation

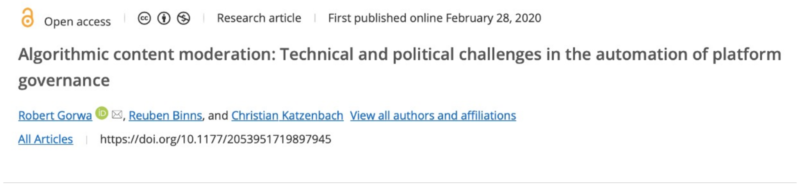


Volume 9, Issue 4

Towards platform observability

Bernhard Rieder, *New Media and Digital Culture, University of Amsterdam, Netherlands, B.Rieder@uva.nl*
Jeanette Hofmann, *Berlin Social Science Center (WZB), Germany*

PUBLISHED ON: 18 Dec 2020 DOI: 10.14763/2020.4.1535

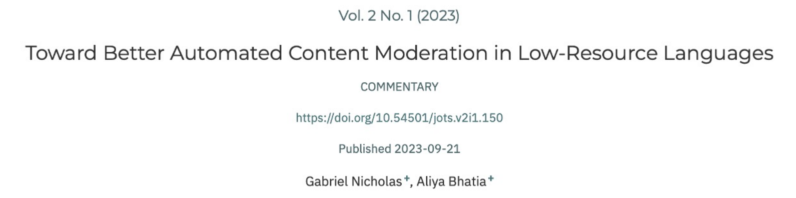


Open access | Research article | First published online February 28, 2020

Algorithmic content moderation: Technical and political challenges in the automation of platform governance

Robert Gorwa, Reuben Binns, and Christian Katzenbach View all authors and affiliations

All Articles | https://doi.org/10.1177/2053951719897945



Vol. 2 No. 1 (2023)

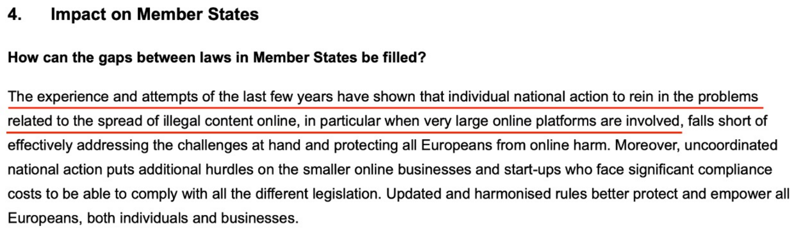
Toward Better Automated Content Moderation in Low-Resource Languages

COMMENTARY

https://doi.org/10.54501/jots.v2i1.150

Published 2023-09-21

Gabriel Nicholas+, Aliya Bhatia+



4. Impact on Member States

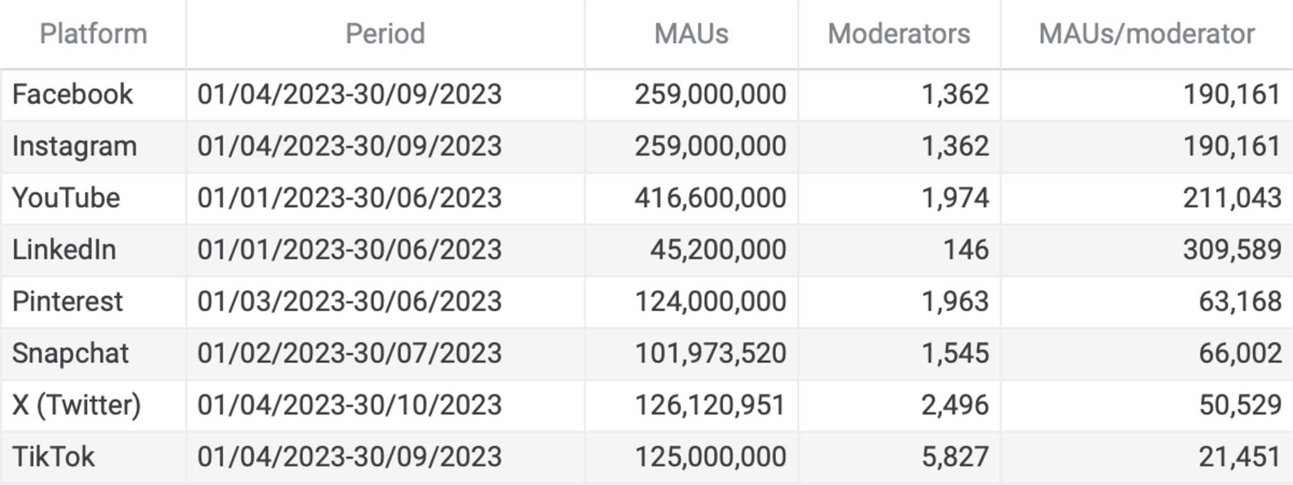
How can the gaps between laws in Member States be filled?

