Ethics and Information Technology _####################_
https://doi.org/10.1007/s10676-024-09795-1

ORIGINAL PAPER

AI content detection in the emerging information ecosystem: new obligations for media and tech companies

Alistair Knott[1,2] · Dino Pedreschi[1,3] · Toshiya Jitsuzumi[1,4] · Susan Leavy[1,5] · David Eyers[1,6] · Tapabrata Chakraborti[1,7,8] · Andrew Trotman[6] · Sundar Sundareswaran[1] · Ricardo Baeza-Yates[1,9] · Przemyslaw Biecek[1,10] · Adrian Weller[1,11] · Paul D. Teal[1,12] · Subhadip Basu[1,13] · Mehmet Haklidir[1,14] · Virginia Morini[1,3] · Stuart Russell[1,15] · Yoshua Bengio[1,16,17]

Accepted: 5 August 2024

Abstract

The world is about to be swamped by an unprecedented wave of AI-generated content. We need reliable ways of identifying such content, to supplement the many existing social institutions that enable trust between people and organisations and ensure social resilience. In this paper, we begin by highlighting an important new development: providers of AI content generators have new obligations to *support* the creation of reliable detectors for the content they generate. These new obligations arise mainly from the EU's newly-finalised AI Act, but they are enhanced by the US President's recent Executive Order on AI, and by several considerations of self-interest. These new steps towards reliable detection mechanisms are by no means a panacea—but we argue they will usher in a new adversarial landscape, in which reliable methods for identifying AI-generated content are commonly available. In this landscape, many new questions arise for policymakers. Firstly, if reliable AI-content detection mechanisms are available, *who should be required to use them?* And how should they be used? We argue new duties arise for media companies, and for Web search companies, in the deployment of AI-content detectors. Secondly, what broader regulation of the tech ecosystem will maximise the likelihood of reliable AI-content detectors? We argue for a range of new duties, relating to provenance-authentication protocols, open-source AI generators, and support for research and enforcement. Along the way, we consider how the production of AI-generated content relates to 'free expression', and discuss the important case of content that is generated jointly by humans and AIs.

Keywords Generative AI · AI-generated content · AI regulation

✉ Alistair Knott
ali.knott@vuw.ac.nz

[1] Social Media Governance Project, Global Partnership on AI, Montreal, Canada

[2] School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

[3] University of Pisa, Pisa, Italy

[4] Chuo University, Tokyo, Japan

[5] Insight SFI Research Centre for Data Analytics, School of Information and Communication, University College Dublin, Dublin, Ireland

[6] School of Computing, University of Otago, Dunedin, New Zealand

[7] Alan Turing Institute, London, United Kingdom

[8] University College London, London, United Kingdom

[9] Institute for Experiential AI, Northeastern University, Silicon Valley, USA

[10] Warsaw University of Technology, Warsaw, Poland

[11] Director of Research, Machine Learning, University of Cambridge, Cambridge, United Kingdom

[12] Victoria University of Wellington, Wellington, New Zealand

[13] Computer Science and Engineering Department, Jadavpur University, Kolkata, India

[14] Artificial Intelligence Institute, Tubitak Bilgem, Gebze, Türkiye

[15] Center for Human-Compatible AI, UC Berkeley, Berkeley, USA

[16] Mila - Quebec AI Institute, Montreal, Canada

[17] University of Montreal, Montreal, Canada

Introduction

The Web, and the world beyond it, are about to be swamped by a wave of AI-generated content. AI text generation systems, such as GPT-4 (OpenAI, 2023) Gemini (Google, 2024), Llama (Touvron et al., 2023), Falcon (UAE TII, 2023) and Mixtral (Jiang et al., 2024), are becoming widely used to produce textual content in a variety of domains, such as news (Newsguard, 2024), business reviews (Berry, 2024), academia (Originality, 2024) and culture (Notopoulos, 2024), in an extensive range of languages (see e.g. Fernandes, 2023). AI image generation systems, such as Dall-E (OpenAI, 2021) and MidJourney (Midjourney, Inc., 2022) are producing huge volumes of AI-generated content online (see e.g. Valyaeva, 2023), and are radically changing workflows for human graphic designers (see e.g. HackerNoon, 2023). Images seem likely soon to be followed by AI video generation, such as Sora (OpenAI, 2024).

The widespread adoption of AI content-generation technologies brings many benefits (see Dell'Acqua et al., 2023; Candelon et al., 2023 for balanced reviews). However, this proliferation of AI-generated content also presents significant challenges. As AI generation systems improve, it will become increasingly difficult for human consumers of content to accurately tell whether an item of content was produced by a person or an AI system, or some combination of the two. This poses a brand new authentication problem: as the differences between AI-generated and human-generated content decrease, it becomes intrinsically harder to adjudicate individual cases.

Why do we need to know whether an item was generated by a person or an AI? Importantly, the reasons don't hinge on the *quality* of the content. Human-generated content and AI-generated content can both vary enormously in quality. In the right contexts, both humans and AIs can produce useful, truthful, informative content; in other contexts, both humans and AIs are capable of producing harmful, misleading, inaccurate content. The reasons rather hinge on the role of AI content generation as a *social practice*. Communication between humans through the creation of enduring content (text, images and other media) is fundamental to the ordering of our societies: human-generated content plays a central role in the creation and enforcement of laws, in education and training, in the dissemination of news and opinion, in the organisation of political debates and democratic processes, in the formation and transmission of culture. In all these contexts, societies have developed resilient institutions that allow citizens to have confidence in human-generated content: from educational providers that certify individuals as reputable content providers in specific domains, to laws governing the broadcasting of content and the functioning of political debates, to conventions about the rule of law. AI-generated content *escapes* many of our existing institutions.

AI content generation escapes existing institutions in two main ways. Firstly, it lets people *deliver content they didn't produce*, and maybe don't even understand. In many cases they may not even have seen or read it. In educational settings, this undermines traditional assessment practices, and disrupts current accreditation systems. It also appears to be impacting academic review processes (see Liang et al., 2024). In the professional world, AI content generation undermines the processes through which people and organisations acquire reputations for reliable work. In all these cases, AI threatens breakdowns of social trust. Secondly, AI lets people *proliferate* content. A single person can produce vastly more content than before, including content carefully tailored to specific audiences. This allows individuals to exert new and unprecedented influences on public discussions. The new influences in political discussions are particularly concerning: the recent deepfake of Joe Biden's voice (NBC, 2024) provides a taste of what is now possible. Organisations can similarly increase their capacity to produce content with generative AI, so organisations also have new powers of influence on public discussions. The fact that public discussions increasingly happen online amplifies the effects of these new abilities to proliferate content, and to add coherently to existing content. And AI-generated content is known to have effects in changing consumers' sentiment; see for instance Jakesch et al. (2023).

