Social Media Governance

Summary of work in 2024

October 2024

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation for Economic Co-operation and Development (OECD) mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

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
Susan Leavy[*], School of Information and Communication Studies, University College Dublin

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; **Tapabrata Chakraborti**[‡], The 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; **Toshiya Jitsuzumi**[*], Chuo University; **David Eyers**[†], School of Computing, University of Otago; **Andrew Trotman**[†], School of Computing, University of Otago; **Adrian Weller**[*], University of Cambridge; **Venkataraman Sundareswaran***, SAAZ Micro; **Paul D. Teal**[†], School of Engineering and Computer Science, Victoria University of Wellington; and **Przemyslaw Biecek**[*], Warsaw University of Technology.

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 **Laëtitia Vu, Camille Séguin,** and **Stephanie King** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Francesca Rossi*** and **Amir Banifatemi***.

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

Citation

GPAI 2024. Social Media Governance: Summary of work in 2024, Report, October 2024, Global Partnership on AI.

Table of Contents

1. Introduction.. 1

2. Topic 1: Transparency for AI-generated content.. 1

2.1. AI-generated content and the problem of attribution..1

2.2. Our proposal: responsibility for detection should be placed with generator providers...........2

2.3. Some concrete policy outcomes on AI-content detection... 3

2.4. Advancing the discussion on AI-content detection: our most recent work.............................4

3. Topic 2: A ‘public science’ of social media platform effects.. 4

3.1. Our GPAI project on recommender systems: A recap.. 5

3.2. Our GPAI project on recommender systems: A recap.. 6

4. Topic 3: Democratic governance of harmful content classifiers..8

4.1. Governance of harmful content classifiers: A new proposal... 8

4.2. A pilot project, focussing on classification of political hate speech in India..........................10

5. Our dataset of Tweets, and our set of Tweet annotators.. 12

References...13

1. Introduction

Social media platforms are one of the main vectors for AI influence in the modern world. In 2024, over 5 billion people were social media users, a number projected to rise to 6 billion by 2028 (Statista, 2024a); these users spent over two hours per day on social media (Statista, 2024b). Social media platforms are largely powered by AI systems, so attention to the AI systems used to drive these platforms is a central strand of any AI governance endeavour.

GPAI has been working on social media governance since its inception: the Social Media Governance project has been running since the first round of GPAI projects in 2020. In this report, we summarise the work of the Social Media Governance project in 2024. The report is structured around the three main influences of AI on social media platforms. **Recommender systems** are AI systems that learn how to push content at platform users, through curation of their content feeds. We will discuss our work on recommender systems in Section 3. **Harmful content classifiers** are AI systems that learn how to withhold content from users, by blocking it or downranking it. We will discuss our work on harmful content classifiers in Section 4. Social media platforms are also a key medium for the dissemination of **AI-generated content**. We begin in Section 2 by discussing our work on AI-generated content, and how it can be identified.

2. Topic 1: Transparency for AI-generated content

2.1. AI-generated content and the problem of attribution

The world is about to be deluged with AI-generated content, for instance in textual news and image domains (Newsguard, 2024; Everypixel, 2024). Much of this content will arrive on social media platforms, because social media increasingly provide the means for citizens to produce and consume content (see e.g. Glucksman, 2017; Hendrickx, 2023). AI-generated content is poised to have major impacts on the world's information ecosystem: it is vital we consider how these impacts can be managed, to derive the most benefits from AI generation tools (Gen AI), and identify and manage problems.

In our view, the area where we have most practical scope for management of AI-generated content is in its *identification*. Clearly, we need AI generators to be accurate, and to generate content that is helpful, and not harmful—and work on these areas of safety is of vital importance. But over and above these requirements for *content safety*, we need reliable ways for consumers to know whether an item of content was created by an AI, or by a person, or by some combination of the two: that is, we need mechanisms to provide *transparency about AI-generated content*.

Content transparency is a brand new form of transparency, which arose with the advent of generative AI. For any AI system, we need transparency about how it operates, about how well it performs, and about the data it is trained on. But for generative AI systems, we also need transparency *about the content it produces*. This content flows out into the world, independently of the system that created it, and might end up anywhere: on a social media site, in a newspaper, in student assignments, in commercial reports. Content consumers often have good reasons to ask whether a given item was generated by AI. These reasons aren't primarily about the quality of the content: generative AI tools can produce very good content. (And humans can produce very bad content.) The reasons hinge more on personal relationships, and on trust. If my business colleague gives me a piece of work, I want to know *what her involvement was* in creating this document, because this will affect my future interactions with that colleague (for instance, the questions I have for her, or my discussions about the work). If an article appears under a journalist's byline, readers have a reasonable interest in knowing whether she wrote the article herself, because news consumers have relationships with the people who provide their news. For content posted on social media, we have particular reasons for wanting to know whether it originates with a person or with AI. Individual consumers expect to be interacting with other people, and can be deceived by generative AI content. At a larger scale, generative AI allows individual actors to produce large amounts of high-quality content in coordinated campaigns, and exert larger influences on the information ecosystem, to their own ends. Social media platforms have strong incentives to identify such uses, and moderate appropriately.

