Ethics and Information Technology (2023) 25:55
https://doi.org/10.1007/s10676-023-09728-4

ORIGINAL PAPER

Generative AI models should include detection mechanisms as a condition for public release

Alistair Knott[1] · Dino Pedreschi[2] · Raja Chatila[3] · Tapabrata Chakraborti[4] · Susan Leavy[5] · Ricardo Baeza-Yates[6] · David Eyers[7] · Andrew Trotman[7] · Paul D. Teal[1] · Przemyslaw Biecek[8] · Stuart Russell[9] · Yoshua Bengio[10]

Accepted: 11 October 2023 / Published online: 28 October 2023

Abstract

The new wave of 'foundation models'—general-purpose generative AI models, for production of text (e.g., ChatGPT) or images (e.g., MidJourney)—represent a dramatic advance in the state of the art for AI. But their use also introduces a range of new risks, which has prompted an ongoing conversation about possible regulatory mechanisms. Here we propose a specific principle that should be incorporated into legislation: that any organization developing a foundation model intended for public use must demonstrate a reliable *detection mechanism* for the content it generates, as a condition of its public release. The detection mechanism should be made publicly available in a tool that allows users to query, for an arbitrary item of content, whether the item was generated (wholly or partly) by the model. In this paper, we argue that this requirement is technically feasible and would play an important role in reducing certain risks from new AI models in many domains. We also outline a number of options for the tool's design, and summarize a number of points where further input from policymakers and researchers would be required.

Keywords Generative AI · AI regulation · AI ethics · AI social impacts · Foundation models

A new content authentication problem, and a proposed solution

The new class of generative AI models, sometimes termed 'foundation models'[1] (FMs), have achieved dramatic advances in AI (Bommasani et al., 2022). Foundation models are trained on very large, domain-general datasets; after training, they have amazing abilities to generate content of the kind they were trained on. For instance, ChatGPT can generate humanlike text and dialogue contributions; MidJourney can generate realistic images. While earlier AI systems were able to generate small amounts of content (for instance, suggesting spelling or style changes to an existing text, or making alterations to images), foundation models can generate high-quality content from scratch, from minimal prompts.

Foundation models also introduce a range of new risks (see again Bommasani et al., 2022). Policymakers and AI researchers are engaged in very active discussions about these risks, and the regulatory measures that might

[1] By 'foundation models', we mean 'systems that use foundation models'. The term 'model' has been adopted in common use, so we use it here.

✉ Alistair Knott
ali.knott@vuw.ac.nz

1 School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

2 Department of Computer Science, University of Pisa, Pisa, Italy

3 Sorbonne University, Paris, France

4 The Alan Turing Institute, University College London, London, UK

5 School of Information and Communication Studies, University College Dublin, Dublin, Ireland

6 Institute for Experiential AI, Northeastern University, Boston, USA

7 School of Computing, University of Otago, Dunedin, New Zealand

8 Warsaw University of Technology, Warsaw, Poland

9 UC Berkeley, Berkeley, USA

10 Mila – Quebec AI Institute, Université de Montréal, Montreal, Canada

practically manage them (see Hurst, 2023 for a recent survey). In this paper, we focus on one key risk, concerning the *provenance* of FM-generated content. Texts or images created by FMs can now readily pass as human-generated (see e.g., Jakesh et al., 2023; Waltzer et al., 2023). As FM-generated content begins to flood the Web and the Apps ecosystem, human consumers of content will be faced with a brand new *authentication problem*: determining whether a given item they encounter was produced by a person or a machine.[2]

