ICLR 2025 Workshop on GenAI Watermarking

Going beyond safety in watermarking research and development

Workshop summary

Watermarking involves embedding a hidden signal into digital media like text, images, and audio to establish ownership or ensure authenticity. It has become increasingly important in the age of generative AI. However, despite its growing significance, watermarking in the AI community often gets lost in broader conversations around adversarial robustness, and general security, and safety. We argue that watermarking needs its own dedicated space in AI conferences for discussion and exploration, where researchers can dig deeper into the technical specifics of this field and build on a foundation of research spanning over 20 years.

The aim of this workshop is to bring together experts from academia, industry, policy and from different communities to discuss advancements and challenges in watermarking technologies. The event will facilitate the exchange of ideas and collaborative problem-solving. Topics of interest include, but are not limited to:

- **Algorithmic Advances:** Multi-modal watermarking, model watermarking, dataset tracing and attribution.
- **Watermark Security:** Theoretical results on strong watermark impossibility, black and white-box adversarial attacks, advanced threat models, open-sourced and publicly detectable watermarking, and zero-knowledge watermarking.
- **Evaluation:** Benchmarks for watermarking, perceptual models and watermark-specific quality evaluation metrics, and bias in watermarking robustness.
- **Industry Requirements:** Large bit watermarking, low FPRs, and complexities of deployment in-the-wild.
- **Policy and Ethics:** Dual use, communication to policy makers, and standards.
- **Explainability and Interpretability:** Understanding how watermarks work and their limitations, human oversight and review, and balancing automation with human judgment.

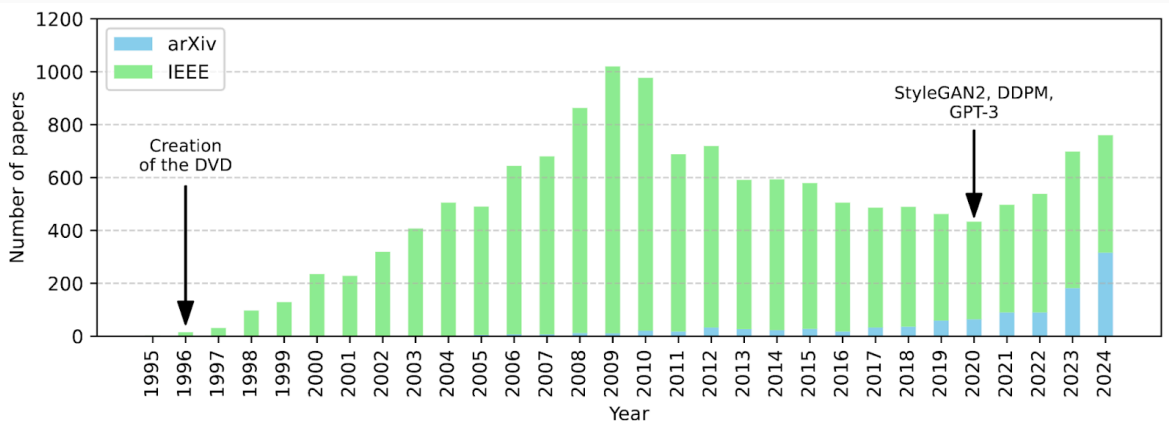
Introduction: Watermarking needs its own place in AI conferences

Watermarking inserts an imperceptible signal into text, image, audio, or a neural network which can be robustly retrieved even after alteration. The best known application of watermarking is now in the context of generative AI, where it helps detect synthetic content, ensuring transparency and responsibility. This area is gaining traction in both academic research and industry, driven by recent regulations on AI, such as the EU AI Act, the US FACT sheet, the Chinese rules on AI, and the Californian bill on Data Provenance. Embedding an invisible trace

into generated content enables orders of magnitude better detection than zero-shot detection, making it a vital tool in the fight against misinformation and deepfakes.