In all these cases, the technical instrument that is needed is a tool that *reliably identifies* AI-generated content. To call for such a tool, or to highlight a need for it, is emphatically not to demonise AI-generated content. The tool would simply provide consumers (and distributors) of content with *information about the origin* of any content item, whether AI-generated or not, that is relevant in different ways in different use cases. The purpose of the tool is to provide *transparency* about AI-generated content, so consumers can take action relevant to their particular context of use.

It's useful to state the purpose of the required tool a little more precisely. As we move forward with Gen AI, much content will be produced as *collaborations* between humans and AIs (see e.g. Srinivasan, 2024). The question of attribution is not a binary one: for a given item of content, the critical question for consumers will be *what degree of human involvement* there was in its creation: the answer can range from '100%' to 'almost none'.

At present, of course, we are faced with a serious problem: there are no reliable tools available to the public for identifying AI-generated content. This is the problem we have tackled in our work.

2.2. Our proposal: responsibility for detection should be placed with generator providers

Our group has argued that reliable AI-content detection tools can realistically be delivered, if the right incentives are placed with the companies that build the generators. It's widely agreed that there's no prospect of building reliable AI-content detection tools that operate by identifying 'signatures' of AI generation in overtly visible content: generators are rapidly eliminating differences of this kind (see e.g. Májovský et al., 2024). However, if companies *instrument* their generators to support content detection, the prospects for reliable detection tools are much greater. There are several conversations here. One is about provenance-authentication mechanisms, such as the C2PA standard, that can indicate in an item's metadata that it was produced by AI—or, conversely, authenticate it as having been produced by a human (see e.g. Bengio et al., 2024). Another conversation is about the use of 'watermarks' which are hidden in an imperceivable way within digital content, but can be identified by a tool developed in tandem with the generator, by the same provider (see e.g. Liu et al., 2024; Aberna and Agilandeeswari, 2024). A third conversation is about 'logging' methods for content detection. In these methods, the generator provider keeps a private log of all content that is generated: the provider can then implement a detection tool as a plagiarism checker on this private log (Krishna et al., 2023; Yang et al., 2023). These latter methods don't receive enough attention in discussions about detection tools; discussions could be usefully moved in this direction.

All of the above detection methods can of course be attacked. Provenance metadata can be eliminated from images by changing format, or by screenshots; watermarks can often be broken by paraphrasing or changing words. Logging methods appear to be somewhat more resilient to paraphrase attacks (Krishna et al., 2023), but further research is urgently needed. (Research into combinations of methods would be particularly useful, as the shortcomings of one method are often covered by others.) Of course, the large generator providers are no strangers to adversarial scenarios of this kind: they can certainly be required to extend detection methods, to cater for known adversarial exploits. Again, those who provide generators are by far the best placed agencies to provide reliable detection tools. In particular, they are the *only* agencies who can deliver detection tools that can crucially identify the extent of human involvement in a piece of generated content—because they are the only agencies who are party to the prompts that are used to create AI content in the first place.

In view of these considerations, our group at GPAI argued for a novel type of law for AI generation. We proposed that an organisation developing an AI generation system *should be required to demonstrate a reliable detection tool for the content their system generates—as a condition of the system's public release*. We presented this argument in a GPAI report last year (GPAI, 2023a). The report received considerable attention in the policymaking community, and was the subject of many discussions. We wrote a second paper, summarising the idea, and picking up on these discussions (Knott et al., 2023). This paper considered many potential problems that arise for our proposal, and attempted to respond to these. (For instance, it considers the proper definition of 'AI generator', and how the proposal might work for 'open-weights' AI generators operating in the public domain; it also considers who should bear the costs for developing detector tools, and how 'user-facing' detector tools could be built from the detectors supplied by providers.) We summarised these ideas in a piece for the OECD's Policy Wonk (Knott and Pedreschi, 2023), which also received lots of attention.

2.3. Some concrete policy outcomes on AI-content detection

We had particular engagement with two groups of policymakers. In the EU, we had discussions with the members of the European parliament (MEPs) who were drafting the text of the AI Act—in particular, while amendments were being created for 'foundation models'. We also had meetings with policymakers in DG-CNECT, where the implications of our proposal were considered in detail. The final text of the AI Act incorporates the substance of our proposal quite effectively: Article 50.2 states that 'Providers of AI systems (...) generating synthetic audio, image, video or text content *shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated'* (our emphasis). In the US, our proposal was also discussed, at last year's Senate Hearing on AI Safety, where two of our co-authors, Yoshua Bengio and Stuart Russell, gave evidence. There are some traces of this discussion in President Biden's Executive Order on AI, in particular in its requirement that the Office of Management and Budget (OMB) make 'recommendations to [executive departments and] agencies regarding (…) reasonable steps to watermark or otherwise label output from generative AI'. It's hard to identify the origins of policy statements in any detail, but our work was certainly part of the discussions that fed into these two concrete policy outcomes.

