

2025 Activity Report

RESEARCH CENTRE: Inria Centre at Rennes University

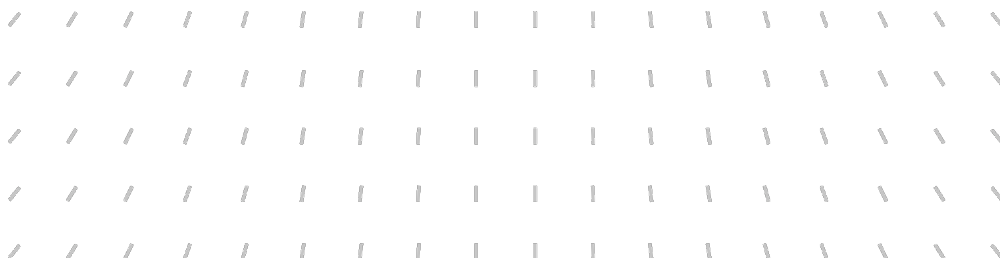
IN PARTNERSHIP WITH: Université de Rennes

Project-Team

ARTISHAU

ARTificial Intelligence: Security, truthHfulness, and
AUDit

In collaboration with Institut de recherche en informatique et systèmes aléatoires
(IRISA)



Project-Team ARTISHAU

Creation of the Project-Team: 2024 October 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

Keywords

Computer sciences and digital sciences

- A4. – Security and privacy
- A9.3. – Signal processing
- A9.11. – Generative AI
- A9.14. – Evaluation of AI models
- A9.17. – Cybersecurity and AI

Other research topics and application domains

- B6.5. – Information systems
- B9.10. – Privacy

Contents

Project-Team ARTISHAU	1
1 Team members, visitors, external collaborators	5
2 Overall objectives	6
2.1 Context	6
2.1.1 AI is scary	6
2.1.2 Grid for reading	7
2.2 Definitions	7
2.2.1 Definition of machine learning security	7
2.2.2 Definition of audits of decision-making algorithms	8
2.2.3 Definition of manipulation of information	8
2.3 Conclusion on the objectives	9
3 Research program	9
3.1 Research axis A: Security of Machine Learning	9
3.1.1 Challenge #1: Protection of the training set \mathcal{D}	9
3.1.2 Challenge #2: Protection of the model θ	10
3.1.3 Challenge #3: Protection of the testing data x	10
3.1.4 Priorities	10
3.2 Research axis B: Audit of black-box AIs	11
3.2.1 Challenge #4: Decidability and conditions for accurate audit tasks	11
3.2.2 Challenge #5: Tracking models evolutions	11
3.2.3 Challenge #6: Auditors coordination and stealth	11
3.2.4 Priorities	12
3.3 Research axis C: Threats from generative models	12
3.3.1 Manipulated data detection	12
3.3.2 Challenge #10: Planting watermarking in generative models.	13
3.3.3 Priorities	13
4 Application domains	13
4.1 Security, cybersecurity, and defense applications	13
4.2 Compliance with coming regulations	14
5 Highlights of the year	14
5.1 Awards	14
5.2 Societal impact	14
6 New results	14
6.1 Security of machine learning	14
6.2 Audit of black-box AIs	17
6.3 Threats from generative models	18
6.4 Miscellaneous	19
7 Bilateral contracts and grants with industry	20
7.1 Bilateral contracts with industry	20
8 Partnerships and cooperations	21
8.1 International initiatives	21
8.1.1 STIC/MATH/CLIMAT AmSud projects	21
8.2 International research visitors	21
8.2.1 Visits of international scientists	21
8.2.2 Visits to international teams	22
8.3 National initiatives	22

8.4	Public policy support	23
9	Dissemination	24
9.1	Promoting scientific activities	24
9.1.1	Scientific events: organisation	24
9.1.2	Scientific events: selection	24
9.1.3	Journal	24
9.1.4	Invited talks	25
9.1.5	Leadership within the scientific community	25
9.1.6	Scientific expertise	25
9.1.7	Research administration	25
9.2	Teaching - Supervision - Juries - Educational and pedagogical outreach	26
9.2.1	Teaching	26
9.2.2	Supervision	26
9.2.3	Juries	27
9.2.4	Educational and pedagogical outreach	27
9.3	Popularization	27
9.3.1	Specific official responsibilities in science outreach structures	27
9.3.2	Productions (articles, videos, podcasts, serious games, ...)	27
9.3.3	Participation in Live events	27
9.3.4	Others science outreach relevant activities	28
10	Scientific production	28
10.1	Major publications	28
10.2	Publications of the year	28
10.3	Cited publications	30

1 Team members, visitors, external collaborators

Research Scientists

- Teddy Furon [Team leader, INRIA, Senior Researcher, HDR]
- Eva Giboulot [INRIA, Researcher]
- Erwan Le Merrer [INRIA, Researcher, HDR]

Faculty Members

- Ewa Kijak [UNIV RENNES, Professor, from Sep 2025, HDR]
- Ewa Kijak [UNIV RENNES, Associate Professor, until Aug 2025, HDR]

Post-Doctoral Fellow

- Ryan Webster [INRIA, Post-Doctoral Fellow]

PhD Students

- Paul Chaurand [INRIA, from Sep 2025]
- Timothee Chauvin [INRIA]
- Adele Denis [INRAE]
- Virgile Dine [INRIA]
- Gautier Evennou [IMATAG, CIFRE]
- Pierre Fernandez [FACEBOOK, CIFRE, until Feb 2025]
- Jade Garcia Bourrée [INRIA, until Sep 2025]
- Enoal Gesny [INRIA]
- Augustin Godinot [INRIA, from Nov 2025]
- Augustin Godinot [UNIV RENNES, until Oct 2025]
- Louis Hemadou [SAFRAN, CIFRE, until Oct 2025]
- Chloé Imadache [INRIA]
- Quentin Le Roux [THALES, CIFRE, until Oct 2025]
- Gurvan Richardeau [PEReN, CIFRE]

Interns and Apprentices

- Paul Chaurand [ENS Rennes, Intern, from Feb 2025 until Jul 2025]
- Malo De Hedouville [CNRS, Intern, from Jun 2025 until Aug 2025]

Administrative Assistant

- Loic Lesage [INRIA]

Visiting Scientist

- Isabela Borlido Barcelos [GOUV BRESIL]

External Collaborator

- Charly Faure [DGA-MI]

2 Overall objectives

2.1 Context

2.1.1 AI is scary

“A picture is worth a thousand words” Confucius, 5th century BC. “I believe only what I see” Saint Thomas at the resurrection of Christ. “The weight of words, the shock of photos” slogan of a well-known french weekly magazine. All these quotations show the importance of images in our civilisation. Every article in the press and every post on social networks is illustrated with photos. But today, Artificial Intelligence is disrupting our relationship with images. Dall-E in 2021, Stable diffusion or MidJourney in 2022, these AIs produce ultra-realistic images. Humans can no longer tell the difference between generated images and real ones. The same applies to text with Large Language Models, such as ChatGPT. AI is nowadays at the root of a crisis of confidence in multimedia data, which is the main content of social networks. These Artificial Intelligences are powerful tools for creating deepfakes, fakenews, and disinformation on a massive scale. This is a major problem in Influence Warfare.

But it’s just one of the many dangers of Artificial Intelligence. Artificial Intelligence is also about analysing data to make decisions. These systems are now in the wild, serving populations in most parts of their online interaction (robots, online curation such as recommendation, pricing or ranking algorithms, self-driving cars, text or image generation, . . .). These systems have demonstrated incredible performances and the market of AI-based systems is awaited to worth hundreds of billions US dollars in the coming years¹. The word ‘performances’ here encompasses the *primary goals* of the algorithms: their ability of performing a given task (accuracy -classification, average error -regression, perplexity -generation), but also their speed (low latency, high throughput), and their frugality (in terms of training samples, memory footprint, and electric power). Yet this undeniable success is hampered by a growing lack of trust in AI and machine learning. These algorithms are scary, and this mixed feeling is fueled by the lack of numerous *secondary properties*: fairness, explicability, plausibility, safety, transparency, truthfulness . . .

Some believe that these problems will be resolved through legislation. Europe has just passed the EU AI Act to ensure that AI deployed in Europe will be safe, transparent, traceable, free from discriminatory bias and environmentally responsible. The Biden-Harris administration states that AI can in no way usurp humans. Recently, the UK backed down: an AI cannot be trained on copyright-protected content without the agreement of the copyright holders. Finally, AI is no longer a simple algorithm that can be benchmarked by measuring its *primary goals* (the probability of giving a correct result, its processing speed, its memory footprint on a GPU). Regulations also impose seemingly *secondary characteristics* on AI that are indeed crucial to its acceptance in society. But what good is regulation if it is not accompanied by technical means of control? This is all the more difficult when we think of the AIs of pure players trained in secrecy and deployed on their clouds. These are veritable black boxes accessible via APIs. Are they compliant with recent legislation? This is the challenge of AI certification where a dishonest AI provider conceals the non-compliance to these standards by preventing or deluding an audit of his system.

Last but not least: AI is increasingly used in critical applications such as cybersecurity where by assumption there exists a malicious person willing to delude the system. Can we trust this new tool? Wouldn’t it be a good idea to check the level of security and privacy of AI before using it in critical applications such as cybersecurity? An example of a vulnerability: if the attacker can modify training data, he can build a backdoor into a model. In other words, the model learned from this poisoned data behaves as expected, but the attacker can control this model in the sense that by modifying the test data in the same way, he can make the model say what he wants.

¹International Data Corporation (IDC)

This is a real problem because there are many models that have already been learned and are available as *open-source* software. Can we trust the integrity of these models? Who can guarantee that they do not contain a backdoor? As another example, machine learning models tend to retain a memory of training data. So, knowing a model, an attacker can, to a certain extent, predict whether a given piece of data was used at training time, and sometimes even reconstruct training data. This represents a threat to data confidentiality and a violation of privacy if the data is sensitive with respect to the GDPR law.

It is then of a societal interest to develop methods to *secure models by design at training time, to audit the compliance of models already deployed online, and to scout out where AI is thwarting our trust. ARTISHAU targets the secondary properties of machine learning algorithms in a hostile environment due to the presence of an attacker.*

2.1.2 Grid for reading

The following list of criteria is here to summarize the main ideas and supporting details thrown in this introduction.

- **Type of AI.** The introduction mentions two types of Artificial Intelligence. *Decision-making AI* analyzes data to take decision, whereas *generative AI* synthesizes data.
- **Access to the model.** One speaks about *white box* when all the internals of the model under scrutiny (*i.e.* architecture, weights and biases) is disclosed. This means that the model is fully reproducible in the lab. *Black box* defines scenarios where one can only query the model and observe its output. The access is granted through an API (a.k.a. MLaaS – Machine Learning as a Service) or the model is embedded in IC (a.k.a. ML-on-Chips).
- **Security issues.** There are of different natures. Either issues stem from intrinsic vulnerabilities of machine learning, or they are posed by a malevolent use of AI. The latter is especially true for generative AI.
- **Goals.** Our goals range from the control and certification of AI (*i.e.* audit) to the publications of recommendations and design of defenses.