An increasingly important area of research within Generative AI

Watermarking for Generative AI is an increasingly important area of research that deserves a dedicated space for discussion and exploration. The number of papers on this topic is growing rapidly, for instance, papers submitted to ICLR have risen significantly over the years: from 7 in 2023 to 23 in 2024, and a substantial jump to 61 in 2025. More broadly, we notice a clear increase in the number of papers since 2020, explained by the new applications in AI.



However, it often gets lost in the broader conversations around adversarial robustness, security, and safety. Many questions surrounding watermarking remain unsolved, particularly when it comes to how to increase its robustness, how to make it secure, or how to evaluate it properly. Moreover, while some may view watermarking uniquely for safety, its applications extend far beyond these areas. Notably, it plays an important role in intellectual property (IP) protection of models and data via imprinting watermarks in training data, or backdoors that allow to claim ownership, for instance.

Given the significance of watermarking and its unsolved problems, it is essential to bring together the community to share their research and discuss the latest developments and challenges in watermarking – even more so as the techniques and challenges used in watermarking are closely linked even when their fields of application vary.

Benefits of one dedicated community

Watermarking involves complex challenges, which require deep knowledge of AI (image/ audio generative modeling, LLMs, detection, segmentation), statistics and probability theory (detection tests), and computer science in general (what does it mean to be imperceptible, how to embed a cryptographically secure message, etc.). Watermarking has been traditionally considered separate from ML, and applied directly to already existing content. However, watermarks for modern generative AI are now intertwined with the generation process (for example when modifying logit distributions or internal representations) and these methods require technical knowledge of Deep Learning common at ICLR.

Dedicated technical discussions are also essential for aligning watermarking practices with evolving global regulations. Dedicated discussions are needed to address complex questions such as the ownership of watermark detectors, accountability for false positives, the legal implications of watermark removal and forgery, or even what to do with the watermark detection/ All these topics are regularly overlooked in papers and regulations. This workshop promotes the development of standards suited for the rapid growth of the industrial applications of watermarking in AI.

A focused workshop in an AI conference finally enables deeper collaboration and could notably attract talents and experts from the digital forensics, information security and data hiding communities, which traditionally submit at signal processing or specialized conferences like ICCV, ICASSP, ICIP, WIFS, etc.

Watermarking from a policy perspective

Labeling AI-generated content is becoming a central theme in global AI policy discussions. The EU AI Act, for instance, mandates AI providers to label content such as text, audio, video, and images in a machine-readable format. Similarly, the G7 encourages the development of watermarks in its International Guiding Principles on Artificial Intelligence, while China has implemented its own requirements for labeling AI-generated content. In the US, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence also emphasizes the importance of labeling synthetic content, including through watermarking.

As watermarking becomes more central to policy debates, fostering a thorough exchange between the research community and policymakers is crucial to ensure alignment between policy requirements and technical advancements. Researchers can play a key role in supporting the global community's needs for transparency in AI systems, but several critical questions remain. Core to the debate will be the state of watermark research and application to models and their reliability to be encoded in regulation. It will be critical to set standards as to who has access to watermark detection and who is liable in the case of false positives, i.e., content being labeled as AI-generated when it is not. The possibility of watermark removal will require discussions between policymakers and developers.

Relationship to other workshops and other watermarking venues

Multiple workshops at AI conferences (ICLR, Neurips, ICML, ACL* and CVPR) are now focusing on trustworthy AI. For example:

- NeurIPS 2024 featured Safe Generative AI, Towards Safe & Trustworthy Agents, and Socially Responsible Language Modelling Research, and the Image Watermarking Competition
- ICLR 2024 hosted Secure and Trustworthy Large Language Models, Reliable and Responsible Foundation Models, and Privacy Regulation and Protection in Machine Learning,

- ICML 2024 hosted Next Generation of AI Safety, Trustworthy Multi-modal Foundation Models and AI Agents and GenLaw

The surge of interest in trustworthy AI has led to a proliferation of workshops at top AI conferences in which the topic of watermarking is continuously present. However, these events often prioritize breadth over depth, covering a wide range of topics without a long term dedicated focus on the technical specifics that this field demands. Meanwhile, established watermarking and forensics venues have a rich history of technical expertise, but may be missing out on the opportunities and challenges presented by the latest developments in generative AI.