2.4. Advancing the discussion on AI-content detection: our most recent work

Our most recent paper (Knott et al., 2024b) picks up from the new policies on AI-content detection that are now enacted. It assumes that when these policies are in force, reliable AI-content detection tools will sometimes be available. In this new scenario, there are new questions to be considered. In particular, who will use these new detection tools? We argue that some organisations (and individuals) will have natural incentives to use them, out of self-interest, to help preserve their reputation, and to help in engagements with collaborators. But these incentives are not present for all organisations. In particular, social media companies are not strongly incentivised to use these tools, because they have no responsibility for the content that is posted on their platforms. In our new paper, we argue that new laws may be needed, that require certain media providers to make use of reliable AI-content detectors, if these are available.

Our intention in this paper was, broadly, to start a new discussion, about the proper use of AI-content detectors. This discussion must tackle questions of how AI-generated content should be properly labelled, and whose job this labelling should be. We raised some interesting possibilities in the paper: for instance, one idea is that on social media platforms, users should be able to choose how much AI-generated content they see. This helps to place decisions about how AI-generated content disseminates with individual content consumers, and gives consumers the ability to choose their own settings.

We also considered the incipient arms race between the providers of AI-content detection tools, and those who seek to evade detection. We concluded our paper by discussing the broader question of how policymakers can influence this arms race, to favour content detection. On that matter, we made a few fairly predictable proposals—governments should support research efforts to build reliable tools, in particular for smaller generator providers, and help with information-sharing. But we also made a more contentious proposal: that governments should consider banning the open-sourcing of ‘frontier’ generative AI models, because the requirement that each generation system is accompanied by a reliable detection tool is much harder to enforce in the open-source world than it is for companies deploying in-house, closed-source systems. In this last proposal, we aligned ourselves with a number of recent commentators, who have argued for a range of reasons that open-sourcing the most powerful generative AI models is an unsafe practice (see e.g. Seger et al., 2023; Harris, 2023).

3. Topic 2: A ‘public science’ of social media platform effects

Social media platforms have effects on their users, and on wider society. Many of these effects arise directly from content posted by other users. But some effects can be attributed to the technologies that drive the way content is disseminated around platforms—which, as we already noted, are largely AI technologies. Recommender systems are a key place where we can look for effects of this kind.

Our GPAI project has been concerned with recommender systems since its inception. We’ll begin by briefly summarising the work we have done in previous years, and then report our new work this year.

3.1. Our GPAI project on recommender systems: A recap

The task of a recommender system is to decide how to prioritise (or ‘rank’) the content items that appear in users’ feeds. These might be news feeds, or feeds of recommended friends, or recommended products: the range of recommender systems is very broad, and growing all the time. For any given feed, a user encounters the items in the feed one-by-one: a recommender system essentially decides which order these items are presented in. The recommender system fulfils the function of a ‘personalised editor’, for each user. This functionality is one of the main novel features of social media platforms, and a key reason for their popularity with users.

Recommender systems are AI systems: their key role is to learn the kinds of content that users like, and then give them more of the same. But the way they learn is subject to certain problems, that are very familiar to AI engineers: the new data they learn from is heavily *influenced* by the things they have already learned. In very concrete terms, recommender systems learn from what items users click on. But by and large, users click *on the items that they are recommended.* (When scrolling through Instagram, we tend to watch what is suggested for us; when searching Google, we tend to choose from the first page of returned results.) What this means is that a recommender system’s learning is to some extent a self-fulfilling prophecy. In AI parlance, we say that the datasets that train recommender algorithms are ‘non-IID’, which means the samples in the dataset are not ‘independent and identically distributed’ (see e.g. Cao, 2016). Learning in such scenarios is inherently unstable (Jiang et al., 2019). Many methods are used to address this instability (see e.g. Cao, 2022); but it creates a general cause for concern, that recommender systems have effects on individual users, and on wider communities, that push towards extremes.

In our first GPAI project report (GPAI, 2021), we documented this cause for concern in some detail, with a focus on effects relating to political extremism. We argued that policymakers need a better understanding of the effects of recommender systems. We reviewed the empirical methods available for studying these effects—and concluded that by far the best methods are those that companies themselves use, and which are not available to external researchers (such as those working in universities, or for public interest groups). The paradigm method used by companies to study the effects of their systems is **A/B testing**. In an A/B test, the company creates two different versions of a given system, and deploys them to two randomly-selected groups of users. It then monitors these user groups over a period of time. If there is any statistically significant difference in the behaviour or experience of the two groups, it can be reliably attributed as a causal consequence of the difference in the systems they are interacting with. An A/B test is an experiment that intervenes in the experience of users, and tests the causal consequences of this intervention. In this sense, it’s like an FDA drug trial: the acid test for examining the causal effects of an intervention. There is no equivalent to A/B tests for researchers working externally to platforms, because only platforms have the ability to intervene in their users’ experiences, and monitor the effects of these interventions.