The experience and attempts of the last few years have shown that individual national action to rein in the problems related to the spread of illegal content online, in particular when very large online platforms are involved, falls short of effectively addressing the challenges at hand and protecting all Europeans from online harm. Moreover, uncoordinated national action puts additional hurdles on the smaller online businesses and start-ups who face significant compliance costs to be able to comply with all the different legislation. Updated and harmonised rules better protect and empower all Europeans, both individuals and businesses.

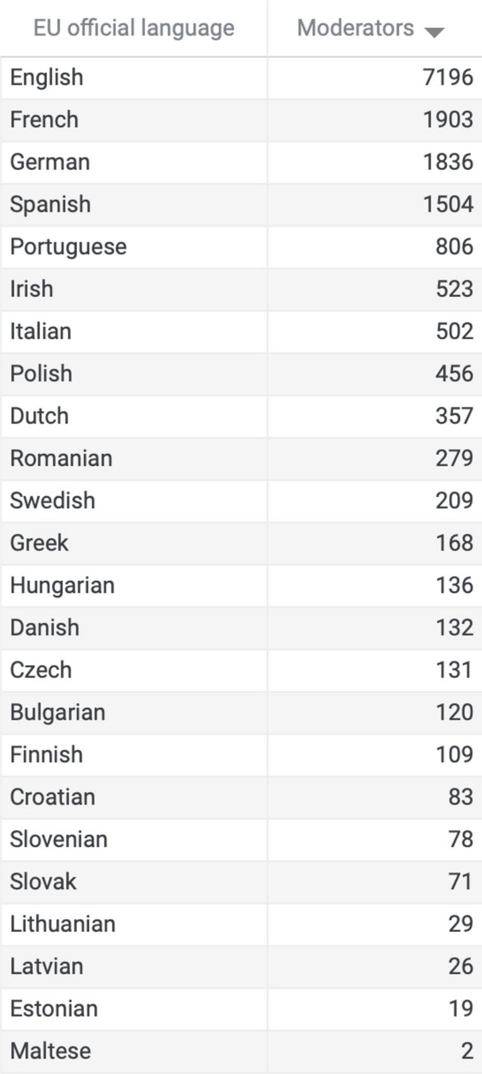
Methods & Data

- Scraped the daily releases of SoRs using R
  - Lack of API & massive bulk .CSV files
- Period (4 months): 25/09/23-25/01/24
- Platforms: TikTok, Facebook, Instagram, LinkedIn, X, Snapchat, Pinterest, YouTube
- **493,135,458 SoRs**
- Also: Collected information from platforms' transparency reports on MAUs & human moderators

