

2025 Activity Report

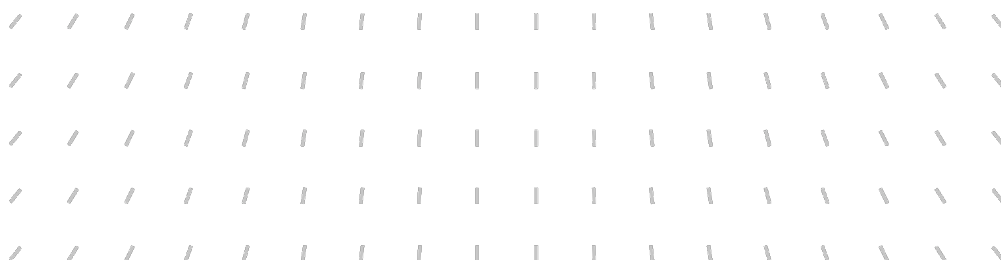
RESEARCH CENTRE: Inria Centre at the University of Bordeaux


Project-Team

REGALIA

Regulations for Artificial Intelligence





Project-Team REGALIA

Creation of the Project-Team: 2025 November 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

- A8.12. – Optimal transport
- A9.2. – Machine learning
- A9.11. – Generative AI
- A9.14. – Evaluation of AI models
- A9.16. – Societal impact of AI

Other research topics and application domains

- B9.6.10. – Digital humanities
- B9.6.12. – Philosophy

Contents

Project-Team REGALIA	1
1 Team members, visitors, external collaborators	4
2 Overall objectives	4
3 Research program	5
4 Application domains	7
5 Social and environmental responsibility	7
5.1 Footprint of research activities	7
5.2 Impact of research results	7
6 Highlights of the year	7
7 Latest software developments, platforms, open data	9
7.1 New platforms	9
8 New results	9
8.1 Evaluation Science	9
8.2 Machine Learning for Trustworthy AI	11
8.3 AI systems Audit	11
9 Bilateral contracts and grants with industry	12
9.1 Bilateral contracts with industry	12
9.2 Bilateral Grants with Industry	12
10 Partnerships and cooperations	13
10.1 International research visitors	13
10.1.1 Other european programs/initiatives	13
10.2 National initiatives	13
10.2.1 ANR Regulia (2023-2027) Coordinator : J-M. Loubes (480 KEuros)	13
10.3 Chair in AI Cluster	14
10.3.1 Head of National Program of Evaluation of AI by Agence des Programmes	14
10.4 Regional initiatives	14
10.5 Public policy support	14
11 Dissemination	15
11.0.1 Scientific events: selection	15
11.0.2 Invited talks	15
11.0.3 Scientific expertise	16
11.0.4 Research administration	16
11.1 Teaching - Supervision - Juries - Educational and pedagogical outreach	17
11.1.1 Teaching	17
11.1.2 Supervision	17
11.1.3 Juries	17
11.2 Popularization	17
11.2.1 Productions (articles, videos, podcasts, serious games, ...)	18
12 Scientific production	18
12.1 Major publications	18
12.2 Publications of the year	18

1 Team members, visitors, external collaborators

Research Scientists

- Jean-Michel Loubes [Team leader, INRIA, Professor Detachement, from Nov 2025]
- Carina Prunkl [INRIA, ISFP, from Nov 2025]
- Benoit Rottembourg [INRIA, Senior Researcher, from Nov 2025]

Technical Staff

- Remy Sourial [INRIA, Engineer, from Nov 2025]

Administrative Assistants

- Flavie Blondel [INRIA]
- Anne-Laure Gautier [INRIA]

2 Overall objectives

Understanding the behavior of AI algorithms and, in particular, machine learning algorithms has become an important challenge for regulation authorities but also for industry players wishing to develop algorithms that comply with legal regulations or ethics requirements. On the one hand, the presence of bias and discrimination is well acknowledged in machine learning and may lead to unwanted behavior and predictions. Rather than providing decisions that appear as sharp and accurate, algorithms may perpetuate or even exacerbate biases in the training data, resulting in decisions that may discriminate part of the users. On the other hand, the opacity and the lack of explainability of such algorithms may lead to behaviors that may be different from the original claims on the purpose of the algorithm, leading to harmful behavior.

The objectives of this project are twofold.

1. we want to develop new tools to decide whether a machine learning algorithm is compliant with norms. The norms can be of legal nature, according to legal or economical regulations, but can also be described by the designers of the algorithm. The research should provide mathematical guarantees to design feasible ways of :
 - deciding whether an algorithm violates legal regulations. In Europe, such regulations are mainly driven on the one hand by regulations from the GDPR, the DSA or the AI Act, including a lack of efficiency or biased behaviour in their moderation or recommendation processes; and on the other hand by regulations from competition policy in economics. Practices like self-preferencing, supplier favoritism and various abusive pricing practices by “gate keepers” are addressed by the DMA. Algorithmic collusion is also a new threat rising from large platforms that have digitized their dynamic pricing processes.
 - verifying that the algorithm complies with the set of requisites or norms that are announced (chosen and communicated in the General Terms and Conditions) by the constructor of the algorithm. Otherwise, the algorithm is disloyal with respect to its own claim. An algorithm must justify its decisions by explaining the principles it uses to make them. Hence, auditing should confront these prior principles to the reality of the decisions that were taken. For example, recommender systems claim that the decision respects the interests of the users while other hidden utility functions, that may favor some retailers with regard to others, may be hidden in the optimization process.

The new tools should be able to detect but also to provide certifiable proofs of compliance or non compliance. This implies that a single measure of bias is not enough but that the deviations with respect to the distributions of the data and the optimization landscape of the algorithm must be controlled.

2. once the evidence of faulty behaviors is obtained, the second aspect of the project will be to provide new methods to build and certify *loyal* algorithms, dealing either with fairness or harmful behavior. This includes the different methodologies that consist in pre-processing the data, controlling the learning phase of the algorithm or post-processing the outputs of the algorithms, depending of the availability of the data and algorithm.