In short, AI content generation systems can pose serious threats to social stability, and especially to political stability. 2024 will see democratic elections taking place across the globe, so these threats are immediate. To counter these threats, we need to *extend* the institutions that currently govern content creation, to make provisions for generative AI. The crucial extension is to provide methods of *reliably identifying AI-generated content*, and reliably distinguishing it from human-generated content. Finding such methods involves tackling several related questions, which bear on technical and legal mechanisms, but also on economics and company incentives, and on the operation of the open-source ecosystem. In two recent papers (GPAI, 2023; Knott et al., 2023) we reviewed these questions, and argued that the best way to obtain reliable mechanisms for detecting AI-generated content is to place responsibility for the provision of these mechanisms with the organisations (principally companies) that build and deploy generative AI tools. Specifically, we proposed that any agency that creates an AI content generator must be required to *demonstrate a reliable detection mechanism* for the content that generator produces, *as a condition of its use by the public*—and to make the detection mechanism publicly available (as a closed-source

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

tool) on its release. See GPAI (2023); Knott et al. (2023) for details of this proposal. (We will discuss what counts as 'reliable' later in the paper.)

Our proposal, along with some allied efforts we will discuss, had good traction with policymakers in the EU and the US: it was influential in shaping some new legal and organisational directives for generative AI providers. In the second section of this paper, we will review these new directives. In the third section, we take stock of the new landscape for AI-generated content detection which these new directives set up. The directives are certainly not a panacea. Instead, we argue they set the stage for an ongoing 'arms race', between *providers* of AI content detectors (both inside and outside generator companies) and actors who seek to *evade* detection. In this new landscape, we expect that reliable methods for discriminating between AI-generated and natural or human-generated content will sometimes—perhaps often—be available.

This analysis prompts two new sets of questions for policymakers. Firstly, if reliable methods exist for identifying AI-generated content, *who should use these methods?* And *how should they be used?* We consider these questions in the fourth section of the paper, and conclude with some recommendations about new rules for media companies, and perhaps for Web search companies. Secondly, what policy steps can be taken to *intervene* in the arms race between providers and evaders of AI-content identification systems, to ensure that reliable identification methods are widely and frequently available? We consider this question in the fifth section of the paper, and conclude with recommendations about several aspects of the broader information ecosystem.

New imperatives on AI providers regarding AI-generated content identification

Obligations imposed by the EU's AI Act

The EU's AI Act, whose final text has recently been agreed (see e.g. EU/FLI, 2024), explicitly recognises the potential of AI-generated content to destabilise society, and the role AI providers should play to prevent this. As stated in Recital 70a:

> A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem (...) In the light of those impacts, (...) it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods (...)

The Act imposes some clear obligations on providers, which are stated in Article 52.1(a):

> Providers of AI systems, including [General-Purpose 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. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate and prosecute criminal offences.

Four comments are useful here. Firstly, obligations about content detection are only imposed for AI systems that generate substantially new content; systems that make minor changes to existing content are sensibly exempted.

Secondly, obligations are subject to considerations of cost and technical feasibility, and reference is made to certain types of content where technical challenges are higher. (Watermarking is more challenging for textual content than for images, for instance, as discussed by Srinivasan, 2024.)

Thirdly, note that the EU directive only refers to specific detection mechanisms (like watermarking) as *examples* of mechanisms that could function to support detection. The directive itself is rightly more general, accommodating the possibility that detection mechanisms may need to change as technology advances. Note that Recital 70a usefully refers to 'logging methods', which are a promising alternative to watermarking, but have received less attention. In these methods, the provider of the AI generator keeps a private log of content it generates (see Krishna et al., 2023 for the original proposal). A detector for the AI-generated content can then be implemented very simply as a plagiarism detector for content in this log, using mature Information Retrieval

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

technology. Further discussion of possible detection mechanisms, along with their pros and cons, is provided in Knott et al. (2023).[1]

Finally, the mechanisms foreseen for detection include mechanisms for *proving provenance* (at least in Recital 70a). The issue of provenance detection is broader than the issue of AI-generated content detection: several groups have suggested that the problems of AI-generated content are best addressed by a broader protocol that allows *human-generated content* to be positively authenticated. That proposal is particularly associated with the Content Authenticity Initiative and Project Origin, whose efforts are unified in the C2PA standard. The aim is that this standard is adopted throughout the ecosystem for capturing or generating, transforming, transmitting and viewing content. The standard could be adopted by camera manufacturers, for instance, to embed information about when and where a photo or video was recorded, or by broadcasters and other media organisations, to retain this embedded information. Of course these wider obligations don't belong in a piece of legislation about AI—but it is useful that the AI Act mentions the provenance-authentication proposal in a recital accompanying obligations on generative AI providers to support detection. We will consider broader legislation supporting provenance-authentication later in this paper. (For now,, we will use the term 'content *identification*' to encompass both focussed AI-content detection and broader provenance-tracking schemes.)

Guidance from biden's executive order on AI

In the US, President Biden issued an Executive Order 'on the Safe, Secure, and Trustworthy Development and Use of AI' in October last year. This order followed a Senate Judiciary Enquiry on 'Oversight of AI', at which two of our co-authors (Yoshua Bengio and Stuart Russell) gave evidence (alongside Dario Amodei from Anthropic). Much of the conversation at this Enquiry was about AI-generated content identification—and again, the methods discussed included mechanisms focussed specifically on AI-generated content detection tools, and broader protocols for tracking the provenance of all content, whether human- or AI-generated. The Executive Order aims to strengthen public trust in the authenticity of government communications, and more generally, to tackle disinformation. To these ends, it asks for a review of work on AI content detection in Sect. 4.5.(a):

> the Secretary of Commerce (...) shall submit a report (...) identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for (…) (ii) labeling synthetic content, such as using watermarking; (iii) detecting synthetic content (...)

and for guidance about both detection and provenance-authentication in Sect. 4.5.(b):

> the Secretary of Commerce, in coordination with the Director of OMB [the Office of Management and Budget], shall develop guidance regarding the existing tools and practices for digital content authentication and synthetic content detection measures (...)

In Sect. 10.1.(b) (viii)(c), the Director of OMB is additionally tasked with making.

