State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release

July 2023

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

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

Acknowledgements

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

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

The report was written by: **Alistair Knott*,** School of Engineering and Computer Science, Victoria University of Wellington**; Dino Pedreschi*,** Department of Computer Science, University of Pisa**; Raja Chatila*,** Sorbonne University**; Susan Leavy*,** School of Information and Communication Studies**,** University College Dublin; **Ricardo Baeza-Yates*,** Institute for Experiential AI, Northeastern University; **Tapabrata Chakraborti†,** University of Oxford and the Alan Turing Institute**; David Eyers†,** Department of Computer Science, University of Otago**; Andrew Trotman†,** Department of Computer Science, University of Otago; **Lama Saouma†,** GPAI's Montreal Center of Expertise - CEIMIA**; Virginia Morini†,** Istituto di Scienza e Tecnologie dell'Informazione, NIRC; **Valentina Pansanella†,** Scuola Normale Superiore, University of Pisa**; Paul D. Teal†,** School of Engineering and Computer Science, Victoria University of Wellington**; Przemyslaw Biecek*,** Warsaw University of Technology**; Ivan Bratko*,** University of Ljubljana**; Stuart Russell*,** UC Berkeley; and **Yoshua Bengio†,** Université de Montréal and Mila – Quebec AI Institute.

We thank Colin Gavaghan† from the University of Bristol Law School and Bristol Digital Futures Institute, and Jordan MacLachlan† and Andrew Lensen† from the School of Engineering and Computer Science, Victoria University of Wellington for comments on drafts of this article. All remaining errors are our own. GPAI would like to acknowledge the efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Responsible AI Working Group, and the dedication of the Working Group Co-Chairs, **Catherine Régis** and **Raja Chatila**.

* Expert of GPAI's Responsible AI Working Group
** Observer at GPAI's Responsible AI Working Group
† Invited specialist
‡ Contracted parties by the CofEs to contribute to projects

Citation

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.

CONTENT

Abstract 4

I. Summary of the proposed pre-release requirement 5

II. Precedent for a pre-release process for foundation models 8

III. How does a pre-release detection mechanism fit with current or imminent legislation? 9

IV. Is a pre-release detection requirement feasible in the current competitive environment? 12

V. What might the details of the pre-release detection rule look like? 13

References 14

Abstract

The new generation of general-purpose ‘foundation AI models’ such as ChatGPT and MidJourney are dramatically more powerful and useful than earlier AI systems. But their use also introduces a range of new risks, which have prompted an ongoing conversation about possible regulatory mechanisms. This paper contributes to this conversation. We propose a specific principle that should be incorporated into legislation—namely, that any organisation developing a new, state-of-the-art foundation model must demonstrate a reliable *detection mechanism* for the content that model produces, as a condition of release. The detection mechanism should be made publicly available in a tool that allows consumers to query, for an arbitrary item of content, whether the item was generated (wholly or partly) by the model. We argue this requirement is technically feasible and would play an important role in reducing certain risks from foundation models in many domains.

Introduction

Foundation models represent a dramatic advance for the state of the art in Artificial Intelligence (AI). In current discussions of AI, a foundation model is defined very generally as an AI model that is trained on large amounts of data, typically using self-supervision, that can be adapted, or 'fine-tuned' to a wide range of downstream tasks (see, e.g., Bommasani et al., 2022).[1] In this paper, we will argue for a specific regulatory mechanism that governs the release of new state-of-the-art foundation models.

For concreteness, our arguments will sometimes make reference to a central ingredient in many current foundation models, namely **large language models** (**LLMs**), which have the ability to generate natural language text as output. Many LLMs are foundation models in their own right: for instance, BERT and GPT-3 are LLMs and also canonical examples of foundation models. The discussions in this paper will sometimes refer to LLMs and the text they generate, to give concrete examples of the content that foundation models can produce and the issues that arise for these models. Our broad argument is about foundation models generally, not just about LLMs. But we will begin by introducing LLMs and then show how LLMs can provide the core of foundation models with wider functionality.

