February 2, 2023,

National Institute of Standards and Technology
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

Via electronic filing

Re: Access Now's Submission to NIST's Request for Information to Support Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Docket Number: 231218-0309)

Access Now appreciates the opportunity to provide input on the National Institute of Standards and Technology's (NIST) Request for Information (RFI), aligning with its responsibilities under the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. While we strongly advocate for a comprehensive federal data protection and privacy law as a cornerstone of robust AI governance, we also recognize the importance of fostering improved processes and approaches within corporate entities and organizations to guard against the misuse and abuse of AI systems and their underlying data. In this regard, we commend and support the initiatives led by the National Institute of Standards and Technology.

Our submission focuses on the RFI's section on developing guidelines, standards, and best practices for AI safety and security, specifically addressing generative AI risk and the efficacy, validity, and long-term stability of watermarking. We express concern regarding the portrayal of digital watermarking as a panacea for AI-generated content and urge a comprehensive examination of its limitations. As a result, we are submitting our discussion paper, diving into the effectiveness of digital watermarks in identifying AI-generated content, assessing their alignment with human rights principles, and presenting provisional policy recommendations.

We firmly believe that relying solely on watermarking will not conclusively resolve the challenges posed by AI-generated content. It fails to address the nuanced questions surrounding the role of generative AI in our society. Access Now commends NIST for its ongoing efforts and urges the agency to shift focus from technology-centric approaches to building a rights-based foundation and establishing people-centered goals for our relationship with emerging technologies. While recognizing the potential value of watermarking in certain contexts, we underline the significance of its thoughtful deployment. Only when the benefits are clear can the risks be properly mitigated, allowing individuals at risk of human rights harm to exert complete control over their interactions with the technology.

Enclosed is our full discussion paper for your consideration.

Best regards,
Willmary Escoto | US Policy Counsel

IDENTIFYING GENERATIVE AI CONTENT: WHEN AND HOW WATERMARKING CAN HELP UPHOLD HUMAN RIGHTS

A DISCUSSION PAPER

SEPTEMBER 2023

Access Now defends and extends the digital rights of people and communities at risk. By combining direct technical support, strategic advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.

This paper is an Access Now publication, written by Gustaf Björksten. We would like to thank the Access Now team members who provided support, including Carolyn Tackett, Daniel Leufer, Eliška Pírková, Fanny Hidvégi, Loren Giordano, and Méabh Maguire. We would also like to thank the partners who have engaged thoughtfully with us on these issues as the landscape around generative AI quickly evolves, and we look forward to receiving further feedback and continuing discussion with a wide range of stakeholders.

For more information, please visit:
https://www.accessnow.org

Published in September 2023

TABLE OF CONTENTS

I. INTRODUCTION

Around the world, policymakers and civil society actors alike are grappling with the proliferation of generative AI tools and the corresponding potential for – and already unfolding – harm to human rights. Many are looking to technological solutions for these complex problems, but all too often with limited understanding of what the technology can or cannot achieve, and overestimating how effective technology alone can be in solving the problems at hand.

As generative AI tools have come into mainstream use, there has been much debate regarding the need to be able to identify content that has been generated by AI. This debate has surfaced many proposals for adopting a mechanism known as "AI watermarking." Various iterations of these proposals have come from governments (including, for example, in the United States and European Union), technologists, industry, civil society groups, and beyond. But many have been ill-informed regarding what is technically possible and how such mechanisms may or may not help to uphold human rights across different use cases.

In a technical sense, watermarking is not well suited for text-based generative AI content. While it is better suited to binary outputs from generative AI – such as for pixel art, video, and audio – there are legal, ethical, and human rights considerations that must be taken into account before determining the potential role of watermarking in addressing the challenges presented by generative AI. Even at its best, digital watermarking on its own will never be a complete "solution" to these challenges, and treating it as such risks layering additional human rights harms on top of the ones we aim to solve.

