Social Media Governance Project

Summary of Work in 2023

October 2023

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on 'Social Media Governance'. The report reflects the personal opinions of the GPAI Experts and External Experts involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the 'Social Media Governance' project, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Responsible AI Expert Working Group. The GPAI Responsible AI Expert Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Alistair Knott[*], School of Engineering and Computer Science, Victoria University of Wellington
Dino Pedreschi[*], Department of Computer Science, University of Pisa

The report was written by: **Alistair Knott**[*], School of Engineering and Computer Science, Victoria University of Wellington; **Dino Pedreschi**[*], Department of Computer Science, University of Pisa; **Raja Chatila**[*], Sorbonne University; **Tapabrata Chakraborti**[‡], Alan Turing Institute, University College London, University of Oxford; **Susan Leavy**[*], School of Information and Communication Studies, University College Dublin; **Ricardo Baeza-Yates**[*], Institute for Experiential AI, Northeastern University; **David Eyers**[†], School of Computing, University of Otago; **Andrew Trotman**[†], School of Computing, University of Otago; **Paul D. Teal**[†], School of Engineering and Computer Science, Victoria University of Wellington; **Przemyslaw Biecek**[*], Warsaw University of Technology; **Stuart Russell**[*], UC Berkeley, and **Yoshua Bengio**[†] Université de Montréal; Mila – Quebec AI Institute.

GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Responsible AI Working Group. We are grateful, in particular, for the support of **Stephanie King** and **Lama Saouma** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Catherine Régis**[*] and **Raja Chatila**[*].

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2023d. Social Media Governance Project: Summary of 2023 Work. Report, October 2023, Global Partnership on AI.

1. Social media platforms are a key vector for AI influence

The founding observation for our project is that *social media platforms are one of the main channels through which AI systems influence people's lives*, and therefore influence countries and cultures. In 2022, the number of social media users worldwide was 4.59 billion (Statista, 2023a); the number is projected to be around 4.89 billion at the current moment, or 59% of the world's population. The average user spent over two and a half hours per day on social media in 2022, a figure which has been rising since 2012 (Statista, 2023b) and is projected to rise further. Crucially, the experience of a social media user, on any given platform, is heavily influenced by AI systems that run on that platform.

- The central component of any social media system is a **content feed** of items presented to the user. The feed for each individual user is curated by a **recommender algorithm**: an AI system that monitors each individual user's interaction with platform content, learns what they like to engage with, and then gives them 'more of the same'. This ability to tailor content to users is what gives social media systems their vast appeal. But the mechanisms through which recommender algorithms learn raise important questions that remain to be answered. Our project has proposed concrete measures for the governance of recommender algorithms, through the study of their effects on platform users. We summarise our work on this issue in Section 2.

- Another pervasive AI influence on social media platforms is in their **content moderation** processes. Content moderation has to be supported by automated tools, given the huge volume of content that is posted, and AI **content classifiers** are the key tools that companies deploy. Again there are concerns about the processes through which these classifiers are trained. Our project is trialling a mechanism for training classifiers outside of companies, in a semi-public domain, which we think may offer a better model for their governance. We summarise our work on this issue in Section 3.

- A final way in which social media companies channel AI influences is through their role as **content disseminators**. The revolutionary advances in generative AI that have taken place over the last year allow ordinary citizens to create, and distribute, AI-generated content, through text generation tools like ChatGPT, and image generation tools like MidJourney. Alongside the many productive uses of such tools, there are concerns that they will facilitate the production of disinformation, potentially destabilising political processes, and other information ecosystems. Our project has recently made a proposal for the governance of generative AI tools, which has had considerable traction within EU policymaking bodies, and also within the US Senate. We summarise this work in Section 4.

2. The governance of recommender algorithms: a summary of our work

The concern motivating our proposal

In very broad terms, a recommender algorithm monitors the behaviour of each user on a social media platform (clicks, comments, 'likes' and so on), and builds a model of what that user likes to engage with, based on this behaviour. The algorithm then prioritises similar content in that user's subsequent feed, to give the user more of what they like. This personal curation of incoming content is what gives social media platforms much of their appeal to users. But it also involves a widely discussed risk. Recommender systems not only learn from the behaviour of users, but also influence this behaviour, because users tend to engage with what they are recommended. This circularity creates a risk that recommender systems may lead users into progressively narrow 'rabbitholes' of content (see Jiang et al., 2019 for a study from DeepMind). Users also have well-studied cognitive biases: for instance, in political domains, to engage with 'moral emotional' content (Brady et al., 2017), content referring to out-groups (Rajthe et al., 2021), and content that is false (Vosoughi et al., 2018); or among teenage girls, to engage with content reflecting ideals of physical beauty (Pedalino and Carmerini, 2022). These biases compound risks arising from circularity, because they suggest users may move preferentially towards certain areas of harmful content: for instance, into domains of political misinformation and extremism, or towards domains promoting eating disorders.