Why is it important to know this? Emphatically not because human-produced content is always 'better' than FM-generated content: this is certainly not the case (Bubeck et al., 2023; Singhal et al., 2023). It is rather that human and FM-generated content need to be *assessed* very differently, because of their very different origin. Consider a piece of text, encountered by a human reader. In many contexts, her assessment of the text will run very differently if she knows it was generated by an AI system. If she is a teacher assessing a piece of student work, she may want to know how engaged the student has been with the text: have they read it closely, has its content been assimilated? How much learning has taken place? If she is an employer assessing a contractor's report, she may want to know how carefully the provider has overseen its generation: how much work has the contractor done in producing the report? If she is assessing the text as a content moderator working in a social media company, she may want to know whether it is part of a larger-scale communication campaign, given that FMs can readily generate personalized communications at scale, including harmful disinformation (e.g., Newsguard, 2023; Tamkin et al., 2021). If she is a citizen receiving the text as professional advice from her doctor or lawyer, she may want to know how thoroughly it has been checked for errors, given the known problems of errors in FM output (e.g., Ji et al., 2023) and overreliance on FM output by human operators (e.g., Wang et al., 2023).[3] In each case, the human assessor *needs to know* whether the text is human- or AI-generated, in order to make a proper assessment. The reasons for this need vary greatly between domains. In professional interactions they are about ensuring accuracy; in education they are about ensuring effective student assessment; in social media contexts they are about ensuring a safe Internet. One might argue that human consumers have a general 'right to know' whether the content they encounter was produced by a person or a machine. In fact we argued for this position in a previous paper (GPAI, 2023). But a more pragmatic argument can also be advanced, that in specific domains and contexts, consumers have specific needs to know about the origin of the content they assess, which justify expenses incurred by the producer in providing this information. This is the kind of argument that justifies laws about labeling the ingredients in food: consumers have no universal right to know what is in the food they eat, but in specific products and sale contexts, their need justifies rules requiring some information to be given (see Messer et al, 2017 for an overview of relevant consumer law).

There are already many actual or proposed laws that require purveyors of AI-generated content to identify it as such. For instance, Article 52.1 of the AI Act being developed by the EU (EU, 2021) requires that AI systems interacting directly with users are clearly identified as AI systems; California's BOT Act already in force (SB1001, 2018) makes similar requirements in specific commercial and political use cases. But these laws do not meet the case we are considering, which is where AI-generated content is disseminated *beyond* the interactive tool through which it was generated, and consumers encounter it 'indirectly', in some arbitrary new online or offline context. Some laws cover this dissemination process, by placing obligations on the disseminator. For instance, the EU's proposed AI Act (Article 52.3) places obligations on people who disseminate one specific type of AI-generated content ('deep fakes') to label this content as AI-generated. This is a useful measure—but consumers cannot rely on disseminators of AI content doing the right thing, even if it is required by law. Regulation must also cater for disseminators who do not disclose the AI origin of the content they spread. We argue that *consumers* should have the ability to determine whether some arbitrary item they see was generated (wholly or partly) by FMs.

The only way we see to meet this consumer need at present is with a tool that allows *automatic detection* of FM-generated content. In the tool we envisage, the user uploads an arbitrary piece of online content, and the tool responds with an analysis of its human or machine provenance.[4] We will discuss this analysis below—for now, our argument is that to help keep FM content generators safe, consumers need access to *another AI tool*, for the *detection* of FM-generated content.[5]

[2] We use the term 'consumers' on occasion in this paper because items of AI-generated content can be thought of as manufactured products as well as as instruments of communication. As yet there are no well-defined terms covering both of these scenarios.

[3] Citizens also have rights to avoid being the subject of fully automated decision-making systems; the EU's GDPR (EU 2016, Article 22) is a case in point.

[4] Content will have to be of a certain size or complexity for the tool to work, as we discuss later.

[5] We will describe the tool as a 'detector' rather than a 'classifier', because it should be able to identify *portions* of an item that are FM-generated, if the item is big enough, rather than just pronounce about the item as a whole. The word 'detector' also captures the function of the mechanism in the large, when deployed by many users on large numbers of content items.

A detection tool for FM-generated content would be valuable for companies that supply content to consumers, as well as for consumers themselves. Keeping social media platforms safe from large-scale disinformation campaigns is a pressing issue which poses considerable threats to democratic processes. A reliable detection tool for FM-generated content could be used by social media companies to detect and defuse such campaigns. The remainder of this paper is concerned with mechanisms that ensure that a detection tool of this kind can be made reliable.

