PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

SAMM

Recueil des résumés

Modèles et apprentissages en sciences humaines et sociales

IA interprétable et responsable en SHS : des modèles aux humanités

Paris, 9–11 septembre 2026

Table des matières

Sessions invitées **3**

L'avenir des maths à l'ère de l'IA générative 4

Intelligence, cognition, connaissance à l'ère de l'IA : évolution ou bouleversement ? 5

IA, géopolitique et souveraineté 6

Algorithmic Fairness:
A Post-Processing Approach 7

Grands modèles de langues : les enjeux éthiques sont des enjeux scientifiques . . 8

Trustworthiness and explainability 9

Concrete measures to operationalize AI ethics principles and tensions 10

IAffectivité : IA, anthropotechnie affective et technomorale 11

Agentic AI: Ethical Issues 12

The use of SOM or Neuron gaz in spatial linear models 13

Handling uncertainty in humanitarian logistic 14

Comparative Analysis of Statistical, Machine Learning, and Deep Learning Time Series Models for Forecasting Dengue, Zika, and Chikungunya Incidence in Cuban Provinces 15

On multiple levels models to design geoparks 16

Vectorial Habitus: Embeddings, Class Language, and the Limits of the Computational Proxy in the Social Sciences 17

Extremal properties of heavy-tailed shot noise processes with applications to food risk assessment 18

Business Intelligence and Business Analytics for supporting decision strategies: facing extreme events 19

Sessions posters **20**

Fondements de l'IA explicable dans les systèmes d'apprentissage adaptatif : vers une compréhension des mécanismes de décision des plateformes éducatives intelligentes (Annulé), Douha El-Otmani 21

Transcrire des doléances au XXIe siècle : le cas des Cahiers Citoyens issus du Grand Débat National, Julie Lenoir [et al.] 22

Reducing Inequalities in Access to Medical Information through AI: Which Data, Which Interpretable Models, and for Which Health Equity? (Annulé), Mouheb Mehdoui [et al.] 23

CRiDiT : Calibrer la confiance humain-IA dans les systèmes à haute responsabilité, Yuntian Ding [et al.] 24

Active Tigger, Adrien Rougier [et al.] 25

Writing and Evaluating Under an Identity: Gender and Affiliation Bias in Scientific Review Generated by Large Language Models, Adam Faci [et al.] 26

A Review of Fine-Tuning for QA and a Proposal for a New Annotated Corpus, Alice Breton [et al.] 27

Identifying EEG signatures of psychological flow state using virtual reality in educational environments, Sadya Martinez 28

Apprentissages et usages de l'intelligence artificielle en archéologie : un retour d'expériences, une mise en perspective et des propositions, Christophe Tufféry . . 29

Does AI Take Her Complaint Seriously? A Counterfactual Audit of Gender Fairness in AI-Generated Hotel Service Recovery, Milena Nikolic [et al.] 30

Liste des auteurs **30**

Sessions invitées

L'avenir des maths à l'ère de l'IA générative

Ivan Nourdin

Université du Luxembourg

Dans cet exposé, je reviendrai sur l'expérience Malliavin-Stein, menée en septembre dernier avec C.-P. Diez et L. Da Maia, afin d'examiner ce que l'IA générative était alors en mesure d'apporter à la recherche mathématique. Nous discuterons ensuite des avancées mathématiques rendues possibles par l'IA depuis lors, avant d'élargir la réflexion à ce que pourrait être l'avenir des mathématiques dans les années à venir.

Intelligence, cognition, connaissance à l'ère de l'IA : évolution ou bouleversement ?

Daniel Andler

Sorbonne Université & École normale supérieure

Le projet conçu et poursuivi avec constance par l'IA est de doter les ordinateurs d'intelligence. La mise en œuvre de ce projet a consisté à leur conférer, pas à pas, les capacités de fournir des solutions à une variété toujours plus grande de problèmes de toute nature. Le succès complet de cet effort conduirait, selon certains, à la réalisation du projet initial : un ordinateur avec une intelligence générale artificielle (IGA) comparable à celle des humains, capable ainsi de se substituer à un humain dans toutes ses activités. En particulier, de faire de la recherche. Le scepticisme qu'a longtemps suscité cette vision a aujourd'hui beaucoup reculé, laissant souvent la place à une profonde inquiétude. L'inquiétude devant les avancées de l'IA est justifiée, mais pas parce que la vision en question serait juste : elle repose sur deux confusions.

