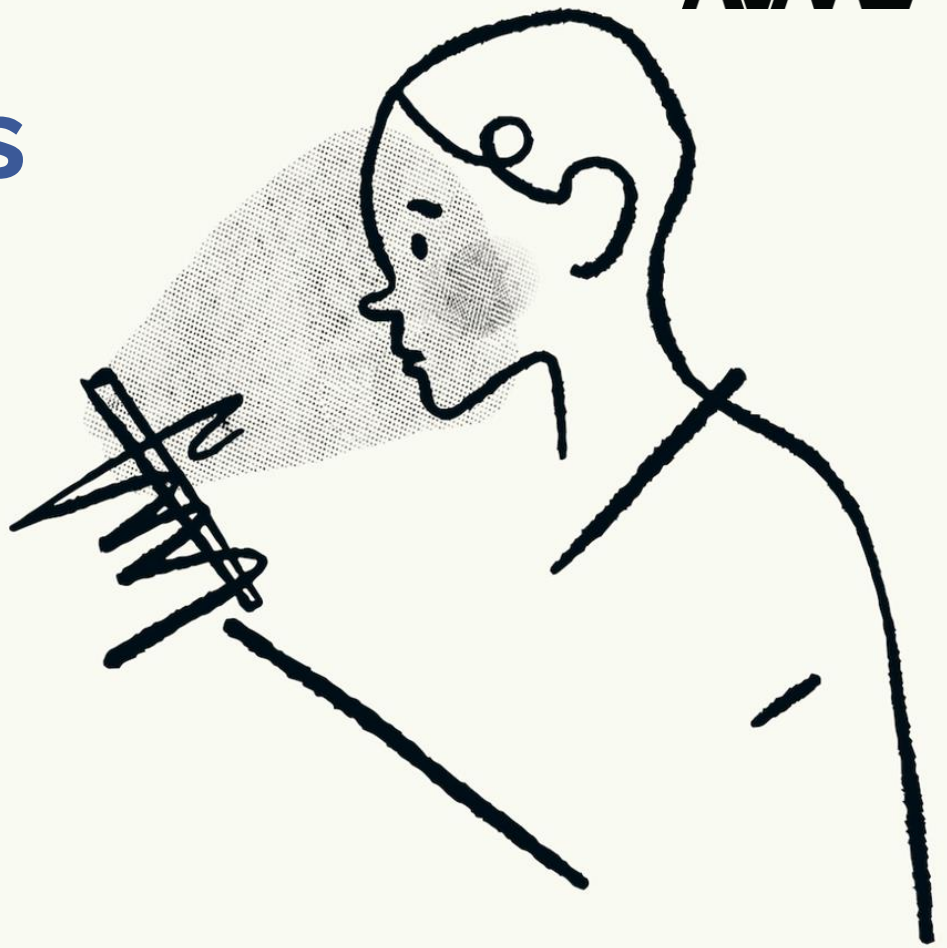
A Human Rights Reboot for Artificial Intelligence



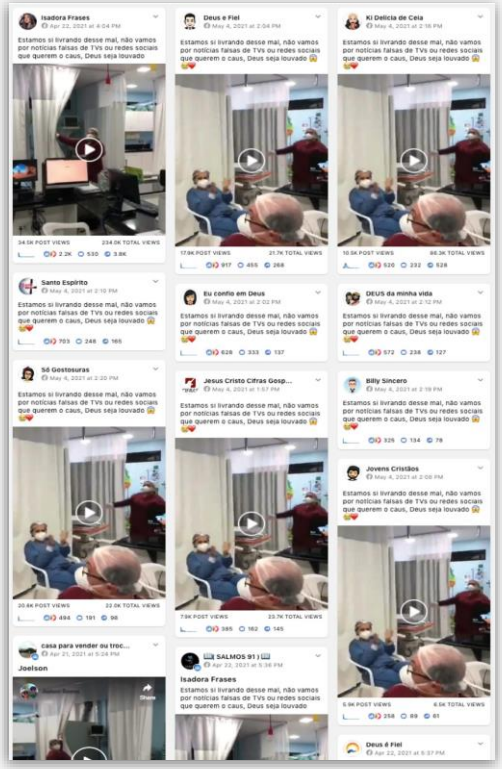
Sarah Andrew
OSCE, Dec 2021

The evidence provided by 2021's global lockdown: AI recommended health disinformation exponentially

In 2020 and 2021, Avaaz investigated the algorithmic acceleration of Covid-19 disinformation.