As is the case with any policymaking around emerging technology, the effective application of digital watermarking in the context of generative AI requires an informed and nuanced approach. Policymakers need to understand in detail the full spectrum of what is and isn't possible, how the capacity to mitigate particular AI problems changes across contexts, and when these tactics are of benefit or detriment to human rights. The following discussion paper addresses each of these questions in detail.[1]

II. WHAT IS GENERATIVE AI?

At their most basic level, generative AI systems are applications that produce unique content in response to prompts from users. Although generative AI only gained mainstream attention in late 2022 with the launch of ChatGPT, the underlying technology has been around for years, primarily in two forms:

[1] Importantly, this paper focused on watermarking at the stage of content creation, and does not go further to cover in detail related proposals regarding watermarking or labeling at the stage of content distribution (e.g. when a media outlet publishes and disseminates content generated using AI tools). We are also exploring similar questions on that front and welcome further discussion across this full range of issues.

- ➔ **Large language models** (LLMs), such as the one that underpins ChatGPT, generate plausible-sounding text in response to a human prompt (e.g. "write a sonnet about the risks of AI in the style of Shakespeare"). The easy-to-use, conversational ChatGPT interface so many people are experimenting with today is merely a refined version of previous iterations of the same technology, such as GPT-3, rather than something radically new or unprecedented.
- ➔ **Multi-modal models**, such as Stable Diffusion, Midjourney, or OpenAI's DALL-E 2, typically take text prompts (e.g. "a purple penguin wearing sunglasses") and generate images as an output. Some models, such as GPT-4, can also take images as input (e.g. a photo of your refrigerator's contents) to produce text as the output (e.g. a recipe for the ingredients you have). Multi-modal models that can generate audio and video outputs are also in development.

To learn more about generative AI, how it works, and how the technology is impacting human rights, read Access Now's responses to frequently asked questions.

III. WHY IDENTIFY AI-GENERATED CONTENT?

To any pursuit where traditionally content has been generated by individuals, often specialists, generative AI comes as a highly disruptive technology. Regardless of whether that content is text, static or moving imagery, audio, or other output formats, generative AI is, or will soon be, at a point where its content is often indistinguishable from human-generated content to the perception of most consumers. Generative AI, like most digital technologies, holds the potential to provide positive benefits to society, alongside the capacity to inflict great harm, especially for people most at risk.

Scenarios resulting from the use of generative AI are requiring us to rethink our traditional approaches to certain situations. For example, when students submit work at school, do we now have to consider if they have circumvented the learning process by utilizing generative AI? When individuals submit work in a competition, what does it mean for fairness of that competition if some have utilized generative AI to construct their entries? When work is conducted by employees, will their use of generative AI introduce issues of correctness, safety, and bias into the outputs? Fake content created using generative AI tools, such as videos purporting to show well-known people saying or doing controversial things, are increasingly difficult to differentiate from real recorded content. How do we differentiate fake from real content, particularly in time-sensitive situations, such as in the lead-up to elections?

For some generative AI models, harm is clearly identifiable, such as the creation of nonconsensual deepfake nude imagery. The problem space is further complicated by legal issues which are yet to be resolved by the courts of individual nations. In a brave new world for copyright, it is yet to be decided who owns the rights to generative AI content. Is it the developer of the AI system, the person crafting the prompt to the AI system, or the collective of human creators whose content was used to train the AI model? There are further possible legal complications for some types of problematic content – for example, AI-generated child sexual abuse material (CSAM), for which it would be difficult to trace back to individual victims whose images were used to train the model. In some jurisdictions such content is still considered illegal, but in other jurisdictions it is not clear if it would be considered illegal or not.

These questions have led some commentators to suggest there are situations where it would be beneficial to society to know if content has been generated using AI, and many are looking to various forms of digital watermarking as a vehicle for achieving that goal.

IV. WHAT IS DIGITAL WATERMARKING?

ATTRIBUTES OF WATERMARKING

Watermarks were first used in Italy in the 13th century to identify the manufacturers of sheets of paper. Since then, watermarks in various forms have been used for purposes including indicating ownership, frustrating attempts to counterfeit, and verifying authenticity. Paper banknotes have watermarks, and polymer banknotes have holograms that serve the same purpose of verifying the legitimacy of the currency and making them hard to counterfeit. In the computer age, digital watermarks have been used in various formats to similarly mark digital content.