2.2 Definitions

2.2.1 Definition of machine learning security

Grid for reading: *Decision-making – Intrinsic vulnerabilities – Design of defenses.*

Revealing the intrinsic security of machine learning is of utmost importance. The main problem of machine learning is that it works great (provided a fair amount of training data and computing power). It works great even under perturbations or light distribution drifts. Yet, this robustness gives a false sense of security making us believe that AI is almighty. Indeed, generalization, robustness, and security are different concepts often confused:

- Generalization is the ability to operate as expected on *unseen data*.
- Robustness is the ability to operate as expected on *noisy data*.
- Security is the ability to operate as expected on *data deliberately perturbed by attackers*, or at least to sense that the conditions are not met for operating safely.

The intention of the attackers but especially the quest of efficient attacks leveraging their knowledge of the targeted system makes a major difference between security and robustness. Recent literature witnesses a flurry of dangers in machine learning. Claiming to be a team expert in machine learning security implies to contribute on the study of all these threats. Yet, we shall first organize this fuzzy ensemble of use cases into a simple vision for the security of machine learning. This vision is the skeleton of the team-project ARTISHAU.

Machine Learning amounts to learn a *model* θ from *training data* \mathcal{D} and to apply this model onto some *testing data* x to output an inference y . Training data \mathcal{D} , model θ , and testing data x are assets needing protection so that the inference y can be trusted. Needing protection means defending some cardinal values which are, according to us, *Confidentiality*, *Privacy*, and *Integrity*.

This makes 3 assets \times 3 values equal 9 scenarios, which are detailed in the sequel of the proposal. On top of this, there is also the huge diversity due to the nature of the data (image, video, audio, or categorical data like text, user profile. . .) and due to the X -learning framework with $X \in \{\text{supervised, unsupervised, continuous, meta, few shot, federated, . . .}\}$.

2.2.2 Definition of audits of decision-making algorithms

Grid for reading: *Decision-making AI – Black box – Certification.*

Computer scientists and engineers are used to design and develop algorithms that process information and that can have an important impact on society. For fine tuning these, developers operate a controlled feedback loop on data fed as inputs to the algorithm, and the corresponding algorithm results (output accuracy for instance). Considering an exterior viewpoint (the viewpoint of users or regulators), that observes or audits the behavior of remote algorithms is less frequent. A so-called black box approach on algorithms can be dated back to Moore's tests black box automata in 1956 [57]. Relatively recent and sporadic works instead placed this viewpoint at the service of algorithmic auditing, in order to allow users to gain some understanding on the algorithmic decisions they are facing. In particular, these nascent forms of algorithmic audits also constitute a prerequisite to enable platform regulation: if a regulation entity wants to enforce some behavior, means for verification are mandatory (as captured by the Russian proverb *trust, but verify*).

In this context, the design of algorithmic audits aims at producing targeted and reproducible methods to verify predicates, or discover some specific behavior of the decision-making algorithm under scrutiny. The link to the security objective of the team should here be clear: if one observer (here an auditor) can extract information from a remote algorithm, then it may constitute a security issue for the platform. This is why the audit axis will benefit from the research done in the other project axes, to fully understand the potential offensive/defensive nature of the task.

2.2.3 Definition of manipulation of information

Grid for reading: *Generative AI – Malevolent Usage – Design of defenses.*

The definitions in sections 2.2.1 and 2.2.2 hold for algorithms analyzing data and taking decisions. Algorithms generating synthetic data are also a source of threats especially in IT fight for influence (in French *Lutte Informatique d'Influence* - L2I). AIs generating or editing images [69, 59, 63, 41] are a great tool for crafting deepfakes, and so are AIs generating texts [31, 67] for fake news.

Whether it is destabilization operations organized on social networks or manipulation of information for the purpose of influencing public opinion, the dangers in the informational sphere are increasingly visible. The attacks in the realm of influence in cyberspace are more and more harmful because AI renders them automatic and massive.

Among numerous examples, we can cite the deepfake video of Facebook CEO Mark Zuckerberg created and shared online on June 2019. The video showed Zuckerberg making comments about Facebook's dominance and power over users' personal data. More recently, AI-generated videos of people expressing support for Burkina Faso's new military junta have recently been shared on social networks, in what may be an attempt to spread pro-military propaganda. The images were generated thanks to *Synthesisia*, a platform that allows to create videos from written texts. Emmanuel Macron was put in several situations by the AI *Midjourney*: in the middle of flames in a Parisian street, fighting with demonstrators or collecting waste.

Apart from artificially generated misinformation, the image repurposing is also widely used for manipulation purposes. As example, a series of photos posted on Facebook in 2021 purport to show French soldiers exploiting Mali's gold resources, rather than fighting alongside the Malian army. But in reality, none of these images were taken in Mali and these soldiers are not French. These pictures are from seizures of gold bullion shipments by U.S. soldiers in Iraq, and they were shot in 2003.

2.3 Conclusion on the objectives

Security is considered in a broad spectrum covering confidentiality, privacy, integrity (first research axis), compliance to regularisation laws and standards (second research axis), and generation of disinformation (third axis). The following section draws a map of this domain composed of three research temporalities.

The short term objective of ARTISHAU is to explore the frontiers in security issues of AI and especially machine learning by developing attacks and defenses and pinpointing who wins the game under which conditions. The main difficulty is to pave all the ten challenges hereafter grouped into three research axes. This exhaustive covering is mandatory to achieve the ambition of ARTISHAU which is to become a team of experts in the security of machine learning at large.

The middle term objective is to make connections between the challenges. Theoretical results will outline *common* key factors impacting the feasibility of attacks and defenses. Practical studies will look at a given defense mechanism from different perspectives. It is paramount to assess that a defense against one attack is not weakening the system against other threats. More likely, studying the improvements in the security level against all types of attack may compensate a loss in utility stemming from one particular defense. The trade-off is more acceptable when considering the full picture. This holistic approach is not yet adopted in the literature.

The long term objective of ARTISHAU is to develop a global procedure assessing the security levels of already deployed machine learning algorithms. It means that we are able to gauge how vulnerable a given model is and potentially to compare to other models in its category. ARTISHAU also aims to issue guidelines and methodologies for security-by-design learning.

3 Research program

3.1 Research axis A: Security of Machine Learning

Grid for reading: Decision-making, Intrinsic vulnerabilities, Design of defenses.

Research axis A describes the 9 scenarios (3 assets \times 3 values) arising from the vision described in section 2.2.1. This tour of the security threats over machine learning shows that this vision is sound because it encompasses 9 scenarios which all make sense². They are here grouped into 3 challenges.

3.1.1 Challenge #1: Protection of the training set \mathcal{D}

The owners of datasets have spent a lot of efforts gathering a large amount of valuable data and annotating them in preparation of supervised learning.

Confidentiality. If training data are sensitive, then their owners simply do not want to disclose them. On the other hand, they are not able to carry up the training procedure that is outsourced to a machine learning expert. Traditional encryption provides protection, but prevents processing the data, whereas *multi-party computation* or *homomorphic encryption* allow both. This amounts to learning a model over encrypted data. Prototypes based on multi-party computation are already running [71] and federated learning receives a lot of attention. On the other hand, there are applications where neither of these solutions is admissible, and homomorphic encryption remains the only option.

Privacy. Model θ is the result of a training over \mathcal{D} , thus it is data dependent and potentially leaks information about \mathcal{D} . A privacy enabling training procedure trades off the utility (*e.g.* the accuracy of the model) for some privacy (lower information leakage). There is an obvious connection with *differential privacy* [32]. Model θ does not leak (or by a small controllable amount) whether a particular data point was part of the training set, even if the attacker knows all the other training data [22]. But, this concept might be too strict as it grants a large advantage (the knowledge of the full dataset except one item) to the attacker. It is more realistic to think in terms of *distributional privacy* [25, 51] or *membership inference* [65]: the attacker is given a finite superset of data, and he/she has to decide which of them were actually used during the training.

²It is almost complete as the only missing use case we are aware of is the accessibility of the model at stake in a deny-of-service attack [43]. This is out of our scope.

Privacy and confidentiality are usually paired especially in collaborative or federated learning where separate parties collaborate and learn from each other’s data [40].

Integrity. The attacker has injected corrupted data in the training set. Poisoning means that this corruption biases the training issuing a model that seamlessly misclassifies a given query [24]. Backdooring generalizes poisoning: by coherently modifying some training data associated to a target class, the model learned to link this modification to the target: any query modified by the same process triggers the target class at the output of the backdoored model [23]. In simple words, the attacker remotely controls the backdoored model.

3.1.2 Challenge #2: Protection of the model θ

Learning over a large training set requires skills and computing power. This motivates the protection of the resulting model θ .

Confidentiality. Deep neural networks run on GPU which are highly computationally efficient but absolutely not secure. It is easy to steal a model by freezing this unit and dumping the parameters. This raises concerns on AI embedded systems be it smartphone applications or military vehicles equipped with sensors and AI. An interesting approach masks the model parameters and sets a protocol between the GPU and the TEE³ in charge of unmasking the final result [68].

Privacy. Is it possible to identify a model enclosed in a black box just by querying it? Black boxes are common: accessing a model via an API (MLaaS⁴) or via an IC (Onboard AI). Our recent work shows that identification is possible to some extent by comparing the fingerprint of models [55, 54]. Moreover, we showed that this piece of information is valuable for the attacker when concocting an evasion attack. The transferability of attacks improves when the target model is disclosed [56]. For the defender point of view, this technique could also scout the theft of models [29].

Integrity. Many learned models are available on the shelf. The user needs guarantees that they are free from backdoors (see section 3.1.1). Trojaning applies very sparse modifications of the weights of a learned model to inject hazardous behaviour [58, 52]. Note that a recent paper claims that it is theoretically possible to plant a backdoor in neural networks without raising suspicion even in a white box setup [39]. Practical implementations are coming [26].

3.1.3 Challenge #3: Protection of the testing data x

Confidentiality. Inferring a query without ‘seeing it’ is feasible thanks to *homomorphic encryption*. This is the dual problem of the above-mentioned scenario (See section 3.1.1) where the query is now encrypted and the model is in the clear. Again, multi-party computation and homomorphic encryption are competing approaches. The first one relies on the assumption that the parties do not collude, whereas the Achilles’ heel of the latter is the amplification of the encryption noise while processing data, which needs to be reset with complex bootstrapping.

Privacy. The user is interested in classifying a query x for a given classification problem. Yet, the user does not want any other information to be extracted from x . Is it possible to sanitize the query by projecting it onto a subspace only containing the information needed for the inference? This approach has been investigated in [36, 60].

Integrity. This is the field of *adversarial examples*. The attacker adds a small perturbation to the query x to delude a DNN classifier. The forgery is perceptually close to the original query or even undetectable. This topic is the tip of the iceberg in machine learning security with more than 6,000 publications in the last four years [30].

3.1.4 Priorities

This first research axis reveals the complexity and the richness of the field. However, this is not exactly a *terra incognita* since the members of ARTISHAU have already cleared some ways before the creation of the team-project in October 2024, notably challenges #2: Privacy, #3: Confidentiality, #3: Integrity.