In our first report, we proposed that companies should collaborate with external researchers, to study potentially harmful effects of their recommender algorithms, using A/B tests. We argued that this could be done without risks to the IP of companies: A/B tests provide transparency about the effects of company algorithms, rather than their internal workings. We also argued that it could be done without any risk of disclosing the personal data of platform users: the results of A/B tests are reported as aggregate measures over large groups of users, and the measures themselves abstract far from the content produced and consumed by users. We concluded by proposing that companies should conduct A/B tests in a particular domain: that of political extremism. At that time, companies had all recently committed to the Christchurch Call to Eliminate Terrorist and Violent Extremist Content (TVEC) Online, so this area seemed a good one to focus on.

In our second year, we engaged with many companies, asking for A/B tests of this kind to be done. Much of this work happened through interactions in a working group at the Global Internet Forum to Counter Terrorism (GIFCT), of which we were a member, along with representatives from many tech companies. This working group released a report on A/B tests, as part of its operation (Thorley et al., 2022). But essentially we could not get any A/B tests approved. We got very close with Twitter: even to the point of an announcement being made at the 2022 Christchurch Call Summit, jointly by Jacinda Ardern and Francois Macron (ChCh, 2022). But by then Elon Musk had taken over the company, and he sacked everyone in the team we were working with at Twitter. All these efforts are documented in our second project report (GPAI, 2022).

At this point, we pivoted from working within the Christchurch Call, which is a voluntary collaboration between companies and governments, to working with the EU, which has developed black-letter law for social media platforms: specifically, the Digital Services Act (DSA). The DSA has interesting provisions that will allow access to 'Very Large Online Platforms' (VLOPs), including the large social media platforms, by external researchers. The DSA has been in force since November 2022, but its provisions for external researcher access are still being formulated at the time of writing: they are contained in a separate Delegated Act. Our focus this year, as we'll discuss in the next section, has been on the details of this Act.

3.2. Our GPAI project on recommender systems: A recap

The DSA provides two new types of access to external parties. One is to auditors, for the purpose of checking compliance with DSA regulations. The rules governing audits are set out in a Delegated Act of the DSA: they give auditors broad powers to access platform data and algorithms, and to conduct experiments. The other type of access is to vetted external researchers. The Delegated Act defining this kind of access is still being finalised, as already noted. But the core purpose of this access is to conduct new studies to foresee 'societal risks' associated with VLOPs, and to gauge the effectiveness of mitigation measures for risks that are identified. Researcher access to VLOPs is crucial for these purposes, because the best ways to study risks and mitigation methods make use of data or methods that are only available within platforms.

A/B tests are a particularly useful method for risk assessment and mitigation, because they allow direct assessment of options that are under the control of companies. It is of course impossible to 'avoid all harms' with technology as pervasive and wide-ranging as a social media platform. But an A/B test can explore the effects of alternative system designs and configurations, and measure which of these results in *least harm*. This is vital information for the governance of platforms—and to our mind, exactly the kind of information that policymakers had in mind when drafting the DSA. Note that in relation to the DSA, the brief for 'external researchers' is necessarily broader than that of 'auditors', because researchers are not assessing compliance against a finite checklist of obligations: their job is to identify risks of VLOPs for users and society, and thus in some sense to *define* the detailed checklist for auditors' assessments.

For the reasons outlined above, our GPAI group wanted to ensure that the researcher access provided by the DSA to VLOPs includes access to companies' experimental platforms—including access to A/B testing protocols. We published two pieces arguing this point. One was a policy brief, released as a GPAI report (GPAI, 2024). The other was a piece directed at researchers working in social media impacts, and disseminated by the Forum for Information and Democracy, coauthored by our GPAI project leads, Stuart Russell and Jonathan Stray (Knott et al., 2024a). We also had many discussions with researchers, to canvass our proposal about the scope of DSA access. We had considerable support for our suggestion; our second piece was accompanied by a list of 39 supporters from the research community, including some very influential voices in AI (Yoshua Bengio, Adrian Weller, Achim Rettinger), tech law and policy (Gillian Hadfield, Rebekah Tromble), tech ethics (Virginia Dignum, Jeroen van den Hoven, Ricardo Baeza-Yates, Raja Chatila), and tech transparency (Brandon Silverman, founder of CrowdTangle). A clear sense is emerging from these initiatives: that the DSA provides a unique opportunity to enable a new *public science* of tech platform impacts, in which the important questions about societal risks and their mitigation are studied by independent researchers, using the best available methods. A/B tests for recommender algorithms provide one example of work conducted in this new paradigm, but there are many others: for instance, work on the performance of AI-content detection tools (see Section 2 of this report), or on the performance of harmful content classifiers (see Section 4 below).

Our work on the DSA's Delegated Act on researcher access to VLOPs has been useful in assembling a community of prominent researchers who are interested in DSA access. The DSA creates a new focus for organisation within the research community, which is manifested in several recent events: for instance, the DSA Stakeholder conference in April 2023, and the events organised by the DSA Observatory at the University of Amsterdam. In collaboration with EU's DG-CNECT, we are in the process of initiating a grouping of researchers who are interested in DSA access, for the purpose of coordinating applications for access, communicating results of work already conducted or currently under way, and conveying information about the methods available within companies—all, naturally, within the guidelines mandated by the DSA. The grouping will be called the **Social Data Science Alliance**.