Data - Transparency Reports

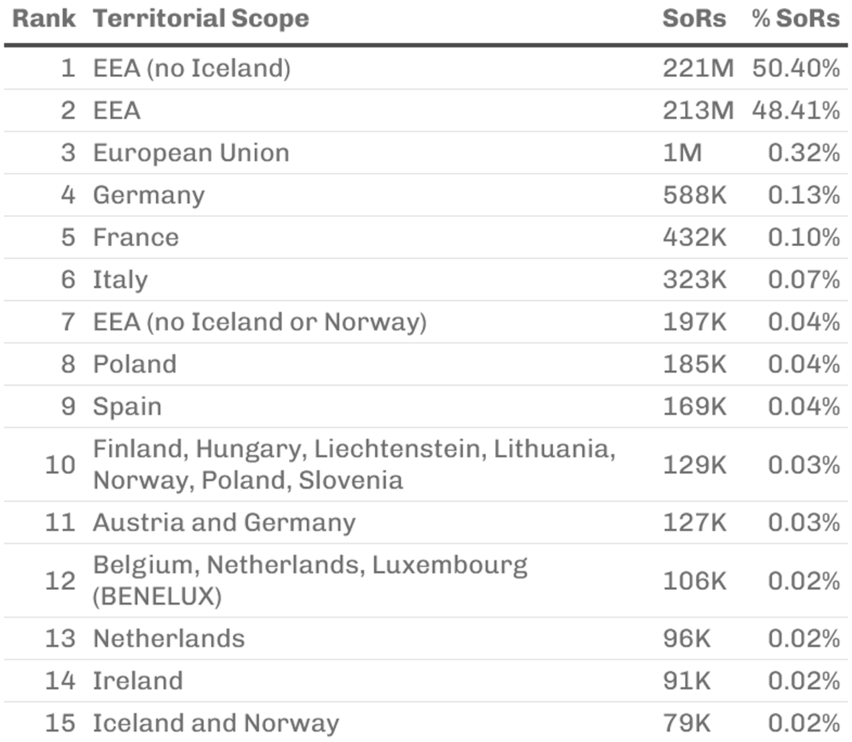


Platform	Period	MAUs	Moderators	MAUs/moderator
Facebook	01/04/2023-30/09/2023	259,000,000	1,362	190,161
Instagram	01/04/2023-30/09/2023	259,000,000	1,362	190,161
YouTube	01/01/2023-30/06/2023	416,600,000	1,974	211,043
LinkedIn	01/01/2023-30/06/2023	45,200,000	146	309,589
Pinterest	01/03/2023-30/06/2023	124,000,000	1,963	63,168
Snapchat	01/02/2023-30/07/2023	101,973,520	1,545	66,002
X (Twitter)	01/04/2023-30/10/2023	126,120,951	2,496	50,529
TikTok	01/04/2023-30/09/2023	125,000,000	5,827	21,451



EU official language	Moderators ▼
English	7196
French	1903
German	1836
Spanish	1504
Portuguese	806
Irish	523
Italian	502
Polish	456
Dutch	357
Romanian	279
Swedish	209
Greek	168
Hungarian	136
Danish	132
Czech	131
Bulgarian	120
Finnish	109
Croatian	83
Slovenian	78
Slovak	71
Lithuanian	29
Latvian	26
Estonian	19
Maltese	2