One of the specificity of the proposed research relates first to its positioning: we shall deal with *a posteriori* auditing of a trained algorithm in the so-called grey or black box models. Actually, audits can take various forms, according to the context of the analysis and the auditor itself. Audit methodologies vary according to the context and ownership of the audits to be performed :

- White box auditing occurs when the auditor in charge of the audit has a full access (both quantitatively and qualitatively) to the training data, objective function and constraints of the algorithm at stake. This happens when the owner itself of the algorithm wants to check the validity of some robustness or fairness assumptions after having developed the algorithm ;
- Black box auditing consists in only monitoring input and output of an algorithm without prior knowledge on its behaviour or on the training data set. Causality in the decisions has to be discovered from the observation on how it operates and then to be inferred from the observations. In those contexts, the auditor might be limited in the amount of queries she can perform on the black box algorithm ;
- Audits can be considered as Grey audit if partial knowledge is known before observation. The knowledge can reside in the objective function, the constraints, hidden variables used by the algorithm or the nature of the algorithm itself.

In this setting, the model and the test sample are either partially known or have to be discovered (by exploration) and thus properly approximated. The difficulty stems from the nature of learning algorithms, whose properties depend on (i) the characteristics of the training data, (ii) the optimizer and the dynamics of learning or fine-tuning, and (iii) the distribution of the data used to evaluate the algorithm. This third aspect is especially salient for large language models, where prompts inject context that conditions the model at inference and shifts its activation patterns. Our approach is innovative: whereas much of the literature focuses solely on properties of learning data or conflates these three sources, we will treat them explicitly.

3 Research program

Here is the outline of the project which is structured into the following work packages WP.

WP 1: Measuring harmful behaviours of machine learning algorithms . This workpackage develop methods to detect, quantify, and explain harmful behaviours in ML algorithms.

1. WP 1.1 [Quantifying the effect of training or fine tuning of ML algorithms]: We will study how optimisation during training and fine-tuning changes model behaviour, with a focus on mechanisms that amplify undesired behaviour, such as hallucination rates, toxic content, or harmful persona adoption. This involves (i) developing new metrics that quantify how optimisation transforms or amplifies properties in the training data, (ii) build a typology of model/optimiser combinations by risk profile, and (iii) develop recommendations to reduce amplification and distortion effects.
2. WP 1.2 [Designing data agnostic metrics]: We will develop evaluation metrics to quantify a model's distributional sensitivity and robustness. This includes considering distributional changes in the sample, including partially or fully unseen regimes.
3. WP 1.3: [Locality of Harmful decisions] we will not only assess the properties on the complete datasets but discover locally the vulnerabilities of the AI algorithm. This will enable to reveal hidden behaviour and may require exploring the set of possible but yet plausible values of the inputs, involving active learning based strategies [AXIS 1 and Axis 2].

WP 2: Auditing general-purpose AI

1. WP 2.1 [Assessing Evaluation of AI and GPAI] : Beyond classical algorithms in supervised and unsupervised machine learning, a particular focus will lie on general-purpose AI (GPAI), including text and image generation. These algorithms are used for recommendation algorithms, pricing algorithms, profiling algorithms based on clustering methods, and usual classification algorithms or general-purpose AI algorithms. We will create a taxonomy of current evaluation methods of GPAIs (including LLMs), especially with regard to bias, robustness, safety, and performance. [Axis 3 and Section 4.5]
2. WP 2.2 [Building Test Benches for AI and GPAI]: stress-testing these evaluation methods, in particular for how appropriate they are, how comprehensive they are, whether they measure what we want them to measure, and whether there's any external validity (and if not, how we can create this). We will in particular construct stress tests based on risk assessments [Axis 1 and Axis 2 and Section 4.5]
3. WP 2.3 [Designing new evaluation methods of GPAI]: Using the metrics of WP1 to develop new methods to assess properties of LLMs that improve upon current methods such as benchmarks and red teaming. [Axis 1, 3 and Section 4.5]

WP 3: Aligning metrics and audit technics with societal needs [AXIS 2]

1. WP 3.1 : [Alignment with regulatory requirements] [Axis 1 and Section 4.5]: the EU AI Act and the GPAI Code of Practice (CoP) require providers/deployers to substantiate safety, transparency, and governance claims with documentation and evidence. This WP develops a validity- and context-aware method for translating legal/CoP requirements into defensible evaluation claims. In particular, we will produce methods for making evaluation claims explicit: what a metric/benchmark does measure, which assumptions it relies on, how it transports across contexts, and how uncertainty should be reported. Outputs include (i) a requirement-to-evidence matrix (moving from obligation to evaluation to design to admissible evidence), (ii) an audit-ready reporting pack (templates with minimal documentation modules), and (iii) an 'evaluation' library with common overclaims and validation checks.
2. WP 3.2 : Link with ethical needs [Axis 1 and Section 4.5]: this WP specifies the normative targets of evaluation and auditing, focusing on agency, meaningful human oversight (Art 14 AI Act), legitimacy, and contestability. For example, we will develop methodologies to effectively operationalise complex notions such as agency, specify what is required for human oversight to be meaningful and effective, and analyse evaluation practices more broadly. This WP feeds into both WP2 and WP1.
3. WP 3.3 : Creation of a toolbox to audit and detect vulnerabilities of ML algorithms, including stress tests and world models [Final Output]: This WP integrates the measurement methods from WP1 and the mitigation/oversight mechanisms from WP2 into a practitioner-facing audit toolbox. It combines (i) stress tests for misuse and distribution shift, (ii) targeted discovery of hidden failure modes via testing, and (iii) evaluations for long-horizon or multi-step interactions. In addition to individual model vulnerabilities, it supports audits of system-level and systemic risks, such as scalable misuse or emergent harmful behaviours.