LLMs are trained on large amounts of human-generated text. After training, a LLM develops certain humanlike linguistic competences: in response to a linguistic 'prompt' posing a question or setting a task, an LLM can autonomously produce a human-like response, which is often hard to distinguish from the response of a real human. LLMs can generate text conditioned on large documents (e.g., OpenAI, 2023b), and LLM-powered systems can make reference to arbitrary material on the World Wide Web and other external sources (e.g., Thoppilan et al., 2023; Microsoft, 2023). In recent systems, input prompts can incorporate images and videos (see, e.g., OpenAI, 2023a). LLMs can also generate computer code, within limits (e.g., Nguyen and Nadi, 2022). They can also produce images from a linguistic prompt (Ramesh et al., 2021; 2022), drawing in particular on diffusion-based methods for image generation (see Zhang et al., 2023). They can even drive goal-directed actions of software agents (Adept, 2022) and robots (Brohan et al., 2022). When LLMs are supplemented with these additional multimodal abilities, they fall within the more general class of foundation models, as defined above. Foundation models, whether purely linguistic or multimodal, are a large leap forward in one of AI's core aims—to simulate and reproduce human abilities.

As foundation models disseminate into society, they are starting to impact people's lives in significant ways, in domains such as education (Kasneci et al., 2023), political processes (Luitse and Denkena, 2021), and the human workplace (Eloundou, 2023). The driving force behind these disruptions is the big tech industry: large multinational companies such as

[1]The terms 'generative AI systems' and 'general-purpose AI systems' are also used, with some relevant overlap. Strictly speaking, LLMs/foundation models are *deployed* in 'systems'; the resulting systems do indeed generate content, and they are often general-purpose. We will use 'foundation model', to avoid proliferating terminology.

Microsoft (Microsoft, 2023, with OpenAI), Google (Pichai, 2023), Meta/Facebook (Meta, 2023), Amazon (Soltan et al., 2022) and Baidu (Baidu, 2023). From a technical perspective, these large companies and their collaborators are by far the best placed to develop foundation models. They can recruit the best AI researchers; they can deploy their models on the most powerful computing infrastructures; and they can train their models on the largest datasets, gathered from interactions with their own users (e.g., Amazon, 2022), from scraping the internet, from privately licensed document collections and from their own crowdsourcing exercises (e.g. Roller et al., 2020, Thoppilan et al., 2022).

The foundation models being produced by tech companies are improving at a startling pace. The pace of progress currently far outruns our ability to assess the impacts they are having on people's lives and on the functioning of society. A wide-ranging debate is under way about how these new models should be assessed and overseen. In this paper, we offer our input into this debate. We propose that when a company develops a foundation model that advances the state of the art, its release should be conditional on demonstration of a detection mechanism that can reliably detect the content it produces. In Section 1 we introduce this idea, and in Section 2 we describe a precedent for it, in the process that led to the release of the early foundation model GPT-2. In Sections 3 and 4, we situate the proposed pre-release requirement within wider regulatory and economic contexts for AI systems. In Section 5, we consider how the requirement for a detection mechanism could be defined in more detail, and we discuss some other pre-release requirements that could be considered in due course.

I. Summary of the proposed pre-release requirement

Our proposal begins from the idea that any organisation that develops a new state-of-the-art foundation model should be required to conduct certain safety processes prior to releasing it to the public. The idea of pre-release obligations on providers is present in several regulatory frameworks, as we'll discuss in Section 3. Our paper focusses on just one pre-release obligation, which we think is both important and practically achievable. We don't mean to suggest there shouldn't be other pre-release obligations–we are just focussing on one important condition for release.

We propose that **a central condition on release of a new state-of-the-art foundation model should be demonstration of a *detection mechanism* that can distinguish content produced by the foundation model from other content, with a high degree of reliability**. On release of the model, this mechanism must be made publicly and freely available in a detector tool that allows consumers to query, for an arbitrary item of content, whether the item was generated (wholly or partly) by the model.[2]

[2] We use the word 'detector' rather than 'classifier' for the tool, because it should be able to identify *portions* of an item that are generated by the model, rather than making a single class decision about the whole item. The word 'detector' also captures the function of the mechanism in the large, as used by a community of users, to identify material generated by the model.

We believe this is a minimal condition needed to ensure the model can be used safely and responsibly. We'll focus on textual content in the current paper. For this content, the detection mechanism could involve a classifier, perhaps making use of watermarks included in the generated text or image (see, e.g., Kirchenbauer et al., 2023; Zhao et al., 2023), or methods exploiting statistical features of generated content (see, e.g., Mitchell et al., 2023). Or it could involve the producer company keeping a log of all texts generated by its LLM and offering a plagiarism detector running on this log, as suggested by Krishna et al. (2023). This latter method appears to be more resilient to adversarial attacks–in particular to the 'paraphrase attacks' discussed by Sadasivan et al. (2023). The detection mechanism could also involve 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. Crucially, it would be for the organisation wishing to release a new foundation model to demonstrate a detection mechanism that is effective in the current adversarial context, and show its practicality, either unilaterally, or in collaboration with other groups. In all cases, the detection mechanism should be freely available to members of the public, subject to measures such as rate limits to counter adversarial use.