A high-level proposal for legislation, and some questions for discussion

There are many tools that attempt to distinguish AI-generated from non-AI-generated content, both for text (e.g., Chaka, 2023) and images (e.g., Stroebel et al., 2023). But as FM generators improve, the ability of detectors to identify FM-generated content *purely from an analysis of the content* is likely to diminish rapidly (see e.g., Thompson & Hsu, 2023). Text generators are producing increasingly humanlike text, and image generators are producing increasingly realistic images: as generators get better at generalizing from their training inputs, the patterns that distinguish FM-generated content from authentic content necessarily become harder to identify. A consensus is emerging that the only way to create a reliable detector for FM-generated content, as generators improve, is to *instrument the generator* in some way, to *support* detection (see e.g., Kirchenbauer et al., 2023a; Tulchinskii et al., 2023). This 'instrumentation' might involve placing hidden patterns or 'watermarks' inside generated content that a detector can identify. But there are other methods too; we will review several options below. For now, the key point is that if reliable detection mechanisms require generators that are configured to support detection, then *responsibility for workable detection mechanisms* ultimately rests with the organizations that build the generators.

We suggest that legislation should recognise this responsibility. Specifically, we propose that any organization that develops a LLM intended for public use should be required by law to demonstrate a reliable detection tool for the content the model generates, as a condition for its release to the public. After release, the detection tool should be freely available to the public.

We made this proposal in an earlier paper (GPAI, 2023),[6] and it has stimulated considerable discussion amongst AI researchers and policymakers. In the remainder of the current paper, we will summarize the main issues that have arisen in this discussion, and our initial thoughts on these. Our focus is on the high-level policy questions that should be resolved before any detailed legislation is drafted.

What generative models are in scope for the proposed rule?

Firstly, our proposal relates only to FMs, not to simpler AI content generation systems, used e.g., for spell checking and image manipulation. (FMs can be used for these tasks too, but it is their ability to produce content de novo that necessitates the proposed rule). Second, our proposal only applies to FMs 'intended for public use'. (FMs developed for a client company, whose content will only be seen within that company, are not in scope, because they do not create the authentication problems we are concerned about.) Third, our proposal does not place obligations on systems that operate 'downstream' of a FM, that use prompts to configure it for a particular task or purpose. (The detection tool for the 'upstream' FM will continue to work for the downstream system's output in these cases.) We are seeing an explosion of systems using FM technology at present (McKinsey, 2023a). But the vast majority of these systems are downstream users of a relatively small number of upstream FMs (McKinsey, 2023b). Our proposal is for the regulation of the upstream systems: a much more manageable prospect.

Some questions about scope remain, however. Should our proposed rule only apply to new generators not yet released, or should it also apply retrospectively to generators already in use? A definition of 'public use' is also needed. Our main intention is to cover generators that are or will be presented to the public as products or services, or as components of products or services: that is, we envisage a scope similar to that envisaged by the EU's proposed AI Act (EU, 2021). (It is also important to cover private generators whose output is intended for public consumption.) But other scopes could also be envisaged. Whether the proposed rule also applies to publicly accessible code repositories, such as code made available on GitHub or Hugging Face, is also a matter for discussion. On this latter question, the wider question of enforcement for open-source AI generators is also relevant,

[6] The authors participate in a project on Social Media Governance run by the Responsible AI Working Group at the Global Partnership on AI (GPAI). The proposal in the current paper differs in some detail from the proposal in our first paper (GPAI, 2023): the current paper reflects our current view. Both papers reflect the personal opinions of the authors and do not necessarily reflect the views of GPAI as a whole, or of its members.

as we discuss below. A final question concerns how complex or realistic a generator needs to be before our rule applies. We suggest realism is a more appropriate criterion than complexity, given the possibility of distilling smaller models from large ones (Hinton et al., 2015). Naturally, the most realistic generators will be the ones most used by the public, so a definition focussing on public use may be sufficient here.