The timeline of the project is the following:

- WP 1 will be the first to be tackled, in parallel with WP 3.1 and WP 3.2. The tasks of WP 1.1 correspond to the work already initiated by J-M. Loubes and his research team in ANITI. A Post-doc is hired on this subject. WP 1.2 is also in progress by the same group. The research in WP 1.3 is led by B. Rottembourg and his engineering team, involving also J-M. Loubes. These tasks will last 1 or 2 years but will then be continued and adapted in WP 2.
- WP 2 is a more prospective research axis. Due to the growing evolution of research in this field and the rapid changes in the algorithms, we will on the one hand generate mathematical sound results which are grounded by theoretical results and provide a better understanding of the behaviours of the algorithms. This work will benefit from results from WP 1 and WP 3. On the other hand we will try to keep the pace and provide auditing methods that use relevant metrics that are computed using an innovative test bench.

- WP 3 is led by C. Prunkl in collaboration with the other researchers. The work on WP 3.1 and 3.2 will build on her previous work to align ethics, regulation and technical constraints. We point out that regulation of AI is a moving environment where norms are not completely written and there is a gap between regulations based on risk assessment and technical ways to prove compliance. We will first work on WP 3.1 trying to fill this gap. Then we will work on future risks of AI in WP 3.1. Finally WP 3.3 which aims to build a library or a software for AI assessment and the mitigation will be continuously developed with different versions. We will first focus on a banking use case and the specific risk analysis.

Collaboration:

- with researchers from the **ANITI's research chair TRIAL** (University from Toulouse in Machine learning F. Iutzeler, C. Lelanne, Toulouse School in Economics in optimisation (J. Bolte), University of Law of Toulouse in legal science (J. Eynard) and University Lille in uncertainty quantification (F. Bachoc))
- with researchers from **LAAS** (Gilles Tredant)
- with researchers from **INRIA Rennes** (E. Le Merrer)
- with researchers from **INRIA Bordeaux** (Marta Avalos and Ariel Adama-Guerra SISTM)

4 Application domains

Our main domain application is the evaluation of AI algorithms. We consider both point of views : the regulation authority who aims at auditing AI models and verify their compliance with respect to some regulations requirements, either driven by regulation on AI or driven by sectorial specifications.

5 Social and environmental responsibility

5.1 Footprint of research activities

Our research on AI does not involve high computational costs similar to the one necessary to train large foundational models. We focus on evaluation and for this provide open source evaluation platform or software for national authorities.

We decided to reduce our carbon footprint by limiting overseas conference travel: we will send only one presenter per conference and avoid long-distance group travel whenever possible.

5.2 Impact of research results

The main research topics of the team contribute to improve robustness, transparency, fairness and privacy in machine learning and AI algorithms. The aim is to provide theoretical foundations for a better trustworthy AI and contribute to promoting guidelines for more compliant AI systems.

In this setting, we promote **metrics** to guarantee conformity and **guidelines** to ensure human oversight of AI systems.

6 Highlights of the year

- **Publication : when majority rules, minority loses.** (Published at Neurips 2025, oral presentation at EUrips 2025)

In this paper, we develop the first theory for bias amplification by machine learning type algorithms. This result is an important step to better understand bias in AI, in particular when the bias is unknown a priori.

When discussing algorithmic bias in machine learning algorithms, we are often conflating three distinct notions: the bias inherent in the training data, the bias present in the data used for evaluation, and

the bias introduced by the training process itself. This last source of bias is particularly significant to understand how so-called AI algorithms are so sensitive to bias. As a matter of fact, in many studies, an algorithm built using machine learning is said not only to learn and reproduce the bias present in the data, but also to amplify it.

We develop a mathematical framework to capture these phenomena and identify the core mechanisms behind bias amplification—specifically, population and variability imbalance, along with their geometric and dynamical implications. We can provide some indications in order to better train the model and to prevent lack of performance due to bias, even in the case of unknown bias source.

We are confident that this is the first step of a series of work to better understand the particular role of the learning part in the apparition (and amplification) of bias.

- **Rapport ANSES : Evaluation des risques sanitaires pour les travailleurs des plateformes numériques de livraison de repas en France** Collaborative work.

The French Agency for Food, Environmental and Occupational Health & Safety (Anses) was commissioned on March 8, 2021, by the Confédération Générale du Travail (CGT) to conduct the following assessment: evaluating the health risks for workers on digital food delivery platforms in France.

CONTEXT AND PURPOSE OF THE REQUEST

With the rise of communication technologies, the facilitated connection between individuals has enabled the development of new economic models, particularly the growth of digital platforms. The types of work offered by these platforms, due to their flexible hours and accessibility (no required level of education), attract many workers. An increasing number of consumers use these interfaces to order goods or services, thereby increasing the demand for workers on these platforms. In light of this situation and the growing number of affected workers, the Confédération Générale du Travail (CGT)—meeting the conditions of Article L.1313-3, paragraph 2, of the Public Health Code—requested that Anses evaluate the health risks for workers on digital food delivery platforms. This evaluation should consider all exposures related to the practice of the activity (accidents, biomechanical constraints, psychosocial risks, air pollution, thermal constraints, etc.), the specific working conditions linked to the organization of the activity, and their relationships with the digital platforms. This assessment was conducted within an evolving regulatory context, both at the French level—such as the adoption of new remuneration rules for delivery workers—and at the European level, with debates surrounding a European directive concerning independent workers. As part of this assessment, Anses aimed to:

- Identify and characterize the digital platforms in France related to the food service market and the workers delivering meals by two-wheeled vehicles;
 - Analyze the associated economic model, the functioning of these platforms, their dynamics, the regulations governing them, and the relationships they establish with delivery workers (contracts, algorithms, etc.);
 - Describe the activities of delivery workers in relation to the characteristics of the work organization implemented by the platforms, particularly the use of technology and algorithmic management;
 - Characterize the risks to workers' health (population characteristics, vulnerability factors, occupational accident/illness statistics, environmental, physical, organizational, social, or psychosocial risk factors, and potential health effects);
 - Identify possible avenues and forms for developing prevention in occupational health and safety.
- **Policyreports: International AI Safety Report First and Second Key Update** In October 2025 and November 2025 we published the First and Second Key Updates of the International AI Safety Report, with a Regalia member as lead writer. The International AI Safety Report is an independent report on the scientific evidence about emerging risks from frontier general-purpose AI systems. More than 100 experts, including nominees from 30+ countries, the EU, the UN, and the OECD have contributed to the report, which is overseen by the Chair, Yoshua Bengio.