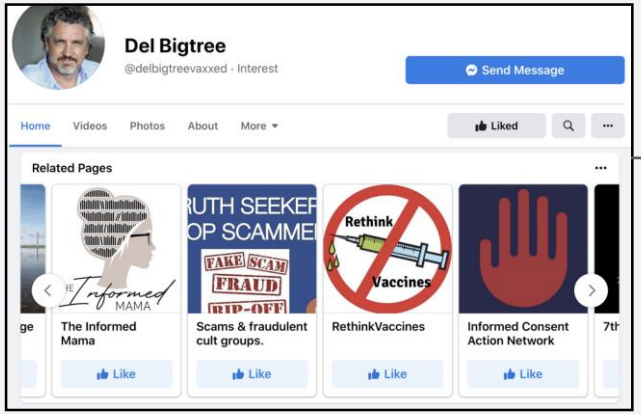
This example had 6.5 million views across Europe.

Spreading in 7 different languages (PT, FR, PL, SR, RO, NL, EN)

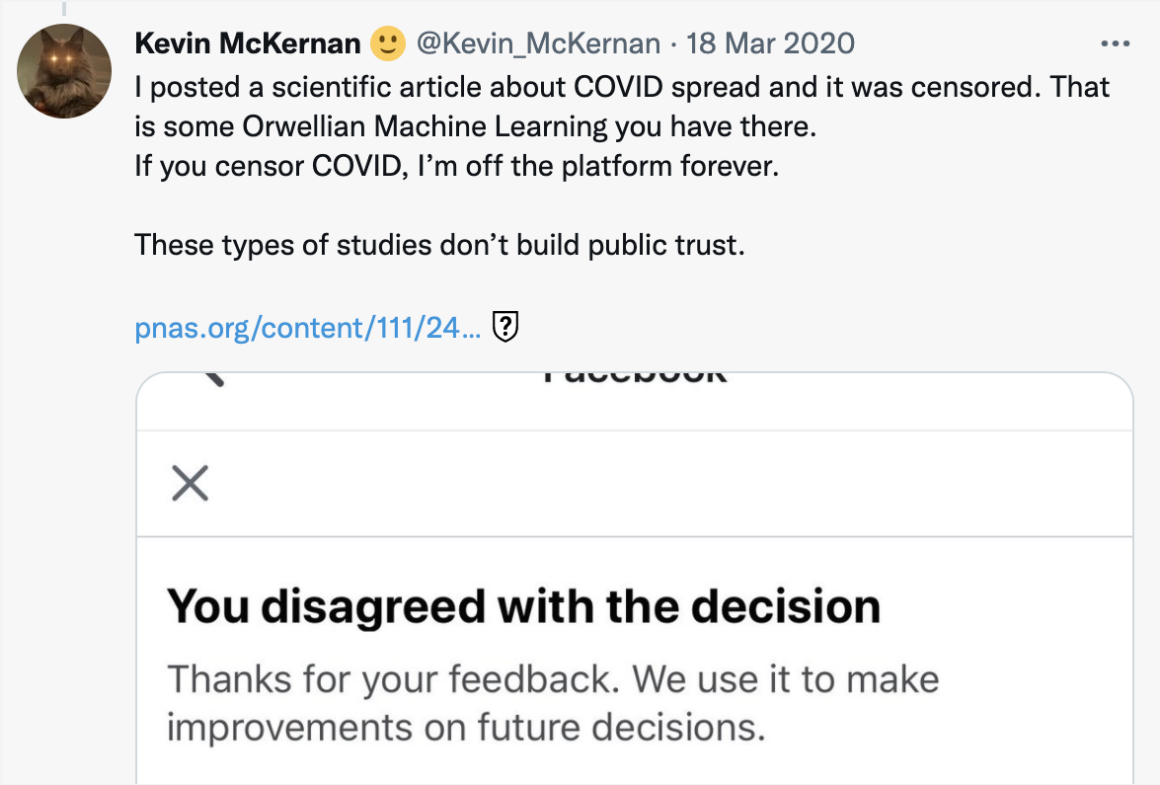


...and recommended funding for the disinformers

Informed Consent Action Network (ICAN) founded by US anti-vaxxer Del Bigtree



...whilst moderation AI censored reliable information



What should the regulatory response be? 1

Accountability

1) **Assess the risks of AI's deployment against risks posed to all fundamental human rights, including:**
    a) **design of the algorithm in the specific context of its deployment**
    b) **training data sets**
    c) **its vulnerability to manipulation or inherent bias**
2) **Mitigate against risks identified before deployment including provision of adequate Human Oversight**

What could accountability do?