> recommendations to [executive departments and] agencies regarding (…) reasonable steps to watermark or otherwise label output from generative AI[.]

These actions don't impose legal obligations on companies, but they directly impact government procurement processes, and create expectations that may have impacts in civil lawsuits.

Obligations arising from the self-interest of AI providers

Alongside external guidance from policymakers, some new research findings give generative AI providers strong incentives of their own to support the detection of AI-generated content. If an AI generator retrains on the content it produced itself, its quality deteriorates substantially: a phenomenon termed 'model', first reported by Shumailov et al. (2023) and now receiving much attention (see e.g. Dohmatob et al., 2024a, 2024b). AI providers therefore have good reason to exclude AI-generated content from their training sets—and thus have good incentives to be able to identify such content reliably. Note that providers also have separate (positive) incentives to identify text from their own generators, to gauge uptake of their systems, which is a commercially important measure of performance.

Of course, companies may not want to impose a blanket ban on AI-generated training items. There are several situations where AI-generated training items can help address issues in the dataset, such as data scarcity and bias (see e.g. de Wilde et al., 2024), and to augment data quality (for instance by removing noise, normalising, or increasing resolution). These *directed* uses of AI-content can be very beneficial; model collapse arises when the model's training set is *indiscriminately* extended with AI-content.

[1] It is worth noting that *combinations* of *different* detection mechanisms are likely to be particularly effective in delivering reliable detectors. Ensemble techniques for classification are likely to be beneficial here, just as they are elsewhere in machine learning (Zhou et al., 2014). We feel such ensemble methods are not yet widely enough discussed in relation to AI-content detection.

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

Summary

Taken together, the new legal requirements about to be imposed in the EU, the recent guidance from Biden's Executive Order, and recently-recognised considerations of corporate self-interest allow us to confidently anticipate new initiatives from companies in support of AI content detection. The very recent 'Munich accord' in which 20 of the leading tech companies pledge to 'work together to detect and counter harmful AI content' in this year's elections (Munich, 2024) is some testament to this. The implementation and enforcement of these new initiatives will of course be challenging: we will review the main challenges in the next section.

Of the obligations discussed in the current section, we should note that by far the most stringent are those imposed by the EU, which require providers operating in the EU market to support detection mechanisms. As an aside, the largest AI generator companies, which will be centre stage for EU regulators, may sometimes deploy the same generators beyond the EU as within it. For detection methods that are built into generators, this may mean that EU-mandated support for detection will naturally extend to jurisdictions outside the EU. We feel there are good prospects for a 'Brussels effect' in this area, as has been found in other areas of EU tech legislation (Bradford, 2020).

The new adversarial landscape for AI content identification

In the previous section, we reviewed a range of new obligations on providers of AI generators, to support reliable methods for identifying the content their systems generate. These obligations should prompt great improvements in the quality of methods for identifying AI-generated content—especially given the 'Brussels effect' we anticipated above. If the big AI companies fully engage with the goal of creating reliable detectors, we can expect reliable detectors to emerge, which are serviceable in the EU and some way beyond. Note that reliable detectors can also be expected to emerge from time to time even without support from providers. For instance, the recent methods for detecting images generated by stable diffusion (see Wang et al., 2023; Zhang and Xu, 2023) are impressively reliable; recent zero-shot methods for detecting LLM-generated text (e.g. Hans et al., 2024; Su et al., 2023) also show some promise, as do models fine-tuned for specific domains (see e.g. Veselovsky et al., 2023).

Of course, these are just the opening moves in a new, and doubtless ongoing, adversarial process. Any reliable method for AI-content detection, whether supported by providers, or developed externally, will trigger responses from actors who wish to *evade* detection. For detectors that rely on finding differences between AI-generated and 'natural' content, there is an obvious point of attack: as noted by Májovský et al. (2024), any identified difference can immediately serve as an error term to train a new generator that eliminates exactly that difference. Detectors can also be attacked by *manipulating AI-generated content*, so it evades detection. For instance, changing some of the words in a generated text can destroy watermarks added by a generator (see e.g. Sadasivan et al., 2023). Automated tools for modifying images, or paraphrasing texts, can likewise defeat detectors.[2] An early summary of this adversarial landscape is given by Crothers et al., (2023); a more recent summary is provided in a recent report by the Forum for Information and Democracy (FID, 2024 Ch1 Sect. 1.5).

Fortunately, the drafters of the AI Act have anticipated these adversarial responses. Article 52.1(a) requires that AI company support for detection mechanisms be adequate *given 'the generally acknowledged state-of-the-art'*, which should certainly be understood to include known adversarial techniques. The AI Act can therefore be seen as defining providers' obligations in the 'arms race' which is now getting underway between the creators of detector tools (both within generator companies and beyond) and those attempting to evade detection. The picture is complicated by actors who are reluctant to comply with existing rules, or unaware of these rules. The open-source software ecosystem poses some special challenges, both for enforcement of rules and in providing platforms for exploring adversarial strategies (as we will discuss further below). Whenever current methods for identifying AI content are defeated, this will prompt the development of improved methods. It may be at certain points that the evaders have the upper hand, and AI providers must work to find new ways of meeting their obligations. (Again, the AI Act provides for this contingency, by making providers' obligations subject to 'technical feasibility'.) Of course, arms races are nothing new for tech companies: Google has an ongoing battle with search engine optimisers (see e.g. Davis, 2006); social media companies have similar battles with purveyors of harmful content (see e.g. Founta et al., 2019). But it is useful to clearly identify the battle that is newly emerging between providers of AI-content detectors and those aiming to evade detection.

In this new adversarial and dynamic context, we foresee several new questions for policymakers. Firstly, if reliable methods for identifying AI-generated content are available at a given moment, *who should make use of them?* And *how should they be properly used?* We will consider those questions in the next section. Secondly, what can policymakers

[2] Logging methods appear more resilient to paraphrase attacks, however, as reported by Krishna et al. (2023).

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

do to *stack* the arms race in favour of reliable detection mechanisms? We will consider that question in the section after that.

When reliable AI-content identification methods become available, who should *make use* of them?

In this section, we will consider a scenario where reliable methods for identifying AI-generated content are available. In this scenario, policymakers need to determine *who should make use of* these reliable methods, and what constitutes their proper use.