The alliance we are initiating is just one element in the new research structures that will need to be developed, to support DSA access by external researchers. Other elements of structure are needed to support the process of vetting and approving applications. Initiatives are under way in this area too: within the European Digital Media Observatory (EDMO), a working group is running, with the goal of creating an **Independent Intermediary Body** (IIB) sitting in between researchers and platforms, with roles in monitoring research (to safeguard its independence), in addressing legal issues that arise around data privacy, and to mediate disputes between companies and researchers. A key function of the proposed body would be to 'vet researchers and their research proposals'. Of course executive decisions in this area will be made by the Digital Services Coordinators (DSCs) of the relevant EU member countries: but of course DSCs will need to be advised; the proposed intermediary body would be able to provide the relevant advice. In this sense, the body would have some of the functions of a grant-awarding institution, assessing the quality of research proposals and research teams. It would also have some of the functions of an ethics committee, assessing projects on ethical criteria, including privacy.

Our proposed Social Data Science Alliance and EDMO's proposed Independent Intermediary Body have distinct roles: our Alliance is a grouping of researchers, coordinating and making proposals under the DSA, while the IIB assesses proposals, and has a role in overseeing projects that are accepted, when they are in operation. We look forward to progressing these two new elements of research structure.

4. Topic 3: Democratic governance of harmful content classifiers

The third area of work for us this year has been a project on harmful content classifiers. This project was initiated last year: again, we'll begin by summarising the motivation for the project, and then review the work we have done this year.

4.1. Governance of harmful content classifiers: A new proposal

All social media platforms must practice content moderation (in one form or another), so that 'harmful content' is identified and removed, or restricted in its circulation (in one way or another). Minimally, there are certain types of harmful content whose possession and distribution are illegal in given jurisdictions, and must be removed to comply with local laws. But beyond these legal limits, social media platforms enforce various moderation policies, to comply with more broadly defined social responsibilities. Of course, content moderation—or censorship, by another name—is a contentious process. The principle of 'removing harmful content to keep platforms safe for users' is in opposition with the principle of 'allowing users freedom of speech': the key question for any moderation process is how to reconcile these two opposing principles.

On social media platforms, content moderation partly involves human workers. But the main platforms deal with so much content that a component of automation is absolutely essential. And this automation component is, again, mostly provided by AI systems: harmful content classifiers. Our interest is in the governance of these classifiers.

Several of the main problems in relation to harmful content classifiers again relate to transparency. One important problem is that we don't have detailed information about the *definitions* of harmful content that are implemented by company classifiers. We have textual definitions, given in general terms. But the *detail* of the definition implemented by a classifier for a given class resides in its training set: specifically, in the set of examples of this class that are identified in the training set. The training set provides a far more nuanced definition of the category that the classifier will learn. And companies disclose very little about how training sets for their classifiers are constructed.

Another problem is that we don't have enough information about how company classifiers *perform* in identifying harmful content. The main measure companies use to report the performance of their classifiers is the 'proactive detection rate', which is the percentage of violating content found by its classifiers before it is reported. This is certainly a useful measure to report; but it would also be useful to know how a classifier performs on a set of 'test' examples held back from training. This is the normal yardstick for evaluating classifiers, used in almost every academic and commercial use case. It is strange that companies don't use it to report performance of their harmful content classifiers.

Two further problems with current practice arise from the fact that companies all build their own 'in-house' datasets to train their classifiers. This problem is partly one of *consistency* across platforms: companies implement their own definitions of harmful content of different kinds, each featuring their own taxonomies—but what counts as 'harmful' (and how it is moderated) arguably shouldn't be something that varies from platform to platform. But the problem is also partly one of *efficiency*. Training sets are expensive to curate, because they require large numbers of annotators to provide labels for large numbers of training items. Larger training sets are better, as a general rule in supervised learning. So the resources deployed by companies to create training sets would be better spent in the construction of a *single* training set, for any given type of harmful content (and locale). The different companies could then train their classifiers on the same set. This scheme would achieve both consistency in category definitions and better efficiency of annotation resources. As an additional benefit, it would instigate a useful new form of competition between companies. They would each train their own classifiers, on the single common training set—and their classifiers could also be evaluated on the same set of held-out training examples. This type of competition through a ‘shared task’ is in fact the way academic research in machine learning has operated for at least the last 20 years: we see no reason why it shouldn’t also operate for harmful content classification by commercial companies.

Our GPAI project has argued for a new way of training harmful content classifiers, that makes use of a shared training set of this kind. Our proposal is outlined in a report for the GPAI summit last year (GPAI, 2023b). In the scheme we envisage, all companies operating in a given region contribute resources to enable the assembly of a single shared training set, implementing a single set of definitions. These definitions would relate to moderation actions, rather than to semantic categories of content. Thus annotators have to indicate whether a given item should be removed from a platform, or downranked in the recommender algorithm, or left alone. These ‘action-based’ definitions force annotators to confront the dilemma between respecting free speech and keeping platforms safe in every annotation they make. (Two annotators may find a given item of content equally harmful, but have different views about the limits on free speech, and consequently provide different annotations.)