In 2019, Avaaz investigated a burgeoning culture of anti-muslim hate in the Indian Region of Assam during the National Register of Citizens

Assam had been identified as an “at risk” region by UN rapporteur Michelle Bachelet, as Assam’s NRC, which had excluded a disproportionate number of Muslims from its citizenship register, had “caused great anxiety to the people of the state”

We found the moderation AI used by Facebook had been built with insufficient data sets in the local languages, so it did not recognise and flag hate-speak against local Bengali Muslims.

The word “miya” is akin to the N word in Assamese when referring to Muslims.

No staff were employed in the region that spoke the relevant dialects.

After our investigation, Facebook introduced a data set of 40,000 new Assamese words into its moderation algorithm.

What should the Regulatory response be? 2

Transparency

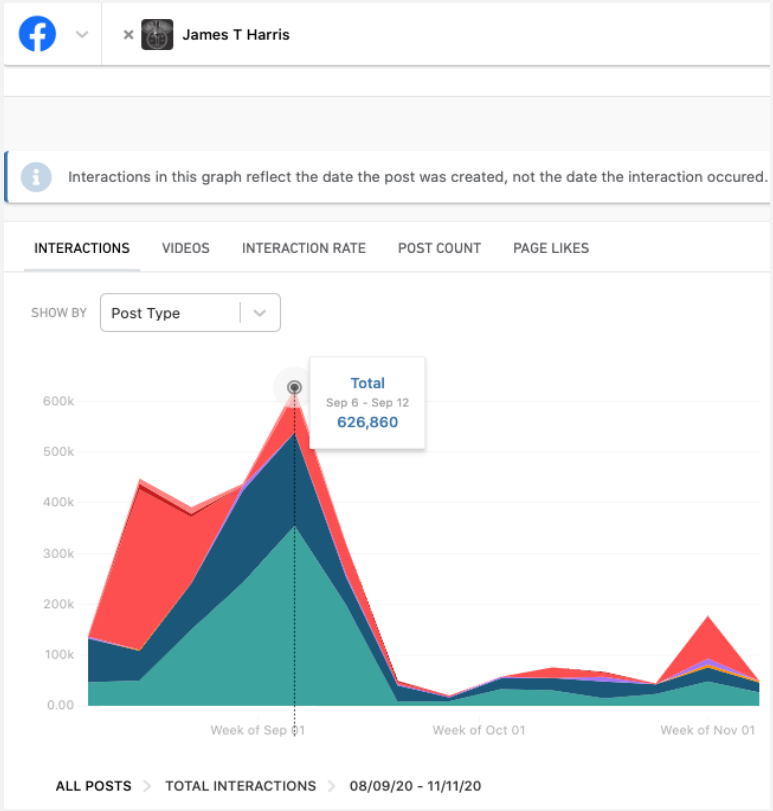
- **For Users**
  - How AI systems collect data and manipulate users' experience, and
  - Outcomes of ongoing Automated decision making
- **For Regulators**
  - Access to risk assessments, and
  - Effective audits of AI operations and claimed mitigation
- **For Civil Society**
  - Access to data underlying the operation of the service in relation to public interest research
- **For Workers**

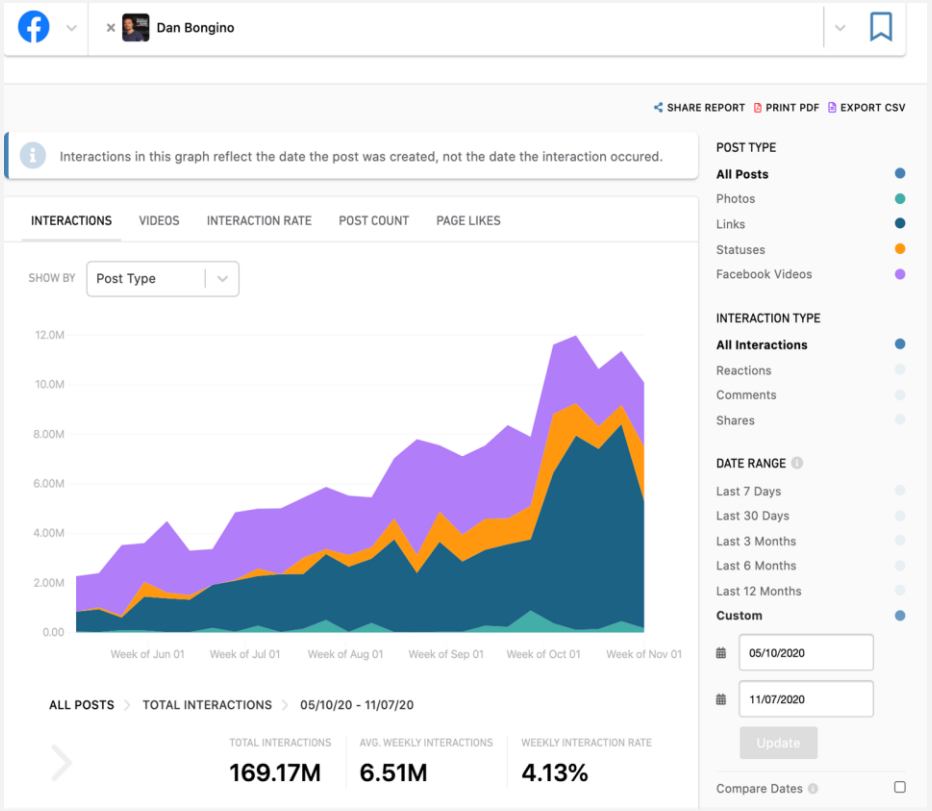
What could transparency do? 1. on content