The intention of all watermarks is to identify something about the content in which the watermark is embedded. Beyond that intention, there are two primary attributes of digital watermarks that should be considered:

Watermarks have varying degrees of invisibility

In contrast to highly visible labels, a general attribute of watermarks is that they are not immediately obvious to individuals interacting with or consuming the content. Watermarks are typically hidden in some way, but the degree to which watermarks are hidden can vary greatly, including:

- ➔ Visible watermarks, such as those used by stock photo company Getty Images;
- ➔ Hidden watermarks not immediately perceptible to the human eye, such as the manipulation of bits in a pixel image, or adding of patterns in text punctuation; and
- ➔ Cryptographically signed content, usually employing some form of Public Key Infrastructure, which identifies the signee, and may include additional metadata relating to the origin of the content. This type of watermarking can be difficult to detect even when actively looking for one.

Watermarks, ideally, are forensically verifiable
To best fulfill its purpose, a watermark should be able to withstand forensic verification. That means without any doubt the watermark is there, could not have been caused by chance, and will hold up in whatever court jurisdiction is necessary, to correctly identify the content in whatever way was originally intended. The process of extracting and verifying the watermark must be repeatable and demonstrable. In some contexts it is also desirable for watermarks to be resilient to removal or tampering.

COMMON PURPOSES OF WATERMARKING

There are many reasons to embed watermarks in content. These purposes need to be understood as it will help in determining both the usefulness of watermarking for generative AI content, as well as for better understanding the human rights implications of some types of watermarking. The main purposes of watermarking are:

Content creator marking
The watermark identifies who created the content. Sometimes this is done with the intention of asserting copyright, but in the context of generative AI, it could also identify the AI model or the user who prompted the AI model to generate a particular piece of content.

Content integrity
Some watermarks can be used to validate the integrity of the content. This means the watermark can be used to determine if the content has been altered after being generated.

Content authentication
This watermark would demonstrate that the content was authorized to be created, and is "genuine" or "valid" in some context – for example, for military allies sharing surveillance content with each other. These surveillance images and videos can include watermarks that verify the content as legitimate surveillance from a trusted partner without necessarily revealing the source.

Content limiting
Watermarks can be used to limit interaction with the content. For example, digital rights management (DRM) media players may treat content differently based on the watermarks extracted from audio and video files. Such mechanisms may allow the content to be region-specific or similar.

Content leak detection
Watermarks unique to each recipient of the content are embedded within the content so that if the content is leaked to an unauthorized party, it can be determined which individual leaked it.

V. CAN DIGITAL WATERMARKS BE USED TO IDENTIFY AI-GENERATED CONTENT?

As described above, content created using generative AI systems comes in one of two forms. For large language models (LLMs) and chat systems this will usually be text, whereas for multi-modal generative AI systems, the generated content will usually be in the form of binary files, such as image, animation, audio, or video files. The ability to successfully watermark AI-generated content is highly dependent on the form of the content output.

Many of the current proposals for identifying AI-generated content rely on various forms of digital watermarking – mostly in the form of algorithmic manipulation of output text for large language models (LLMs), but also cryptographic watermarking for binary content. Sometimes proposals combine those watermarks with a log of the users' interactions with the AI system, recording their prompts and potentially even a stored copy of the subsequently AI-generated content itself.

WATERMARKING TEXT-BASED OUTPUTS

To get a watermark into generated text, the words chosen, sentence structure, punctuation, or a combination of those components have to be modulated to contain patterns extractable and interpretable as a watermark. This kind of watermarking relies on algorithmic manipulation rather than cryptographic watermarking, because typically the representation of text content on computers has no extra binary bits where a cryptographic watermark could be stored. While text content can be stored in binary formats where cryptographic watermarks could be added, if that text is cut and pasted to another application, then the cryptographic watermark within the binary would be lost. This makes cryptographic watermarking not viable in most contexts for text content. Likewise, digital watermarking through algorithmic manipulation of text content is not viable in most contexts due to a number of concerns in its implementation.

Digital watermarking is generally incompatible with delivering high-quality text-based AI-generated content

First of all, modulating the text to create the watermark pattern affects the value proposition of the AI model. The value in LLM AI is in its ability to both predict the next most appropriate word in a sentence and to add a randomness factor. Having to generate an identifiable pattern-based watermark undermines both the prediction and the randomness the AI model would otherwise employ, reducing the quality of the output. Additionally, for generative AI systems specifically designed to create text "in the voice" of the user, attempting to overlay a watermark modulation is antithetical to the system's goal and would negatively affect the result.