A reliable automated detection mechanism is essential for several reasons. Some of these reasons are specific to particular domains. For instance, in educational domains, and again focussing on textual content—while LLMs will certainly create many new teaching paradigms (e.g., McKnight, 2023), and their proper use still demands good writing skills (Villasenor, 2023), teachers still need ways of differentiating between text generated by students and text generated by AI systems (Bommasani et al., 2022, Section 3.3). Tools that can detect LLM-generated text are currently in high demand and are likely to be of practical use to educators for the foreseeable future. In social media domains, the users are likely to be companies as well as individuals. A social media company needs efficient ways of determining if a post is generated by an AI model, particularly in areas relating to disinformation, so it can take effective action against large-scale disinformation campaigns. If companies can't identify AI-generated content quickly and efficiently, foundation models are likely to be used to cause widespread disruptions to democratic processes (see, e.g., Heikkilä, 2022; Goldstein et al., 2023; Ordonez et al., 2023). In the law enforcement domain, police need methods for countering sophisticated spoofing and phishing operations (see, e.g., Europol, 2023). In commercial domains, consumers need ways of knowing if their information or product advice comes from a person or a machine. (This right is enforced, for instance, by California's 'BOT' Act (SB2001, 2018).) Companies that develop LLMs would also arguably benefit from a detection mechanism, for use when training a new language model, to ensure they don't train on text the model produced itself, or on text produced by other LLMs. Either scenario which could lead to degenerate performance over time. The detection mechanism would play a variety of important roles in different domains, in different ways, for a variety of actors. Emphatically, requiring a detection mechanism does *not* presuppose that using AI to aid in the production of content is necessarily wrong. AI-aided generation will of course be very valuable in many, many situations. The important thing is that a detection mechanism is available, *for judicious use*, *if needed*, in ways appropriate for the domain at hand.

Stepping back from specific domains, we also suggest there are *general* reasons why mechanisms for detecting AI-generated content are essential. Foundation models are improving rapidly, and it is increasingly difficult to tell whether content is produced by a person or a machine. LLMs can produce human-like text rapidly, and at scale: so inevitably, the world is about to be flooded with a huge amount of human-like, machine-generated text. In the absence of detection mechanisms, humans will face a completely new *attribution problem*: they won't know whether text they see comes from people or machines. If a human reader interprets an LLM-generated text as coming from a person, they are being fundamentally misled. People intuitively understand what communication *between people* is: it is a practice which governs our whole lives. The consumption of machine-generated content is a brand-new phenomenon: we don't know anything about the effects it will have, especially at large scales. Human communication is fundamental to the organisation of our societies: we need to protect this institution, by requiring that machine-generated content be identifiable as such. Reasoning of this kind has led to proposals that people have a *right to know* whether the content they receive comes from a human or a machine (see, e.g., IBM, 2021), and that transparency legislation should provide this information (e.g., Engler, 2020). We subscribe to this idea too. This is a general reason for our argument that new foundation models should not be released without evidence for a detection mechanism for the content they produce. Without a detection mechanism, the enforcement of this newly posited right would be difficult, if not impossible.

All the reasons above are given extra weight by the fact that we don't know how much better foundation models are likely to get in the coming years. The performance of new technologies is often describable by a logistic 'S-curve', beginning with a rapid rise, and culminating with a plateau (see e.g. Modis, 2007). We don't yet know where we are on the curve for foundation models—but it is possible we are at a very early point, and that considerable improvements are still to come. If this is the case, there are particular reasons for a pre-release detection mechanism. Many of the risks of foundation models (Weidinger et al., 2021) can be expected to increase as their performance improves. If foundation models are able to match, or even exceed, human performance in language-processing tasks, then it will be particularly important to keep track of how they contribute to products and services, and more broadly to public debates and decision-making.