In the run-up to the 2020 Presidential election, Facebook downranked this story in its recommender algorithm.

Facebook referred to a policy stating that "if we have **signals** that a piece of content is false, we temporarily reduce its distribution pending review by a third-party fact-checker". It never disclosed what those signals were.

What could transparency do? 2. on actors





Source: https://secure.avaaz.org/campaign/en/facebook_fact_check_failure/

The DSA and AIA Proposals: A Quick Comparison

DSA

- **Two-Part Structure**

1) Notice and Takedown provisions for illegal activities.

2) Due diligence framework for service risks and mitigation.

- Cross-sectoral regulation
- Enhanced obligations for very large online platforms ("VLOPs").

EU AI Act

- **Four levels** of risk identified: prohibited (Title II), high risks (Title III), risks of manipulation (Title IV), and all other AI systems.
- Compulsory risk assessment and audits apply only to Title III uses in 8 fixed sector categories only.
- Bans apply to Title 1 uses and voluntary assessment for "all other".

For the DSA (as at 7 December 2021)

ACCOUNTABILITY

Risk Assessment, **Article 26:** risk assessments should extend cross-sectorally to AI systems on Very Large Platforms as against all fundamental rights defined by the Charter, to include risks inherent in the functioning of the services as well as risks through its intentional manipulation.

Risk Mitigation, Article 27: reasonable, transparent, proportionate and effective mitigation measures should include content moderation, algorithms, or recommender systems including their decision-making processes, the design, the features or functioning of their services, their advertising model and their terms and conditions.

For the DSA (as at 7 December 2021)

TRANSPARENCY

For users:

- data use
- algorithmic design with easily usable choices over the values/data use of algorithmic recommenders and
- mitigation actions

For Regulators: **Article 24a/29** in addition to data required to ensure the obligations under Articles 26 and 27 are met, transparency requirements on recommender systems should apply to services beyond the VLOPs

For Civil Society: **Article 31** - the extension of the transparency measures allowing access to data to research organisations to *civil society organisations with an established track record or public interest research in the field their request relates to*

For the AI Act

Accountability

Risk Assessment (Article 7): extend Title III assessment cross-sectorally to the deployment of all AI systems that pose risks of harm to health and safety or a risk of adverse impact on fundamental rights.

Harmonise human oversight (Article 14): AI is as good as its coding and data sets, so human oversight is needed on all AI systems that pose risks of harm to health and safety or a risk of adverse impact on fundamental rights.

Transparency (Articles 13, and 52,1):

Information to users and regulators on how AI systems collect data and manipulate users' experience, with a parallel right of access to civil society as provided under the DSA.

Extend Article 5 - Precautionary bans on:

- ➔ **Harmful social profiling** practices to also include private actors
- ➔ **“Real-time” remote biometric categorisation systems** to track or categorise people in publicly accessible spaces and/or online based on any protected characteristics, or gender identity criteria
- ➔ **Emotion recognition systems** to infer people’s emotions and mental states from behavioural, as well as biometric data
- ➔ **The use of AI systems by law enforcement and criminal justice authorities for the purpose of predicting crimes**
- ➔ **The use of AI systems for immigration enforcement purposes to** profile or risk-assess natural persons or groups in a manner that **restricts the right to seek asylum and/or prejudice the fairness of migration procedures**

Enhance workers rights

- ➔ **Trade Union consultation** on the use of high risk and intrusive forms of AI in the workplace
- ➔ **Transparency** to ensure that workers are aware of the AI systems at the workplace
- ➔ **Human review** of decisions made by AI systems about them, and to be able to "switch off" from work - so workers can have proper downtime
- ➔ **Annual conformity assessment** is needed for workplace based AI to address recruiting and management bias

Contact: sarah.andrew@avaaz.org