Longer text length can help solve for quality issues, but uniqueness is still a challenge
The larger the length of generated text content, the easier it is to successfully embed a pattern modulation of the word choices/sentence structure/punctuation. If the generative AI model is outputting a text novel of average novel length, then watermarking is in the realm of possibility. The shorter and more rigidly structured the output, the less possible to successfully add a watermark. For example, if we ask the LLM to give us a haiku on a particular topic, there simply is not enough text nor enough flexibility in word, sentence, or punctuation structure to be able to watermark.

Past a certain threshold for minimum length, we could arrive at statistical certainty that a watermark embedded in the text is sufficiently unique and there is no longer a reasonable probability of a collision (for example, a human authoring a piece of text that contains a similarly identifiable pattern). However, in practical terms, given the most common ways AI-generated text is currently utilized in real-world scenarios, the length of generated text is almost always too short to consistently ensure a watermark can be properly identified and will appear uniquely only where intended.

Digital watermarking in AI-generated text is easily removed and difficult to enforce
We cannot talk about what is possible for watermarking in text without also talking about what is possible in terms of defeating watermarks in AI-generated content. Modulation signatures in AI-generated text are trivial to defeat by putting the content through another AI model that does not use modulation signatures, as the text will be reconstructed with words, sentence structure, and punctuation altered.

It is also important to acknowledge that the mandating of modulation signatures in generative AI content is unenforceable. LLMs that do not add modulation watermarks will exist somewhere on the internet outside any jurisdiction that attempts to enforce generative AI text watermarking. For example, the attempted ban of ChatGPT in Italy did not stop people in the country from using the service, with a reported 400% increase in the use of virtual private networks (VPNs), most likely to circumvent the ban, and use of ChatGPT for coding back to normal within two days of implementation of the ban.

Non-watermark cryptographic provenance is possible but not broadly applicable for text
Cryptographic provenance mechanisms using private and public key pairs, such as PGP/GPG, can be used to identify the creator of text output. Unlike watermarks, the signatures created to prove the provenance are not hidden. The full value of generative AI can be maintained as cryptographic provenance does not alter and create patterns in the words or punctuation chosen. However, cryptographic provenance is not maintained if the text is cut and pasted to a new document or context, and therefore is only useful under limited circumstances.

WATERMARKING BINARY OUTPUTS

The size and nature of binary files – such as images, videos, and audio files – provide significant scope for watermarking. There are many redundant data bits in binary files that can be utilized to hide watermarks, and therefore the full range of cryptographic and non-cryptographic digital watermarking techniques are available for AI-generated binary content. However, choices about where and how digital watermarking is applied carry significant implications for privacy and other fundamental rights. The most robust and appropriate forms of cryptographic provenance in this context utilize Public Key Infrastructure, but as is often the case, the devil is in the details. The primary factor to consider is whether the private keys in the mechanism belong to and identify the AI model itself or the user prompting it.

VI. DOES DIGITAL WATERMARKING IN AI-GENERATED CONTENT RESPECT HUMAN RIGHTS?

Pairing watermarking with logging of users, user-submitted inputs, and/or AI-generated outputs is a serious threat to privacy and free expression

Watermarking in both text-based and binary AI-generated content most often aims to identify either the AI model itself or to more specifically identify the user prompting the output. This can also be paired with server-side logging, where the identity of the person using the AI model, the prompt they submit, and even a copy of the generated output could be logged. This could be keyed to a watermark in the generated content and allow the lookup of who asked the AI to generate the content. However, both the use of cryptographic user keys to identify individuals prompting binary AI-generated content and server-side logging of people's identities, prompts, and outputs for text-based AI-generated content present serious human rights risks and should never be mandatory.

Any implementation of AI-generated content watermarking that utilizes user and prompt logging would undermine user privacy and freedom of expression. Those using LLMs or any other type of generative AI system to create content for private communications with other individuals, or just for their own use, would be subjected to having the thought process in the creation prompt revealed to third parties. This is not in alignment with human rights principles.

As WITNESS Executive Director Sam Gregory explains, "People using generative AI tools to create audiovisual content should not be required to forfeit their right to privacy to adopt these emerging technologies. Personally-identifiable information should not be a prerequisite for identifying either AI-synthesized content or content created using other digital processes. The 'how' of AI-based production elements is key to public understanding; this should not require a correlation to the identity of 'who' made the content or instructed the tool."