- IH&MMSec Information Hiding and Multimedia Security Workshop is a +20 years old workshop addressing topics like forensics, steganography, and watermarking.
- IWDW The International Workshop on Digital Watermarking is a +20 years old workshop that focuses on multimedia security, digital forensics, and watermarking.
- WIFS IEEE International Workshop on Information Forensics and Security is a 16 years old workshop on topics like digital forensics, privacy, cybersecurity, and watermarking.
- MWSF Media Watermarking, Security, and Forensics is a 15 years old conference focusing on advancements and discussions around watermarking, security measures, and forensic techniques for media content.

Our proposed workshop seeks to address this disconnect by establishing a venue that brings together researchers from both communities to share knowledge, expertise, and best practices in watermarking.

Workshop program

A technical-first workshop dedicated to watermarking

This workshop puts a particular emphasis on in-depth technical advancements and applications in Generative AI, data tracing, and IP protection. Publications to the workshop may cover research areas separated by the following high level topics:

Algorithmic advances (e.g., better robustness, capacity or speed)

- Multi-modal watermarking, applied to image/audio/text [Cox et al. 1997, Zhu et al. 2018, Kirchenbauer et al. 2023, San Roman et al. 2024]
- Generative watermarking [Venugopal et al. 2007, Kirchenbauer et al. 2023, Wen et al. 2024, Fernandez et al. 2023]
- Model watermarking [Uchida et al. 2017, Adi et al. 2018]
- Dataset tracing and attribution [Sablayrolles et al. 2020, Asnani et al. 2024]
- Intellectual property protection (e.g., protect models against distillation) [Zhao et al. 2023, Sander et al. 2024]
- Industry-grade watermarking: MVP for watermarks, large bit size watermarks, high-efficiency watermarks, and robustness to compression [SynthID]

- New applications: detecting deepfakes, authenticity verification, and watermarking in emerging domains (e.g., 3D models, virtual reality, augmented reality [Li et al. 2023])
- Watermarking and cryptography [Cox et al. 2006, Christ et al.2024]

Watermark security

- Theoretical results on strong watermark impossibility [Zhang et al. 2024]
- Black and white-box adversarial attacks [Zhao et al. 2023]
- Advanced threat models (e.g., API jailbreaking), and countermeasures
- Open-sourced and publicly detectable watermarking [Fairoze et al. 2023]
- Zero-knowledge watermarking [Adelsbach et al. 2001]

Evaluation

- Benchmarks for watermarking [An et al. 2024]
- Perceptual models & Watermark-specific quality evaluation metrics (beyond usual domain-specific metrics, e.g. PSNR) [Czolbe et al. 2020]
- Bias in watermarking robustness: modality-dependent bias (code/math, text generation, language, etc.), dataset bias, and diversity [Lee et al. 2024]

Industry requirements and complexities of deployment in-the-wild

- Large bit watermarking, low FPRs
- Collaborative watermarking (multi-stakeholder approaches to watermark development and deployment) [Martínez et al. 2018]

Policy and Ethics of Watermarking

- Dual use of watermarking
- Communication to policy makers: standards, interoperability, transparency, and regulatory frameworks for watermarking [Fernandez et al. 2024]
- Explainability and interpretability: understanding how watermarks work and their limitations, balancing automation with human judgment [Leibowicz 2023]

Invited speakers and panelists

We will have a diverse group of researchers from both academia and industry speaking at our workshop, each bringing their unique perspective on the latest developments in watermarking. Our goal is to bring established researchers in both traditional and newer watermarking, as well as industrial and policy actors to broaden the discussions. Our policy discussions will feature a range of global perspectives as well as local perspectives from South East Asia. We'll be sharing the titles and abstract of their talks before the event.