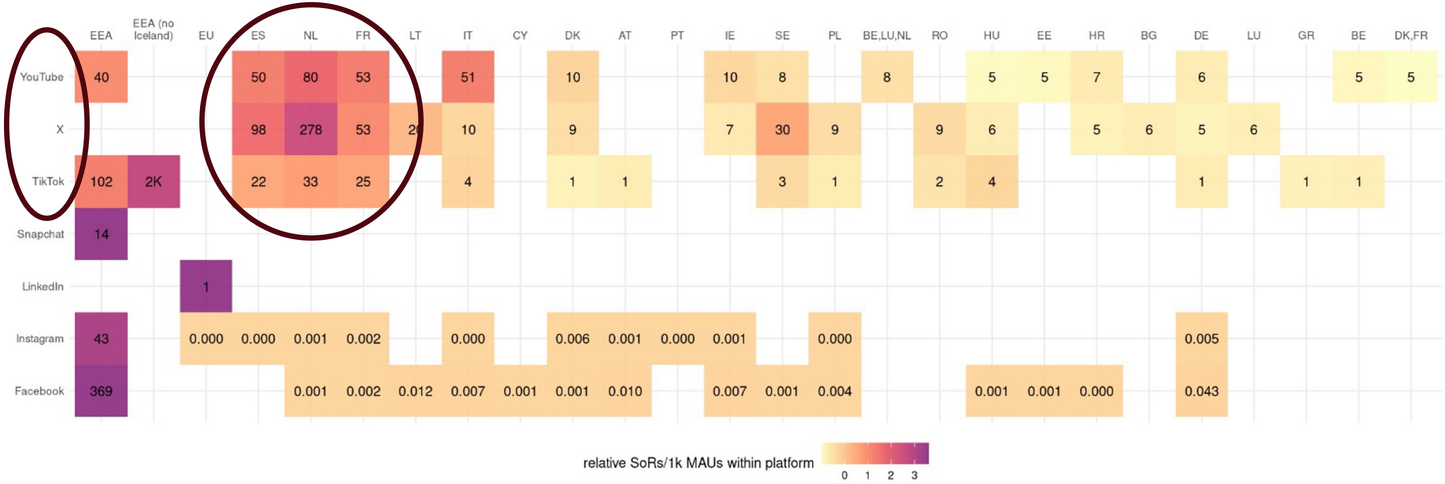
Preliminary findings - Harmonisation



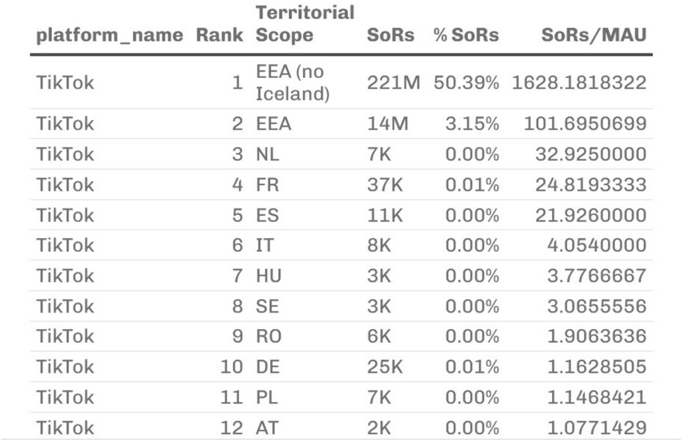
Rank	Territorial Scope	SoRs	% SoRs
1	EEA (no Iceland)	221M	50.40%
2	EEA	213M	48.41%
3	European Union	1M	0.32%
4	Germany	588K	0.13%
5	France	432K	0.10%
6	Italy	323K	0.07%
7	EEA (no Iceland or Norway)	197K	0.04%
8	Poland	185K	0.04%
9	Spain	169K	0.04%
10	Finland, Hungary, Liechtenstein, Lithuania, Norway, Poland, Slovenia	129K	0.03%
11	Austria and Germany	127K	0.03%
12	Belgium, Netherlands, Luxembourg (BENELUX)	106K	0.02%
13	Netherlands	96K	0.02%
14	Ireland	91K	0.02%
15	Iceland and Norway	79K	0.02%