A key consideration for policymakers relates to the *incentives* that ensure the proper use of identification methods within the information ecosystem. We begin by arguing that many organisations in society will naturally adopt reliable methods as they become available, as an organic extension of their existing mechanisms for maintaining reputation and trustworthiness amongst those they interact with. We then consider the case of media organisations. We argue that some of these organisations aren't naturally motivated to adopt systematic AI-generated content identification policies, and hence should be required to do so by law. We consider various ways media companies could moderate the AI-generated content they detect. We conclude by surveying the many risks that arise in the process of identifying and moderating AI-generated content, and consider how policies can balance these against the risks arising from proliferation of AI content.

Free-market incentives to use reliable AI-content identification methods

As we discussed in the first section, AI content generation lets people deliver work that is not their own, that they may have had minimal involvement in, and may not have thoroughly checked. (We are thinking particularly here of AI-generated *text*, where the process of checking or vetting can require a considerable amount of human work.) This creates potential accountability gaps in any organisation where content is to be produced. For instance, in educational institutions, students can deliver work they didn't produce or don't fully understand, which threatens the accreditations these institutions provide. In the professional world, workers can likewise deliver content they didn't produce, and can't fully vouch for, which threatens to undermine the credibility of individuals, and more importantly of whole organisations.

These problems are exacerbated by the tendency of AI generators to 'hallucinate' (see e.g. Rawte et al., 2023). This tendency can be mitigated in various ways (see e.g. Tonmoy et al., 2024), but it is still an inherent feature in systems that are optimised on the surface form of training items, rather than on more abstract measures of meaning. But even disregarding hallucinations, there is a deeper problem: AI content generation potentially lets human providers 'fall out of the loop' in a professional relationship (see e.g. Zerilli et al., 2019). There is no guarantee that services are being provided by the people or companies who are contracted to do the work. Again, this leads to a huge accountability gap.

If reliable ways of identifying AI-generated content become available, we believe the principles that govern competition in free market economies will suffice to lead many institutions to adopt them.[3] Schools and universities will make use of them in certain assessment contexts. Companies that believe that the involvement of human beings has a significant impact on the quality of their output will use them in new vetting procedures. Of course, AI content generators will continue to be *used* in all institutions: they provide a myriad of new productivity-enhancing methods. AI-generated content identifiers will simply be incorporated into institutions' existing methods for creating trust and preserving reputation. For instance, if a student submits work that is identified as AI-generated, the teacher may engage in additional interactions with the student, to check the content is understood; if a professional submits work identified as AI-generated, the assessor may likewise ask further questions. The key idea is simply that AI-generated content must be treated in certain special ways, befitting its origin.

Proposed rules for media companies

As we also discussed in the first section, AI content generation also allows people to *proliferate* content more than was previously possible, allowing content that is untethered from traditional human production processes to flow in large volumes into society. The mechanisms for disseminating content in society can be thought of as the 'media', very broadly speaking, so we believe these organisations have important new roles in deploying reliable AI-generated content identifiers, if these are available. We will consider 'mainstream media' and 'social media' separately. We will also consider Web search companies, which are also involved in disseminating information.

Mainstream media companies

Mainstream media companies include traditional newspapers and radio and TV broadcasters. AI-generated content is finding its way into these venues in various forms: for instance in print articles (see e.g. Farhi, 2023), photos (see

[3] We must of course ensure that identification methods are affordable. We discuss the cost of identification methods later in the paper.

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

e.g. Oremus & Verma, 2023), and even video and audio content (see e.g. Stokel-Walker, 2023).

Mainstream media providers' business models certainly rely on reputation and trust, and we presume most such providers only include AI-generated content unintentionally. These providers certainly have an interest in using reliable AI-generated content identifiers if they are available. But many mainstream media providers are proving to be slow in adapting to the new AI world, and could benefit from guidance. Given that these providers disseminate content in large volumes to the wider public, we suggest they have a moral duty to use reliable content identifiers when these are available—and to use them systematically, so that *all* content they disseminate is checked. If content identifiers are affordable and run automatically, this filter should be minimally intrusive for companies—and would help to preserve their reputation in a world where AI-generated content is proliferating.

In most cases, we think it should be possible for media companies to disseminate AI-generated content, *if this is clearly flagged as such*. A flag would indicate, minimally, that the media outlet is *aware* that the flagged content is AI-generated, and can therefore be expected to have undertaken the kind of actions needed to preserve its reputation as a trustworthy provider. In fact there are some new companies that explicitly position themselves as providers of AI-generated content—in particular for local news: see for example NewsCorp's Data Local (Meade, 2023) and the UK's Radar News. The important thing is that these companies indicate clearly to their consumers that their content is AI-generated. The obligation to treat this content with due caution then falls on those who consume this content.

There may be some types of AI content where stronger obligations are appropriate. For instance, the Paris Charter on AI and Journalism (PAIJ, 2023) takes a stronger line on multimodal content 'mimicking real-world captures and recordings or realistically impersonating actual individuals'. The Charter recommends that outlets should *refrain* from using content of this kind. This proposed policy draws a very clear line between authentically captured content and synthetically created content. We feel that stronger moderation policies may indeed be required for AI content that convincingly appears to have been recorded directly from the world.

If media providers have a moral duty to check for and appropriately moderate AI-generated content, we can ask whether this duty should also be encoded in law. It is likely that different jurisdictions will take different approaches here. For instance, US law places strong emphasis on freedom of the press, while laws in European countries often define conditions on this freedom (see e.g. Tenorio, 2013). But the practical outcomes of press regulation are often more similar across jurisdictions than one might think (see e.g. Heller & van Hoboken, 2019): for instance, child pornography is illegal everywhere. Clearly, the category of AI-generated content would require a much more nuanced moderation policy. Nonetheless, we believe there may be mechanisms in many jurisdictions for encoding rules about AI-generated content, and we recommend policymakers consider such rules.

In relation to existing rules: the EU's AI Act does in fact envisage a 'disclosure obligation' on the publishers of 'AI-generated or manipulated text' (in Recital 70b). This obligation appears to be waived if the AI content 'has undergone a process of human review or editorial control and a natural or legal person holds editorial responsibility for the publication of the content'. We think even in this case, there should be an obligation of some kind (whether legal or ethical) to explicitly flag AI-generated content. This is partly because 'human review' is an imprecise concept: it's hard to know how engaged the human reviewer was in the process, especially if large amounts of AI content are to be reviewed, because of the risk of 'automation bias' (see again Zerilli et al., 2019). But we also feel consumers have a right to know how much AI-generated content they are seeing: in other words, to know what the editorial practices on this matter are, for a given outlet.