It's useful to situate our proposal in relation to other recent calls for oversight of foundation models. There have been several recent suggestions about how LLMs should be audited. Mökander et al. (2023) discuss many types of audit; they usefully distinguish between ex-ante audits that occur prior to release and ex-post audits that occur afterwards. They advocate both types of audit, but their emphasis is on ex-post audits, while our proposal focuses on pre-release processes. There have also been many suggestions about specific safety mechanisms that could or should be applied to LLMs prior to release. For instance, several methods have been suggested for preventing LLMs producing offensive content (e.g., Perez et al., 2022), assessing bias (Tamkin et al., 2021) or improving factual accuracy (e.g., Nakano et al., 2022); all of these could potentially be required pre-release. Askell et al. (2019) review the importance of pre-release processes in their general characterisation

of ‘responsible AI’. Brundage et al.’s (2020) discussion of how manufacturers can make verifiable ‘claims’ about an AI system also highlights pre-release processes. But none of these discussions suggest that workable detection mechanisms should be a condition for release.

A very recent call for oversight comes from the Future of Life Institute (FOL, 2023), in an open letter calling for a 6-month moratorium on releases of all new AI systems ‘more powerful than GPT-4’. During the moratorium, companies should develop suites of safety protocols and a broad ‘governance system’ for AIs should be instituted to control their use. This ‘governance system’ includes ‘provenance and watermarking systems’ on AI system outputs, to ‘distinguish real from synthetic’. Our proposed detection mechanism rule is one possible aspect of this governance system.

II. Precedent for a pre-release process for foundation models

The recent history of LLM development contains one prominent example of the kind of pre-release process we advocate: it was carried out by OpenAI in 2019.

OpenAI has been a leader in the creation of LLMs over the last four years. A key milestone for this company (and the field as a whole), was GPT-2, released in 2019 (Radford et al., 2019). GPT stands for Generative Pretrained Transformer, the breakthrough machine learning technology behind LLMs (Vaswani et al., 2017), itself based on the introduction of attention mechanisms (Badhanau et al., 2014) in deep learning. At the time GPT-2 was created, OpenAI presented itself as a ‘non-profit artificial intelligence research company’, with the goal of ‘advanc[ing] digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return’ (OpenAI, 2015). OpenAI also had a mission to share cutting-edge AI technologies: its researchers were strongly encouraged to publish their work, and it undertook to share any patents it obtained. Given this mission, the company faced a challenge when it produced GPT-2, which was significantly better at generating human-like text than its predecessors: researchers were concerned about the impacts the system might have if released. Accordingly, the system’s release was ‘staged’: a restricted version of GPT-2 was released in February 2019, a larger version in August, and the full version in November.[3] During this period, OpenAI collaborated with several external partners to conduct risk and benefit analyses for GPT-2, and to study how the initial releases of the system were used. The studies conducted were described in a report whose final version was published when the full version of the system was released (Solaiman et al., 2019). The report essentially described a due diligence process, undertaken to ensure that the full version of GPT-2 could be released without undue risk, and that consequences of release had been carefully thought through.

[3] The staged versions of GPT-2 differed in their size; earlier versions used fewer parameters than the full version released in November.

A focus for assessing risk was the possibility that GPT-2 could be 'misused' by malicious actors, motivated by 'the pursuit of monetary gain, a particular political agenda, and/or a desire to create chaos or confusion' (Solaiman et al. 2019:9). The key scenario considered was one where such actors produced large volumes of texts, flooding social media and other Internet fora. Studies focussed on methods for detecting text produced by versions of GPT-2 tailored to these malicious domains. Some studies involved measuring the ability of human judges to detect GPT-generated text; others involved training AI classifiers to detect GPT-generated text in various ways and evaluating their performance. Human judges were found to be quite bad at identifying GPT-generated text, especially if the texts they saw were manually 'cherry picked' to select the most human-like ones. Custom-built classifiers performed better, but were still fallible, with accuracy as low as 74% for full versions of GPT-2 customised on a specific domain. Crucially, detection accuracy increased as the texts being judged grew longer: for texts of around 200 words in length, accuracy was around 95%.

OpenAI's staged release process also included surveys of how the early versions of GPT-2 were used 'in the wild'. These early surveys were also useful, but it's hard to use such surveys to make predictions about future use, as Solaiman et al. concede. The central focus of OpenAI's pre-release process, therefore, was on the issue of detection mechanisms. Solaiman et al. stopped short of saying that demonstrating a reliable detector was a precondition for release of their model. In fact, they were ambivalent about whether automatic detection would remain feasible as LLMs improve. They made no explicit statement about the conditions under which it is acceptable to release a language model. But their demonstration of a workable detection mechanism was strongly implicit in supporting their decision to release the full GPT-2.