Mandating digital watermarking that identifies users could lead to serious human rights abuses. These mechanisms can be used to identify dissidents and suppress dissent – for example, the extrajudicial persecution of an individual creating a parody image of a powerful person.

Whistleblowers are not likely to be impacted by proposals targeting the identification of AI-generated content, but other forms of AI-enabled watermarks may put them at risk

One concern that has been raised in relation to the mandated watermarking of AI-generated content is the possibility of those watermarks being used to reveal the identity of whistleblowers exposing the content. In most scenarios of watermarking, the watermark identifies the model or user prompting the model to generate the content, and the watermark is consistent across all distributed copies of the content. Assuming the whistleblower is not seeking to expose content they generated themselves, watermarking is not likely to put them at risk unless it was somehow made unique to each recipient of the content after it had been initially generated. AI could be specifically utilized to create such unique watermarks, both in binary and LLM-generated text content as well as content created by a human, but this is a whole different purpose of watermarking that does not have crossover with watermarking that identifies the content as AI-generated. Where whistleblowers seek to expose issues with the AI model itself and are relying on their own user prompts to do so, the same concerns addressed in the section above about mandating user-linked watermarks apply.

Digital watermarking can have benefits as a non-mandated feature

There may be times when someone would want to robustly identify themselves as the prompter of the AI, and such functionality could be offered as a feature for use at the user's discretion. In most instances, generative AI systems are not operating by themselves in a vacuum, but as a result of being asked to produce content to a specification set out by a human. Therefore there is some artistry in the use of a generative AI system, either in the prompt alone, or a combination of the prompt and content supplied as an input to the AI model (e.g. pieces of original artwork, existing writing samples, etc.). Given this artistry, the user may want to have the content identifiable as having been prompted by them, establishing credit and attribution for the ways in which they were able to leverage the AI tool to generate a particular output. In such cases, having the option of watermarking the content, according to their specifications, may be useful. Further, if the legal system rules that copyright is held by the user prompting generative AI models, then there will be an incentive for the user to utilize watermarking to identify themselves as the copyright holder.

People using generative AI tools can also benefit from the option to include a digital watermark identifying the AI model in the output. Likewise, AI model developers may benefit from the general use of digital watermarks, including as a mechanism for protecting against false attribution where content did not come from their model. However, these features should always be off by default, and end users of generative AI systems should always retain the final choice of whether to include watermarks in content they have prompted, since it is most likely the user, not the developer, that will be impacted by negative consequences arising from the presence of watermarks.

Mandating digital watermarking that identifies the AI model can still lead to discrimination
Identifying the model, rather than the prompting user, likely carries less potential dangers for the user, but that does not mean it should ever be mandated. As noted above, there is artistry in the use of generative AI, and in some scenarios having the final content watermarked as coming from the AI model would be disingenuous and harmful. People with disabilities and non-neurotypical people have been early adopters of generative AI such as ChatGPT. For example, people with dyslexia have fed their own content to ChatGPT to have the AI make their authored content more understandable to a general audience. The content, with all its intention and substance, was created by a human, not the AI, but reformatted by the AI to be more understandable. People using generative AI in this way would be discriminated against if their content was subsequently labeled as being generated by AI. The dyslexic person, utilizing generative AI to level the playing field, would be punished and robbed of their authorship by mandated watermarking. While it could be argued that more "information-rich" watermarking or labeling could add nuances here to distinguish such accessibility uses from others, this would inevitably lead to the storing of huge amounts of personal information, including historical prompts, and is thus not advisable.

SUMMARY OF USE CASES

AI-generated content	What digital watermarking is possible?	Is this approach desirable?
Short text	Word-punctuation modulation	**No.** Not possible to effectively modulate a signature with acceptable misidentification error rates.
Highly structured text	Word-punctuation modulation	**No.** Not possible to modulate a signature without compromising the structure of the text.
Long loosely structured text	Possible to use modulation signatures	**No.** Modulation signatures are trivial to defeat using a second AI model.
Binary content (static images, moving images, video, sound, etc.)	Cryptographic provenance/watermarking with AI model keys	**Yes, in some contexts,** but never where broadly mandated and always at the discretion of the user. Can protect the AI model developer from fake content purported to have come from the AI model.

AI-generated content	What digital watermarking is possible?	Is this approach desirable?
	Cryptographic watermarking with user keys	**Not in the general case,** never where broadly mandated, and always at the discretion of the user. Privacy concerns. Could protect the rights of the copyright holder.
All of the above	User-prompt-output logging	**No.** Privacy concerns.