Possible detection methods

There are several ways of instrumenting an AI content generator to support detection. One is to include watermarks in the generated content. This method has been demonstrated for text and image generators (see, e.g., Kirchenbauer et al., 2023a, 2023b; Zhao et al., 2023). Other methods involve exploiting statistical features of FM-generated content (see, e.g., Mitchell et al., 2023 for a method operating on text content). A final method, which we feel needs more attention, is for the producer organization simply to keep a (private) *log* of all the content it generates—a detector tool can then be implemented as a regular plagiarism detector operating on this log. This method was recently demonstrated for text by Krishna et al. (2023). A plagiarism detector is essentially an information retrieval (IR) device: the companies at the forefront of FM content generation also have huge expertise in this area, and would be very well placed to provide detectors of this type. Other better methods may well be discovered as research advances. To future-proof legislation, it should avoid mention of particular methods, and simply require 'a reliable detection mechanism'.

The detector's response format

What information would the detector return, when given an input document? As a concrete basis for discussion: for textual input, we currently envisage an analysis similar to that given by plagiarism detectors such as TurnItIn (TurnItIn, 2021). For a short text, the tool returns a probability (with some confidence interval) that it was generated by a FM. It may refrain from any output for very short texts, where confidence is necessarily low. For a longer text, it might identify specific segments that have some super-threshold probability of being FM-generated, again with confidence intervals. (Current commercial detectors such as GPTZero and open-source detectors such as GLTR have some of this functionality.) In cases where small FM-generated 'suggestions' are interleaved throughout a document, we envisage the tool should treat the text as human-generated if these are sparse, and AI-generated if they are dense. Images can similarly be analyzed as wholes or by parts. (FM generators can be asked to produce a specified region of an image, and humans can also post-process certain parts of an image.)

Aggregation of detectors in a user-facing tool

In the proposed rule, an organization providing an FM generation system must make available a detector for content produced *by that system*. Users obviously need a tool that calls detectors for *all* generation systems in common public use, and aggregates their responses. Clearly, an aggregator can only target the most commonly used generators, if it is to be practical. But the market share for generators is likely to be heavily skewed towards a few 'winning' systems at any time (see Hefti & Lareida, 2021 for a recent analysis), so a focus on commonly used generators will still provide reasonable coverage. Who should provide this aggregator? There are various possibilities. It could be a commercial company, or a non-profit organization (academia, user group), or an international regulator of some kind. It might also be the FM-generation companies themselves. Note these companies have their own pressing commercial needs for a tool detecting FM-generated content, so they can avoid the 'model collapse' that may potentially occur when a content generator is iteratively retrained on its own output (see Shumailov et al., 2023).

Resistance to adversarial attacks

Any detector tool will naturally be attacked by people seeking to evade detection. For texts, the most commonly discussed attack method at present is by passing the text through an automated 'paraphrasing' system, which changes its form but retains its meaning. Sadasivan et al. (2023) note this method is quite effective against watermarking schemes. (Other methods for evading watermarking schemes are discussed by Jiang et al., 2023; Shi et al., 2023.) Krishna et al.'s logging scheme appears more resistant to paraphrase attacks. But here too, we should anticipate effective attacks in due course. An arms race will naturally play out between detection methods and evasion methods, whether or not detection methods are mandated by law. If there is a law, as we propose, it should require a detector that is reliable 'in the current adversarial context', whatever that is. As evasion methods mature, it may be that detection methods require broader systems for guaranteeing the provenance of content: for instance, agreements to track and share the provenance of identifiable source material through, and onwards from paraphrasing products. (These systems could also provide methods for authenticating the human origin of content.) Organizations would have to collaborate in developing systems of this kind. (Again, given companies' shared interest in workable detection systems to prevent 'model collapse', such collaboration is likely a viable proposition.) Crucially, it would be for the agency developing a new FM generator to demonstrate a detection method that is effective in the current adversarial context, and show its practicality, either

unilaterally, or in collaboration with other groups. Naturally, each new detection method will elicit new attacks: so our proposed rule will not lead to a perfect detection system for consumers. But it will help to keep consumers safe.