Since GPT-2, OpenAI have released GPT-3 (Brown et al., 2020), GPT-3.5 (ChatGPT: OpenAI, 2022) and GPT-4 (OpenAI, 2023a). Strikingly, the staged release process undertaken for GPT-2 was not retained for these later systems. Pre-release versions of GPT-3 onwards were given to some collaborating companies and groups, and some consultation was undertaken about the capabilities and risks of these systems (see, e.g., Tamkin et al., 2021; OpenAI, 2023c), but there was no systematic documentation of the pre-release process akin to that for GPT-2. In this paper, we argue that a formal pre-release process should be re-instituted.

III. How does a pre-release detection mechanism fit with current or imminent legislation?

While some companies may voluntarily supply a detection mechanism for their foundation models on release, we don't anticipate that all companies will do so. So a detection mechanism requirement would likely have to be imposed by law. How would this requirement fit with existing laws for AI? It's vital that laws for AI are collectively coherent and that redundancy is avoided, just as for any area of legislation (see, e.g., LDAC, 2021).

EU legislation around AI is a key area to discuss, as it is further advanced than legislation in many other jurisdictions. The EU's AI Act (EU, 2021) is particularly relevant. It places particular obligations on the developers of 'high-risk' AI systems, and these include an 'ex ante conformity assessment' prior to release, to check for various risks defined in Articles 8–15. The Act has recently been amended by the European Parliament to include provisions about foundation models (EU, 2023a). The amendments don't explicitly class foundation models as high-risk AI systems, but they do define specific requirements for the developers of foundation models: in particular, they specify (in an amendment to Recital 60g) that 'Generative foundation models should ensure transparency about the fact the content is generated by an AI system, not by humans'. This amendment appears to make reference to the kind of detection mechanisms we are calling for here. But if this is the intention, the requirement could be more precisely defined. We suggest the requirement should make explicit mention of a reliable detection mechanism, tested prior to release, and made available to the public.

The EU's Directive on Adapting Civil Liability Rules to AI (EU, 2022) also fails to cover the case we are advocating. This Directive focuses on cases where an AI product causes harm after release, while our proposal is for processes that occur prior to release.

The OECD and UNESCO AI principles provide broad context for our proposal. They have quasi-legal status as all member states have agreed to them; they arguably give general grounds for our proposed detection mechanism. The OECD's principles (OECD 2022) under Clause 1.3 (Transparency and Explainability) require AI actors 'to provide meaningful information (...) to make stakeholders aware of their interactions with AI systems'. The UNESCO AI principles (UNESCO, 2022) under 'Communication and Information' state (Clause 113) that AI actors should 'respect and promote (...) access to information with regard to automated content generation', and notes that 'appropriate frameworks, including regulation, should enable transparency of online communication'. Either of these principles could be argued to support a detection mechanism requirement for foundation models. But further detail would be needed to clarify obligations on companies.

In Canada, AI legislation has been presented to parliament (the Artificial Intelligence Data Act, AIDA 2022), which states high-level principles that future adaptive regulation should attempt to achieve. Although the law was drafted over a year ago and thus has no explicit elements regarding LLMs, the associated regulation can be adapted, which is an important feature given the rate of advances in AI and the likelihood that new nefarious uses will come up. It would thus be good if that upcoming Canadian regulation includes elements such as those discussed here.

It's useful to mention other legislation that applies to 'direct' providers of LLM content, such as providers of chatbots. For instance, California's bot disclosure law, SB1001 (2018), prohibits the use of AI 'bots' with the intent to mislead users into thinking they are human, in order to incentivise a purchase or to influence a political vote. Article 52(1) of the EU's AI Act (EU, 2021) goes further: providers of AI bots must make their human users aware of the

bot's AI status in all domains of interaction ('unless this is obvious from the circumstances'), not just in commercial and political contexts. These are useful provisions. They overlap with broader attempts to identify and/or ban 'inauthentic accounts' on social media platforms; and conversely, to positively authenticate bona fide human users (for instance, Twitter's 'verified users' scheme; Twitter, 2023). Schemes for positively authenticating human users may usefully complement detection methods for LLMs. But detection methods are still needed for AI systems that engage in 'direct' interactions with users, even if these are clearly labelled as coming from AIs. The key point is that textual or multimedia content produced by AI systems in direct interaction with users can readily be copied and disseminated further, *without* the labels signalling its origin. A human user could readily generate and disseminate AI-generated content, without identifying it as such. There may be cases where this unattributed use of AI content is harmless or legitimate--but sometimes it can be harmful and deceptive. A detection mechanism is necessary for these latter 'indirect' uses of the AI-generated content.