Our proposed method for addressing the concern

Whether recommender systems cause users to move towards domains of harmful content is ultimately an empirical question. In our first report (GPAI 2021), we reviewed the methods that have been used to address this question by researchers working 'externally' to social media companies. We found that these techniques all have serious methodological problems–for instance, in sampling bias, confounding variables, or simulation validity. Most importantly, these 'external' methods cannot test hypotheses about causal effects of recommender systems. Assessing causal effects involves intervening in users' recommender system experiences and measuring user behaviour after different interventions. Crucially, this is exactly what companies themselves do to measure the effects of their recommender systems on their users. The key experiments are **A/B tests**, that deploy different versions of the recommender system to different randomly selected groups of users, and then measure the behaviour of users in the different groups, on various metrics. Any significant differences in user behaviours across groups can be *causally attributed* to the version of the recommender system they were given. Companies use various metrics to measure user behaviour in these experiments. They are particularly interested in measures of 'user engagement', because of the significance of those measures for company revenues. Our proposal is that these same A/B tests can also be used to ask whether manipulating the recommender system affects the amount of harmful content seen or engaged with, by users in different groups. To be clear: we are not proposing the running of new A/B experiments on users. There would be ethical problems with doing this, if the experimental manipulations were done explicitly to test hypotheses about what causes exposure to harmful content. Our proposal is just to re-analyse the logs of user behaviour gathered in A/B tests already conducted by companies, as part of their regular ongoing commercial operation.

Our GPAI group has engaged with many companies, seeking a way to carry out our proposed study, as we describe in our second report (GPAI, 2022). We have worked within the Christchurch Call

(where we helped to define the Initiative on Algorithmic Outcomes that was announced at last year's Christchurch Call Summit). We have worked in the Global Internet Forum to Counter Terrorism, (in particular on a report on evaluation methods for recommender algorithms, GIFCT 2022). But we have had no success with these voluntary collaborations. This year we switched our attention to processes that are newly enabled by EU law, in particular the EU's Digital Services Act (DSA). We received approval from the head of DG CNECT, Roberto Viola, to engage with the European Centre for Algorithmic Transparency, the body which will administer auditing and research processes sanctioned by the DSA. We are in discussions with this body at present, and have submitted a proposal about A/B tests.

3. The governance of harmful content classifiers: a summary of our work

The concern motivating our proposal

Companies are doing plenty of good work in the area of content moderation. But we have two concerns about the AI content classifiers that companies deploy to identify harmful content.

One concern relates to **transparency**. We (the public) don't know very much about how companies create and evaluate their classifiers. We don't know what methods are used to create the datasets they are trained on, or how the annotation process works. We don't know much about how content items are selected for training sets, and how annotators are selected to label these items. We don't know how training sets are kept up to date. Most importantly, we don't know how well the trained classifiers *perform*. It is standard practice for classifiers to be evaluated on examples held out from training: a classifier's performance on unseen examples can be reported as a percentage accuracy. But companies don't report this fundamental evaluation metric.

The other concern relates to **consistency across platforms**. Different platforms have different definitions of harmful content, and have different policies on what to do with such content. At one extreme, some platforms (such as Parler and Gab) have no content moderation at all. But we believe that some moderation is essential to create workable information ecosystems–and this belief is shared by all the mainstream social media companies. In fact, the mainstream platforms adopt somewhat similar definitions of harmful content categories, and are likely to implement somewhat similar content moderation policies on these categories. (Of course we don't know the details, because classifiers are built behind closed doors, as already noted.)

Our proposed method for addressing the concern

The proposal we explore in our project is that companies should all use *the same* definitions of harmful content, in some given locality or jurisdiction. Our specific proposal is that in a given locality or jurisdiction, and for a given category of harmful content, social media companies operating in that jurisdiction should all use *the same training set* to train their classifiers for this content category. In our proposed system, this training set should be developed externally to companies, in a semi-public domain. The public should be informed about the processes by which the training set is created, and updated: specifically, they should know how content items are chosen for inclusion in the training set, and how annotators are selected to perform the labelling. There are nonetheless limits to what the public should know in our proposed scheme. They should not know the identity of the annotators–because this would make annotators vulnerable to coercion of various kinds. They should not have access to the annotated dataset–because this would allow 'adversarial' methods to

be deployed, by purveyors of harmful content who wish to avoid detection. (Such actors could train their own classifiers, and use these to finesse content items that escape detection.)

Our proposal has a number of advantages, which we will now enumerate. For the sake of concreteness, we are focussing on one particular category of harmful content: **hate speech**.