VII. PROVISIONAL POLICY RECOMMENDATIONS

We have examined the application of watermarking to both text and binary content generated by AI. We note that while many implementations of watermarking are neither a technical nor a functional fit to solve perceived problems arising from the use of generative AI content, there are some watermarking schemes that could have benefit in limited scenarios. Given how quickly developments in generative AI are moving, and the lack of clear evidence about the real impact of these technologies, we wish to propose three provisional policy recommendations:

1. Governments should not legally mandate, or otherwise recommend through soft law instruments, the default watermarking of AI-generated content by generative AI systems.[2]
2. Any creation and adoption of watermarking features, should they be offered by the AI developer, must be off by default and opt-in at the discretion of the users.
3. For any watermarking feature offered for an AI model, the user must be able to set the level of identification, either identifying the AI model, or the user themselves, and the implications of each must be made clear to the user.

We welcome feedback on these recommendations, and remain open to adapting them in light of emerging evidence and technological developments.

[2] We wish to note that this does not exclude the possibility that AI-generated content could be subject to watermarking or identification obligations in certain use cases, or by certain actors, such as media organizations being obliged to identify or add watermarks to AI-generated content in some scenarios. Any such obligation differs from an obligation to have watermarks embedded by default in all content generated by AI models or applications.

VIII. CONCLUSION: BEYOND SOLELY TECHNICAL APPROACHES

Attempting to rely on watermarking to solve problems arising from generative AI content is an ill-conceived technosolutionist approach to a complex human-centric problem. While it is one tool at our disposal that can be useful in certain contexts, it must be thoughtfully deployed only in the specific instances where the benefits are clear, the risks can be properly mitigated, and the individuals most at risk of human rights harms are placed in full control of their interaction with the technology.

It is more valuable to identify trusted sources

To more fully address the issues arising from a rapidly shifting information landscape, more emphasis should be placed on systems to verify the provenance of human-generated content and the recordings of people, actions, and events that actually occurred, rather than prioritizing identifying content produced through a specific branch of generative AI models. The current approach to watermarking AI-generated content supports the assumption that any content without an AI watermark is real, human-generated, and trustworthy – which is both inaccurate and dangerous. Watermarks – or their absence – will never be able to effectively prove generative AI was not involved in the production of a piece of content, and attempting to use them in this way pulls focus away from other more meaningful pathways to establishing trusted sources.

Sources of content in general should be encouraged, but not forced, to add watermarks. In a future where a large proportion of content is AI-generated, it will be more important to use watermarking or provenance systems to identify human-generated content, rather than focusing on watermarking AI-generated content. This alternative approach of encouraging the normalization of voluntary watermarking of content means the absence of a watermark would become cause for further investigation, rather than relying on the flawed approach of attempting to enforce watermarking only on AI-generated content. We also remain open to the exploration of watermarking or labeling at the point of distribution, holding media outlets and others accountable for identifying where they are knowingly passing on AI-generated content.

Trust must be earned

Technological solutions cannot create trust, but rather can only preserve trust that already exists. Trust in the watermark in any given piece of AI-generated content is only as good as the trust one has in the conduct of the company that developed and maintains the AI model. At its most basic, watermarking of AI-generated content asks us to trust the developers to deploy and maintain reliable, effective, privacy-respecting mechanisms for embedding those watermarks, and that trust cannot be taken for granted. Taking a step back, there is also an underlying ask to trust that watermarks signal attribution properly. However, people have consistently raised concerns about the ways in which developers have trained AI models on human-generated content without the original creators' consent. Labeling content as "originating" from a particular AI model sidesteps crucial questions around authorship and ownership, and enables these companies to evade accountability for their harmful and extractive practices. Given this context, it is essential to take any policymaking decisions around specific technologies in their broader context, and with the history of their development in mind.

Defining the terms of a productive and rights-respecting relationship with generative AI

Ultimately no amount of watermarking will solve the challenges presented by AI-generated content, nor will it relieve us of the need to answer difficult questions around what an appropriate and rights-respecting role for generative AI in our society will look like. We must focus not on the technologies themselves, but on people-centered goals we aim to achieve and the human rights we must uphold, and build upon that foundation to define our relationship with new technologies as they emerge.