The First Key Update (published in October 2025) addresses general-purpose AI capability developments since January 2025, as well as their risk implications. Notable developments include major advances

in certain domains such as coding, mathematics, and scientific research (although reliability challenges persist); improvements were more driven post-training methods such as inference-time scaling and fine-tuning; the implementation of additional safeguards by major AI developers after pre-deployment testing found out they couldn't rule out their models helping novices to develop biological weapons; the fact that aggregated labour effects remain limited despite widespread adoption; and the fact that some systems exhibited strategic behaviour in test settings, raising potential oversight challenges.

The Second Key Update focussed on recent developments in risk management and technical safeguards. It included an overview over new regulatory and international initiatives, such as the EU AI Code of Practice and the OECD Incident Reporting Frameworks. It discusses various safeguards across different stages of the AI development lifecycle, such as watermarking, input filters, and other monitoring and intervention techniques.

- **Presentation at the OECD GPAI Plenary** Presentation of the above mentioned Key Updates at the OECD GPAI Plenary, with subsequent discussion with member states.
- **Recorded lecture at the College de France** Recorded lecture by Carina Prunkl as part of the "AI Days" at the College de France.

AI-driven decision-making is often framed as an antidote to the flaws of human judgment. It is praised for improving efficiency, ensuring consistency, and reducing personal biases. Critics, on the other hand, argue that AI systems - like humans - can replicate biases and that they lack mechanisms of accountability that apply to human decision-makers. Most of these discussions compare AI-driven decisions to individual decision-making, discussing how AI either improves or replaces human judgment. This talk challenges that framing, arguing that the more appropriate comparison is with bureaucratic decision-making. AI systems exhibit deep structural similarities with bureaucracies, including rule-based standardisation, depersonalised authority, task specialisation, and a strong reliance on quantification. Recognising this parallel shifts the focus from AI as a substitute for human judgment to AI as a system that mirrors bureaucratic decision-making, with its own institutional logic and constraints.

7 Latest software developments, platforms, open data

7.1 New platforms

Participants: Benoit Rottembourg, Jean-Michel Loubes.

Development of a platform for AI evaluation. **Testing Platform** The platform operates in open source conformity verification for robustness and bias of supervised machine learning algorithms. The platform performs a detection of vulnerabilities of supervised algorithms and provide a full analysis of its performance, robustness and the possible biases. This platform is open source on request and is designed to be tested by regulation authorities.

This platform will be maintained and enriched with additional tools, such as automated tracking of local bias.

8 New results

8.1 Evaluation Science

Participants: Jean-Michel Loubes.

When majority rules, minority loses: bias amplification of gradient descent.

Despite growing empirical evidence of bias amplification in machine learning, its theoretical foundations remain poorly understood. We develop a formal framework for majority-minority learning tasks, showing how standard training can favor majority groups and produce stereotypical predictors that neglect minority-specific features. Assuming population and variance imbalance, our analysis reveals three key findings: (i) the close proximity between “full-data” and stereotypical predictors, (ii) the dominance of a region where training the entire model tends to merely learn the majority traits, and (iii) a lower bound on the additional training required. Our results are illustrated through experiments in deep learning for tabular and image classification tasks.

Fairness is in the details: Face Dataset Auditing

Auditing involves verifying the proper implementation of a given policy. As such, auditing is essential for ensuring compliance with the principles of fairness, equity, and transparency mandated by the European Union’s AI Act. Moreover, biases present during the training phase of a learning system can persist in the modeling process and result in discrimination against certain subgroups of individuals when the model is deployed in production. Assessing bias in image datasets is a particularly complex task, as it first requires a feature extraction step, then to consider the extraction’s quality in the statistical tests. This paper proposes a robust methodology for auditing image datasets based on so-called “sensitive” features, such as gender, age, and ethnicity. The proposed methodology consists of both a feature extraction phase and a statistical analysis phase. The first phase introduces a novel convolutional neural network (CNN) architecture specifically designed for extracting sensitive features with a limited number of manual annotations. The second phase compares the distributions of sensitive features across subgroups using a novel statistical test that accounts for the imprecision of the feature extraction model. Our pipeline constitutes a comprehensive and fully automated methodology for dataset auditing.

Fairness-aware grouping for continuous sensitive variables: Application for debiasing face analysis with respect to skin tone

Within a legal framework, fairness in datasets and models is typically assessed by dividing observations into predefined groups and then computing fairness measures (e.g., Disparate Impact or Equality of Odds with respect to gender). However, when sensitive attributes such as skin color are continuous, dividing into default groups may overlook or obscure the discrimination experienced by certain minority subpopulations. To address this limitation, we propose a fairness-based grouping approach for continuous (possibly multidimensional) sensitive attributes. By grouping data according to observed levels of discrimination, our method identifies the partition that maximizes a novel criterion based on inter-group variance in discrimination, thereby isolating the most critical subgroups. We validate the proposed approach using multiple synthetic datasets and demonstrate its robustness under changing population distributions - revealing how discrimination is manifested within the space of sensitive attributes. Furthermore, we examine a specialized setting of monotonic fairness for the case of skin color. Our empirical results on both CelebA and FFHQ, leveraging the skin tone as predicted by an industrial proprietary algorithm, show that the proposed segmentation uncovers more nuanced patterns of discrimination than previously reported, and that these findings remain stable across datasets for a given model. Finally, we leverage our grouping model for debiasing purpose, aiming at predicting fair scores with group-by-group post-processing. The results demonstrate that our approach improves fairness while having minimal impact on accuracy.