In the short term, the priority is given to the following topics where we miss expertise:

³the Trusted Execution Environment chip, highly secure but not computationally efficient

⁴Machine Learning as a Service, *i.e.* on the cloud [62]

- **Challenge #1: Privacy.** A previous collaboration with our colleague Yufei Han from the PIRAT team has discovered that membership inference attacks are over-claimed. They do work but only under specific lab conditions, notably overfitting models. This year, equipped with this new expertise, we explore the topic of *machine unlearning* under the new convention DGA-MI - Inria.
- **Challenges #1 and #2: Integrity.** These two challenges are known as backdoor detection by analysing either the training data (#1) or the learnt model (#2). The Cifre Ph.d. thesis with THALES puts the priority on challenge #2, especially on a black-box model [64, 50].

3.2 Research axis B: Audit of black-box AIs

Grid for reading: Decision-making AI, Black box, Certification

3.2.1 Challenge #4: Decidability and conditions for accurate audit tasks

While model creators have by definition a complete (*i.e.*, white-box) access to their models for improving them, it is tempting for external auditors or users to try and infer some properties of these models from their (black-box) standpoint.

This challenge questions the conditions for correctness of a given audit task. We indeed know from *active property testing* that for some task assumptions on the model *symmetry* are taken. The models we are targeting are of a much higher complexity in their input cardinality and domain; we will most likely have to restrict the tasks one wants to perform correctly. This challenge thus tackles the conditions under which some tasks can be performed (*i.e.*, decided). Proving that some tasks are not feasible under certain conditions (see [49, 27]) is also of importance in order to delimit what is auditable or not, and then clarify some areas where empirical work is often engaged.

3.2.2 Challenge #5: Tracking models evolutions

Most research effort are devoted to AI systems *in vitro*, *i.e.* under restricted lab conditions, oversimplified assumptions, and a static ML pipeline. Machine Learning Operations (MLOps) is a new paradigm considering the life cycles of real-world and large-scale AI systems with continuous model updates. What is the impact for auditors?

Several research works [54, 45] show that one auditor can measure a distance between models. One straightforward use is to track the changes of several versions of an audited model. This requires successive queries with specific test data x 's (see *e.g.* [46]). A more advanced challenge is to find out if a consistent change (*i.e.*, evolution) of a model for a precise reason can also be captured by such distances under the form of an also consistent direction. An auditor would then be able to track a model evolution to assess whether or not a platform is updating a model in a direction deemed suitable. This example application further underlines the link between audits (for regulators) and security (*i.e.*, potential leaks of company related operational secrets for instance).

3.2.3 Challenge #6: Auditors coordination and stealth

After the conditions for correct audits are better understood (from challenge #4), and some concrete application of audits are designed (challenge #5), the question of efficiency in audits, by means of collaboration can arise. Indeed, a collaboration between multiple auditors interested in inferring a given model property may bring significant improvements for the accuracy of audits. Since the queries by the auditors are most likely part of the normal operation of model θ , letting these observations (queries/results) be shared in a collaborative audit environment might be of interest. This is to be opposed to a single auditor setup, where queries have the very purpose of checking properties (*i.e.*, not using the service *per se*). We are thus interested in studying the benefit of collaboration on diminishing the number of non service related queries, or to perform audits in a stealthy mode.

This axis also relates to coordination techniques that can be borrowed from the field of distributed computing (and expertise from the WIDE team).

3.2.4 Priorities

The priorities are listed from short to long term research investigations.

- **Challenges #4:** Explainability in a black-box setup cannot imply trust when conditions for an audit are insufficient. Finding the minimal set of assumptions for an auditor, that makes it possible to design efficient audit algorithms (in other words that would permit a gain over purely randomized queries to a platform) is our first priority and currently pursued.
- **Challenges #5:** Prior to the creation of the team-project, we designed a method for computing distances between models. Yet, this only operates for predictive AI models, *i.e.* classifiers. Our priority is to tackle generative AI models, especially LLM [48].
- **Challenges #6:** An informal collaboration has already started within the associated team with EPFL's SaCS team to work on parallel audits from several agents with various individual objectives [70]. Another priority is the assessment of stealthiness of the auditor, and if not, the possibility for the model owner to cheat.

3.3 Research axis C: Threats from generative models

Grid for reading: Generative AI, Malevolent Usage, Design of defenses.

The development of AI-based image editing techniques makes tampered images and facial manipulation (commonly referred to as deepfakes) widely available and more realistic [21]. A new step has been taken with modern AI generative technologies that now make it possible to alter content by means of a simple textual instruction [28, 42]. Many detection methods rely on the general assumption that any falsification, or synthesis process, introduces low-level artifacts. While these approaches can be effective, they suffer from three major weaknesses: (i) poor generalization ability [72]; (ii) the need for a large annotated dataset, (iii) a lack of robustness when the data have been compressed or rescaled several times, as is the case for images circulating on social networks.

3.3.1 Manipulated data detection

Challenge #7: Person-centric deepfake detection.

Concerning facial manipulation, one strategy is to address the above limitations by proposing person-centric deep fake detection, that involves the learning of behavioural-signatures representing enrolled subjects. By overcoming the method used to generate the deepfakes, this approach improves robustness to low quality videos and allow to handle a wide range of modifications, from lip-synchronization to face swapping. We intend to explore metric learning, as well as one-class learning. Metric learning enables the maximization of feature-wise distances between real and manipulated frames, while minimizing the feature-wise distances between frames obtained from real videos. In one-class learning, deepfake detection can be formulated as an anomaly detection problem. The distribution of non-manipulated face images would be modelled, aiming to identify manipulated face images as anomalies w.r.t. this model. The associated challenges are to determine discriminative and robust behavioral characteristics for an individual, and to limit the number of videos needed to learn the model of an individual.

Challenge #8: Tampered images detection. A possible solution to increase the robustness of tampered image detection is to search for similar candidates in a trusted database or in an open database, in order to obtain additional information to the simple statistics of the content. Nevertheless, a simple comparison of images is not sufficient. The diversity of modifications made to images in real case scenario (low quality, cropping, editing operations) makes the search results very noisy. Most importantly, to be used in practice, a method must not only detect a manipulation, but also differentiate a malicious modification from an editing operation that does not change the meaning of the image.

Challenge #9: Out-of-context images detection.

Image reuse is one of the easiest, and therefore most common methods of spreading false information, in which an (unmodified) image is reused to illustrate a different story. In this case, the falsified part is the link between the image and the text. This makes detection very difficult, as it requires a semantic understanding of

the text, the image and their relation. Detecting this link requires text and image alignment, which is one of the modern multimodal tasks. Despite recent advances in vision-language models, the detection capabilities of these methods are limited [53, 44]. We are interested in improving these models to better take into account the detection of named entities, paraphrases, negations. As for tampered image detection, the use of external open-domain resources (typically the web) is a way to improve the detection. This involves cross-modal search and fine-grained semantic understanding of text and images. The challenges are related to the use of a noisy open domain, as well as extraction and processing of evidences. Generative methods can be used as levers to improve cross-modal search and text and image alignment methods.

3.3.2 Challenge #10: Planting watermarking in generative models.

Watermarking is a well-known means to prove source of a content (image, audio, video, source code). It has witnessed a revival thanks to deep learning [47] where the model extracts robust features from content which carry the watermark signal [35, 34].

Several governments agree that key players in the field of AI should protect their models and also offer means to detect that a piece of content has been generated [61]. Their AI should not usurp humans. For both applications, watermarking is one approach, so-called active in opposition to challenges #8 (Sect. 3.3.1) and #2 (Sect. 3.1.2) based on passive forensics.

Watermarking is used to embed a signature onto the generated pieces of content. A first distinction is whether the model is private (like chatGPT, Midjourney, . . .) or publicly available like Stable Diffusion. In the latter case, the challenge is to merge generation and watermarking, so that the model generates natively watermarked content. Another question is whether versioning is possible. The user downloads a generative model which embeds an identification signature into generated content. The second difficulty is who is in charge of the watermark detection. A public detector would open the door to oracle attacks. This topic is highly trendy in image and especially text generation.

3.3.3 Priorities

- **Challenges #8:** The detection and characterization of modified images already started. The goal is to develop a vision-language model reporting the differences in between two images together with a global measurement of the change of the semantic. Embedding semantic information within the content is one possibility [33].
- **Challenge #10:** Watermarking for AI generated content is our priority due to the agenda of the EU AI Act Art.50. This includes watermarking of text generated by LLMs [38, 66]. Also, the question of the security of these new watermarking primitives is of utmost importance [37].
- **Challenges #7:** Deepfake detection will focus on face reenactment, which is the most difficult case, but the least studied. After identifying the weaknesses of current methods for this type of modification, the aim is to develop a person-centric approach. The robustness and the amount of training data required will be the key issues to be resolved.

4 Application domains

4.1 Security, cybersecurity, and defense applications

The main utility of AI in defense applications is the processing of information. Security and defense is facing a deluge of data, in terms of quantity and resolution. AI is here to help exploiting that data, be they hot (freshly captured in real-time) or cold (archived). Exploiting means helping the operator to analyse hot data and to navigate among cold archives in order to discover events and to raise alarms instrumental during the decision-making process.

Specificities concern data and the operation mode. Data is plentiful, large, heterogeneous, but also possibly confidential, sensitive, with an access forbidden to anyone not having “*le droit d’en savoir*”. Data originate from sources varying in trust. AI for defense requires reliability in operation mode. This requirement is set to an extreme level: missing an alarm may endanger the life of civilians, soldiers or cause irreparable damages.

The existence of very serious adversaries fundamentally differentiates defense and security applications from any other context. Adversaries can attack systems and manipulate data in order to cause e.g. false negatives (missing an event) or false positives (raising a wrong alarm), overall reducing the performance of an AI-based decision process. Defense also means coalition with allies. This spurs interoperability, but with restrictions. Collaborating does not mean sharing all data and knowledge. France may grant allies access to AI systems while preventing them from stealing technology, or inferring about the training data. Another specificity is purely technological: AI systems for defense might be considered as weapons, and as such they cannot rely on any untested outsourced technology.

Cybersecurity also encompasses the fight against disinformation. Though information operations are as old as war itself, armed forces have had to adapt their strategies to establish themselves in cyberspace. Belligerent states have been particularly active in this field of cyberspace information operations. AI generation or modification of content is disrupting information warfare. This transformation poses many challenges to the armed forces as well as to our industries.

4.2 Compliance with coming regulations

National and European regulations (EU AI Act) classifies AI systems according to their risks. The majority of obligations fall on providers of high-risk AI systems. All providers must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections. Yet, even limited-risk AI systems are subject to light transparency obligations. Deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).

Beyond self assessments, governments and the European Commission have recently created entities (INESIA in France, AI Office in Europe) whose goal is to edit a Code of Practice detailing the technical implementation of the related legal concepts, to assess the compliance of deployed models with respect to this Code through black-box audit, and to investigate present and future systemic risks.

5 Highlights of the year

5.1 Awards

- Best Paper Award at SRDS 2025 (Symposium on Reliable Distributed Systems) to "*Robust Fingerprinting of Graphs with FinG*" [7] co-authored by Jade Garcia Bourrée and Erwan Le Merrer.
- Spotlight paper at ICML 2025 (International Conference on Machine Learning) to "*Robust ML Auditing using Prior Knowledge*" [4] co-authored by Jade Garcia Bourrée, Augustin Godinot, and Erwan Le Merrer.
- Teddy Furon received the *Prix Innovation Inria-Dassault System de l'Académie des Sciences*.