Cost of providing a detection tool

A detection tool has a certain cost, both in its development and in its deployment to users. But we should note that AI-generated content detection is emerging as a commercial field in its own right (see e.g., Marshall, 2023). While companies would provide their detector free of charge to users in our proposed scheme, they could likely generate revenue through advertising. Smaller companies should be able to build on open-source detector tools, which will help limit costs. State agencies could also fund research on detection tools, which then could be made available to companies; arguably states have some responsibility in providing AI safety 'infrastructure' of this kind, especially if they enact rules that require such infrastructure. When considering cost, it is also important to bear in mind the cost of *not* having a reliable detection tool, both on individual users in specific domains (e.g., the additional costs for teachers, in checking for AI-generated work) and more general on society (e.g., the destabilization of democracies through AI-generated disinformation).

What counts as a reliable tool?

Any detector tool can be expected to make errors, both false positives (identification of human-generated text as AI-generated) and false negatives (identification of AI-generated texts as human-generated). Decisions will have to be made as to what level of these errors is acceptable. These decisions should be part of the interpretation of the law, rather than the law itself, as they may also change as technology and adversarial methods advance. But the basic evaluation principle can be clearly stated: a classifier's performance must be tested on a sample of AI-generated and human-generated content unseen during its training.

Enforcement for open-source generator models

Providers of open-source FMs would also have to comply with our proposed rule, and to supply detector mechanisms for the content their models produce. But enforcing this compliance is likely to be harder for open-source providers than for other providers, because versions of open-source software can proliferate more readily. Nonetheless, there is some useful structure to this proliferation. Within the open-source world, the vast majority of FMs are built as modifications of a small set of high-profile core models (see e.g., Gao & Gao, 2023, for evidence from Hugging Face's language model collection). If the core models comply when first released, and include licenses that require compliance to be maintained, this should provide some support for compliance in the open-source ecosystem. It may also be possible to make the compliance code hard to remove—for instance, by 'obfuscating' it (see Goldwasser & Rotblum, 2007 and subsequent work). Independently of this, any open-source generators that attract a large user base will necessarily become visible to enforcement agencies. But generators used by smaller groups (for instance, state-sponsored bad actors) are likely to be harder to find. Of course actors of this kind won't comply with our proposed law, and regular policing methods for identifying the origin of illegal content will have to be used.

Current initiatives by companies and legislators

Several of the large AI companies have recently announced an initiative to include watermarks in AI-generated audio and visual content (see White House, 2023). This is a good initiative, but it is some way from the scheme we are proposing. For one thing, our proposal extends to FM-generated text as well as audio and visual content. But more importantly, our suggested rule makes reference to an *objective*—a reliable detection tool—rather than to a specific mechanism such as watermarking. On the legislation front, the EU Parliament has made some reference to our proposal in the amendments it recently agreed to its proposed AI Act catering specifically for FMs (EU, 2023). An amendment to Recital 60 g states that generative foundation models 'should ensure transparency about the fact the content [they produce] is generated by an AI system, not by humans'. This amendment is pushing in the right direction. But again, we suggest this requirement should be stated more precisely, by making reference to a workable detection tool. And the intention behind the recital should also be fully reflected in the Act's Articles—most likely in Article 28b (obligations on distributors) and/or Article 52 (transparency obligations).

We look forward to a productive discussion with legislators, companies and other stakeholders about these open questions.

Acknowledgements We are grateful to the Global Partnership on AI (GPAI) and the Montreal International Center of Expertise in Artificial Intelligence (CEIMIA) for supporting this project. Yoshua Bengio thanks CIFAR and NSERC; Stuart Russell thanks Open Philanthropy Foundation for support of the Center for Human-Compatible AI at UC Berkeley. Dino Pedreschi has been supported by the European Commission under the NextGeneration Programme PE0013 - 'FAIR - Future Artificial Intelligence Research' Spoke 1, 'Human-Centered AI', and under the EU H2020 ICT-48 Network of Excellence n.952026 'Human-AI net'. We would like to thank colleagues at the EU (Lucilla Sioli's group at DG-CNECT) and the OECD (Sebastian Hallensleben's group) for useful discussions, and Marcin Betkier and Rebecca Downes for comments on an earlier draft of this article. We also thank the two

anonymous reviewers for their useful comments. The views expressed here, and any remaining errors, are of course our own.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, MS., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson E., & Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, p., Tat Lee, Y., Li, Y., Lundberg, S., Nori, H., Palangi, H., Tulio Ribeiro, M., Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*. https://doi.org/10.37074/jalt.2023.6.2.12