Our final proposal is that for some types of content, annotations should be made by *a representative sample of the public*, rather than by a dedicated in-house team of annotators. This scheme is certainly not appropriate for some forms of harmful content—in particular, content featuring violence, abuse or sexual material. But for other forms, it is an interesting idea to explore. In particular, we suggest that the training set for a ‘hate speech’ classifier might be usefully assembled from annotations provided by sampling the public. This is the idea we have focussed on in our work. We’ll describe what we have done so far in Section 4.2.

Before we continue, we should stress that while we are exploring the idea of constructing training sets for harmful content classifiers ‘outside companies’, we are certainly not envisaging that the training sets themselves will be publicly available. That, of course, would open up opportunities for adversarial operators to evade classification. The shared training set would be private. But, crucially, the processes through which the training set is constructed, and maintained, would be transparent to the public.

Our suggestion is that this proposed scheme might lead to harmful content classifiers whose decisions are more accountable. Their training sets are built by annotators who are explicitly considering the trade-offs between harmful content moderation and free speech maintenance. There is transparency about how they are built. The classifiers on different platforms all implement the same definitions. And they are transparently evaluated, on a level playing field, in ways that promote healthy competition between companies.

In the case of classifiers that are trained by sampling public opinion, there’s an additional element of accountability. Annotations are gathered through what is essentially a *democratic process*, reflecting the opinions of the people in the location where the classifier operates. (In fact, we plan to use opinion polling methods too, in future iterations of this work.) Of course there will be disagreement between annotators. But disagreement can be very useful, and can be readily accommodated in classifier training. Neural network classifiers can be trained to return ‘the majority verdict’ for a given training item, and ignore disagreement between annotators. But they can also be trained to return the complete distribution over recorded annotations. (The AI paradigm here, often referred to as ‘soft labels’, is very easy to implement: the standard loss function for classifiers is a cross-entropy term, which is easily used to train a classifier to produce a probability distribution over output categories: see Uma et al., 2021 for an introduction to this method). If our classifier can produce a probability distribution over possible categories for a given item, we can measure the entropy of that distribution, to estimate how much disagreement there is between annotators for items of this type. We can then modify our moderation action based on this disagreement—for instance, by erring towards less moderation in cases where there is high disagreement, to retain an existing plurality of opinion.

Of course, all of these ideas must be put to the test. Our group is piloting them in a study running in India, on political hate speech. We will describe that study in the next section.

4.2. A pilot project, focussing on classification of political hate speech in India

Our decision to focus on India for our pilot study is motivated from several considerations. Firstly, hate speech is a serious problem in Indian politics, which often leads to violent incidents and even fatalities (see e.g. Mirchandani, 2018). Secondly, there is a tendency for hate speech researchers (and social media platforms) to devote more effort to Western languages, and less effort to so-called ‘low-resource’ languages. Finally, India is the world’s largest democracy (and the GPAI chair nation for 2024), and offers interesting opportunities to explore democratic consultation. Our current project is a very small-scale pilot, which only consults a small group of respondents, because our main focus is on establishing a workable proof of concept protocol and analysis methods. But India offers rich opportunities to scale up the current pilot, if its evaluation yields promising results.

There were two ‘phases’ of annotations in the process we are piloting. In Phase 1, annotators made *categorical* decisions about the content items (Tweets and memes) they saw on a web based annotation platform built in-house for this project. Should a given item be removed, or downranked, or retained without any action? The dataset created by these annotators would be suitable for training (or fine-tuning) a *classifier*, which learns to return a probability distribution over these three discrete alternative categories, using a ‘soft-labels’ loss term as described above.

In Phase 2, annotators made decisions that will be used to inform the downranking process, for items where downranking is the appropriate action. Downranking items in a recommender algorithm happens on a *continuous* scale, rather than a discrete scale. To inform downranking, we need a way of placing content items on a continuous scale of harmfulness (or ‘hatefulness’, in the case of our chosen domain). Our approach to this task is to ask annotators to make choices between pairs of items—which is ‘worst’? That is, which should be disseminated less on a platform? We chose this method because there’s good evidence that when people are asked for judgements relating to a continuous scale, they are better at deciding on the ordering of pairs of items than they are at placing individual items at absolute positions on the scale (see e.g. Goffin and Olson, 2011). For instance, teachers mark most consistently if they work by placing assignments in ranked order, rather than by directly assigning marks. There are many methods for converting data about preference orderings over pairs of items into valuations of individual items on a continuous scale.