5.2 Societal impact

Teddy Furon co-funded the startup [Label4.ai](#) in January 2025. This company offers means to detect AI generated content. It is based on the forensics analysis of content (research transfer from CNRS and Univ. Napoli) or on watermarking techniques labelling AI content at the generation step (research transfer from Inria). This covers four modalities: image, video, audio, text. Label4.ai received the award *Trophée Start-Up Numérique 2025* by IMT Starter.

6 New results

6.1 Security of machine learning

On the Vulnerability of Retrieval in High Intrinsic Dimensionality Neighborhood

Participant: Teddy Furon.

This work investigates the vulnerability of the nearest neighbors search, which is a pivotal tool in pattern analysis, data science, and machine learning. The vulnerability is gauged as the relative amount of perturbation that an attacker needs to add to a dataset point in order to modify its proximity to a given query. The statistical distribution of the relative amount of perturbation is derived from simple assumptions, outlining the key factor that drives its typical values: The higher the intrinsic dimensionality, the more vulnerable is the nearest neighbors search. Experiments on six large-scale datasets validate this model up to some outliers, which are explained as violations of the assumptions. Related publication [3].

Robust Fingerprinting of Graphs with FING

Participants: Odysseas Drosis (*EPFL*), Jade Garcia Bourrée, Anne-Marie Kermarrec (*EPFL*), Erwan Le Merrer, Othmane Safsafi (*EPFL*).

Graphs have become fundamental for carrying invaluable insights into numerous scientific disciplines. Controlling if they are further shared and modified is essential when sharing such graphs. This control is typically achieved using digital watermarking by embedding identification information in the graph structure. In this work, we propose the first approach to fingerprinting graphs by associating a characteristic signature of these graphs that can be extracted later as proof of ownership. This work provides the same guarantees as watermarking while avoiding the need to modify the graph, instead by exporting the fingerprint to an external timestamped database. We present the novel fingerprinting scheme FING. FING relies on the Factor- r Sum Subsets problem to create a digital fingerprint. This problem is NP-hard, so it is easy to create and extract for the graph originator while being intractable for an attacker. We provide an analysis of the robustness of FING facing a wide range of attacks that aim at removing or extracting the fingerprint. Finally, we empirically show FING's scalability. A fingerprint can be created in around four minutes on a single core for 10 million node graphs and is robust against attacks removing thousands of edges, for instance. Related publication [7].

Queries, Representation & Detection: The Next 100 Model Fingerprinting Schemes

Participants: Augustin Godinot, Erwan Le Merrer, Camilla Penzo (*PEReN*), François Taïani (*Inria WIDE*), Gilles Trédan (*CNRS LAAS*).

The deployment of machine learning models in operational contexts represents a significant investment for any organisation. Consequently, the risk of these models being misappropriated by competitors needs to be addressed. In recent years, numerous proposals have been put forth to detect instances of model stealing. However, these proposals operate under implicit and disparate data and model access assumptions; as a consequence, it remains unclear how they can be effectively compared to one another. Our evaluation shows that a simple baseline that we introduce performs on par with existing state-of-the-art fingerprints, which, on the other hand, are much more complex. To uncover the reasons behind this intriguing result, this work introduces a systematic approach to both the creation of model fingerprinting schemes and their evaluation benchmarks. By dividing model fingerprinting into three core components – Query, Representation and Detection (QuRD) – we are able to identify around 100 previously unexplored QuRD combinations and gain insights into their performance. Finally, we introduce a set of metrics to compare and guide the creation of more representative model stealing detection benchmarks. Our approach reveals the need for more challenging benchmarks and a sound comparison with baselines. To foster the creation of new fingerprinting schemes and benchmarks, we open-source our fingerprinting toolbox. Related publication [10].

BAIT: A new DNN backdoor attack using inpainted triggers

Participants: Quentin Le Roux, Yannick Teglia (*THALES*), Eric Bourbao (*THALES*), Philippe Loubet-Moundi (*THALES*), Teddy Furon.

Backdoor attacks compromise deep neural networks by injecting them with covert, malicious behaviors during training, which attackers can later activate at test-time. As backdoors become more sophisticated, defenses struggle to catch up. This paper introduces a simple yet effective Backdoor Attack using Inpainting as a Trigger, dubbed BAIT. The attack’s trigger relies on a randomly-drawn polygonal patch, filled via inpainting with an off-the-shelf generative adversarial network. Using BAIT, we show that several defenses, including common test-time input purification methods, can be bypassed by patch-based backdoors. To counter this, we propose four targeted defense strategies. Related publication [14].

Survivability of Backdoor Attacks on Unconstrained Face Recognition Systems

Participants: Quentin Le Roux, Yannick Teglia (*THALES*), Teddy Furon, Eric Bourbao (*THALES*), Philippe Loubet-Moundji (*THALES*).

The widespread use of deep learning face recognition raises several security concerns. Although prior works point at existing vulnerabilities, DNN backdoor attacks against real-life, unconstrained systems dealing with images captured in the wild remain a blind spot of the literature. This work conducts the first system-level study of backdoors in deep learning-based face recognition systems. This work yields four contributions by exploring the feasibility of DNN backdoors on these pipelines in a holistic fashion. We demonstrate for the first time two backdoor attacks on the face detection task: face generation and face landmark shift attacks. We then show that face feature extractors trained with large margin losses also fall victim to backdoor attacks. Combining our models, we then show using 20 possible pipeline configurations and 15 attack cases that a single backdoor enables an attacker to bypass a system’s entire function. Finally, we provide stakeholders with several best practices and countermeasures. Related publication [20].

Task-Agnostic Attacks Against Vision Foundation Models

Participants: Brian Pufler (*Univ. Geneva*), Yury Belousov (*Univ. Geneva*), Vitaliy Kinakh (*Univ. Geneva*), Teddy Furon, Slava Voloshynovskiy (*Univ. Geneva*).

The study of security in machine learning mainly focuses on downstream task-specific attacks, where the adversarial example is obtained by optimizing a loss function specific to the downstream task. At the same time, it has become standard practice for machine learning practitioners to adopt publicly available pre-trained vision foundation models, effectively sharing a common backbone architecture across a multitude of applications such as classification, segmentation, depth estimation, retrieval, question answering and more. The study of attacks on such foundation models and their impact to multiple downstream tasks remains vastly unexplored. This work proposes a general framework that forges task-agnostic adversarial examples by maximally disrupting the feature representation obtained with foundation models. We extensively evaluate the security of the feature representations obtained by popular vision foundation models by measuring the impact of this attack on multiple downstream tasks and its transferability between models. Related publication [17].

Multi-modal Identity Extraction

Participants: Ryan Webster, Teddy Furon.

The success of multi-modal foundational models is partly attributed to their diverse, billions-scale training data. By nature, web data contains human faces and descriptions of individuals. Thus, these models pose potentially widespread privacy issues. Recently, identity membership inference attacks (IMIAs) against the CLIP model showed that membership of an individual’s name and image within training data can be reliably inferred.

This work formalizes the problem of identity extraction, wherein an attacker can reliably extract the names of individuals given their images only. We provide the following contributions (i) we adapt a previous

IMIA to the problem of selecting the correct name among a large set and show that the method scales to millions of names (ii) we design an attack that outperforms the adapted baseline (iii) we show that an attacker can extract names via optimization only. To demonstrate the interest of our framework, we show how identity extraction can be used to audit model privacy. Indeed, a family of prominent models that advertise blurring faces before training to protect privacy is still highly vulnerable to attack. Related publication [16].

Improving Unlearning with Model Updates Probably Aligned with Gradients

Participants: Virgile Dine, Teddy Furon, Charly Faure.

We formulate the machine unlearning problem as a general constrained optimization problem. It unifies the first-order methods from the approximate machine unlearning literature. This work then introduces the concept of feasible updates as the model’s parameter update directions that help with unlearning while not degrading the utility of the initial model. Our design of feasible updates is based on masking, i.e. a careful selection of the model’s parameters worth updating. It also takes into account the estimation noise of the gradients when processing each batch of data to offer a statistical guarantee to derive locally feasible updates. The technique can be plugged in, as an add-on, to any first-order approximate unlearning methods. Experiments with computer vision classifiers validates this approach. Related publication [6].

6.2 Audit of black-box AIs

Robust ML Auditing using Prior Knowledge

Participants: Jade Garcia Bourrée, Augustin Godinot, Sayan Biswas (*EPFL*), Anne-Marie Kermarrec (*EPFL*), Erwan Le Merrer, Gilles Trédan (*CNRS LAAS*), Martijn de Vos (*EPFL*), Milos Vujasinovic (*EPFL*).

Among the many technical challenges to enforcing AI regulations, one crucial yet underexplored problem is the risk of audit manipulation. This manipulation occurs when a platform deliberately alters its answers to a regulator to pass an audit without modifying its answers to other users. In this work, we introduce a novel approach to manipulation-proof auditing by taking into account the auditor’s prior knowledge of the task solved by the platform. We first demonstrate that regulators must not rely on public priors (e.g. a public dataset), as platforms could easily fool the auditor in such cases. We then formally establish the conditions under which an auditor can prevent audit manipulations using prior knowledge about the ground truth. Finally, our experiments with two standard datasets illustrate the maximum level of unfairness a platform can hide before being detected as malicious. Our formalization and generalization of manipulation-proof auditing with a prior opens up new research directions for more robust fairness audits. Related publication [4].

P2NIA: Privacy-Preserving Non-Iterative Auditing

Participants: Jade Garcia Bourrée, Hadrien Lautreite (*Univ. Québec*), Sébastien Gambs (*Univ. Québec*), Gilles Trédan (*CNRS LAAS*), Erwan Le Merrer, Benoît Rottembourg (*Inria Bordeaux*).

The emergence of AI legislation has increased the need to assess the ethical compliance of high-risk AI systems. Traditional auditing methods rely on platforms’ application programming interfaces (APIs), where responses to queries are examined through the lens of fairness requirements. However, such approaches put a significant burden on platforms, as they are forced to maintain APIs while ensuring privacy, facing the possibility of data leaks. This lack of proper collaboration between the two parties, in turn, causes a significant challenge to the auditor, who is subject to estimation bias as they are unaware of the data distribution of the platform. To address these two issues, we present P2NIA, a novel auditing scheme that proposes a mutually beneficial collaboration for both the auditor and the platform. Extensive experiments demonstrate

P2NIA’s effectiveness in addressing both issues. In summary, our work introduces a privacy-preserving and non-iterative audit scheme that enhances fairness assessments using synthetic or local data, avoiding the challenges associated with traditional API-based audits. Related publication [9].

6.3 Threats from generative models

Reframing image difference captioning with BLIP2IDC and synthetic augmentation

Participants: Gautier Evennou, Antoine Chaffin (*LightOn*), Vivien Chappelier (*Imatag*), Ewa Kijak.