EU (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

EU (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. 21 April 2021 (original proposed Act). https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

EU (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065&qid=1666857835014

EU (2023). Artificial Intelligence Act: Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

Gao, S. and Gao, A. (2023). On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models. https://arxiv.org/pdf/2307.09793.pdf

Goldwasser, S., & Rothblum, G. N. (2007). On best-possible obfuscation. In Theory of Cryptography: 4th Theory of Cryptography Conference, TCC 2007, Proceedings 4 (pp. 194–213). Springer

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

Hefti, A., & Lareida, J. (2021). Competitive attention, superstars and the long tail. University of Zurich, Department of Economics, Working Paper, (383).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

White House (2023). Voluntary AI Commitments. Joint statement from Seven Leading AI Companies released by the White House. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

Hurst, A. (2023). How generative AI regulation is shaping up around the world. Information Age, July 2023. https://www.information-age.com/how-generative-ai-regulation-shaping-up-around-world-123503911/

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading watermark based detection of AI-Generated content. arXiv preprint arXiv:2305.03807.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., Goldstein, T. (2023). On the reliability of watermarks for large language models. arXiv preprint arXiv:2306.04634.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. arXiv preprint arXiv:2301.10226.

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

Marshall, J. (2023). As AI cheating booms, so does the industry detecting it: 'We couldn't keep up with demand'. The Guardian, July 2023. https://www.theguardian.com/technology/2023/jul/05/as-ai-cheating-booms-so-does-the-industry-detecting-it-we-couldnt-keep-up-with-demand.

McKinsey (2023a). The state of AI in 2023. Generative AI's breakout year. McKinsey report. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

McKinsey (2023b). Exploring opportunities in the generative AI value chain. McKinsey report. https://www.mckinsey.com/capabilities/quantumblack/our-insights/exploring-opportunities-in-the-generative-ai-value-chain.

Messer, K. D., Costanigro, M., & Kaiser, H. M. (2017). Labeling food processes: The good, the bad and the ugly. *Applied Economic Perspectives and Policy, 39*(3), 407–427.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. arXiv preprint arXiv:2302.08500.

Newsguard (2023). Despite OpenAI's promises, the company's new AI tool produces misinformation more frequently, and more persuasively, than its predecessor. https://www.newsguardtech.com/misinformation-monitor/march-2023/

OpenAI (2015). Introducing OpenAI. https://openai.com/blog/introducing-openai.

OpenAI (2022). Introducing ChatGPT. OpenAI blog post. https://openai.com/blog/chatgpt.

OpenAI (2023a). GPT-4 Technical Report. arXiv:2303.08774v2.

OpenAI (2023b). GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI blog post. https://openai.com/product/gpt-4.

OpenAI (2023c). Our approach to AI safety. OpenAI blog post. https://openai.com/blog/our-approach-to-ai-safety.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? arXiv preprint arXiv:2303.11156.

SB1001 (2018). Bolstering Online Transparency ('BOT') Act. California legislation. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K. W., & Hsieh, C. J. (2023). Red teaming language model detectors with language models. arXiv preprint arXiv:2305.19713.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

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

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.

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

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., Burnaev, E. (2023). Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. arXiv preprint arXiv:2306.04723.

TurnItIn (2021). Understanding the Turnitin Similarity Report. https://help.turnitin.com/Resources/PDF/understanding_the_turnitin_similarity_report-a_student_guide.pdf

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies*. https://doi.org/10.1155/2023/1923981

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. *Journal of Medical Internet Research, 25*, e48009.

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N. M., & Lin, M. (2023). A recipe for watermarking diffusion models. arXiv preprint arXiv:2303.10137