There are already some proposed laws that address 'indirect' uses of AI of the kind just mentioned. For instance, Article 52(3) of the EU's AI Act (EU, 2021) requires 'users' of an AI system that generates audiovisual content to *disclose* that this content was artificially generated, if the content 'would falsely appear to a person to be authentic'. (The provision implicitly covers the case where users *share* the generated content with other people.) The recent amendments by the EU Parliament extend this requirement to include systems that generate textual content, and detail what form disclosures should take (see again Bertuzzi, 2023). These provisions expressly cover the 'indirect' uses of AI-generated content that we are concerned about. We think these are useful provisions. But again, we don't think imposing obligations on *users who disseminate* AI-generated content obviates the need for companies to provide a detection mechanism, so that human consumers have agency of their own. Human *consumers* of content can't rely on human disseminators of AI-generated content to comply with the law: they need a tool they can use themselves, for their own purposes, in a variety of contexts. We see a detection mechanism as being of great use to individual consumers of online content, to apply at their own discretion. Of course, a detection mechanism is also helpful in enforcing laws about AI-generated content—and in enforcing more informal policies (for instance in educational settings). Laws and informal policies about AI-generated content are likely to be very context-specific: our proposed detection mechanism is of use in all such cases.

Existing legislation and ethical guidelines in China are also relevant to note. A 2022 position paper from the Chinese government on 'Strengthening Ethical Governance of AI' (PRC, 2022a) notes that 'Governments should strengthen pre-use study and evaluations of AI products and services', which is consistent with the pre-release process we are advocating. A law enacted last year (PRC, 2022b) imposes many requirements on 'deep synthesis Internet information services', which appear to subsume our definition of foundation models. Notably, providers of such services must 'employ technical measures to attach symbols to information content produced or edited by their services': this presumably amounts to a kind of watermarking scheme. This requirement is reiterated in draft legislation released very recently (PRC, 2023), which goes further, requiring the content produced by generative AI

systems to meet certain concrete standards: for instance, it should be ‘true and accurate’ (Article 4), and not contain ‘discriminatory content’ (Article 12). There is an overriding requirement that the content should ‘reflect the core values of socialism, and must not contain subversion of state power’, which obviously imposes a level of state control far beyond what is acceptable in the West. But some of the articles in this draft law resonate with our proposed detection mechanism rule. For instance, Article 6 requires a ‘security assessment’ of an AI system before it is released to the public.

IV. Is a pre-release detection requirement feasible in the current competitive environment?

At present, tech companies are competing fiercely amongst themselves to produce foundation models. Foundation models are a new commercial frontier, with applications in many markets, from entertainment to customer service to web search. Competition also exists between countries, with many countries vying to dominate these markets. These competitive pressures create a difficult backdrop for safety processes, as well documented by Askell et al. (2019). How would our proposed pre-release detection mechanism requirement play out in this competitive environment?

Our proposed mechanism may be easier to incorporate into legislation that already has an international focus–for instance, legislation within the EU. But given existing competition between countries, the mechanism would be best introduced by a broader international agreement, along the lines of the OECD and UNESCO principles or EU legislation mentioned in Section 3. A fully international agreement would of course be challenging to obtain–but there are at least some points of commonality between proposals in the EU, the US and China, as again discussed in Section 3. In other areas of science, there have certainly been workable international agreements. For instance, the International Atomic Energy Agency (IAEA) provides some measure of international governance for nuclear safety. Several commentators have called for an IAEA-like organisation to oversee AI internationally, most recently the CEO of OpenAI (Hern, 2023). (Other commentators have called for a 'CERN-like' organisation (Marcus, 2017; LAION, 2023), including very recently the CEO of Google (Milmo, 2023), but IAEA is more relevant as a regulatory model.) Elsewhere the agreement banning human cloning (Macintosh, 2022) is well supported internationally. Obviously there are many differences between these technologies and foundation models. But there are also important similarities: in each case, the technology would open up huge and lucrative new markets, but in each case, there are concerns with the technology that go deeply to the issue of the value of human lives and human activity. There are also analogies with international agreements about climate change, that are grounded in science but regulate aspects of countries’ economic activity. Finally, there are possible analogies with international treaties in arms control/nonproliferation, such as the bans on chemical weapons, cluster munitions, land mines, and the treaties restricting deployment of nuclear weapons. These are clearly further removed: foundation models don’t pose direct risks to life and limb. But the international ‘arms race’ in AI certainly has a

strategic component as well as a commercial one, so arms control treaties are worth mentioning in this context.