The rise of the generative models quality during the past years enabled the generation of edited variations of images at an important scale. To counter the harmful effects of such technology, the Image Difference Captioning (IDC) task aims to describe the differences between two images. While this task is successfully handled for simple 3D rendered images, it struggles on real-world images. The reason is twofold: the training data-scarcity, and the difficulty to capture fine-grained differences between complex images. To address those issues, we propose a simple yet effective framework to both adapt existing image captioning models to the IDC task and augment IDC datasets. We introduce BLIP2IDC, an adaptation of BLIP2 to the IDC task at low computational cost, and show it outperforms two-streams approaches by a significant margin on real-world IDC datasets. We also propose to use synthetic augmentation to improve the performance of IDC models in an agnostic fashion. We show that our synthetic augmentation strategy provides high quality data, leading to a challenging new dataset well-suited for IDC named Syned. Related publication [8].

Evaluating the security of public surrogate watermark detectors

Participants: Chloé Imadache, Eva Giboulot, Teddy Furon.

The omnipresence of generated content has led to an increasing need of multimedia content traceability. Watermarking techniques have been proven to provide both detection guarantees and robustness. However, widespread use of such methods would require disclosing the watermark detector to the public. Such access breaches the watermark security: end users with unlimited access to the detector could easily craft adversarial examples, through white-box and black-box attacks.

To circumvent this issue, we suggest providing to the public a surrogate, less accurate detector. Calls to the private detector would be reserved for important or anomalous cases. This paper studies the potential leakage of information from the surrogate detector. We first create a wide panel of images adversarial to the surrogate detector. The efficiency of the private detector is then assessed on this data. This allows us to introduce a metric of the transferability of these attacks from the surrogate to the private detector. Through this metric, we evaluate the security of different designs of surrogate detectors. Related publication [13].

Watermark anything with localized messages

Participants: Tom Sander (*Meta FAIR*), Pierre Fernandez, Alain Durmus (*CMAE Ecole Polytechnique*), Teddy Furon, Matthijs Douze (*Meta FAIR*).

Image watermarking methods are not tailored to handle small watermarked areas. This restricts applications in real-world scenarios where parts of the image may come from different sources or have been edited. We introduce a deep-learning model for localized image watermarking, dubbed the Watermark Anything Model (WAM). The WAM embedder imperceptibly modifies the input image, while the extractor segments the received image into watermarked and non-watermarked areas and recovers one or several hidden messages from the areas found to be watermarked. The models are jointly trained at low resolution and without perceptual constraints, then post-trained for imperceptibility and multiple watermarks. Experiments show

that WAM is competitive with state-of-the-art methods in terms of imperceptibility and robustness, especially against inpainting and splicing, even on high-resolution images. Moreover, it offers new capabilities: WAM can locate watermarked areas in spliced images and extract distinct 32-bit messages with less than 1 bit error from multiple small regions -no larger than 10% of the image surface -even for small 256×256 images. Related publication [15].

6.4 Miscellaneous

Algorithmic curation of News on YouTube: Evidence from the 2022 French Presidential Campaign

Participants: Julien Figeac (CNRS), Erwan Le Merrer, Marie Neihouser (Université Paris 1 Panthéon-Sorbonne), Gilles Trédan (CNRS LAAS).

Debate is growing over how algorithmic recommendations influence news visibility, particularly regarding interest and ideological bias. This work examines YouTube's News Recommender System (NRS), focusing on the French "News" section of the homepage through a large-scale audit using automated browsing agents. Our findings show that the NRS prioritises platform-native creators over established news outlets, favouring content that aligns with YouTube's features rather than traditional editorial standards. Politically charged, opinion-based videos, especially those from prominent figures affiliated with extreme political parties, receive ongoing algorithmic promotion. While centrist and moderate political figures remain underrepresented, the algorithm boosts their visibility once users interact with this type of content. This two-part mechanism, which amplifies already prominent content and appears to overcompensate for rare content, does not just reflect engagement-based optimisation, but is driven by the algorithm's tendency to maintain coherence in a highly unbalanced content landscape. However, this compensatory logic does not counteract the algorithm's broader tendency to promote content from radical political figures while marginalising institutional news outlets. Through this process, the NRS actively reshapes news exposure within the "News" section, privileging political expression over journalistic authority and reproducing structural hierarchies of visibility. Related publication [19].

Fast & Fourier: spectral graph watermarking

Participants: Jade Garcia Bourrée, Anne-Marie Kermarrec (EPFL), Erwan Le Merrer, Othmane Safsafi (EPFL).

We address the problem of watermarking graph objects, which consists in hiding information within them, to prove their origin. The two existing methods to watermark graphs use subgraph matching or graph isomorphism techniques, which are known to be intractable for large graphs. To reduce the operational complexity, we propose FFG, a new graph watermarking scheme adapted from an image watermarking scheme, since graphs and images can be represented as matrices. We analyze and compare FFG, whose novelty lies in embedding the watermark in the Fourier transform of the adjacency matrix of a graph. Our technique enjoys a much lower complexity than that of related works (i.e. in $O(N^2 \log N)$), while performing better or at least as well as the two state-of-the-art methods. Related publication [5].

Adapting Without Seeing: Text-Aided Domain Adaptation for Adapting CLIP-like Models to Novel Domains

Participants: Louis Hemadou, Hélène Vorobieva (Safran Tech), Ewa Kijak, Frédéric Jurie (CNRS GREYC).

This work addresses the challenge of adapting large vision models, such as CLIP, to domain shifts in image classification tasks. While these models, pre-trained on vast datasets like LAION 2B, offer powerful visual representations, they may struggle when applied to domains significantly different from their training data,

such as industrial applications. We introduce TADA, a Text-Aided Domain Adaptation method that adapts the visual representations of these models to new domains without requiring target domain images. TADA leverages verbal descriptions of the domain shift to capture the differences between the pre-training and target domains. Our method integrates seamlessly with fine-tuning strategies, including prompt learning methods. We demonstrate TADA’s effectiveness in improving the performance of large vision models on domain-shifted data, achieving state-of-the-art results on benchmarks like DomainNet. Related publication [11].

Cross-task knowledge distillation for few-shot detection.

Participants: Louis Hemadou, Hélène Vorobieva (*Safran Tech*), Ahmed Nasreddine Benaichouche (*Safran Tech*), Frédéric Jurie (*CNRS GREYC*), Ewa Kijak.

While powerful pretrained visual encoders have advanced many vision tasks, their knowledge is not fully leveraged by object detectors, especially in few-shot settings. A key challenge in transferring this knowledge via cross-task distillation is the semantic mismatch between outputs: classifiers produce clean probability distributions, while detector scores implicitly encode both class and objectness. To address this, we propose a lightweight fine-tuning strategy guided by a novel, correlation-based distillation loss. This loss aligns the detector’s relative class preferences with those of a strong image classifier, effectively decoupling the learning of class semantics from objectness. Applied to a state-of-the-art detector, our method consistently improves performance in a low-data regime, demonstrating an effective way to bridge the gap between powerful classifiers and object detectors. Related publication [12].

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

CIFRE PHD: Certification of Deep Neural Networks

Participants: Quentin Le Roux, Teddy Furon.

Duration: 3 years, ended in November 2025, Partner: THALES

This is a CIFRE PhD thesis project aiming at assessing the security of already trained Deep Neural Networks, especially in the context of face recognition.

CIFRE PHD: Watermarking Generative AIs

Participants: Pierre Fernandez, Teddy Furon.

Duration: 3 years, ended in February 2025, Partner: Meta FAIR

This is a CIFRE PhD thesis project aiming at designing watermarking techniques dedicated to generative AIs (text to image, text to speech, LLM).

CIFRE PHD: Domain generalization exploiting synthetic data

Participants: Louis Hemadou, Ewa Kijak.

Duration: 3 years, defended in December 2025, Partner: Safran Tech

This is a CIFRE PhD thesis project aiming at exploiting synthetic data to be able to perform transfer learning in presence of very few or inexistent real data in the context of image detection or classification tasks.

CIFRE PHD: Detection and explanation of semantic manipulations in multimedia content

Participants: Gautier Evennou, Ewa Kijak.

Duration: 3 years, started in September 2023, Partner: Imatag

This is a CIFRE PhD thesis project aiming at detecting and explaining semantic manipulations in multimedia content, in the context of misinformation.

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 STIC/MATH/CLIMAT AmSud projects

Isabela Borlido Barcelos

Status PhD

Institution of origin: Pontifical Catholic University of Minas Gerais

Country: Brazil

Dates: January to December 2025

Project: STIC-AMSUD GiMMD (Graph-based Analysis and Understanding of Image, Video and Multimedia Data)

Partners: Universidad de la República de Uruguay, Pontifical Catholic University of Minas Gerais, Inria

Context of the visit: Feature learning from image markers by using graph analysis

Mobility program/type of mobility: research stay

8.2 International research visitors

8.2.1 Visits of international scientists

Other international visits to the team

Martijn de Vos, Sayan Biswas, Milos Vujanovic

Status PhD and post-Docs

Institution of origin: EPFL, **SACS team**

Country: Switzerland

Dates: 13-17, January 2025

Context of the visit: preparation of a submission for ICML 2025

Mobility program/type of mobility: research stay

8.2.2 Visits to international teams

Research stays abroad

Augustin Godinot

Visited institution: Vector Institute for Artificial Intelligence and University of Toronto

Country: Canada

Dates: 6 months from March to September 2025

Context of the visit: Visit to Nicolas Papernot, Chaire Inria International

Mobility program/type of mobility: Internship

8.3 National initiatives

PEPR Cybersécurité projet COMPROMIS

Participants: Teddy Furon, Eva Giboulot, Ewa Kijak, Enoal Gesny, Chloé Imadache, Ryan Webster, Paul Chaurand.

Duration: 4.5 years, started April 2024

The COMPROMIS project is based on a modern vision of multimedia data protection, with deep learning at its heart. This project defends the idea that the protection of multimedia data must necessarily be associated with the security of the tools that analyse this data, i.e. these days Artificial Intelligence (AI). The observation is simple: the protection of multimedia data is undoubtedly the area of cybersecurity that has benefited most from AI, but it has neglected to check the level of security of this new tool. AI has become one of the weak links in the protection of multimedia data. The scientific hurdles thus concern both the classic applications of multimedia data protection and the emerging field of deep learning.

DGA-Inria collaboration: Machine Unlearning

Participants: Virgile Dine, Teddy Furon, Charly Faure (AMIAD).

Duration: 3 years, started in October 2024, Partner: AMIAD

The project aims at developing algorithms to make computer unlearn. From a model trained over a training dataset, we aim at deriving a second model ignoring some training samples, or some classes of samples without retraining it from scratch.

MinArm-Inria collaboration: EVE4

Participants: Eva Giboulot, Teddy Furon.

Duration: 3 years, ended in April 2025. Partners: MinArm, CRIStAL Lille, LIRMM, Univ. Tech. Troyes, Univ. Paris Saclay

Teaching and technology survey on steganography and steganalysis in the real world.

MinArm-Inria collaboration: EVE5

Participants: Eva Giboulot.

Duration: 18 months, started in December 2025. Partners: MinArm, CRISAL Lille, GREYC

Teaching and technology survey on steganography, steganalysis, watermarking and forensics analysis of multimedia content in the real world.