When mitigating bias is unfair: multiplicity and arbitrariness in algorithmic group fairness Most research on fair machine learning has prioritized optimizing criteria such as Demographic Parity and Equalized Odds. Despite these efforts, there remains a limited understanding of how different bias mitigation strategies affect individual predictions and whether they introduce arbitrariness into the debiasing process. This paper addresses these gaps by exploring whether models that achieve comparable fairness and accuracy metrics impact the same individuals and mitigate bias in a consistent manner. We introduce the FRAME (Fairness Arbitrariness and Multiplicity Evaluation) framework, which evaluates bias mitigation through five dimensions: Impact Size (how many people were affected), Change Direction (positive versus negative changes), Decision Rates (impact on models’ acceptance rates), Affected Subpopulations (who was affected), and Neglected Subpopulations (where unfairness persists). This framework is intended to help practitioners understand the impacts of debiasing processes and make better-informed decisions regarding model selection. Applying FRAME to various bias mitigation approaches across key datasets allows us to exhibit significant differences in the behaviors of debiasing methods. These findings highlight the limitations of current fairness criteria and the inherent arbitrariness in the debiasing process.

8.2 Machine Learning for Trustworthy AI

Participants: Jean-Michel Loubes.

Improved learning theory for kernel distribution regression with two-stage sampling The distribution regression problem encompasses many important statistics and machine learning tasks, and arises in a large range of applications. Among various existing approaches to tackle this problem, kernel methods have become a method of choice. Indeed, kernel distribution regression is both computationally favorable, and supported by a recent learning theory. This theory also tackles the two-stage sampling setting, where only samples from the input distributions are available. In this paper, we improve the learning theory of kernel distribution regression. We address kernels based on Hilbertian embeddings that encompass most, if not all, of the existing approaches. We introduce the novel near-unbiased condition on the Hilbertian embeddings that enables us to provide new error bounds on the effect of the two-stage sampling, thanks to a new analysis. We show that this near-unbiased condition holds for three important classes of kernels, based on optimal transport and mean embedding. As a consequence, we strictly improve the existing convergence rates for these kernels. Our setting and results are illustrated by numerical experiments.

Hoeffding decomposition of functions of random dependent variables

Hoeffding's functional decomposition is the cornerstone of many post-hoc interpretability methods. It entails decomposing arbitrary functions of mutually independent random variables as a sum of interactions. Many generalizations to dependent covariables have been proposed throughout the years, which rely on finding a set of suitable projectors. This paper characterizes such projectors under hierarchical orthogonality constraints and mild assumptions on the variable's probabilistic structure. Our approach is deeply rooted in Hilbert space theory, giving intuitive insights on defining, identifying, and separating interactions from the effects due to the variables' dependence structure. This new decomposition is then leveraged to define a new functional analysis of variance. Toy cases of functions of bivariate Bernoulli and Gaussian random variables are studied.

On the Private Estimation of Smooth Transport Maps

Estimating optimal transport maps between two distributions from respective samples is an important element for many machine learning methods. To do so, rather than extending discrete transport maps, it has been shown that estimating the Brenier potential of the transport problem and obtaining a transport map through its gradient is near minimax optimal for smooth problems. In this paper, we investigate the private estimation of such potentials and transport maps with respect to the distribution samples. We propose a differentially private transport map estimator achieving an error of at most up to poly-logarithmic terms. We also provide a lower bound for the problem.

8.3 AI systems Audit

Participants: Benoit Rottembourg.

P2NIA : Privacy-Preserving Non-Iterative Auditing 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), in which 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.

Participants: Jean-Michel Loubes.

Exposing the illusion of fairness: Auditing vulnerabilities to distributional manipulation attacks

Proving the compliance of AI algorithms has become an important challenge with the growing deployment of such algorithms for real-life applications. Inspecting possible biased behaviors is mandatory to satisfy the constraints of the regulations of the EU Artificial Intelligence’s Act. Regulation-driven audits increasingly rely on global fairness metrics, with Disparate Impact being the most widely used. Yet such global measures depend highly on the distribution of the sample on which the measures are computed. We investigate first how to manipulate data samples to artificially satisfy fairness criteria, creating minimally perturbed datasets that remain statistically indistinguishable from the original distribution while satisfying prescribed fairness constraints. Then we study how to detect such manipulation. Our analysis (i) introduces mathematically sound methods for modifying empirical distributions under fairness constraints using entropic or optimal transport projections, (ii) examines how an auditee could potentially circumvent fairness inspections, and (iii) offers recommendations to help auditors detect such data manipulations. These results are validated through experiments on classical tabular datasets in bias detection.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Collaboration with French company EASYCASH

Participants: Benoit Rottembourg.

Auditing the Buybox mechanism of EASYCASH marketplace. EASYCASH is a group of 150 franchised point of sales selling used goods like phones and computers with 300 M€ annual income.

The purpose of the audit was to assess the fairness of the buybox algorithm electing the product to be pushed to the customer’s page. Our conclusions, in October 2025, helped to improve the algorithm and illustrate the mechanics in place in order to facilitate the understanding for the franchised.

Contract with Moroccan company OCP

OCP Group (formerly Office Chérifien des Phosphates) is a Moroccan state-owned phosphate rock miner, phosphoric acid manufacturer and fertilizer producer. Founded in 1920, the company has grown to become the world’s largest producer of phosphate and phosphate-based products and it is one of the largest phosphate, fertilizer, chemicals, and mineral industrial companies in the world by revenue. The Group employs nearly 17,000 people in Morocco. In 2024, it generated revenues of US \$9.76 billion

Initiated in partnership with INOCS team at Inria Lille, the purpose of the mission is to develop a new strategic stock allocation mechanism assigned to the regions and markets of OCP.

The contract has been launched in December 2025 and will last one year.

9.2 Bilateral Grants with Industry

Participants: Jean-Michel Loubes.

CIFRE Phd with L’Oréal and Artefact The contract has been started in 2024 for 3 years, along with a Phd funding. The topic of the contract is about fairness for beauty recommendation and 3D face generation.

Participants: Jean-Michel Loubes.

CIFRE Phd with EDF The contrat has started in 2024 with a Phd funding. The topic is explainability of AI decisions.

10 Partnerships and cooperations

10.1 International research visitors

Other international visits to the team

David Rodriguez-Vitores

Status (PhD)

Institution of origin: University of Valladolid

Country: Spain

Dates: 1/04/2025-1/07/2025

Context of the visit: work on fairness and privacy

Mobility program/type of mobility: funded by University of Valladolid (research stay)

10.1.1 Other european programs/initiatives

Participants: Jean-Michel Loubes.