Returning to commercial competition: it is important to note that requiring new foundation models to be accompanied by reliable detectors would institute a new form of competition between companies and other stakeholders, *in developing innovative methods for detecting the use of foundation models*. Requiring companies to compete in developing technologies that directly further the public good, such as detectors for LLMs, is a way for commercial competition to directly serve the general public.

V. What might the details of the pre-release detection rule look like?

There are many questions about how our proposed rule would be implemented. We will conclude by mentioning a few of these.

Some questions relate to the administration of the mechanism. Which organisations should have to demonstrate a detection mechanism? Companies of all sizes, or just the largest ones? (The EU's Digital Services Act requires additional processes for the largest companies.) What about university research labs, or government labs? What kinds of AI models should have their release conditioned on a detection mechanism? How exactly should 'foundation models' be defined? And how should a 'state-of-the-art' model be defined?

Some questions relate to the detection mechanism itself. What level of accuracy would be required for this mechanism? How should the mechanism cater for AI-generated content that is post-processed by a person? (It must be able to provide useful feedback about content that is partly machine-generated and partly human-generated. A large amount of content is likely to be of this hybrid type as foundation models become more widely adopted.) How should the detection mechanism be defined, so it effectively covers different modalities of content, such as text, images, video and audio? (Are separate definitions needed for different modalities?) What conditions should govern use of the detection mechanism? If it's developed in-house, should access be free? How could adversarial use of the mechanism be prevented? Finally, end-users are likely to work with a detector that aggregates results from many model-specific detectors, so multiple possible origins can be considered simultaneously. What should this aggregation mechanism look like, and who would administer it?

Some questions relate to other checks that could be included in the pre-release process. Maybe there could be requirements for functionality that reports the system's confidence in its own output (Wang et al., 2022), or which document sources were used to produce the output (see, e.g., Thoppilan et al., 2022, for one evaluation scheme). Or requirements to document harmful biases found in the system (for instance, biases in relation to gender,

sexuality or ethnicity), or to provide transparency about the training data used (see, e.g., Bender et al., 2021, for proposals in both areas). Some of these additional checks might require experimental deployments of the system pre-release, to test functionality and behaviour with users in controlled environments, possibly through the 'regulatory sandboxes' discussed in the EU's AI ACT and elsewhere (see, e.g., Ranchordas, 2021).

Some final questions concern how a requirement for detection mechanisms could be enforced. Enforcement would likely operate at a country level, through national laws reflecting an international agreement. (Even the EU's AI Act will be enforced this way.) Enforcement will be easiest if the ability to produce state-of-the-art foundation models is restricted to the largest commercial or public actors, because these are readily identifiable. But it is also possible that smaller open-source developers can create their own high-quality models (Patel and Ahmad, 2023). If a larger variety of groups have the ability to develop powerful foundation models, then enforcement of a detection mechanism requirement will of course be more challenging. But models that gain a large user base will necessarily become visible to enforcement agencies. And to pursue 'private' developers of non-compliant models, there are established methods for identifying sources of illegal online content (see, e.g., Rid and Buchanan, 2015).

We look forward to a discussion of these, and other, questions about our proposed detection mechanism requirement.

References

Adept (2022). ACT-1: Transformer for Actions – Adept." https://www.adept.ai/blog/act-1.

AIDA (2022). The Artificial Intelligence Data Act - Companion Document. Canadian Government. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

Amazon (2022). How Amazon protects customer privacy while making Alexa better. Amazon blog post. https://www.aboutamazon.com/news/devices/how-amazon-protects-customer-privacy-while-making-alexa-better

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. arXiv preprint arXiv:1907.04534.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473.

Baidu (2023). ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology. Baidu research blog. http://research.baidu.com/Blog/index-view?id=183

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

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

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

CERN (2023). The CERN Nuclear Safeguards Policy. https://edms.cern.ch/ui/file/2812766/LAST_RELEASED/Nuclear_Safeguards_Policy_docx_cpdf.pdf.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv preprint arXiv:2303.10130.