ANR PACMAM (ANR-24-CE23-7787)

Participants: Erwan Le Merrer, Timothée Chauvin.

Duration: 42 months, started in 2024. Partners: PEReN, LAAS-CNRS

The PACMAM project seeks to increase the transparency of algorithmic decisions by laying the foundations for efficient black-box auditing of large-capacity models under budget constraints. The project will focus on active learning strategies for auditing that have recently been introduced in the literature, yet whose applicability to concrete cases remains uncertain. The proposed research is organized in three work packages (WPs), each of which addresses a fundamental challenge in this research area. WP1 aims to understand how audit efficiency is affected by a model's capacity, leveraging measures such as VC dimension and Rademacher complexity. This information will help auditors strike a balance between query budget and accuracy. Building on WP1, WP2 focuses on making active auditing practical for large-capacity models by identifying efficient ways to select optimal inputs and determining what auditors need to know about audited models to succeed. Finally, WP3 explores how models that are frequently updated can be monitored efficiently. The goal is to reduce the query budget needed to continuously monitor an evolving model. Overall, PACMAM will thus provide the foundation for the efficient auditing of evolving high-capacity models. The project will ensure that any developed solutions is implemented rapidly thanks to the involvement of PEReN, the French government's department in charge of algorithmic regulation. More details are available at [Website](#).

8.4 Public policy support

COFRA-funded Ph.D. thesis with PEReN

Participants: Gurvan Richardeau, Erwan Le Merrer.

Duration: 3 years, started in 2024.

PEReN, the Center of Expertise for Digital Platform Regulation, is an interministerial office with national competence placed under the joint authority of the ministers responsible for the economy, culture and digital affairs. This collaboration deals with the audit of LLMs, especially the fingerprinting of the models. It amounts to identify the model in a black box with the minimum number of interactions.

EU Artificial Intelligence Act

Participants: Teddy Furon.

Teddy Furon participates to the Transparency Working Group in charge of publishing the *Code of Practice* related to the Art. 50 of the EU AI Act, under the supervision of the EU AI office. This working group

gathers industrials, academics, and NGOs. It deals with the future obligation for the generative AI providers in the EU to offer means to mark and detect AI generated content. It specifies the technical solutions and the way to audit them.

Expertise for the Department of Justice

Participants: Erwan Le Merrer.

Erwan Le Merrer is a technical expert for the Department of Justice on some undergoing investigations that cannot be publicly disclosed.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

General chair, scientific chair

- Teddy Furon is the general chair of ESSAI, European Symposium on Security of Artificial Intelligence.

Member of the organizing committees

- Teddy Furon was a member of the organizing committee of the workshop on *GenAI watermarking* at ICLR 2025.
- Eva Giboulot co-organized the French workshop *Detection of IA generated content* (GdR IASIS & GdR SI).
- Erwan Le Merrer was a member of the scientific committee for the organization for the Inria/DFKI IDESSAI European Summer School on AI.

9.1.2 Scientific events: selection

Member of the conference program committees

- Ewa Kijak is a member of the steering committee of IEEE International Conference on Content-Based Multimedia Indexing (CBMI).

Reviewer

- Teddy Furon was a reviewer for ICLR 2026, NeurIPS 2025, ICCV 2025, ICML 2025, IEEE ICASSP 2025, IEEE WIFS 2025, Workshop on GenAI watermarking at ICLR 2025.
- Erwan Le Merrer was a reviewer for ECAI 2025, AAAI 2026, ECML/PKDD 2026.
- Ewa Kijak was a reviewer for CBMI 2025.

9.1.3 Journal

Reviewer - reviewing activities

- Teddy Furon was a reviewer for IEEE Transactions on IFS, IEEE Transactions on Multimedia.
- Ewa Kijak was a guest editor for Multimedia Tools and Applications.
- Ewa Kijak was a reviewer for Multimedia Tools and Applications.

9.1.4 Invited talks

- Teddy Furon participated to the Winter School of PEPR Cybersécurité.
- Teddy Furon and Charly Faure invited by the ComCyber to the panel *Enjeux de sécurité des systèmes d'information intégrant de l'IA*.
- Ewa Kijak was invited speaker at the French workshop *Detection of IA generated content* (GdR IASIS & GdR SI).
- Ewa Kijak was invited speaker at the 1st edition of the scientific days of INESIA (Institut National de l'Evaluation et de la Sécurité de l'IA).
- Ewa Kijak was invited speaker at the Academic Day of the Union of Physics and Chemistry Teachers.
- Teddy Furon participated to the event *De pixels à Perception* organized by Campus Innovation – Université de Rennes.

9.1.5 Leadership within the scientific community

- Erwan Le Merrer is the recipient of a chair of the SEQUOIA cluster.
- Teddy Furon received the Prix Innovation Inria-Dassault System de l'Académie des Sciences.
- Jade Garcia Bourrée and Erwan Le Merrer received the best paper award at SRDS 2025 for paper [7].
- Spotlight paper at ICML (International 2025 to "Robust ML Auditing using Prior Knowledge" [4] co-authored by Jade Garcia Bourrée, Augustin Godinot, and Erwan Le Merrer.

9.1.6 Scientific expertise

- Teddy Furon is the scientific advisor of the startup [Label4.ai](#).
- Erwan Le Merrer is an expert for the *Crédit d'Impôt Recherche* funding program at the *Direction Générale des Finances Publiques*.
- Erwan Le Merrer is an expert for the *thèse CIFRE* funding program at the *Association Nationale de la Recherche et de la Technologie* (ANRT).

9.1.7 Research administration

- Erwan Le Merrer is the president du conseil scientifique de la *Société Informatique de France*.
- Ewa Kijak is a member of the executive committee of the IA Cluster SequoIA.
- Teddy Furon was a member of the jury *Prix de thèse GdR Sécurité Informatique*.
- Teddy Furon is the president of the *Commission des Délégations du Centre Inria de l'Université de Rennes*.
- Teddy Furon is a member of the *Commission du Personnel du Centre Inria de l'Université de Rennes / IRISA*.
- Teddy Furon participates to the coordination of the call for proposals on the research effort related to INESIA and managed by Inria *Agence de programmes*.

9.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

9.2.1 Teaching

- Eva Giboulot, Rare Event Simulations, 40h, M2, INSA Rennes
- Ewa Kijak is head of the Image engineering track (M1-M2) of ESIR, Univ. Rennes
- Ewa Kijak, Information retrieval and Multimodal applications, 24h, M2, ESIR
- Ewa Kijak, Deep Learning for Vision, 12h, M2, ESIR
- Ewa Kijak, Supervised machine learning, 20h, M1R, ENS Rennes
- Ewa Kijak, Machine learning, 12h, M1, ESIR
- Ewa Kijak, Image processing, 45h, M1, ESIR, Univ. Rennes

9.2.2 Supervision

- PhD Pierre Fernandez, Watermarking Generative AI, defended January 2025, with Teddy Furon
- PhD Louis Hemadou, Domain generalization exploiting synthetic data, defended December 2025, with Ewa Kijak
- PhD Quentin Le Roux, Backdoors in DNN applied to face recognition systems, defended November 2025, with Teddy Furon
- PhD Jade Garcia Bourrée, Trust but verify: robust statistical auditing of ML black-boxes, defended October 2025, with Erwan Le Merrer
- PhD in progress: Adele Denis, IA-based automated detection and behavior analysis among piglets. Started September 2024, with Ewa Kijak, Caroline Clouard (INRAE) and Céline Tallet (INRAE)
- PhD in progress: Virgile Dine, Machine Unlearning. Started October 2024, with Teddy Furon
- PhD in progress: Enoal Gesny, Watermarking of Generative AI. Started November 2024, with Eva Giboulot and Teddy Furon
- PhD in progress: Chloé Imadache, Security of Deep Learning based Watermarking. Started December 2024, with Eva Giboulot
- PhD in progress: Gautier Evennou, Detection and explanation of semantic manipulations in multimedia content. Started in September 2023, with Ewa Kijak
- PhD in progress: Augustin Godinot, Tools for machine learning audits in the presence of deceptive model providers. Started in 2021, with Erwan Le Merrer
- PhD in progress: Gurvan Richardeau, Audit of evolutions between LLMs. Started in 2024, with Erwan Le Merrer
- PhD in progress: Paul Chaurand, Zero-shot IA-manipulated content detection. Started September 2025, with Ewa Kijak
- PhD in progress: Timothee Chauvin, Auditing LLM-based Agents with Computer Interaction Capabilities. Started in 2024, with Erwan Le Merrer

9.2.3 Juries

- Teddy Furon was president of the PhD jury of Benoit Coquerel, Univ. Rennes, December 2025
- Teddy Furon was a reviewer for the HDR of Cédric Gouy-Pailler, CEA, May 2025
- Teddy Furon was a reviewer for the PhD of Mohamed Lansari, Univ. Brest, December 2025
- Teddy Furon was a reviewer for the PhD of Wassim Bouaziz, Institut Polytechnique de Paris, December 2025
- Teddy Furon was a reviewer for the PhD of Lucas Gnecco Heredia, Université Paris Sciences et Lettres, May 2025
- Ewa Kijak was a reviewer for the HDR of Camille Guinaudeau, LIMSI, Paris-Saclay University, November 2025
- Ewa Kijak was a reviewer for the PhD of Theo Gigant, CentraleSupélec, Paris-Saclay University, October 2025
- Teddy Furon was a jury member for the PhD of Matthieu Serfaty, ENS Paris-Saclay, December 2025
- Ewa Kijak was a jury member for the HDR of Petra Gomez-Kramer, La Rochelle University, June 2025
- Ewa Kijak was a reviewer for the PhD of Felipe Belem, Gustave Eiffel University, February 2025
- Erwan Le Merrer was an invited member of the PhD defense of Jade Garcia-Bourrée, October 2025

9.2.4 Educational and pedagogical outreach

- Teddy Furon presented ‘*What is it to be a researcher in computer science?*’ to 5 highschool classes (program *Chiche!*).

9.3 Popularization

9.3.1 Specific official responsibilities in science outreach structures

- Erwan Le Merrer leads the scientific board of the *Société Informatique de France* in 2025.

9.3.2 Productions (articles, videos, podcasts, serious games, ...)

- Erwan Le Merrer published four interviews with the *Société Informatique de France* this year on the impact of AI on jobs, on *Binaire* magazine.
- We have welcomed in our team the author / illustrator Marie Spenale who drew our storie on [Instagram](#) and [LinkedIn](#).
- Teddy Furon participated to the proposal *Projet de pôle territorial DEMOCRAT’ICC*.

9.3.3 Participation in Live events

- Ewa Kijak participated as panelist to the Conference on Generative AI and Disinformation organized by the Sorbonne Center for Artificial Intelligence (SCAI) as part of the AI Action Summit, January 2025.
- Ewa Kijak participated to the *Procès de l’IA* as part of the West Data Festival, in Laval, March 2025.
- Ewa Kijak participated to the *Fête de la Science*, Rennes, October 2025.