Academic perspectives

- Mauro Barni (confirmed): professor at the University of Siena, with over 20 years of research experience in digital image processing and information security and over 100 publications on watermarking. Mauro has made significant contributions to the field of digital watermarking and has been recognized as a fellow member of the IEEE and AAIA.
- John Collomosse (confirmed): professor of Computer Vision and AI at the University of Surrey, and Principal Scientist at Adobe Research, where John chairs cross-industry

task forces on these topics and leads research for Adobe's Content Authenticity Initiative (CAI).

- Claire Leibowicz (pending) is the Head of the AI and Media Integrity Program at the Partnership on AI, where she addresses the challenges posed by synthetic media and deepfakes. A doctoral candidate at the University of Oxford studying AI governance and synthetic media, Claire has a background in psychology and computer science and has worked extensively on issues related to AI, media, and society. Claire is a sought-after commentator on AI, media, and society, featured in outlets such as The New York Times, MIT Tech Review, and WIRED.
- Tom Goldstein (confirmed): professor of computer science at Maryland University. Tom has recently made significant contributions to the AI safety and watermarking research, and notably won an outstanding paper award for “A Watermark for Large Language Models” at ICML 2023.

Industry perspectives

- Zohaib Ahmed (confirmed): CEO of Resemble AI, a voice cloning platform that generates hyper-realistic audio replicas from minimal voice samples, enabling various applications such as personalized messaging, voiceovers, and virtual assistants.

Policy perspectives

- Melissa Omino (pending): Director at the Center for Intellectual Property and Information Technology Law (CIPIT) specializing in Intellectual Property and Trade Law, with a research focus on IP provisions in international trade agreements involving Global South countries. Dr. Omino is an Advocate of the High Court of Kenya, Commissioner of Oaths, and Notary Public.
- Adina Yakefu (confirmed): China AI expert at Hugging Face, helps to build the Chinese open-source AI community and provides insights and comments on the Chinese AI policy.
- A representative of Singapore AISI (pending - in talks): The Singapore's AI Safety Institute, which was established in May 2024, aims to do research on AI safety to provide science-based input for AI governance.

Tentative schedule

Morning

09:00 – 09:15	Introduction and opening remarks
09:15 - 09:45	Invited Talk 1: *Watermarking Then and Now**
09:45 - 10:15	Invited Talk 2: *Emerging Trends in Watermarking: Recent Advances and Future Directions**
10:15 - 10:30	Oral 1
10:30 - 10:45	Oral 2
10:45 - 11:00	Coffee Break
11:00 - 12:00	Poster Session 1
12:00 - 13:30	Break

Afternoon

13:30 - 14:30	Poster Session 2
14:30 - 15:00	Invited Talk 3: *Watermarking against Voice Cloning - An Industrial Perspective**
15:00 - 15:30	Invited Talk 4: *Watermarking Techniques for Content Authenticity**
15:30 - 16:00	Invited Talk 5: *A Policy Perspective on AI Transparency and Content Provenance**
16:00 - 16:15	Coffee break
16:15 - 17:15	Panel Discussion: *The Role of AI Watermarking Researchers in Providing Actionable Policy Decisions*
17:15 - 17:30	Closing Remarks

**Please note that the titles listed are preliminary and subject to change, but they do reflect the general area of expertise and topics that our speakers will be addressing.*

Organizers and biographies

Lucie-Aimée Kaffee is EU Policy Lead & Applied Researcher at Hugging Face. She works on the intersection of policy and AI technology and is highly involved in the discussions around the EU AI Act in Brussels. Her research aims to harness AI for supporting online communities with an emphasis on lower-resourced language communities. She holds a PhD from the University of Southampton and postdocs from the University of Copenhagen and HPI Potsdam.
She was a co-organizer of the first *Wikipedia for NLP* workshop at EMNLP 2024, *Knowledge Graphs and LLMs* workshop at ACL 2024, *Wiki-M3L: Wikipedia and Multi-Modal & Multi-Lingual*

Research workshop at ICLR 2022, the tutorial *Wikimedia Data How-To* at ICWSM 2024, and organized the *Wikidata Workshop* 2020 - 2023.