Social media companies

Social media companies' business model is different from that of mainstream media companies. They both have incentives to maximise the viewer/user base; but social media companies have less incentive to present themselves as trusted information providers. Famously, under Section. 230 of the US Communications Decency Act, social media companies are not responsible for the content they disseminate: rather, platform users have responsibility for the content they post. Individual users have incentives to disseminate AI-generated content, to increase the volume of content they produce. This could be motivated on financial grounds, to increase revenue from advertising, or simply through a desire to reach a large audience, to promote a political message, for instance. Reputation for individual users in this latter case is less of an issue, because users on social media are somewhat anonymous: it is easy for an individual to create multiple accounts, or to migrate between accounts, even if these practices are discouraged by most platforms. This means that large volumes of AI-generated content are likely to proliferate on social media platforms, as uptake of generators becomes a common public practice.

These considerations again lead us to recommend that social media companies should be *required* to use reliable AI-generated content identifiers when these are available, to systematically vet all content posted on their platforms, and moderate AI-generated content appropriately when it is found. We believe this is a crucial new regulatory

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

requirement, with an important role in preventing the dissemination of content that is unconnected to traditional human production mechanisms, and an important role in extending society's existing mechanisms for regulating human communication into the new domain of AI-generated content.

Web search companies

Another important type of AI-content provider is 'fully AI-generated' websites. These are websites which are set up to cheaply disseminate information, in the interest of attracting users visiting from search engines (see e.g. Ryan-Mosley, 2023). They exist independently on the Web, rather than within a social media platform. The relevant actors for *identifying* AI-generated content in this case are Web search companies.

It is important that search engines deploy any reliable AI identification methods that exist, to systematically look for AI-generated sites, and inform their users of any sites that are found, whether by flagging identified sites or downranking them in search results. We believe that the search engine companies are intrinsically motivated to do this, to retain the trust of their users. In this sense, the free market creates incentives to use AI-content identifiers, as in the cases discussed above. But competition among search engines is not always strong; Google is still the dominant market leader (Oberlo, 2024). So we suggest policymakers should monitor whether free market considerations are sufficient to motivate search companies to make good use of AI content-identification resources. The EU's Digital Markets Act (EU, 2022) should enable this kind of monitoring, at least for search companies operating within the EU.

How should media companies moderate the AI-generated content they identify?

Moderation methods are different for different types of media provider, so we will consider them separately. But we suggest one general rule for all providers: any content that is disseminated (or linked) that is identified as AI-generated should be clearly flagged as such.

Mainstream media companies

For mainstream media companies, the decision to publish a piece of AI-generated content will be taken by a human editor. Editors should certainly be able to run AI-generated content if they choose, as already noted. The key question is how to flag such content when it is published. There are various options to be explored. A textual flag could suffice, provided it is presented prominently enough to alert the consumer. A graphical flag could also be designed, that conventionally denotes AI-generated content: perhaps an image of a robot with a pen.

Social media companies

For social media companies, decisions in relation to AI-generated content fall within the domain of content moderation. Content moderation methods on social media platforms involve many automated classifiers, looking for content of different kinds. Some moderation actions are taken automatically; others are passed to human moderators for final decisions. We recommend that AI-content detectors are incorporated into these moderation processes, to implement the following policy.

In the case where a single individual or group creates multiple accounts ('burner accounts'), that all disseminate AI-generated content pursuing a single goal, we recommend the appropriate moderation action is to remove this coordinated set of accounts altogether. This already seems to be standard policy for several social media platforms, such as Meta (see e.g. Facebook, 2023). Obviously the usual provisions for challenges and transparency should apply in such cases, as they do whenever an account is deleted.

In the case where a single user posts AI-generated content, we suggest the content can always be left in place, provided it does not violate other company policies. But it should again be clearly flagged as AI-generated. For users who are posting large amounts of AI-generated content, for the sole purposes of increasing user engagement and advertising revenue, we suggest a further measure: content from such users should be downranked in platform recommender algorithms, so it disseminates less rapidly than other types of content. The amount of downranking of content from a given user could be a function of the amount of AI-generated content they are posting. (More generally, there could be limits imposed on the volume of AI-content disseminated by the platform as a whole, similar to the limits on the amount of pollution that can be produced by heavy industry.)

In addition to the above moderation policies (or perhaps instead of them), we suggest social media users should have broader agency of their own in relation to AI-generated content. We suggest users should be able to configure settings for their own account so they can opt out of receiving *any* content that has been reliably identified as AI-generated, whatever its source. An alternative measure would be to allow users to *opt in* to receiving AI-generated content, so the default policy is that they receive none. The right choices here will depend on balancing the risks inherent in AI content moderation against those resulting from the unmoderated dissemination of AI content. We discuss how to approach this in the next subsection.

Finally, we suggest that social media companies have certain new obligations in their reports to the general public,

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

if reliable AI content detection methods exist. They should report the overall amount of AI-generated content on their platforms, as part of regular transparency reporting. They should also report fluctuations in this amount, which may be linked to elections or other political events. And they should report the proportion of AI-generated content they removed—as well as the proportion of users who opted in (or out) of receiving AI-generated content, if these options are available. These reports are important in timely identification of risks arising from misinformation.

Web search companies

Web search companies already have mature policies that withhold or downrank content from untrusted providers. We suggest that AI-generated content should feature within these policies. In particular, websites that provide large amounts of AI-generated content, and do not clearly identify this content as AI generated, should be withheld from search results.[4] Websites which occupy the 'borderline' on this criterion should be downranked in the search results. Google's current stated policy is to rank content by quality, without regard for its human or AI origin (see e.g. Schwartz, 2024; Tucker, 2024). But there are likely already penalties for AI content that is presented deceptively as human-generated. If there aren't, we suggest there should be.

In order to have some oversight over policies of this kind, as with social media companies, we also suggest that search companies should be required to report the overall amount of AI-generated content they identify on the Web, as part of their regular transparency reporting. Again, the EU's Digital Markets Act may provide helpful mechanisms of overseeing this reporting.

Communication when AI-content detection is unreliable

In all the above policies, it is important to cater for circumstances when reliable AI-content detection mechanisms are not available. In such contexts, the absence of an 'AI-generated' flag on a piece of content does not positively indicate it is human-generated—and consumers need to know this. We suggest that in such situations, media companies display a general message for users, indicating that normal methods for moderating AI-generated content are not running, or are impaired. This may be presented in some prominent place in a newspaper, or on the user's app screen.