- Erwan Le Merrer proposed and won a funding for 1artist1scientist event for the 50 years of the IRISA laboratory. An art performance was built in the form of a vending machine embedding an AI, to question the choice of users accessing AI-enhanced services. This work was presented at *Fête de la science* in Rennes for 2 days, and then at the laboratory party. An intern from INSA was supervised and participated in the work. Several interviews followed, including by a regional TV ; these are available [here](#).
- Eva Giboulot participated to a round table at the *Société des auteurs dans les arts graphiques et plastiques* (ADAGP) on the subject of *Intelligences artificielles génératives et traçabilité*, January 2025.

9.3.4 Others science outreach relevant activities

- Erwan Le Merrer was interviewed by the french newspaper *Le Télégramme*, about our research on YouTube (3 articles), February 2025.
- Erwan Le Merrer was interviewed for the Data Skeptic podcast on “LLMs hallucinate graphs too”, January 2025.
- Teddy Furon was interviewed by the think tank *Villa Numeris*.

10 Scientific production

10.1 Major publications

- [1] J. G. Bourrée, A. Godinot, S. Biswas, A.-M. Kermarrec, E. L. Merrer, G. Tredan, M. de Vos and M. Vujasinovic. ‘Robust ML Auditing using Prior Knowledge’. In: ICML 2025 - 42nd International Conference on Machine Learning. Vancouver, Canada: arXiv, July 2025, pp. 1–17. DOI: [10.48550/arXiv.2505.04796](https://doi.org/10.48550/arXiv.2505.04796). URL: <https://hal.science/hal-05268400>.
- [2] J. Garcia Bourrée, H. Lautreite, S. Gambs, G. Tredan, E. L. Merrer and B. Rottembourg. ‘P2NIA: Privacy-Preserving Non-Iterative Auditing’. In: ECML-PKDD 2025 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Porto, Portugal, Sept. 2025, pp. 1–18. DOI: [10.48550/arXiv.2504.00874](https://doi.org/10.48550/arXiv.2504.00874). URL: <https://hal.science/hal-05268379>.

10.2 Publications of the year

International journals

- [3] T. Furon. ‘On the Vulnerability of Retrieval in High Intrinsic Dimensionality Neighborhood’. In: *IEEE Transactions on Information Forensics and Security* 20 (2025), pp. 3576–3586. DOI: [10.1109/TIFS.2025.3553067](https://doi.org/10.1109/TIFS.2025.3553067). URL: <https://hal.science/hal-05012282> (cit. on p. 15).

International peer-reviewed conferences

- [4] J. G. Bourrée, A. Godinot, S. Biswas, A.-M. Kermarrec, E. L. Merrer, G. Tredan, M. de Vos and M. Vujasinovic. ‘Robust ML Auditing using Prior Knowledge’. In: ICML 2025 - 42nd International Conference on Machine Learning. Vancouver, Canada: arXiv, July 2025, pp. 1–17. DOI: [10.48550/arXiv.2505.04796](https://doi.org/10.48550/arXiv.2505.04796). URL: <https://hal.science/hal-05268400> (cit. on pp. 14, 17, 25).
- [5] J. G. Bourrée, A.-M. Kermarrec, E. Le Merrer and O. Safsafi. ‘Fast & Fourier: spectral graph watermarking’. In: *ALGOTEL 2025 – 27èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications*. ALGOTEL 2025 – 27èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications. Saint Valery-sur-Somme, France, 2025, pp. 1–5. URL: <https://hal.science/hal-05016042> (cit. on p. 19).

- [6] V. Dine, T. Furon and C. Faure. ‘Improving Unlearning with Model Updates Probably Aligned with Gradients’. In: *Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec ’25)*. AISec ’25: 18th ACM Workshop on Artificial Intelligence and Security. Taipei, Taiwan: ACM, 22nd Nov. 2025, pp. 4889–4891. DOI: [10.1145/3733799.3762975](https://doi.org/10.1145/3733799.3762975). URL: <https://hal.science/hal-05247290> (cit. on p. 17).
- [7] *Best Paper*
O. Drosis, J. Garcia Bourrée, A.-M. Kermarrec, E. Le Merrer and O. Safsafi. ‘Robust Fingerprinting of Graphs with FING’. In: SRDS 2025 - Symposium on Reliable Distributed Systems. Porto, Portugal, 2025, pp. 1–11. URL: <https://hal.science/hal-05267625> (cit. on pp. 14, 15, 25).
- [8] G. Evennou, A. Chaffin, V. Chappelier and E. Kijak. ‘Reframing Image Difference Captioning with BLIP2IDC and Synthetic Augmentation’. In: *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision. WACV 2025 - IEEE/CVF Winter Conference on Applications of Computer Vision*. Tucson (Arizona), United States, 2025, pp. 1–11. DOI: [10.1109/wacv61041.2025.00143](https://doi.org/10.1109/wacv61041.2025.00143). URL: <https://hal.science/hal-04889899> (cit. on p. 18).
- [9] J. Garcia Bourrée, H. Lautreite, S. Gambis, G. Tredan, E. L. Merrer and B. Rottembourg. ‘P2NIA: Privacy-Preserving Non-Iterative Auditing’. In: ECML-PKDD 2025 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Porto, Portugal, Sept. 2025, pp. 1–18. DOI: [10.48550/arXiv.2504.00874](https://doi.org/10.48550/arXiv.2504.00874). URL: <https://hal.science/hal-05268379> (cit. on p. 18).
- [10] A. Godinot, E. L. Merrer, C. Penzo, F. Taïani and G. Tredan. ‘Queries, Representation & Detection: The Next 100 Model Fingerprinting Schemes’. In: *Proceedings of the AAI Conference on Artificial Intelligence*. AAI 2025 - 39th Annual AAI Conference on Artificial Intelligence. Vol. 39. 16. Philadelphia (Pennsylvania), United States, 11th Apr. 2025, pp. 16817–16825. DOI: [10.1609/aaai.v39i16.33848](https://doi.org/10.1609/aaai.v39i16.33848). URL: <https://inria.hal.science/hal-05093903> (cit. on p. 15).
- [11] L. Hémadou, H. Vorobieva, E. Kijak and F. Jurie. ‘Adapting Without Seeing: Text-Aided Domain Adaptation for Adapting CLIP-like Models to Novel Domains’. In: *Proceedings of 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2025 - IEEE International Conference on Acoustics, Speech, and Signal Processing*. Hyderabad, India: IEEE, 6th Apr. 2025, pp. 1–5. DOI: [10.1109/icassp49660.2025.10889681](https://doi.org/10.1109/icassp49660.2025.10889681). URL: <https://hal.science/hal-04889885> (cit. on p. 20).
- [12] L. Hémadou, H. Vorobieva, A. Nasreddine Benaïchouche, F. Jurie and E. Kijak. ‘Cross-task knowledge distillation for few-shot detection’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2025. ICCVW 2025 - International Conference on Computer Vision Workshops*. Honolulu, Hawaii, United States: IEEE, 2025, pp. 1–6. URL: <https://hal.science/hal-05268562> (cit. on p. 20).
- [13] C. Imadache, E. Giboulot and T. Furon. ‘Evaluating the security of public surrogate watermark detectors’. In: *Proc. of the IEEE ICASSP. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. DOI: [10.1109/ICASSP49660.2025.10889821](https://doi.org/10.1109/ICASSP49660.2025.10889821). URL: <https://hal.science/hal-05168353> (cit. on p. 18).
- [14] Q. L. Roux, Y. Teglia, E. Bourbao, P. Loubet-Moundi and T. Furon. ‘BAIT: A new dnn backdoor attack using inpainted triggers’. In: *Proc. of the IEEE ICIP. ICIP 2025 - IEEE International Conference on Image Processing*. Anchorage (AK), United States: IEEE, Sept. 2025, pp. 1–6. URL: <https://hal.science/hal-05168358> (cit. on p. 16).
- [15] T. Sander, P. Fernandez, A. Durmus, T. Furon and M. Douze. ‘Watermark anything with localized messages’. In: *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR’25)*. International Conference on Learning Representations - ICLR 2025. Singapore, Singapore, Apr. 2025. URL: <https://hal.science/hal-04970818> (cit. on p. 19).
- [16] R. Webster and T. Furon. ‘Multi-modal Identity Extraction’. In: *Proc. of the ICCV. ICCV 2025 - International Conference on Computer Vision*. Honolulu, Hawaii, United States, 2025, pp. 1–9. URL: <https://hal.science/hal-05168368> (cit. on p. 17).

Conferences without proceedings

- [17] B. Puffer, Y. Belousov, V. Kinakh, T. Furon and S. Voloshynovskiy. ‘Task-Agnostic Attacks Against Vision Foundation Models’. In: 5th Workshop of Adversarial Machine Learning at CVPR 2025. Nashville, United States, 5th Mar. 2025, pp. 1–18. URL: <https://hal.science/hal-05172315> (cit. on p. 16).

Reports & preprints

- [18] T. Chauvin, E. Le Merrer, F. Taïani and G. Tredan. *Log Probability Tracking of LLM APIs*. 3rd Dec. 2025. URL: <https://hal.science/hal-05421014>.
- [19] J. Figeac, E. Le Merrer, M. Neihouser and G. Trédan. *Algorithmic curation of News on YouTube: Evidence from the 2022 French Presidential Campaign*. 2025. URL: <https://shs.hal.science/halshs-05398989> (cit. on p. 19).
- [20] Q. L. Roux, Y. Teglia, T. Furon, P. Loubet-Moundi and E. Bourbao. *Survivability of Backdoor Attacks on Unconstrained Face Recognition Systems*. 2nd July 2025. URL: <https://hal.science/hal-05168341> (cit. on p. 16).

10.3 Cited publications

- [21] K. W. (TechCrunch). *Deepfakes for all: Uncensored AI art model prompts ethics questions*. Ed. by TechCrunch. 2022. URL: <https://techcrunch.com/2022/08/24/deepfakes-for-all-uncensored-ai-art-model-prompts-ethics-questions> (cit. on p. 12).
- [22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang. ‘Deep learning with differential privacy’. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318 (cit. on p. 9).
- [23] M. Barni, K. Kallas and B. Tondi. ‘A new backdoor attack in cnns by training set corruption without label poisoning’. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 101–105 (cit. on p. 10).
- [24] B. Biggio, B. Nelson and P. Laskov. ‘Poisoning attacks against support vector machines’. In: *Proceedings of the 29th International Conference on Machine Learning*. 2012, pp. 1467–1474 (cit. on p. 10).
- [25] A. Blum, K. Ligett and A. Roth. ‘A learning theory approach to noninteractive database privacy’. In: *Journal of the ACM (JACM)* 60.2 (2013), pp. 1–25 (cit. on p. 9).
- [26] M. Bober-Irizar, I. Shumailov, Y. Zhao, R. Mullins and N. Papernot. ‘Architectural backdoors in neural networks’. In: *arXiv preprint arXiv:2206.07840* (2022) (cit. on p. 10).
- [27] J. G. Bourrée, E. L. Merrer, G. Tredan and B. Rottembourg. ‘On the relevance of APIs facing fairwashed audits’. In: *arXiv preprint arXiv:2305.13883* (2023) (cit. on p. 11).
- [28] T. Brooks, A. Holynski and A. A. Efros. ‘InstructPix2Pix: Learning to Follow Image Editing Instructions’. In: *arXiv preprint 2211.09800* (2023) (cit. on p. 12).
- [29] X. Cao, J. Jia and N. Z. Gong. ‘IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary’. In: *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 2021, pp. 14–25 (cit. on p. 10).
- [30] N. Carlini. *A Complete List of All (arXiv) Adversarial Example Papers*. <https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>. (Visited on 2022) (cit. on p. 10).
- [31] ChatGPT. *website: https://chat.openai.com*. URL: <https://chat.openai.com> (cit. on p. 8).
- [32] C. Dwork, A. Roth et al. ‘The algorithmic foundations of differential privacy’. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407 (cit. on p. 9).