Email: lucie.kaffee@huggingface.co
Google Scholar: https://scholar.google.com/citations?user=xiuGTq0AAAAJ&hl=de
Homepage: https://luciekaffee.github.io/

Hady Elsahar is a research scientist at Meta FAIR. His research specializes in developing safe multilingual, multimodal generative models. Central to his efforts is a focus on removing biases and ensuring safety in AI, balancing cutting-edge innovation with ethical responsibility. He notably developed the audio watermarking algorithm used in FAIR's public research demos like Audiobox and Seamless, and in production at Meta (e.g., in the Instagram auto-dubbing feature). Before that, he was a research scientist at Naver labs Europe and holds a PhD from Université de Lyon.
He was a co-organizer of workshops in ICLR2021, ICLR2022, ICLR2023, EACL 2021 and NAACL 2021.

Email: hadyelsahar@meta.com
Google Scholar: https://scholar.google.com.eg/citations?user=SbcM6bsAAAAJ
Homepage: https://www.hadyelsahar.io/

Pierre Fernandez is a research scientist at Meta FAIR Paris. His research focuses on content protection in machine learning, particularly watermarking generative models to ensure AI-generated content is traceable and identifiable. During his PhD at FAIR and Inria Rennes, he notably developed the image watermarking algorithm used for Imagine (Meta's image generation assistant), as well as influential papers on watermarking. He holds degrees in computer science and mathematics from École polytechnique and from Paris-Saclay University.

Email: pfz@meta.com
Homepage: https://pierrefdz.github.io/
Google Scholar: https://scholar.google.com/citations?user=osCX1YQAAAAJ

Teddy Furon is a director of research at Inria, where he leads a team on ML safety and security. He has more than 20 years of expertise in digital watermarking, fingerprinting (traitor tracing), and security. He holds a PhD in signal and image processing from the Telecom ParisTech. He has also worked as a research engineer at THOMSON multimedia and at the Security Labs of Technicolor from 2008 to 2010.

Email: teddy.furon@inria.fr
Homepage: https://people.rennes.inria.fr/Teddy.Furon/website/Welcome.html
Google Scholar: https://scholar.google.fr/citations?user=aLUbWzAAAAAJ

Jonas Geiping is a researcher at the ELLIS Institute Tübingen and Max-Planck Institute for Intelligent Systems, where he leads a joint research group for safety and efficiency-aligned learning. His research focuses on understanding and improving the safety and efficiency of modern AI systems, particularly in natural language processing. His research on watermarks for LLMs has received a best-paper award at ICML 2023. He has previously worked at the University of Maryland and the University of Siegen, where he holds his PhD from.

Email: jonas.geiping@gmail.com
Homepage: https://jonasgeiping.github.io/
Google Scholar: https://scholar.google.de/citations?user=206vNCEAAAAJ

Nikola Jovanović is a PhD student at the SRI Lab at ETH Zurich, advised by Prof. Dr. Martin Vechev and Prof. Dr. Florian Tramèr. His research is centered around safe and trustworthy AI, with the current focus on watermarking for large language models, where he aims to understand and help bridge the gap between research and practical applicability. He has previously also worked on topics related to ML fairness, (certified) adversarial robustness, and privacy of federated learning.