- Firstly, our proposal implements a fundamental democratic principle: definitions of harmful content in a given place should be determined by 'the people' in that place. Consider hate speech, for instance. Our scheme presupposes that the experts in deciding what counts as hate speech in a given place are the people in that place. (We additionally explore the idea that those who are the *targets* of hate should have a particular say in defining what counts as hate.) Often, local expertise involves linguistic expertise too: each language needs its own classifiers, each with its own dedicated training set. Hate speech often also references particular groups in a given locality: many of the relevant groups are local to a given country or region. Our proposal basically implements a principle of *local governance* of social media platforms, at least as regards harmful content moderation, that embodies direct participatory democracy.

- Secondly, our proposal provides a far subtler way of articulating definitions of harmful content than the definitions provided in black-letter law, or company policy. These latter definitions are textual documents, that express certain high-level generalisations. Such definitions leave many open questions about specific content items, and much room for interpretation. An annotated dataset, on the other hand, precisely provides detailed assessment of many specific items: that is, it *includes* the interpretation step. (Note that there is no requirement that annotators agree on their interpretations and assessments. In fact, we argue disagreement can helpfully *inform* content moderation practices, and provide a quantitative basis for some of these.) There is even an interesting argument to be made that a law on harmful content can *reside* in an annotated training set, which we are considering, and would be happy to discuss.

- Thirdly–and moving to more pragmatic considerations–creating a single training set for use by all companies is an *efficient use of resources*. The more items a training set contains, the better the trained classifier will be. If companies can pool resources in the creation of a training set, the result will be a better classifier. The resourcing argument is of particular relevance to the (many) jurisdictions where companies are not able to devote large resources to content moderation, for instance for regions of the developing world, or for indigenous languages.

- Fourthly—and centrally—our proposal makes the process of training content classifiers much more *transparent* than it currently is. There are limits to this transparency, as already noted, but the method whereby training sets are created would be a matter of public record. The method could even be specified by law, in a given jurisdiction.

- Finally, our proposal makes it possible for companies' harmful content classifiers to be more directly *evaluated* than is currently the case. We envisage that companies will continue to build their own content classifiers: these would remain behind closed doors, as companies' private IP. But in our proposed scheme these classifiers would all be evaluated *on the same dataset*. This paradigm for building and evaluating classifiers has in fact been central to all AI machine learning research for the last 30 years. In a given domain, a 'shared task' is defined, for which a training set is made available. A *competition* is then organised around this shared task: research teams compete to build the best system for that task, that learns

most effectively from the shared training set. In competition, trained systems are assessed on unseen data held out from the training set. If the training set for a harmful content classifier is curated in a public domain, and used by all companies, this will allow companies to compete against one another *on the quality of their harmful content classifiers*–that is to say, on a metric that goes directly to the public good.

We are currently running a small pilot of our proposed model for training harmful content classifiers in India, in collaboration with an academic research lab based at Jadavpur University, Kolkata. Our chosen domain is hate speech: specifically, sectarian hate speech that occurs in political discussions. We are using a dataset of tweets posted in the leadup to recent local and national elections in India. We present the results of this pilot study in another report delivered at the GPAI Summit (GPAI, 2023c).

4. The governance of foundation AI models: a summary of our work

The concern motivating our proposal

A final strand of our work at present relates to governance of the new generation of generative AI tools, often termed **foundation models** (FMs; Bommasani et al., 2021). As these models become widely used, the Internet will be flooded with AI-generated content. This will lead to a completely new problem of **attribution**: consumers of content (citizens, communities and companies) will have to consider the new possibility that the content items they encounter (texts, images, videos) are produced, wholly or partly, by AI systems.

This is a serious problem in many domains. Consider a piece of text, encountered by a human reader. In many contexts, her assessment of the text will run very differently if she knows it was generated by an AI system. If she is a teacher assessing a piece of student work, she may want to know how engaged the student has been with the text: have they read it closely, has its content been assimilated? How much learning has taken place? If she is an employer assessing a contractor's report, she may want to know how carefully the provider has overseen its generation: how much work has the contractor done in producing the report? If she is assessing the text as a content moderator working in a social media company, she may want to know whether it is part of a larger-scale communication campaign, given that FMs can readily generate personalised communications at scale, including harmful disinformation. If she is a citizen receiving the text as professional advice from her doctor or lawyer, she may want to know how thoroughly it has been checked for errors, given the known problems of errors in FM output (e.g. Ji et al., 2023) and overreliance on FM output by human operators (e.g. Wang et al., 2023). In each case, the human assessor *needs to know* whether the text is human- or AI-generated, in order to make a proper assessment.