Preliminary findings - Harmonisation

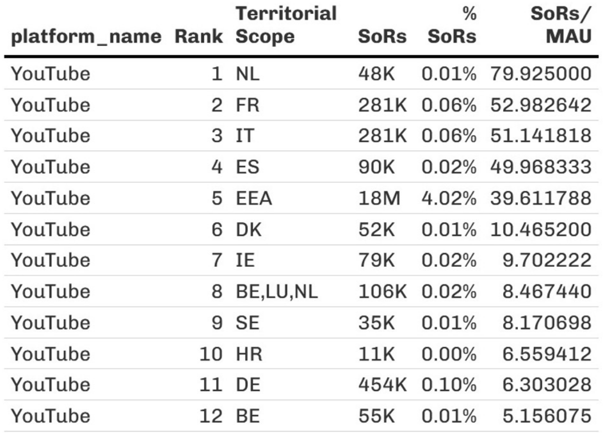
- Darker colors indicate more SoRs per 1,000 users on that platform.
- Outliers: Spain, Netherlands, France.
- Most useful platforms to explore for our RQs: YouTube, X, TikTok



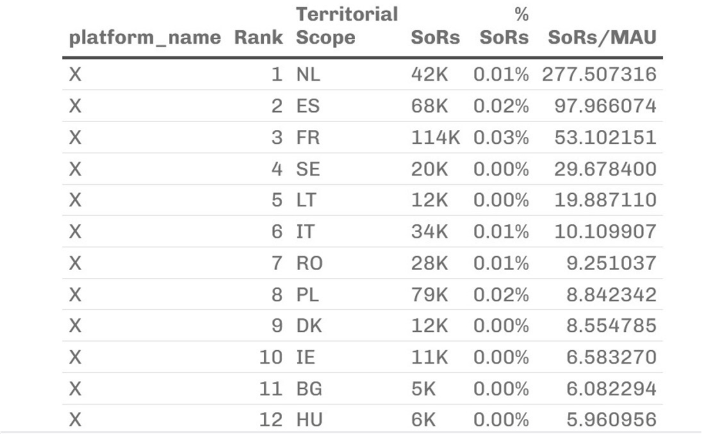
	EEA	EEA (no Iceland)	EU	ES	NL	FR	LT	IT	CY	DK	AT	PT	IE	SE	PL	BE,LU,NL	RO	HU	EE	HR	BG	DE	LU	GR	BE	DK,FR
YouTube	40			50	80	53		51		10			10	8		8		5	5	7		6			5	5
X				98	278	53	20	10		9			7	30	9		9	6		5	6	5	6			
TikTok	102	2K		22	33	25		4		1	1			3	1		2	4				1		1	1	
Snapchat	14																									
LinkedIn			1																							
Instagram	43		0.000	0.000	0.001	0.002		0.000		0.006	0.001	0.000	0.001		0.000							0.005				
Facebook	369				0.001	0.002	0.012	0.007	0.001	0.001	0.010		0.007	0.001	0.004			0.001	0.001	0.000		0.043				



platform_name	Rank	Territorial Scope	SoRs	% SoRs	SoRs/MAU
TikTok	1	EEA (no Iceland)	221M	50.39%	1628.1818322
TikTok	2	EEA	14M	3.15%	101.6950699
TikTok	3	NL	7K	0.00%	32.9250000
TikTok	4	FR	37K	0.01%	24.8193333
TikTok	5	ES	11K	0.00%	21.9260000
TikTok	6	IT	8K	0.00%	4.0540000
TikTok	7	HU	3K	0.00%	3.7766667
TikTok	8	SE	3K	0.00%	3.0655556
TikTok	9	RO	6K	0.00%	1.9063636
TikTok	10	DE	25K	0.01%	1.1628505
TikTok	11	PL	7K	0.00%	1.1468421
TikTok	12	AT	2K	0.00%	1.0771429



