Should the United States or the European Union Follow China's Lead and Require Watermarks for Generative AI?

Peter Henderson

Link: https://gjia.georgetown.edu/2023/05/24/should-the-united-states-or-the-european-union-follow-chinas-lead-and-require-watermarks-for-generative-ai/

Abstract*: AI-generated content is becoming increasingly prevalent and realistic, leading to concerns about its potential misuse. China has been a fast mover in regulating AI and recently implemented requirements to label and watermark AI-generated content. But watermarks for text-based generative AI have many nuances so US and EU policymakers should proceed cautiously as they consider implementing similar regulations.*

Content generated by AI is everywhere. You might open social media to see the Pope wearing a stylish puffer coat or Donald Trump being chased through the streets by police—all of which look real, but are AI-generated deepfakes. A teacher might receive an A-grade essay from a student that is AI generated. Or a chatbot might feel so real that it causes real-world harms to people, such as encouraging an individual to take his life to help stop climate change. Creating high-quality, realistic content with AI no longer demands specialized skills or equipment. Advanced machine learning models can be accessed through convenient web interfaces ("It takes a few dollars and 8 minutes to create a deepfake"), consumer-grade laptops, or even a Raspberry Pi. A single actor can use the models for mass-posting to internet forums or targeted influence campaigns. In fact, a single machine learning researcher trained a model on 4chan data and used it to post over 30,000 times to the forum in the span of a few days before it was taken offline. With the increasing abilities of single actors to wield mass influence, it is only logical for governments to identify ways to regulate this technology and mitigate potential risks.

One mechanism to do this is via watermarking: creating specialized mechanisms embedded into AI-generated outputs that would identify them as automated generations. Private entities have long used watermarks to track violations of intellectual property. Film studios watermark movies sent to critics to identify the culprit if the movie leaks online. Stock image companies (e.g., Getty Images) will visibly watermark their photos to identify unauthorized uses. In recent litigation, GettyImages even pointed out that AI systems trained on their images regenerate the company's watermark, proving that their intellectual property appeared in the training data.

The Chinese government has gone a step further by *requiring* watermarking of AI-generated outputs. China's Cyberspace Administration (CAC) issued regulations requiring that generative AI providers mark generated content without affecting user usage (Article 16). If the generated content could mislead the public or cause confusion, then a prominent label must be placed near the content (Art. 17). And it is illegal to delete, alter, or conceal these watermarks or labels (Art. 18). All of these requirements would apply to a wide range of generative AI systems, including text generation, question-and-answering systems, and chatbots (Art. 23).

This is one of the first laws requiring watermarks for generative content. Other countries are considering similar mechanisms for regulating AI-generated content. In fact, in recent US Senate committee hearings, Senator Sinema emphasized the need for transparency in generative AI, including by using watermarks. A key question is whether the watermarking component of the CAC regulation is a good model for tackling the same issues. Watermarking text-based AI-generated content is certainly desirable, potentially helping to identify the prevalence and origin of AI-generated disinformation and more. But when it comes to text-based generative content, like content created by ChatGPT, the picture is not so clear. Text-based watermarks in their current form are easily manipulated, and there are risks that those watermarks and the regulations around them can be misused. As such, policymakers and legislators should proceed cautiously and understand the nuances of text-based AI watermarks.

Problems with Watermarks for Text-Based Content

In image-based systems, watermarks function by adding imperceptible noise to an image (for example, changing every seventh pixel slightly) to create a cryptographic marker. However, text-based watermarks are more difficult to create since there are limited ways to perturb text without changing the underlying meaning.

With a bit of craft, embedding detectable markers in text is possible. In recent litigation, Genius.com sued Google for scraping song lyrics off of its website. To prove this, Genius replaced certain apostrophes in the lyrics on its site with curly and straight apostrophes. This series of curly and straight apostrophes would spell out “REDHANDED” in morse code. According to the lawsuit, this pattern then appeared on Google’s platform, proving it had scraped Genius.com.

Recent work from Kirchenbauer et al., OpenAI, and many others has generated similar approaches for discreetly watermarking text generated by AI systems. This typically works by adjusting the pattern of words that the AI generates so that it creates a unique identifiable signature, like in the case of Genius.com. This signature can be detected later and traced back to the AI model. Ideally, the pattern would be imperceptible and would not affect the model's capabilities (or the user experience). In essence, this would likely comply with Article 16 of the CAC’s new regulations.