The Bradley-Terry model is commonly used for this purpose, for instance (see Firth, 2005 for an introduction). In our analyses, we use a Bayesian version of the Bradley-Terry model (Caron and Doucet, 2012), which will enable us to associate each content item in the training set with a Normal distribution on the continuous scale of harmfulness, and compare the same against a vanilla Bradley Terry approach. Items where there was high disagreement between annotators (identified within the Bradley-Terry model) have distributions with higher variance. Again, we can train a neural network model to reproduce these distributions, so that it captures a measure of annotator disagreement. We can then use the variance of a predicted distribution to modify the moderation action that is taken, to take account of disagreement.[1]

In our chosen domain of Indian political hate speech, we have completed pilots of both phases of annotation. Phase 1 is described in GPAI (2023b), and further in a conference paper (Bhattacharya et al., 2024). Phase 2 has just been completed; our initial description will appear at this year's GPAI Summit (GPAI, in press). In each case, we explored annotations of content items in two modalities: firstly Tweets, short texts, gathered from Indian political discussions relating to the national election in 2019, the state elections in 2022 and this year's national election; and secondly graphical 'memes', scraped from Google images but using search keywords similar to the hashtags used for the Tweet dataset.

Unlike the standard Bradley-Terry method, the Bayesian approach yields a much higher correlation, primarily because it incorporates prior information—in this case, the discrete labels assigned to the Tweets. As a result, Tweets that appear in fewer pairwise comparisons, which accounts for the majority, tend to align more closely with the prior information.

We also tested three recent LLM models as content classifiers on the annotated Tweet dataset. They were all variants of the BERT model (Devlin et al., 2019): RoBERTa (Liu, 2019), ALBERT (Lan, 2019) and DistilBERT (Sanh et al., 2019). Our results are summarised below.

Methods	**Accuracy**		**Precision**		**Recall**		**F1 score**	
	4 labels	2 labels	4 labels	2 labels	4 labels	2 labels	4 labels	2 labels
RoBERTa	0.7842	0.8151	0.8729	0.8529	0.8213	0.8017	0.7677	0.8477
ALBERT	0.7408	0.7862	0.7221	0.7523	0.6915	0.7011	0.6872	0.7527
DistilBERT	0.7241	0.7759	0.7722	0.7616	0.7213	0.8011	0.6436	0.7471

Table 1. Summary of results from fine-tuning three LLM content classifiers on our annotated Tweet dataset.

[1] As an aside—it happens that preference decisions about the relative harmfulness of pairs of content items also feature prominently in another area of AI at present—'AI alignment'. This active area of research is concerned with fine-tuning large-scale generative AI models to encourage them to produce 'helpful' outputs, and discourage them to produce 'harmful' outputs. The datasets used to perform this 'alignment' often involve annotators' preferences over pairs of alternative outputs, generated for a given prompt—and typically make use of the Bradley-Terry model to interpret annotator preferences (see e.g. Rafailov et al., 2024, and much subsequent work). There are interesting potential commonalities between this work in content *generation*, and work in content *interpretation* like ours, that seeks to place existing content items on a continuous scale of harmfulness. Both tasks seem likely to make use of the same conceptions of harmfulness. And both tasks make use of similar datasets and methods. But at present, these two areas of work operate rather separately. We are currently exploring ways of bringing them closer together.

5. Our dataset of Tweets, and our set of Tweet annotators

We are looking forward to continuing our work on the governance of AI systems deployed in social media platforms. This work has rich connections with AI governance work being done elsewhere in GPAI, and in the OECD. Our work touches on several aspects of Generative AI safety: our proposals about ways of ensuring that AI-generated content can be reliably detected (see Section 2) address an overarching safety concern; and our methods for placing content on a continuous scale of harmfulness have interesting connections with methods used in LLM alignment (see Section 4.2). These issues are being explored in both GPAI and OECD. Our work on harmful content classifiers has important connections to GPAI's project on diversity and gender equality, because hate speech is often directed at groups defined on the dimensions discussed in this project. This same work is also connected with GPAI's project on scaling responsible AI solutions, because the pilot project described in Section 4 is now precisely at a point where it can be scaled up and replicated. Our proposals for DSA-enabled A/B tests exploring recommender system effects engage with many of the data and privacy issues explored in GPAI's Data Governance working group, and OECD's Data and Privacy expert group. At the same time, the topic of AI governance in social media is a central topic in AI governance in its own right. As we noted at the outset, social media platforms are one of the main vectors for AI influence in the modern world, that influence the lives of 5 billion people—that is, over 60% of the world's population. There are many aspects of AI deployment in social media that need their own domain-specific analysis and governance methods; the governance of recommender algorithms (see Section 3) is a case in point, as is the issue of political content moderation and free speech (Section 4). We are keen to continue our work in these important areas in 2025.

References

Aberna, P., & Agilandeeswari, L. (2024). Digital image and video watermarking: methodologies, attacks, applications, and future directions. Multimedia Tools and Applications, 83(2), 5531-5591.

Bengio, Y., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y. et al. (2024). International Scientific Report on the Safety of Advanced AI. DSIT research paper series number 2024/009.

Bhattacharya, A., Chakrabarti, T., Basu, S., Knott, A., Pedreschi, D., Chatila, R., ... & Biecek, P. (2024). Towards a crowdsourced framework for online hate speech moderation-a case study in the Indian political scenario. In Companion Publication of the 16th ACM Web Science Conference (pp. 75-84).