Our proposed method for addressing the concern

The only way we see to meet this consumer need at present is with a tool that allows *automatic detection* of FM-generated content. In the tool we envisage, the user uploads an arbitrary piece of online content, and the tool responds with an analysis of its human or machine provenance. Note that a detection tool for FM-generated content would be particularly valuable for the content

moderation teams in social media companies, to detect and defuse disinformation campaigns, and help keep platforms safe.

There are already many tools that attempt to distinguish AI-generated from non-AI-generated content, both for text (e.g. Chaka, 2023) and images (e.g. Stroebel et al., 2023). But as FM generators improve, the ability of detectors to identify FM-generated content *purely from an analysis of the content* is likely to diminish rapidly (see e.g. Thompson and Hsu, 2023). A consensus is emerging that the only way to create a reliable detector for FM-generated content, as generators improve, is to *instrument the generator* in some way, to *support* detection. This instrumentation might involve placing hidden patterns or 'watermarks' inside generated content, that a detector can identify (see e.g. Kirchenbauer et al., 2023). But there are other methods too; methods that involve logging generated content and then running plagiarism detectors on these logs are particularly promising (see in particular Krishna et al., 2023). For now, the key point is that if reliable detection mechanisms require generators that are configured to support detection, then *responsibility for workable detection mechanisms* ultimately rests with the organisations that build the generators.

Our group argues that legislation should recognise this responsibility. Specifically, we propose that **any organisation that develops a foundation AI model intended for public use should be required by law to demonstrate a reliable detection tool for the content the model generates, as a condition of its release to the public. After release, the detection tool should be freely available to the public.**

We have written two papers this year presenting our proposal in more detail (GPAI 2023a; 2023b). Our proposal has had good traction with policymakers. The EU is in the process of negotiating amendments to its AI Act catering for foundation models; our proposal was adopted (in modified form) by the EU Parliament in its proposed amendments. These amendments are still under discussion with the EU Commission and Council of Ministers; we have had productive discussions with the team in DG CNECT led by Lucilla Sioli, whose comments informed our second paper (GPAI 2023b). Our proposal was also discussed in the US, at a recent Senate Judiciary Hearing on AI Oversight. Two of the three AI experts who gave evidence at this hearing were coauthors on both of our papers: Yoshua Bengio and Stuart Russell. Their advocacy for our proposal is telling, given their expertise in the relevant technologies. Our proposal is also being discussed in the group implementing the G7's Hiroshima AI Process (G7, 2023), whose declared purpose is to cooperate in promoting project-based activities with international organizations, explicitly including GPAI (G7, 2023b). Among the objectives of these project-based activities the latter G7 document specifies: 'advancing research and understanding of state-of-the-art technical capabilities for distinguishing AI-enabled mis/disinformation'. Our proposal about detection mechanisms fits squarely within this category.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313–7318.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 6(2).

G7 (2023a). G7 Hiroshima Leaders' Communiqué. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique/

G7 (2023b). G7 Hiroshima AI Process: G7 Digital & Tech Ministers' Statement. https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf

GIFCT (2022). Methodologies to Evaluate Content Sharing Algorithms & Processes. GIFCT Technical Approaches Working Group report. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TA-Evaluate-1.1.pdf

GPAI (2021). Responsible AI for Social Media Governance: A proposed collaborative method for studying the effects of social media recommender systems on users. Report, November 2021, Global Partnership on AI.

GPAI 2022. Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions. Report, November 2022, Global Partnership on AI.

GPAI (2023a). State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release. Report, 2023, Global Partnership on AI.

GPAI (2023b). Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology (in press).

GPAI (2023c). Crowdsourcing the curation of the training set for harmful content classifiers used in social media: A pilot study on political hate speech in India. Report, November 2023.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. Science, 381(6656), 398-404.

Huszár, F., Ktena, S., O'Briena, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on twitter. PNAS, 119(1), e2025334119.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 383–390).

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., ... & Goldstein, T. (2023). On the Reliability of Watermarks for Large Language Models. arXiv preprint arXiv:2306.04634.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv preprint arXiv:2303.13408.

Pedalino, F., & Camerini, A. L. (2022). Instagram use and body dissatisfaction: The mediating role of upward social comparison with peers and influencers among young females. International Journal of Environmental Research and Public Health, 19(3), 1543.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. Proceedings of the National Academy of Sciences, 118(26).

Statista (2023a). Number of social media users worldwide from 2017 to 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Statista (2023b). Daily time spent on social networking by internet users worldwide from 2012 to 2023. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. Journal of Cyber Security Technology, 7(2), 83-113.

Thompson, S. and Hsu, T. (2023). How Easy Is It to Fool A.I.-Detection Tools? New York Times, June 2023. https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical Considerations of Using ChatGPT in Health Care. Journal of Medical Internet Research, 25, e48009.