Balancing the risks of AI-content moderation against the risks of AI-content proliferation

In any discussion of automated tools for identifying AI-generated content, it is vital to consider the effects of *errors* in tool performance. We are aiming for 'reliable' tools, but in practice errors will always occur, and they can be harmful. False positives, where human-generated content is wrongly identified as AI-generated, are particularly harmful—at least, in that they create harms to the reputation of individual human generators of content, and may also infringe their rights to free expression, if identification triggers moderation actions. False negatives are also harmful, of course in misleading content consumers. How can these harms be balanced against the risks of unmoderated proliferation of AI-generated content? We suggest the main focus should be on minimising false positives. It will also be important to check for biases in false positives: we do not want to see more false positives for some demographic groups than others. There is clearly a need for discussion between agencies and providers as to what counts as a 'reliable' identification method. In relation to the EU's AI Act, this will likely be decided as a technical standard, rather than in black-letter law, because the appropriate definition is likely to change as technologies advance.

Another important question concerns what stance to take for content that is generated partly by humans and partly by AI. For instance, if a user writes a text then asks GPT to 'tidy it up', we would not want this to be identified as a piece of 'AI-generated content'. It is difficult to identify mixed human-LLM text using a classifier running externally to the provider company (see e.g. Gao et al., 2024). Detection methods that rely on company support have a strong advantage here, because they can make reference to the context in which the content was generated, including (crucially) the prompt history that led to the generated item. For instance, a company can choose to omit the identifying watermark or provenance metadata in cases where the human had a sizeable role in creating the content—or to omit the generated content from the logged content, if a log-based detector is implemented.

A final important consideration in any discussion of content moderation is freedom of speech. As a general rule, moderating content provided by a person infringes their right to freedom of expression if he/she does not give clear consent to the moderator. This is a fundamental human right—though of course, the right to freedom of expression often trades off against other human rights (see e.g. Heyman, 1998). But in the case of AI-generated content,

[4] A more far-reaching idea, which goes beyond the scope of the current paper, is that a cap could be imposed on the amount of AI-generated content a single provider can make available. The idea of capping 'volume' of content has precedents in other areas of regulation—for instance, in the regulation of polluters. A rule of this kind may be useful in addressing wider problems of information overload (see e.g. Holyst et al., 2024). Such a rule could potentially make use of an AI content detection tool—but it might more practically be enforced by restrictions on compute resources allocated to companies (see Sastry et al., 2024 for a relevant proposal).

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

some completely new considerations may arise. If Joe posts a piece of content that was produced (from scratch) *by an AI system*, and this content is moderated, is Joe's right to free expression in any way being curtailed? Ex hypothesi, Joe did not *express* the content. Joe *disseminated* it (by posting it), but he didn't create it. Of course, there are gradations of human involvement in AI content generation, as just discussed: the more involved Joe is in the process, the more rights he has. The act of posting content can likewise involve gradations of human involvement. Nonetheless, the concept of freedom of expression may apply somewhat differently to AI-generated content—arguably removing some of the difficult issues that arise in most content moderation. The strong moderation actions we recommended above for media companies all apply in cases where the human provider is minimally involved, or not involved at all, and particularly if the provider is anonymous.

Support for reliable identification mechanisms in the wider tech world

In the previous section, we asked how reliable methods for identifying AI-generated content should be deployed, if they are available. But as discussed in the section before that, we find ourselves in a new adversarial situation, in which some actors have incentives to defeat the dominant identification methods. In this section, we conclude by considering what policies would help give identification methods the upper hand in this new arms race. Of course, we can learn a lot from long-running arms races in other areas—for instance, relating to search engine optimisation or malicious content detection. In particular, techniques for identifying coordinated malicious efforts (see e.g. Pacheco et al., 2021) will readily extend to AI-fuelled disinformation campaigns. But the AI-content-detection arms race also offers new technical opportunities for interventions, because the adversarial content in this case is all AI-generated. In this section, we review these new opportunities.

Regulation on provenance-authentication protocols

As we noted earlier, requiring the providers of AI content generators to support detection only covers *one* method for identifying AI-generated content. Another method involves establishing broader protocols for provenance authentication, that apply to human-generated content as well as AI-generated content. Through these protocols, trusted providers of content, whether AI-generated or human-generated, can positively identify the content they provide. Content whose provenance is *not* authenticated can then be regarded with more caution, and perhaps moderated accordingly. The details of a workable provenance-authentication scheme still remain to be worked out: implementing such a scheme is a long term project. In particular, it is important to implement a way of authenticating content as produced by an individual person, without disclosing this person's identity. (A system such as that used for German ID cards is one possibility here; see e.g. Poller et al., 2012.)

We also noted earlier that provenance authentication mechanisms require support throughout the information ecosystem, from creation and capture, through transmission and modification, to final display. So if there is to be regulation in this area, it must be separate from regulation focussed narrowly on AI providers. In this section, we will consider possible regulatory actions relating to provenance-authentication.

Our main point is that rules requiring AI providers to support content detection and rules requiring the wider ecosystem to adopt provenance methods should not be seen as alternatives to one another. We see roles for both types of rule. Crucially, neither type of rule provides a failsafe method for the identification of AI-generated content, in the arms race we are embarking on. As we already stressed above, the rules in the AI Act will sometimes be defeated by adversaries, will be flatly ignored by malicious actors, and will not thoroughly permeate the open-source generator ecosystem. A provenance scheme provides a good supplement to detector tools. Conversely, a provenance-authentication scheme is also fallible, and has important limits. For instance, authentication information can often be removed or changed if a piece of content is copied. It will also be difficult to instrument every device that can manipulate content.

As already noted, voluntary schemes for adopting provenance protocols are already beginning to infiltrate the tech world. But widespread adoption is necessary to ensure the success of a provenance scheme. We believe this will only be possible if broader legislation supporting provenance-authentication is enacted. But crucially, this broader legislation should complement legislation requiring providers of AI content generators to support detection mechanisms.

Once again, the EU's AI Act is very well formulated to accommodate provenance authentication schemes. Recital 70a, which states the context for rules on content identification, makes reference to provenance schemes as well as to detection methods. But Article 52.1(a), which states the obligations on AI providers, refers only to support for detection methods. The Act would therefore dovetail well with additional broader rules about provenance authentication. Biden's Executive Order also envisages a division of labour between detection schemes and provenance authentication schemes.

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

Regulation preventing the open-sourcing of 'frontier' AI models

Enforcing regulations on AI systems is harder in the open-source world than for proprietary commercial systems. For instance, as we discussed earlier in the paper, the rule that AI providers must support detection mechanisms is harder to enforce for open-source AI generators than for commercial generators. Copies of open-source generators can proliferate, existing code supporting detection can be modified or removed. Open-source generators are also helpful to actors looking for ways to evade detectors elsewhere in the ecosystem: they provide a platform for exploring evasion methods.