There is a catch, however. These text-based watermarks are imperfect. Sadasivan et al. used readily-available open-source paraphrasing systems to overcome text-based watermarks, dropping the accuracy of detecting the watermark. They also claim—with some key assumptions—that as AI system capabilities approach human performance, it will be increasingly difficult to distinguish between the two, even with watermarks. Regardless of whether this result holds, the current state of affairs is that text-based watermarking systems are imperfect, resulting in false positives and negatives. The cat-and-mouse game between new ways to bypass watermarks and new watermarking mechanisms will continue for years to come.

Challenges for Pursuing AI Watermark Regulations in the United States or the European Union

Requiring that companies institute a mechanism to label or watermark AI-generated content is not necessarily harmful on its own. Instituting mechanisms to implement best-in-practice watermarking more broadly could be helpful for downstream research on the effects of models. For example, one might try to measure how content from a given service spreads through internet platforms—though in practice, such estimates would have to incorporate the likelihood of error in watermark detection mechanisms.

How could the United States or the European Union institute such reporting requirements? California's Social Media Accountability and Transparency Act could be viewed as a comparative model. In that law, social media companies must provide aggregate statistical data about the number of posts flagged by content moderation policies, how many users viewed flagged posts, and more. To facilitate the understanding of the risks of AI-generated models, a new law could require that AI companies watermark generated content, and then social media companies could report the prevalence of watermarked content on their platform, adjusting for potential errors. However, such a law would have potential risks.

Significant challenges come with enforceability and penalties on individuals. Watermarking content and providing imperfect tools to detect AI-generated content can encourage institutions to create real harms when detection tools falsely flag human-generated content as AI-generated. Students, for example, might be harmed if they are accused of academic misconduct—potentially affecting their career—if their work is incorrectly flagged by such a detection tool. Researchers even showed that current detection tools "consistently misclassify non-native English writing samples as AI-generated," creating the potential for disparate impact. To be clear, legal systems have already dealt with watermarking and standards of proofs in a wide range of cases, particularly in intellectual property rights settings. But the risk is that there may not be the same due process rights or understanding that text-based watermarking tools are imperfect in private contexts.

False positives are also problematic when adversaries purposefully mimic a watermark, as Sadasivan et al. described. One could imagine a scenario where a country wishes to create certain pretenses: to block a company's product in its country, to accuse a company of election interference, etc. It could mimic a watermark and then use it for its disinformation campaigns, creating a false trail.

Penalizing users for removing or tampering with watermarks (Art. 17 of the CAC regulations) can also be problematic in the text-based setting. How does one reliably prove that a text-based watermark was *removed*? This carries with it all the potential harms of false positives but with even more uncertainty in many cases. In the United States, the First Amendment would also likely make such regulations on individual speech difficult to enforce, if not untenable.

Finally, one might argue that perhaps hidden watermarks aren't necessary and that there should be a simple requirement for labeling AI-generated content prominently. This might be viewed like Article 17 of the CAC regulations—akin to food labels or health warnings on tobacco products. But how should websites handle user-contributed content whose provenance is unknown? Would individual users be liable in such a scenario if they upload AI-generated content? Would the company? If so, how would you prove that they used AI? The result would likely require using AI-detectors to identify watermarks or other patterns associated with AI-generated text, resulting in all of the aforementioned misidentification challenges.

While the story may differ for image-based content, all the uncertainties of text-based watermarking and detection mean that the United States and the European Union should be cautious of following in the CAC's footsteps in adopting watermarking legislation. Watermarking is a useful tool, but its risks must also be considered and mitigated.

Peter Henderson is a JD-PhD (Computer Science, Artificial Intelligence) candidate at Stanford University. He is also an OpenPhilanthropy AI Fellow, adjunct technical advisor for the Institute for Security and Technology, and Graduate Student Fellow at the Regulation, Evaluation, and Governance Lab. His work at the intersection of AI, law, and policy is regularly covered by the popular press, including TechCrunch, Science, The Wall Street Journal, Bloomberg, and more.