- [33] G. Evennou, V. Chappelier, E. Kijak and T. Furon. ‘SWIFT: Semantic Watermarking for Image Forgery Thwarting’. In: *Proc. of IEEE WIFS*. IEEE, Roma, Italy: IEEE, Dec. 2024, pp. 1–6. URL: <https://hal.science/hal-04728070> (cit. on p. 13).
- [34] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou and M. Douze. ‘Tatouage Numérique d’Images dans l’Espace Latent de Réseaux Auto-Supervisés’. In: *GRETSI 2022 - Colloque Francophone de Traitement du Signal et des Images*. Nancy, France, Sept. 2022, pp. 1–4. URL: <https://hal.archives-ouvertes.fr/hal-03696016> (cit. on p. 13).
- [35] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou and M. Douze. ‘Watermarking Images in Self-Supervised Latent Spaces’. In: *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*. Ed. by IEEE. IEEE, Singapore, Singapore: IEEE, May 2022, pp. 1–5. URL: <https://hal.inria.fr/hal-03591396> (cit. on p. 13).
- [36] C. Feutry. ‘Two sides of relevant information : anonymized representation through deep learning and predictor monitoring’. PhD thesis. Université Paris-Saclay, 2019 (cit. on p. 10).
- [37] E. Gesny, E. Giboulot and T. Furon. ‘When does gradient estimation improve black-box adversarial attacks?’ In: *Proceedings of IEEE WIFS 2024*. Ed. by IEEE. Roma, Italy: IEEE, Dec. 2024, pp. 1–6. URL: <https://hal.science/hal-04728275> (cit. on p. 13).
- [38] E. Giboulot and T. Furon. ‘WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off’. In: *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. Vancouver, Canada, Dec. 2024, pp. 1–34. URL: <https://hal.science/hal-04766606> (cit. on p. 13).
- [39] S. Goldwasser, M. P. Kim, V. Vaikuntanathan and O. Zamir. ‘Planting undetectable backdoors in machine learning models’. In: *arXiv preprint arXiv:2204.06974* (2022) (cit. on p. 10).
- [40] A. Grivet Sébert, R. Pinot, M. Zuber, C. Gouy-Pailler and R. Sirdey. ‘SPEED: Secure, PrivatE, and Efficient Deep learning’. In: *Machine Learning* 110.4 (2021), pp. 675–694 (cit. on p. 10).
- [41] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan and B. Guo. ‘Vector quantized diffusion model for text-to-image synthesis’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10696–10706 (cit. on p. 8).
- [42] I. Han, S. Yang, T. Kwon and J. C. Ye. ‘Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion’. In: *arXiv preprint 2303.08767* (2023) (cit. on p. 12).
- [43] S. Hong, Y. Kaya, I.-V. Modoranu and T. Dumitras. ‘A Panda? No, It’s a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference’. In: *International Conference on Learning Representations*. 2020 (cit. on p. 9).
- [44] M. Huang, S. Jia, M.-C. Chang and S. Lyu. ‘Text-Image De-Contextualization Detection Using Vision-Language Models’. In: 2022, pp. 8967–8971 (cit. on p. 13).
- [45] H. Jia, H. Chen, J. Guan, A. S. Shamsabadi and N. Papernot. ‘A Zest of LIME: Towards Architecture-Independent Model Distances’. In: *International Conference on Learning Representations*. 2021 (cit. on p. 11).
- [46] E. Le Merrer and T. Gilles. ‘Tampernn: efficient tampering detection of deployed neural nets’. In: *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2019, pp. 424–434 (cit. on p. 11).
- [47] E. Le Merrer, P. Perez and G. Trédan. ‘Adversarial frontier stitching for remote neural network watermarking’. In: *Neural Computing and Applications* 32 (2020), pp. 9233–9244 (cit. on p. 13).
- [48] E. Le Merrer and G. Trédan. ‘LLMs hallucinate graphs too: a structural perspective’. In: *complex networks*. Istanbul, Turkey: Springer, Dec. 2024. URL: <https://hal.science/hal-04684742> (cit. on p. 12).
- [49] E. Le Merrer and G. Trédan. ‘Remote explainability faces the bouncer problem’. In: *Nature Machine Intelligence* 2.9 (2020), pp. 529–539 (cit. on p. 11).
- [50] Q. Le Roux, K. Kallas and T. Furon. ‘A Double-Edged Sword: The Power of Two in Defending Against DNN Backdoor Attacks’. In: *EUSIPCO 2024 - 32nd IEEE European Signal Processing Conference*. Lyon, France: IEEE, Aug. 2024, pp. 2007–2011. DOI: [10.23919/EUSIPCO63174.2024.10715340](https://doi.org/10.23919/EUSIPCO63174.2024.10715340). URL: <https://hal.science/hal-04850574> (cit. on p. 11).

- [51] Z. Lin, S. Wang, V. Sekar and G. Fanti. ‘Distributional Privacy for Data Sharing’. In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*. 2022. URL: <https://openreview.net/forum?id=6oVAzFsHlFK> (cit. on p. 9).
- [52] Y. Liu, X. Ma, J. Bailey and F. Lu. ‘Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks’. In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm. Cham: Springer International Publishing, 2020, pp. 182–199 (cit. on p. 10).
- [53] G. Luo, T. Darrell and A. Rohrbach. ‘NewsCLIPPings: Automatic Generation of Out-of-Context Multimodal Media’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 6801–6817 (cit. on p. 13).
- [54] T. Maho, T. Furon and E. Le Merrer. ‘Empreinte de réseaux avec des entrées authentiques’. In: *Conference on Artificial Intelligence for Defense*. Actes de la 4ème Conference on Artificial Intelligence for Defense (CAID 2022). DGA Maîtrise de l’Information. Rennes, France, Nov. 2022. URL: <https://hal.archives-ouvertes.fr/hal-03879849> (cit. on pp. 10, 11).
- [55] T. Maho, T. Furon and E. L. Merrer. *FBI: Fingerprinting models with Benign Inputs*. submitted to IEEE Trans. on Information Forensics and Security. 2022. DOI: [10.48550/ARXIV.2208.03169](https://doi.org/10.48550/ARXIV.2208.03169). URL: <https://arxiv.org/abs/2208.03169> (cit. on p. 10).
- [56] T. Maho, S.-M. Moosavi-Dezfooli and T. Furon. ‘How to choose your best allies for a transferable attack?’ In: *IEEE Int. Conf. on Computer Vision, ICCV*. 2023 (cit. on p. 10).
- [57] E. F. Moore. ‘Gedanken-Experiments on Sequential Machines’. In: *Automata Studies. (AM-34), Volume 34*. Ed. by C. E. Shannon and J. McCarthy. Princeton: Princeton University Press, 1956, pp. 129–154. DOI: [doi:10.1515/9781400882618-006](https://doi.org/10.1515/9781400882618-006). URL: <https://doi.org/10.1515/9781400882618-006> (cit. on p. 8).
- [58] A. S. Rakin, Z. He and D. Fan. ‘Tbt: Targeted neural network attack with bit trojan’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13198–13207 (cit. on p. 10).
- [59] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen. ‘Hierarchical text-conditional image generation with clip latents’. In: *arXiv preprint arXiv:2204.06125* (2022) (cit. on p. 8).
- [60] B. Razeghi, F. P. Calmon, D. Gunduz and S. Voloshynovskiy. ‘Bottlenecks CLUB: Unifying Information-Theoretic Trade-offs Among Complexity, Leakage, and Utility’. In: *arXiv preprint arXiv:2207.04895* (2022) (cit. on p. 10).
- [61] M. H. (T. Review). *How to create, release, and share generative AI responsibly*. 2023 (cit. on p. 13).
- [62] M. Ribeiro, K. Grolinger and M. A. Capretz. ‘MLaaS: Machine Learning as a Service’. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. 2015, pp. 896–902. DOI: [10.1109/ICMLA.2015.152](https://doi.org/10.1109/ICMLA.2015.152) (cit. on p. 10).
- [63] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer. ‘High-resolution image synthesis with latent diffusion models’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695 (cit. on p. 8).
- [64] Q. L. Roux, E. Bourbao, Y. Teglia and K. Kallas. ‘A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems’. In: *IEEE Access* 12 (2024), pp. 47433–47468. DOI: [10.1109/ACCESS.2024.3382584](https://doi.org/10.1109/ACCESS.2024.3382584). URL: <https://hal.science/hal-04850549> (cit. on p. 11).
- [65] A. Sablayrolles, M. Douze, C. Schmid and H. Jégou. ‘Radioactive data: tracing through training’. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8326–8335 (cit. on p. 9).
- [66] T. Sander, P. Fernandez, A. Durmus, M. Douze and T. Furon. ‘Watermarking Makes Language Models Radioactive’. In: *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. Vol. Spotlight. Vancouver, Canada, Dec. 2024, pp. 1–35. URL: <https://hal.science/hal-04766621> (cit. on p. 13).
- [67] C. Stokel-Walker. ‘AI bot ChatGPT writes smart essays-should academics worry?’ In: *Nature* (2022) (cit. on p. 8).
- [68] Z. Sun, R. Sun, C. Liu, A. R. Chowdhury, S. Jha and L. Lu. ‘ShadowNet: A secure and efficient system for on-device model inference’. In: *arXiv preprint arXiv:2011.05905* (2020) (cit. on p. 10).

- [69] Y. Viazovetskyi, V. Ivashkin and E. Kashin. ‘Stylegan2 distillation for feed-forward image manipulation’. In: *European conference on computer vision*. Springer. 2020, pp. 170–186 (cit. on p. 8).
- [70] M. de Vos, A. Dhasade, J. Garcia Bourrée, A.-M. Kermarrec, E. Le Merrer, B. Rottembourg and G. Trédan. ‘Fairness Auditing with Multi-Agent Collaboration’. In: *Frontiers in Artificial Intelligence and Applications*. Frontiers in Artificial Intelligence and Applications. Santiago de Compostela, Spain: IOS Press, Oct. 2024, pp. 1–14. DOI: [10.3233/FAIA240604](https://doi.org/10.3233/FAIA240604). URL: <https://laas.hal.science/hal-04800328> (cit. on p. 12).
- [71] S. Wagh, S. Tople, F. Benhamouda, E. Kushilevitz, P. Mittal and T. Rabin. ‘Falcon: Honest-Majority Maliciously Secure Framework for Private Deep Learning’. In: *Proceedings on Privacy Enhancing Technologies* 1 (2021), pp. 188–208 (cit. on p. 9).
- [72] Y. Wang and A. Dantcheva. ‘A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes’. In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 2020 (cit. on p. 12).