A debate is emerging between groups seeking to promote the practice of open-sourcing generative AI models (such as the AI Alliance) and groups seeking to prevent the practice: see Bommasani et al. (2023) for a good overview. In relation to detection of AI-generated content, we see considerable risks in the practice of open-sourcing generative AI models—especially for the 'frontier' models with the best performance, created by the best-resourced providers. In this sense, we align ourselves with the recent stance of Seger et al. (2023), who argue persuasively that many risks arise from the open-sourcing of these frontier models. We suggest that regulation that prevents the open-sourcing of new frontier models (or in Seger's terms, 'highly capable' AI models) will do a great deal to stack the playing field in favour of reliable AI-content identification mechanisms. (A recent analysis by Kapoor et al., 2024 also summarises risks of open-source foundation models, but is more equivocal in its conclusions.)

Support for applied research in detection mechanisms

In the adversarial climate we sketched above, new or extended detection mechanisms for AI-generated content will always be needed. This research could come from academia or from industry: in either case, there is a good argument that governments should actively support such research. Results from this research should perhaps be kept out of public venues, if this would make it harder for new schemes to be attacked.

Support for compliance with identification schemes

Rules requiring provenance-authentication schemes and rules requiring AI providers to support detection schemes obviously need to be enforced, in jurisdictions where they apply. In these contexts, policymakers also have a role in resourcing compliance and enforcement efforts, and making enforcement as efficient as possible.

As regards compliance, it is vitally important to consider the financial costs of complying with mandated detection or provenance-authentication schemes—especially given the importance of making identification methods available at low costs (which we have already emphasised). We might imagine governments bearing some of these costs—especially for smaller companies, for whom they would be particularly burdensome. At a national level, institutions like the UK's new AI Safety Institute may have a role to play here. International bodies could also have a role; for instance, the EU's newly formed AI Office.

As regards efficiency, there are two useful directions. Firstly, large providers of AI generators which are not providing all possible support for detection tools should be a focus for enforcement. Part of the effort should be to disseminate good information about the best available tools to providers. Providers in the open-source community may be a particular focus here. Secondly, certain links in the information ecosystem have particular roles in attacks on AI-content detection methods. For instance, as we have already discussed, systems that paraphrase text or alter images can be used to evade detection. It is particularly important that these content-modification systems adopt provenance protocols, to provide relevant information to content consumers.

Summary

In this paper, we have sketched the problems that are likely to arise if AI-generated content disseminates into society on a large scale without appropriate checks and balances. We have summarised some recent policy initiatives in the EU and US that address this scenario, by requiring AI providers to support mechanisms that allow reliable identification of AI-generated content. We applaud these new initiatives. They are not a panacea, but we judge that they will apply a consistent impetus on AI providers, to create reliable detection mechanisms. They create a new dynamic context, in which policymakers can consider some new questions.

Our paper considers what new options there are for policymakers in this new dynamic context. Our recommendations are of two types. Firstly, we recommend some new rules about who should *use* reliable AI-content detectors, when these are available, and how they should be used. Our proposals here focus on new obligations for media companies. We make different recommendations for mainstream media companies, social media companies and Web search companies. Secondly, we recommend some new rules that will help create an environment where reliable AI-generated content identification methods exist. We suggest a variety of different rules: rules instituting broad protocols for

Journal : Large 10676 | Article No : 9795 | Pages : 13 | MS Code : 9795 | Dispatch : 15-8-2024

provenance-authentication throughout the digital information ecosystem; rules preventing the open-sourcing of new 'frontier' generative AI models; policies supporting applied research in AI-generated content detection; and policies supporting compliance with identification schemes, including through assistance with costs of compliance.

Data availability No datasets were generated or analysed in the study reported in this paper.

References

Berry, S. (2024). Fake Google restaurant reviews and the implications for consumers and restaurants. PhD dissertation, William Howard Taft University. https://arxiv.org/pdf/2401.11345.pdf

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.

Candelon, F., Krayer, L., Rajendran, S. and Zuluaga Martínez, D. (2023). How People Can Create—and Destroy—Value with Generative AI. BCG Henderson Institute report. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access, 11*, 70977–71002.

Davis, H. (2006). Search engine optimization.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L. Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24–013).

de Wilde, P., Arora, P., Buarque de Lima Neto, F., Chin, Y., Thinyane, M., Stinckwich, S., Fournier-Tombs, E., & Marwala, T. (2024). *Recommendations on the use of synthetic data to trainAI models*. United Nations University Policy Guideline.https://collections.unu.edu/eserv/UNU:9480/Use-of-Synthetic-Data-to-Train-AI-Models.pdf

Dohmatob, E., Feng, Y., & Kempe, J. (2024a). Model Collapse Demystified: The Case of Regression. arXiv preprint arXiv:2402.07712.

Dohmatob, E., Feng, Y., Yang, P., Charton, F., & Kempe, J. (2024b). A Tale of Tails: Model Collapse as a Change of Scaling Laws. arXiv preprint arXiv:2402.07043.

EU (2022). Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)". EUR-Lex.

EU/FLI (2024). EU Artificial Intelligence Act. The Act Texts. Resources provided by the Future of Life Institute. https://artificialintelligenceact.eu/the-act/

Facebook (2023). Account integrity and authentic identity. Facebook Transparency Center. https://transparency.fb.com/en-gb/policies/community-standards/account-integrity-and-authentic-identity/

Farhi, P. (2023). A news site used AI to write articles. It was a journalistic disaster. Washington Post, January 2023. https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/

Fernandes, F (2023). Mapped: Interest in Generative AI by Country. Visual Capitalist blog post. https://www.visualcapitalist.com/cp/mapped-interest-in-generative-ai-by-country/

FID (2024). AI as a Public Good: Ensuring Democratic Control of AI in the Information Space. Report by the Forum for Information and Democracy. https://informationdemocracy.org/2024/02/28/new-report-of-the-forum-more-than-200-policy-recommendations-to-ensure-democratic-control-of-ai/

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019, June). A unified deep learning architecture for abuse detection. In Proceedings of the 10th ACM conference on web science (pp. 105–114).

Gao, C., Chen, D., Zhang, Q., Huang, Y., Wan, Y., & Sun, L. (2024). LLM-as-a-coauthor: The challenges of detecting LLM-human mixcase. arXiv preprint arXiv:2401.05952.