Member of CEN-CENELEC Commission JTC 21 on the Europeans Norms for AI.

Participants: Jean-Michel Loubes, Benoit Rottembourg.

Participation to European Comminssion workshop on regulations of AI october 2024.

10.2 National initiatives

10.2.1 ANR Regulia (2023-2027) Coordinator : J-M. Loubes (480 KEuros)

Participants: Jean-Michel Loubes.

The objective of this project is to audit machine-learning algorithms and repair them when necessary.

New legislation (the GDPR and the European AI Act) provides a legal framework that will govern the practical deployment of algorithms. It sets out a number of recommendations that algorithms must follow. In particular, these algorithms should not behave differently across user subgroups unless those subgroups are identified in advance and such differences are justified. They must also clearly state what they are designed for and must not mislead users.

A large body of research already exists on assessing bias in machine learning and on explaining algorithmic decisions. First, this work should be pursued further in order to better understand these issues and to develop

procedures that can certify the presence or absence of bias. Moreover, the difficulty of auditing algorithms largely stems from the fact that the relevant measurements depend on the sample distribution. But in a “black-box” audit setting—i.e., when only the algorithm’s outputs on a dataset selected or provided by the auditor are available—it is also necessary to account for the variability of the algorithm with respect to the underlying distributions themselves.

Our goal in this project is therefore to develop new ways to define, detect, and control the effects of bias, in a uniform and robust manner, when the data-generating distribution is only partially known. Our approach is multidisciplinary: it draws on robust statistics and machine learning (mathematics and computer science) to define properties that hold over distributional neighborhoods; on Gaussian processes to build optimal experimental designs for discovering informative observations; and on optimization to make it possible to construct practical algorithms.

10.3 Chair in AI Cluster

Participants: Jean-Michel Loubes.

Regulating, auditing, and improving AI systems poses an unprecedented scientific challenge. Rising to it demands collaboration across many disciplines—computer science, mathematics, law, economics, and more. Our team is part of this research movement: we bring together expertise in machine learning, law, mathematics, optimization, computer science, and economics to create a fertile environment for reflection that addresses today’s grand challenges of trust and regulation. Our goal is to help lay the groundwork for a sustainable, ethical, and responsible future for AI technology, transforming the way these systems are managed and governed.

- Principa Investigator in Chair of AI Institute ANITI, responsible of synergy chair Trustworthy AI (TRIAL)
- Member of Scientific Direction of ANITI (6 members)

10.3.1 Head of National Program of Evaluation of AI by Agence des Programmes

Participants: Jean-Michel Loubes.

10.4 Regional initiatives

Participants: Jean-Michel Loubes.

Member of the Project "Occitanie Sandbox" The project aims at building a sandbox in Toulouse and Montpellier in order to prove conformity of AI algorithms in health and for transport. Occitanie Région is leading the consortium made by University of Montpellier, CHU Montpellier, ANITI. The piloting comitee is composed of 6 persons.

10.5 Public policy support

Participants: Jean-Michel Loubes.

Direction Projet National

- Member of Scientific and Direction Board of INESIA (National Institute for Evaluation and Security of AI)

I am in charge of the axis 2 of INESIA : systemic risks of IA. Participation aux colloques et formations:

- Training to DSI of CNRS : conformity of the AI algorithms (3 × 2 days november 2025)
- AI Training to french airforce officers (8 hours may 2025)

Participants: Benoit Rottembourg.

Coordination et expertise

- Coordinator for the scientific collaboration Inria-CNIL (French Commission Nationale de l'Informatique et des Libertés)
- Expert for the Tribunal judiciaire de Paris (Cybercrime Service)

Training and vulgarization

- Sénat Colloquium "Aligner l'IA : Faire d'un impératif une opportunité stratégique pour l'Europe" (Sénator Paoli-Gagin).
- Training to GDPE, "Le biais dans les algorithmes d'IA" (with Jean-Michel). Avril 2025
- Training to la Cour des Comptes. "FORMATION COUR DES COMPTES : AUDIT of LLM" (avec Jean-Michel Loubes)

11 Dissemination

11.0.1 Scientific events: selection

Chair of conference program committees

- Chair of "Pricing Algorithms 2025" conference program comity. Bordeaux, November 2025. . Benoit Rottembourg.

Member of the conference program committees

Participants: Carina Prunkl.

- CHI 2026 programme committee
- Undone Computer Science 2026 (organised by people from CNRS, Inria, and the University of Luxembourg), programme committee

11.0.2 Invited talks

- OECD GPAI Plenary (Nov 2025), Carina Prunkl
- Workshop on the Philosophy of AI, University of Louvain (February 2025) Carina Prunkl
- Juelich Speaker Series on the Philosophy of Technology, Juelich Forschungszentrum (June 2025) Carina Prunkl
- Seminar IHPST Paris 1 (September 2025) Carina Prunkl