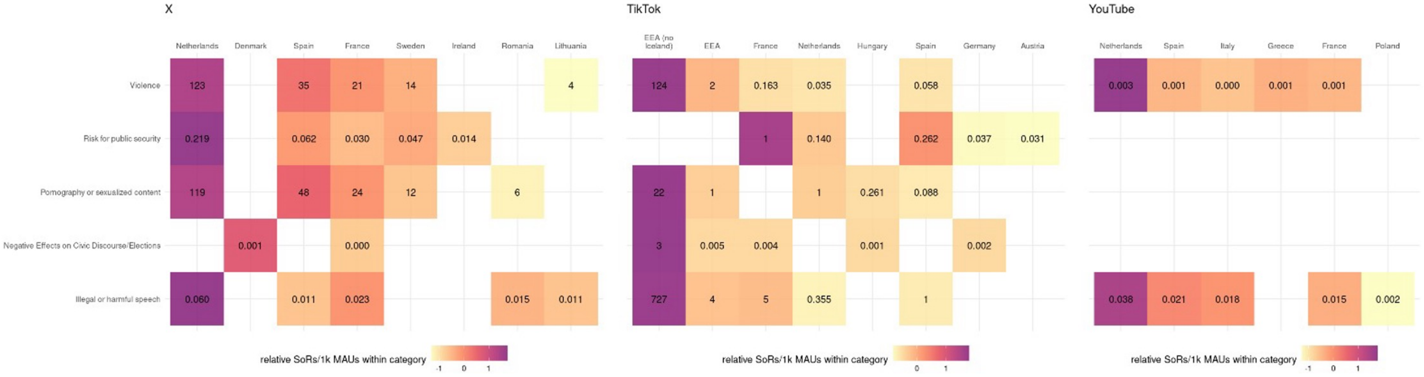
platform_name	Rank	Territorial Scope	SoRs	% SoRs	SoRs/ MAU
YouTube	1	NL	48K	0.01%	79.925000
YouTube	2	FR	281K	0.06%	52.982642
YouTube	3	IT	281K	0.06%	51.141818
YouTube	4	ES	90K	0.02%	49.968333
YouTube	5	EEA	18M	4.02%	39.611788
YouTube	6	DK	52K	0.01%	10.465200
YouTube	7	IE	79K	0.02%	9.702222
YouTube	8	BE,LU,NL	106K	0.02%	8.467440
YouTube	9	SE	35K	0.01%	8.170698
YouTube	10	HR	11K	0.00%	6.559412
YouTube	11	DE	454K	0.10%	6.303028
YouTube	12	BE	55K	0.01%	5.156075



platform_name	Rank	Territorial Scope	SoRs	% SoRs	SoRs/MAU
X	1	NL	42K	0.01%	277.507316
X	2	ES	68K	0.02%	97.966074
X	3	FR	114K	0.03%	53.102151
X	4	SE	20K	0.00%	29.678400
X	5	LT	12K	0.00%	19.887110
X	6	IT	34K	0.01%	10.109907
X	7	RO	28K	0.01%	9.251037
X	8	PL	79K	0.02%	8.842342
X	9	DK	12K	0.00%	8.554785
X	10	IE	11K	0.00%	6.583270
X	11	BG	5K	0.00%	6.082294
X	12	HU	6K	0.00%	5.960956