Google (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

HackerNoon (2023). AI Design Tools That are Changing How Graphic Designers Work. https://hackernoon.com/ai-design-tools-that-are-derailing-how-graphic-designers-work

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. arXiv preprint arXiv:2401.12070.

Heller, B., & van Hoboken, J. (2019). Freedom of expression: A comparative summary of United States and European law. *Available at SSRN 4563882*. https://doi.org/10.2139/ssrn.4563882

Heyman, S. J. (1998). Righting the balance: An inquiry into the foundations and limits of freedom of expression. *BUL Rev, 78*, 1275.

Hołyst, J. A., Mayr, P., Thelwall, M., Frommholz, I., Havlin, S., Sela, A., & Sienkiewicz, J. (2024). Protect our environment from information overload. *Nature Human Behaviour, 8*, 402–403.

Jakesch, Maurice, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–15. Hamburg, Germany: ACM.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., & Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*. https://doi.org/10.48550/arXiv.2401.04088

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., Song, D., Storchan, V., Zhang, D., Ho, D., Liang, P., Narayanan, A. (2024). On the Societal Impact of Open Foundation Models. Stanford University Center for Research on Foundation Models. https://crfm.stanford.edu/open-fms/paper.pdf

Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Eyers, D., Trotman, A., Teal, P. D., Biecek, P., 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.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2303.13408

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., ... & Zou, J. Y. (2024). Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. arXiv preprint arXiv:2403.07183.

Májovský, M., Černý, M., Netuka, D., & Mikolov, T. (2024). Perfect detection of computer-generated text faces fundamental challenges. *Cell Reports Physical Science, 5*(1), 101769.

Meade, C. (2023). News Corp using AI to produce 3,000 Australian local news stories a week. The Guardian, July 2023. https://www.theguardian.com/media/2023/aug/01/news-corp-ai-chat-gpt-stories

Munich (2024). Tech Accord to Combat Deceptive Use of AI in 2024 Elections. Pledge made at the Munich Security Conference, February 2024. https://securityconference.org/en/aielectionsaccord/

NBC (2024). Fake Joe Biden robocall tells New Hampshire Democrats not to vote Tuesday. NBC News. https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984.

Newsguard (2024). Tracking AI-enabled Misinformation: 702 'Unreliable AI-Generated News' Websites (and Counting). https://www.newsguardtech.com/special-reports/ai-tracking-center/

Notopoulos, K. (2024). Women laughing alone with AI-generated content spam. Business Insider https://www.businessinsider.com/the-hairpin-blog-ai-spam-content-farm-cybersquatting-2024-1

Oberlo (2024). Search Engine Market Share in 2024. https://www.oberlo.com/statistics/search-engine-market-share

OpenAI (2021). DALL·E: creating images from text. Retrieved from https:// openai.com/research/dall-e (accessed 19 March 2024).

OpenAI. (2023). GPT-4: Scaling up deep learning. Retrieved from https://openai.com/research/gpt-4

OpenAI. (2024). Sora: Creating video from text. Retrieved from https://openai.com/sora

Oremus, W and Verma, P. These look like prizewinning photos. They're AI fakes. Washington Post, November 2023. https://www.washingtonpost.com/technology/2023/11/23/stock-photos-ai-images-controversy/

Originality (2024). Ai-generated Research Papers Published On Arxiv Post Chatgpt Launch. Originality.AI blog post. https://originality.ai/blog/ai-generated-research-papers

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 455–466.

PAIJ (2023). Paris Charter on AI and Journalism. https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20Charter%20on%20AI%20and%20Journalism.pdf

Poller, A., Waldmann, U., Vowé, S., & Türpe, S. (2012). Electronic identity cards for user authentication-promise and practice. *IEEE Security & Privacy Magazine, 10*(1), 46–54.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv: 2309.05922*. https://doi.org/10.48550/arXiv.2309.05922

Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, Percy Liang (2023). Considerations for Governing Open Foundation Models. Stanford University Center for Research on Foundation Models.

Ryan-Mosley, T. (2023). Junk websites filled with AI-generated text are pulling in money from programmatic ads. MIT Technology Review. https://www.technologyreview.com/2023/06/26/1075504/junk-websites-filled-with-ai-generated-text-are-pulling-in-money-from-programmatic-ads/

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv preprint arXiv: 2303.11156*. https://doi.org/10.48550/arXiv.2303.11156

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., & Coyle, D. (2024). Computing power and the governance of artificial intelligence. *arXiv preprint arXiv: 2402.08797*. https://doi.org/10.48550/arXiv.2402.08797

Schwartz, B. (2024). Google Responds To Claims Of Google News Boosting Garbage AI Content. Search Engine Roundtable, Jan 2024. https://www.seroundtable.com/google-responds-garbage-ai-content-in-google-news-36757.html

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.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv: 2305.17493*. https://doi.org/10.48550/arXiv.2305.17493

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/

Stokel-Walker, C. (2023). TV channels are using AI-generated presenters to read the news. The question is, will we trust them? BBC News, January 2024. https://www.bbc.com/future/article/20240126-ai-news-anchors-why-audiences-might-find-digitally-generated-tv-presenters-hard-to-trust

Su, J., Zhuo, T. Y., Wang, D., & Nakov, P. (2023). DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. arXiv preprint arXiv:2306.05540.

Tenorio, P. (2013). Freedom of Communication in the US and Europe. *ICL Journal, 7*(2), 150–173.

UAE TII. Falcon-180b: A 180 billion token language model. https: // huggingface.co/tiiuae/falcon-180B, 2023.

Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Tucker, E. (2024). New ways we're tackling spammy, low-quality content on Search. Google blog post, March 2024. https://blog.google/products/search/google-search-update-march-2024/

Valyaeva, I (2023). AI Has Already Created As Many Images As Photographers Have Taken in 150 Years. Statistics for 2023. EveryPixel Journal. https://journal.everypixel.com/ai-image-statistics

Veselovsky, V., Ribeiro, M. H., & West, R. (2023). Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv preprint arXiv:2306.07899.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hezhen, Hu., Chen, H., & Li, H. (2023). DIRE for diffusion-generated image detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023*, 22445–22455.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic DECISION-MAKING AND THE CONTROL PROBLEM. *Minds and Machines, 29*, 555–578.

Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. arXiv preprint arXiv:2312.02625, 2023.

Zhou, Z. H. (2014). Ensemble methods. *Combining pattern classifiers* (pp. 186–229). Wiley.