- Paris Conference on AI and Digital Ethics, Paris (June 2025) Carina Prunkl
- Google DeepMind Seminar on AI and Society, (April 2025) Carina Prunkl
- New Directions Conference for the Philosophy of Physics, Slovenia (May 2025) Carina Prunkl
- PAISS Grenoble (September 2025) Carina Prunkl
- Seminar Radboud University (October 2025) Carina Prunkl
- **Journées de Statistique mathématique**(IHP Paris) January 2025 Jean-Michel Loubes
- Congrès Mondial de l'IA (Saclay) 2025 :Jean-Michel Loubes
- Séminaire Columbia University (April 2025: Jean-Michel Loubes)
- SCOR Foundation Workshop | Confidence and Fairness: Scientific Foundations in AI and Risk invited talk Jean-Michel Loubes
- **CONFIANCE AI**. Invited talk on fairness evaluation Jean-Michel Loubes (May 2025)
- Premières journées scientifiques INESIA (July 2025) Jean-Michel Loubes
- Optimal Transport Workshop Sociedad Espanola de Matematicas, workshop Castro Urdiales (November 2025) Jean-Michel Loubes
- **L'IA sous surveillance : biais et atteinte aux droits** INRIA Alumni (December 2025) J-M. Loubes & Benoit Rottembourg.
- Interdisciplinary Colloquium called Observatoire de la Surveillance en Démocratie. "La liberté de choix à l'heure des algorithmes : influences et surveillances" Benoit Rottembourg. (Feb 2025)
- Paris-Saclay Submit Choose Science "Sommes-nous sous la coupe des algorithmes ?" Benoit Rottembourg. (Feb 2025)"
- AI-product Day "Les biais dans les produits à base d'IA : risques & mitigation" Benoit Rottembourg. (March 2025)"
- Revenue Makers Day "Quand les algorithmes trichent : What could go wrong with AI pricing? " Benoit Rottembourg . (March 2025)"
- L'IA en question, questions à l'IA Debate at Centre Pompidou. "L'IA PEUT-ELLE VOUS REMPLACER DANS VOTRE TRAVAIL ?" Benoit Rottembourg. (May 2025)"
- Franco-German Chambre de commerce. "Faire danser un éléphant dans un magasin de porcelaine La fonction RH face à l'IA " Benoit Rottembourg. (June 2025)"
- Colloquium of Loria, Nancy. "Approches pour la détection de biais dans les recommandations algorithmiques " Benoit Rottembourg. (July 2025)"

11.0.3 Scientific expertise

- Formation for administrative staff of Toulouse School of Economics : AI for dministration june 2025 Jean-Michel Loubes
- Member of Scientific Comitee AI for Health of région Occitanie Jean-Michel Loubes

11.0.4 Research administration

- Jean-Michel Loubes is Member of Scientific Comitee of ANITI (6 members)

11.1 Teaching - Supervision - Juries - Educational and pedagogical outreach

11.1.1 Teaching

- Course on Trustworthy AI at Basq center for Excellence BECAM and University of Pamplona (Spain) (12 hours) Jean-Michel Loubes
- Course on Fairness and Bias Mitigation at University of Medellin (Colombia) (8 hours) Jean-Michel Loubes
- Course on AI Evaluation (Projet année M2 SID Toulouse) Jean-Michel Loubes
- Projets Recherche X 3A Jean-Michel Loubes

11.1.2 Supervision

- Post Doc : 01-08 2025 : Bilel Bensaid supervised by J-M. Loubes. Analysis of learning tasks collapse in overparametrized models.
- Post Doc : 09/2025 09/2026 : Clément Lezanne supervised by J-M. Loubes. Role of optimisers in bias amplification second order methods.
- Doc : Gayane Taturyan defended on september 2025 co-supervised by J-M. Loubes and M. Hebiri (Université G. Eiffel)
- Doc : V. Lafargue (01/20025 ...) cosupervised by J-M. Loubes and E. Claeys (IRIT Toulouse) Auditing Generative AI models funded by ANR Regulia
- Doc : Mahdi Tavassoli (08/2023-08/2026): New methods for assessing fairness on graph supervised by J-M. Loubes, funded by TUPLES European Project.
- Doc : Gabriel Ferrere (01/2024...) cosupervised with F. Gamboa and N. Bousquet (EDF) : new methods for explainability of AI algorithms based on Hoeffding decomposition funded by EDF CIFRE grant
- Doc : Veronika Suvorika (2023-...) cosupervised with L. Risser : fairness on face generation funded by Loreal CIFRE Grant.

11.1.3 Juries

- Phd Reviewer : Antoine Barczewski (Jan Ramon INRIA Magnet) defended in september 2025 Jean-Michel Loubes
- Phd Reviewer : ATIENZA Nicolas (M. Sebag) defended in may 2025 Jean-Michel Loubes
- Phd Reviewer : Mathieu Molina (Crest G. Loiseau et Vianney-Perchet) defended september 2025 Jean-Michel Loubes
- Phd Reviewer : Marco Favier (Toon Calders University of Antwerp) defended in october 2025. Jean-Michel Loubes

11.2 Popularization

- **Interview** of Jean-Michel Loubes at France Info
- **Regional Press** La Dépêche du Midi de Jean-Michel Loubes
- TV Show. "IA : l'Europe peut-elle garder la main ?" **Novembre 2025**. Benoit Rottembourg.
- Collège de France, "L'IA en tant que bureaucratie" Implications philosophiques d'IA, (May 2025) Carina Prunkl

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

- Benoit Rottembourg "IA et Pricing : Comment les algorithmes fixent nos prix". Podcast « [Changement d'époque en cours](#) » Avril 2025
- Op-Ed (Transformer), (with Y. Bengio and S. Clare), [AI is advancing far faster than our annual report can track](#)

12 Scientific production

12.1 Major publications

- [1] F. Bachoc, L. Béthune, A. González-Sanz and J.-M. Loubes. *Improved learning theory for kernel distribution regression with two-stage sampling*. 19th Feb. 2025. URL: <https://hal.science/hal-04956911>.
- [2] F. Bachoc, J. Bolte, R. Boustany and J.-M. Loubes. *When majority rules, minority loses: bias amplification of gradient descent*. 15th May 2025. URL: <https://hal.science/hal-05072199>.
- [3] 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://arxiv.org/abs/10.48550/arXiv.2504.00874). URL: <https://hal.science/hal-05268379>.
- [4] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss and J.-M. Loubes. 'Hoeffding decomposition of functions of random dependent variables'. In: *Journal of Multivariate Analysis* 208 (Apr. 2025), p. 105444. DOI: [10.1016/j.jmva.2025.105444](https://doi.org/10.1016/j.jmva.2025.105444). URL: <https://hal.science/hal-04233915>.
- [5] V. Lafargue, E. Claeys and J.-M. Loubes. *Fairness is in the details: Face Dataset Auditing*. 23rd Sept. 2025. URL: <https://ut3-toulouseinp.hal.science/hal-05374875>.
- [6] C. Lalanne, F. Iutzeler, J.-M. Loubes and J. Chhor. 'On the Private Estimation of Smooth Transport Maps'. In: ICML 2025 - 42nd International Conference on Machine Learning. Vancouver (BC), Canada, 13th July 2025, 30 p. URL: <https://hal.science/hal-04923578>.
- [7] A.-M. Nicot, V. Angel, I. Biletta, S. Boarini, I. Daugareihl, J. Duguépéroux, L. Fontana, P. Rème-Harnay, E. Leblanc, F. Lemozy et al. *Evaluation des risques sanitaires pour les travailleurs des plateformes numériques de livraison de repas en France*. Saisine n°2021-SA-0045. Anses, 26th Mar. 2025, 245 p. URL: <https://anses.hal.science/anses-05049790>.