Caron, F., & Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. Journal of Computational and Graphical Statistics, 21(1), 174-196.

Cao, L. (2016). Non-IID recommender systems: A review and framework of recommendation paradigm shifting. Engineering, 2(2), 212-224.

Cao, L. (2022). Beyond IID: Non-IID thinking, informatics, and learning. IEEE Intelligent Systems, 37(4), 5-17.

ChCh (2022). Christchurch Call Initiative on Algorithmic Outcomes. New Zealand Government announcement, September 2022.
https://www.beehive.govt.nz/release/christchurch-call-initiative-algorithmic-outcomes

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

Everypixel (2024). People Are Creating an Average of 34 Million Images Per Day. Statistics for 2024. https://journal.everypixel.com/ai-image-statistics. Accessed October 2024.

Firth, D. (2005). Bradley-Terry models in R. Journal of Statistical software, 12, 1-12.

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

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. GPAI Social Media Governance Project. November 2022.

GPAI (2023a). State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release. GPAI Social Media Governance Project. June 2023.

GPAI (2023b). Crowdsourcing the curation of the training set for harmful content classifiers used in social media: A pilot study on political hate speech in India. GPAI Social Media Governance Project. November 2023.

GPAI (2024). How the DSA can enable a public science of digital platform social impacts. Social Media Governance project Policy Brief. May 2024.

GPAI (in press). Crowdsourcing annotations for harmful content classifiers: an update from GPAI's pilot project on political hate speech in India. GPAI Social Media Project Report (to be presented at this year's Summit).

Glucksman, M. (2017). The rise of social media influencer marketing on lifestyle branding: A case study of Lucie Fink. Elon Journal of undergraduate research in communications, 8(2), 77-87.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. Perspectives on Psychological Science, 6(1), 48-60.

Harris, D.E. (2023). How to regulate unsecured "Open-Source" AI: No exemptions. Tech Policy Press, December 2023.
https://www.techpolicy.press/how-to-regulate-unsecured-opensource-aino-exemptions/

Hendrickx, J. (2023). The rise of social journalism: An explorative case study of a youth-oriented Instagram news account. Journalism Practice, 17(8), 1810-1825.

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).

Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., ... Russell, S. & Bengio, Y. (2023). Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology, 25(4), 55.

Knott, A. and Pedreschi, D. (2023). Human, or human-like? Transparency for AI-generated content. OECD AI Policy Wonk.
https://oecd.ai/en/wonk/human-or-human-like-transparency-for-ai-generated-content

Knott, A., Pedreschi, D., Stray, J. and Russell, S. (2024a). The EU's Digital Services Act must provide researchers access to VLOPs' experimental protocols. Forum for Information and Democracy post.
https://informationdemocracy.org/wp-content/uploads/2024/06/The-EUs-Digital-Services-Act-must-provide-external-researchers-access-to-companies-experimental-platforms-2024.pdf

Knott, A., Pedreschi, D., Jitsuzumi, T., Leavy, S., Eyers, D., Chakraborti, T.,… Russell, S. & Bengio, Y. (2024b). AI content detection in the emerging information ecosystem: new obligations for media and tech companies. Ethics and Information Technology 26:63.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 27469-27500).

Lan, Z. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942(2019).

Liu, Yinhan (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, A., Pan, L., Lu, Y., Li, J., Hu, X., Zhang, X., ... & Yu, P. (2024). A survey of text watermarking in the era of large language models. ACM Computing Surveys.

Májovský, M., Černý, M., Netuka, D., & Mikolov, T. (2024). Perfect detection of computer-generated text faces fundamental challenges. Cell Reports Physical Science, 5(1)

Mirchandani, M. (2018). Digital hatred, real violence: Majoritarian radicalisation and social media in India. ORF Occasional Paper, 167, 1-30.

Newsguard (2023). Rise of the Newsbots: AI-Generated News Websites Proliferating Online. https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/. Accessed October 2024.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., and Finn, C., Direct Preference Optimization: Your Language Model is Secretly a Reward Model, arXiv: 2305.18290 (2024).

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Gupta, A. (2023). Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks. Benefits, and Alternative Methods for Pursuing Open-Source Objectives.

Srinivasan, S. (2024). Detecting AI fingerprints: A guide to watermarking and beyond. Brookings Institute report.
https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/

Statista (2024a). Number of social media users worldwide from 2017 to 2028. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Accessed October 2024.

Statista (2024b). Daily time spent on social networking by internet users worldwide from 2012 to 2024. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/. Accessed October 2024.

Thorley, T., Llansó, E., & Meserole, C. (2022). Methodologies to evaluate content sharing algorithms & processes. GIFCT Technical Approaches Working Group report.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from disagreement: A survey. Journal of Artificial Intelligence Research, 72, 1385-1470.

Yang, X., Pan, L., Zhao, X., Chen, H., Petzold, L., Wang, W. Y., & Cheng, W. (2023). A survey on detection of LLMs-generated content. arXiv preprint arXiv:2310.15654.