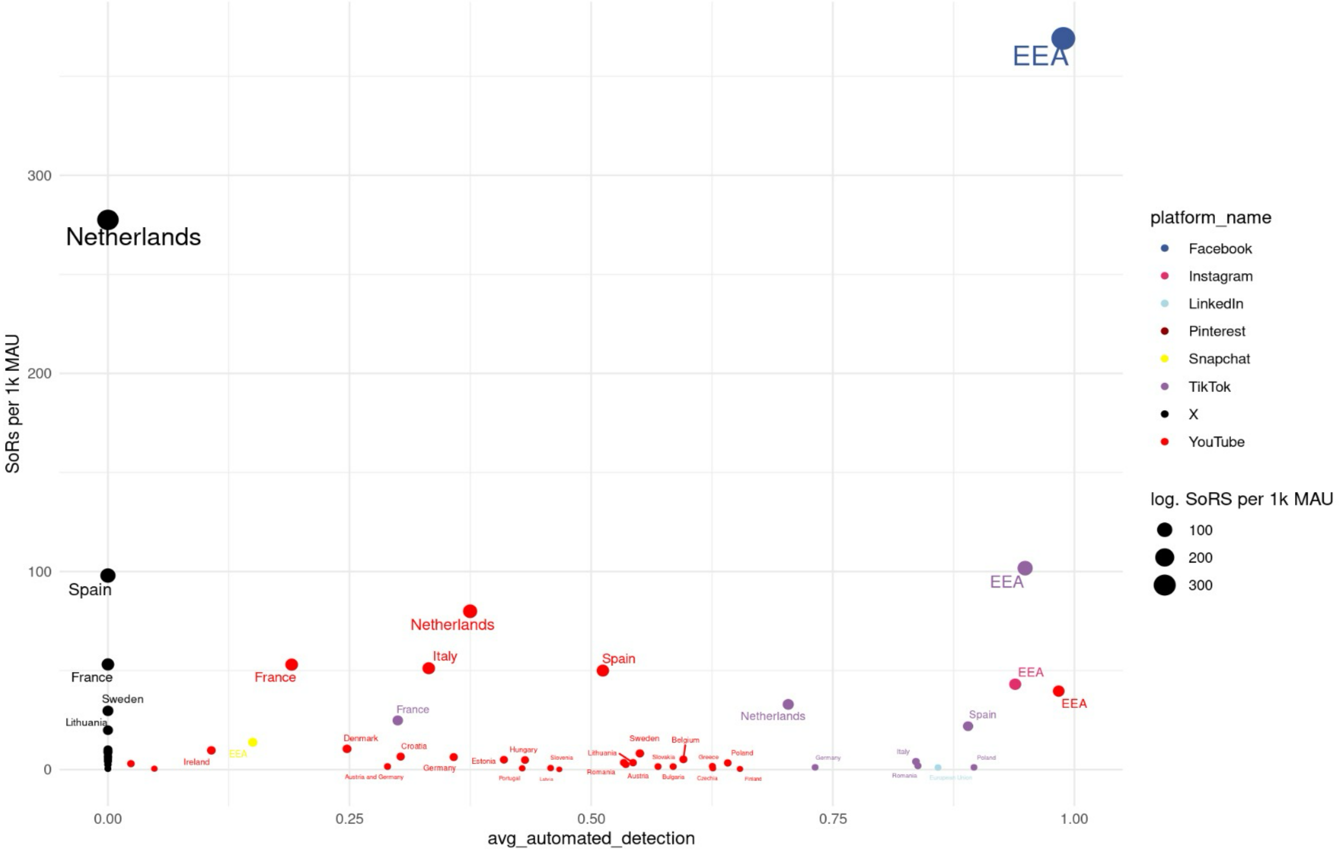
Preliminary findings - Harmonisation

X:



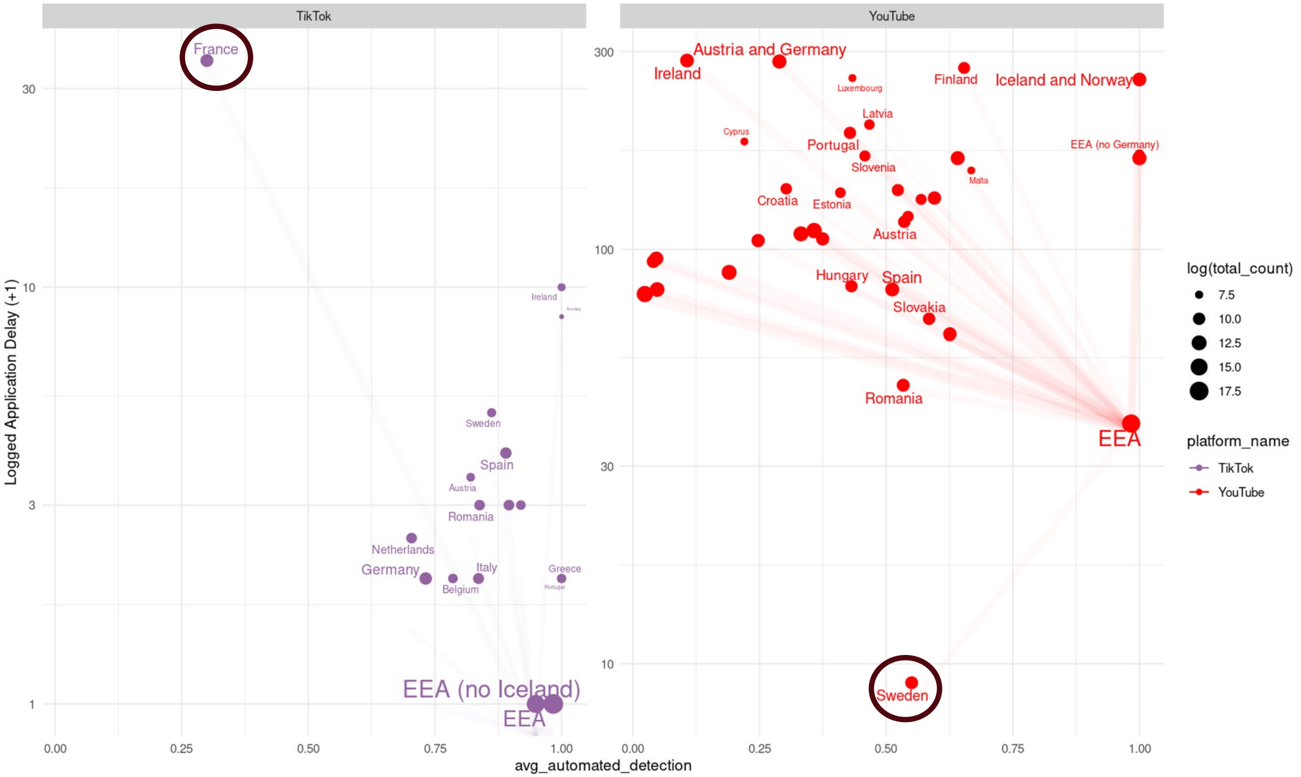
Preliminary findings - Automation

- VLOPs
  - YouTube: mostly combination of automatic detection
  - X: manual detection
  - TikTok: mostly automatic detection
- Variation between territorial scope
- Automation most likely to apply same across EEA and manual on specific member-states
- Countries
  - NL, FR, ES



Preliminary findings - Automation/Harmonisation

- TikTok: When content is automatically detected it is likely to be dealt with swiftly
- YouTube: it can take up to a month (consistent with Trujilo et al., 2024)
- When not automated, it's more likely to get more time to be dealt with (correlation btn automated detection-enforcement Drolsbach & Pröllochs, 2023; DSA SoR Database)
- In cases of decisions with specific territorial scope, decisions take longer (e.g., FR)



Conclusions

- Observability
  - The DSA/SoR Database does allow for a continuous observation of platform behaviour & overall is a positive step towards observability; but has several technical/architectural shortcomings
- Automation
  - Variation across countries and platforms with respect to detection and enforcement
  - X's numbers are dubious
- Harmonisation
  - Broadly exists? (mainly for broad categories and automated moderation)
  - Variation across countries and platforms (especially for manual moderation)
- Limitations
  - Lack of language of content per platform
  - Self-reporting is not enough as platforms lack consistency & might be unreliable
- Thoughts & future research
  - Not 'actionable' observability
  - DSA Art. 40 and access to data might allow us to enrich our understanding

Thank you for your attention!

Charis Papaevangelou

Postdoctoral Researcher

AlgoSoc, UvA - IViR
c.papaevangelou@uva.nl

Fabio Votta

Postdoctoral Researcher

AlgoSoc, UvA - ASCoR
f.a.votta@uva.nl