Engler, A. (2020). The case for AI transparency requirements. Brookings Institute report. https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/.

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. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

EU (2022). Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive). https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf.

Europol (2023). The criminal use of ChatGPT—A cautionary tale about large language models. https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models.

EU (2023a). 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.

Future of Life (2023). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv preprint arXiv:2301.04246.

Heikkilä, M. (2022). How AI-generated text is poisoning the internet. MIT Technology Review. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/.

Hern, A. (2023). OpenAI leaders call for regulation to prevent AI destroying humanity. The Guardian, 24 May. https://www.theguardian.com/technology/2023/may/24/openai-leaders-call-regulation-prevent-ai-destroying-humanity.

IBM (2021). What an AI Bill of Rights Should Look Like. IBM Policy Lab blog post. https://www.ibm.com/policy/what-an-ai-bill-of-rights-should-look-like/#:~:text=An%20individual%20has%20a%20right,AI%20must%20be%20explainable.

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.

LAION (2023). Petition for keeping up the progress tempo on AI while ensuring its transparency and safety. Large-Scale AI Open Network post. https://laion.ai/blog/petition/.

LDAC (2021). How new legislation relates to the existing law. New Zealand Legislation Design Advisory Committe guidelines Ch3. http://www.lac.org.nz/guidelines/legislation-guidelines-2021-edition/early-design-issues-2/chapter-3/.

Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. Big Data & Society, 8(2), 20539517211047734.

Macintosh, K. L. (2022). Human Cloning. In Elgar Encyclopedia of Human Rights. Edward Elgar Publishing.

Marcus, G. (2017). AI is stuck. Here's how to move it forward. New York Times. https://www.nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html

McKnight, L. (2023). Eight ways to engage with AI writers in higher education. Times Higher Education Campus. https://www.timeshighereducation.com/campus/eight-ways-engage-ai-writers-higher-education.

Meta (2023). Introducing LLaMA: A foundational, 65-billion-parameter large language model. Meta AI blog. https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

Microsoft (2023). Introducing the new Bing. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Milmo, D (2023). Google chief warns AI could be harmful if deployed wrongly. The Guardian, 17 April. https://www.theguardian.com/technology/2023/apr/17/google-chief-ai-harmful-sundar-pichai

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.

Modis, T. (2007). Strengths and weaknesses of S-curves. Technological Forecasting and Social Change, 74(6), 866-872.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... & Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. Red teaming language models with language models, 2022. URL https://arxiv. org/abs/2202.03286.

Nguyen, N., & Nadi, S. (2022). An empirical evaluation of GitHub copilot's code suggestions. In Proceedings of the 19th International Conference on Mining Software Repositories (pp. 1-5).

Ordonez, V., Dunn, T. and Noll, E. (2023). OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this'. ABC News. https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122

OECD (2022). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, 2022. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

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.

PRC (2022a). Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI) https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html

PRC (2022b). Provisions on the Administration of Deep Synthesis Internet Information Services. Translation by China Law Translate. https://www.chinalawtranslate.com/en/deep-synthesis/.

PRC (2023). Notice of the Cyberspace Administration of China on Public Comments on the "Administrative Measures for Generative Artificial Intelligence Services (Draft for Comment)". http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm.

Patel, D. and Ahmad, A. (2023). We Have No Moat, And Neither Does OpenAI. Leaked Google internal memo. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

Pichai, S (2023). An important next step on our AI journey. Google blog. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. International Conference on Machine Learning (pp. 8821-8831).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.

Ranchordas, S. (2021). Experimental regulations for AI: Sandboxes for morals and mores. University of Groningen Faculty of Law Research Paper, (7).

Rid, T., & Buchanan, B. (2015). Attributing cyber attacks. Journal of strategic studies, 38(1-2), 4-37.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

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

SB2001 (2018). Bolstering Online Transparency ('BOT') Act. California legislation. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., ... & Natarajan, P. (2022). Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. arXiv preprint arXiv:2208.01448.

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.

Twitter (2023). How to get the blue checkmark on Twitter. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

UNESCO (2022). Recommendation on the Ethics of Artificial Intelligence. UNESCO report. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L and, Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762.

Villasenor, J. (2023). How ChatGPT Can Improve Education, Not Threaten It. Scientific American. https://www.scientificamerican.com/article/how-chatgpt-can-improve-education-not-threaten-it/.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). Text-to-image Diffusion Model in Generative AI: A Survey. arXiv preprint arXiv:2303.07909.

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