Email: nikola.jovanovic@inf.ethz.ch
Homepage: https://www.sri.inf.ethz.ch/people/nikola
Google Scholar: https://scholar.google.com/citations?user=nE1czKQAAAAJ

Anticipated audience size & Plan to get an audience

Our goal is to reach more than 30 paper submissions and a target participation of 100 researchers and practitioners in the field of GenAI watermarking between virtual and in-person participation. To achieve this, we will implement the following dissemination strategy:

Pre-Workshop Promotion (Dec 2024 - Feb 2025)

1. Call for Papers: We will issue three calls for papers:
    - Dec 2, 2024: First call for papers upon acceptance
    - Jan 23, 2025: Second call for papers, aligning with ICLR decision notifications targeting authors of watermarking related papers.
    - Feb 15, 2025: Last call for papers, two weeks before the deadline
2. Community Outreach: We will reach out to collaborative AI research and OSS communities that the OC has close ties with, including Hugging Face, Deep Learning Indaba, Eluther AI, and Stable Diffusion community.
3. Invited papers and Targeted Promotions: We will target research groups and publication authors from previous and concurrent events, ensuring that our workshop reaches the most relevant audiences.

To promote the workshop, we will utilize a range of channels, including:

- Relevant academic and industry mailing lists
- Targeted email announcements to organizer institutions, the wider privacy research community, and traditionally underrepresented institutions
- Social media platforms (Twitter, LinkedIn) through both company accounts and personal networks of organizers
- Personal invitations to authors of relevant papers at other conferences
- Media coverage opportunities through contacts with journalists, such as Melissa Heikkilä from the MIT Tech Review.

Diversity commitment

We commit to diversity and inclusivity, ensuring that a range of perspectives and backgrounds are represented. We believe that a diverse mix of ideas enriches discussions and drives innovation. To achieve this, our focus includes:

- **Diverse policy perspectives:** We are inviting speakers and panelists who can share insights on intellectual property policies from different regions, including Singapore, South East Asia, Africa, and European perspectives. This allows us to explore the unique challenges and opportunities in these regions and how research can be tailored to them.
- **Balance between industry and research:** Our speaker lineup includes 2 industry experts from Adobe and Resemble AI who will share their practical experiences and insights, as well as several researchers from academia who will present the latest advances in watermarking.
- **Old and new techniques:** with speakers like Mauro Barni sharing historical overview and insights on the development of the watermarking field.
- **Diversity in speakers and organizers:** We did our best to create a diversity in terms of ethnicity, gender, and seniority among our speakers and organizing committee. Our organizers and speakers come from a variety of institutions, including open source organizations, universities, and industry leaders, such as Meta, Hugging Face, Inria, ETH, and Tübingen.
- **New and returning organizers:** Our organizing committee is a diverse group of researchers with varying levels of experience. We have a mix of returning organizers, including Lucie-Aimée Kaffee and Hady Elsahar, who have previously co-organized several workshops at top conferences such as ICLR, EMNLP, and ACL as well as new organizers who bring fresh perspectives and expertise to the table. In terms of seniority, our team includes both senior researchers who have over 20 years of experience in digital watermarking, as well as Phd researchers.

Virtual access to workshop materials and outcome

1. **Workshop Website**: We will create a dedicated workshop website as a central hub for information and updates. This webpage will include:
    - Instruction to authors and suggested example papers
    - Accepted papers and abstracts of each talk
    - Schedule of talks and poster sessions
2. **Slack Channel for Q&A**: We are setting up a Slack channel for attendees to ask questions related to organization.
3. **Collect Panel Questions in Advance**: To facilitate discussion during the panel session, we will use online platforms to collect questions from attendees before and during the workshop. This will allow us to prioritize the most relevant and interesting questions and make the most of our panelists' expertise.

4. **Pre-Workshop Materials**: Before the workshop, we will make available on the workshop webpage:
    - Accepted papers and abstracts of each talk
    - Schedule of talks and poster sessions
5. **Recorded Talks and Posters**: After the workshop, we will add direct links to the ICLR website's recorded talks page to ease finding them.
6. **Video Presentations for Posters**: To enhance the visibility of authors' work beyond the physical attendance, we will encourage accepted poster authors to submit an additional 10-minute YouTube video presenting their work. These videos will be uploaded to the workshop website and organized into a single playlist, allowing researchers to browse all videos together in one go. This approach has been successful in our previously organized workshops by some of the OC members.