12.2 Publications of the year

International journals

- [8] F. Bachoc, L. Béthune, A. González-Sanz and J.-M. Loubes. 'Improved learning theory for kernel distribution regression with two-stage sampling'. In: *Annals of Statistics* 53.4 (2025). URL: <https://hal.science/hal-04956911>.
- [9] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss and J.-M. Loubes. 'Hoeffding decomposition of functions of random dependent variables'. In: *Journal of Multivariate Analysis* 208 (Apr. 2025), p. 105444. DOI: [10.1016/j.jmva.2025.105444](https://doi.org/10.1016/j.jmva.2025.105444). URL: <https://hal.science/hal-04233915>.

International peer-reviewed conferences

- [10] D. Gala, M. Phillips-Brown, N. Goel, C. Prunkl, L. Alvarez Jubete, M. Corcoran and R. Eitel-Porter. 'Fairness-Aware Interactive Target Variable Definition'. In: IJCAI 2025 - Thirty-Fourth International Joint Conference on Artificial Intelligence. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 16th Aug. 2025, pp. 11048–11052. DOI: [10.24963/ijcai.2025/1260](https://doi.org/10.24963/ijcai.2025/1260). URL: <https://hal.science/hal-05459967>.

- [11] 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>.
- [12] N. Krčo, T. Laugel, V. Grari, J.-M. Loubes and M. Detyniecki. ‘When Mitigating Bias is Unfair: Multiplicity and Arbitrariness in Algorithmic Group Fairness’. In: *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Copenhagen, Denmark: IEEE, 9th Apr. 2025, pp. 735–752. doi: [10.1109/SaTML64287.2025.00046](https://doi.org/10.1109/SaTML64287.2025.00046). URL: <https://hal.science/hal-05444953>.
- [13] S. Ozgunay, L. Travé-Massuyès, J.-M. Loubes and R. Sena Ferreira. ‘Detecting Anomalies Using Graph Neural Networks: A Review’. In: *RJCIA 2025 - 23es Rencontres des Jeunes Chercheurs en Intelligence Artificielle*. Dijon, France: Association Française pour l’Intelligence Artificielle, July 2025. URL: <https://hal.science/hal-05002338>.

Conferences without proceedings

- [14] F. Bachoc, J. Bolte, R. Boustany and J.-M. Loubes. ‘When majority rules, minority loses: bias amplification of gradient descent’. In: *Conference on Neural Information Processing Systems*. San Diego (CA), United States, 2nd Dec. 2025. URL: <https://hal.science/hal-05072199>.
- [15] C. Lalanne, F. Iutzeler, J.-M. Loubes and J. Chhor. ‘On the Private Estimation of Smooth Transport Maps’. In: *ICML 2025 - 42nd International Conference on Machine Learning*. Vancouver (BC), Canada, 13th July 2025, 30 p. URL: <https://hal.science/hal-04923578>.

Reports & preprints

- [16] Y. Bengio, S. Clare, C. Prunkl, M. Andriushchenko, B. Bucknall, P. Fox, N. Maslej, C. McGlynn, M. Murray, S. Rismani et al. *International AI Safety Report 2025 Second Key Update: Technical Safeguards and Risk Management*. Mila - Quebec AI Institute; UK AI Safety Institute, 25th Nov. 2025. URL: <https://hal.science/hal-05459391>.
- [17] Y. Bengio, S. Clare, C. Prunkl, S. Rismani, M. Andriushchenko, B. Bucknall, P. Fox, T. Hu, C. Jones, S. Manning et al. *International AI Safety Report 2025 First Key Update: Capabilities and Risk Implications*. UK AI Security Institute; Mila - Quebec AI Institute, 15th Oct. 2025. URL: <https://hal.science/hal-05459345>.
- [18] E. Claeys, E. Kerjean and J.-M. Loubes. *Buzz, Choose, Forget: A Meta-Bandit Framework for Bee-Like Decision Making*. 18th Oct. 2025. URL: <https://ut3-toulouseinp.hal.science/hal-05374841>.
- [19] B. Ferrere, N. Bousquet, F. Gamboa, J.-M. Loubes and J. Muré. *Multivariate Bernoulli Hoeffding Decomposition: From Theory to Sensitivity Analysis*. 7th Oct. 2025. URL: <https://hal.science/hal-05301780>.
- [20] V. Lafargue, E. Claeys and J.-M. Loubes. *Fairness is in the details: Face Dataset Auditing*. 23rd Sept. 2025. URL: <https://ut3-toulouseinp.hal.science/hal-05374875>.
- [21] V. Lafargue, A. L. Monteiro, E. Claeys, L. Risser and J.-M. Loubes. *Exposing the Illusion of Fairness: Auditing Vulnerabilities to Distributional Manipulation Attacks*. 28th July 2025. URL: <https://ut3-toulouseinp.hal.science/hal-05374858>.
- [22] A.-M. Nicot, V. Angel, I. Biletta, S. Boarini, I. Daugareihl, J. Duguépéroux, L. Fontana, P. Rème-Harnay, E. Leblanc, F. Lemozy et al. *Evaluation des risques sanitaires pour les travailleurs des plateformes numériques de livraison de repas en France*. Saisine n°2021-SA-0045. Anses, 26th Mar. 2025, 245 p. URL: <https://anses.hal.science/anses-05049790>.
- [23] D. Rodríguez-Vitores, C. Lalanne and J.-M. Loubes. *Learning with Differentially Private (Sliced) Wasserstein Gradients*. 19th May 2025. URL: <https://hal.science/hal-04923829>.