IA, géopolitique et souveraineté

Francesca Musiani

Centre Internet et Société, CNRS

L'intelligence artificielle est aujourd'hui au cœur des recompositions géopolitiques mondiales. Au-delà de la compétition technologique, elle soulève des enjeux de souveraineté liés à la maîtrise des données, des infrastructures, des modèles et des normes qui structurent l'espace numérique. Cet exposé propose d'examiner les liens entre IA et rapports de puissance entre États, entreprises et organisations internationales, et analyser les conceptions concurrentes de la souveraineté numérique qui émergent en lien avec ces enjeux. Il s'appuiera notamment sur les travaux des projets de recherche DIGISOV, consacré à la gouvernance numérique et à la circulation des normes de souveraineté, et ClaimSov, qui analyse la manière dont la souveraineté numérique est revendiquée, définie et mise en œuvre dans les discours, les politiques publiques et les pratiques. L'objectif est de mieux comprendre les tensions entre innovation, dépendances technologiques et autonomie stratégique dans un monde numérique de plus en plus fragmenté.

Algorithmic Fairness: A Post-Processing Approach

Christophe Denis

Université Paris 1 Panthéon-Sorbonne

Algorithmic fairness has become a major area of machine learning, aiming to mitigate the impact of biases present in data and algorithmic decision-making. In this talk, we consider post-processing methods for achieving group fairness in the settings of multiclass classification and regression. We derive explicit expressions for the corresponding optimal fair predictors and present a plug-in post-processing algorithm to estimate them. Finally, we discuss both the theoretical guarantees of the proposed approach and its empirical performance through numerical experiments.

Grands modèles de langues : les enjeux éthiques sont des enjeux scientifiques

Karën Fort

Université de Lorraine

En l'espace de moins d'une décennie, les grands modèles de langues de type ChatGPT ont révolutionné le domaine du traitement automatique des langues. Ces modèles ont non seulement submergé le domaine, mais également nos sociétés, avec des impacts sociétaux auxquels il est urgent de s'intéresser et qu'il est impératif de mesurer. Dans cette intervention, je présenterai les problèmes réels que posent ces systèmes, loin des prédictions de prise de pouvoir par l'IA. Je montrerai également les différents outils que nous avons développés pour mieux évaluer les grands modèles de langues et favoriser ainsi une prise de décision informée sur leur utilisation.

Trustworthiness and explainability

Barbara Hammer

Bielefeld University

Explainable AI (XAI) aims at the integration, extension, or substitution of otherwise black-box machine learning models by components, which are comprehensible for humans. Albeit many XAI methods exist, it is unclear, what information exactly they reveal, why different methods lead to different results, and in which sense humans can trust the results. Within the talk, I will give examples how XAI technologies can be used in applications relating to critical infrastructures, and which challenges such as uniqueness, plausibility, or vulnerability exist. Further, I will present first results how to target some of these challenges: Specifically, I will have a closer look at feature-based explanations and how to extend those to incorporate the effect of interactions, leading to more reliable results, for explaining multimodal foundation models, for example.

Concrete measures to operationalize AI ethics principles and tensions

Alexei Grinbaum

CEA

Ten use cases across different industrial sectors have been studied in the AIOLIA project. We gained important insights into AI ethics principles, specifically their interlinkages and contextual components. Our findings align well with the overarching HLEG ALTAI framework but also show important deviations. The status and meaning of ethics principles may be much more fluid in actual design practice than in the abstract AI ethics frameworks. We have also elicited and narrativized many tensions between principles. The core of AIOLIA findings is a portfolio of technical and organisational measures to implement each principle and address trade-offs. Based on these empirical measures, AIOLIA is now building pedagogical materials and offering pilot training sessions in AI ethics.

IAffectivité : IA, anthropotechnie affective et technomorale

Nadia Guerouaou

Centre Internet et Société & PEPR ensEMBLE

De plus en plus, nous confions à l'IA générative le soin de calibrer nos émotions : un centre d'appels qui efface la colère d'un client, un chatbot qui adoucit le ton d'un e-mail, un « sourire vocal » injecté en visioconférence. Ces dispositifs, comme « Emotion Cancelling » du géant japonais SoftBank, promettent d'optimiser nos interactions en filtrant nos émotions « négatives ». Certaines études vont dans ce sens : lors de sessions de brainstorming, un sourire artificiel généré par visioconférence suffirait à augmenter le nombre d'idées produites. Mais que masquent ces bénéfices immédiats ?

Longtemps, la voix ou le visage étaient considérés comme miroirs de nos émotions « authentiques ». Aujourd'hui, l'IA générative permet de fabriquer ces expressions de toutes pièces – le masque social devient alors algorithmique, sans conscience ni transparence sur les normes qu'il applique. Pourtant, cette délégation de notre expressivité émotionnelle à l'IA reste un sujet peu traité, alors qu'on interroge déjà beaucoup celle de nos capacités cognitives (mémoire, langage, résolutions de problèmes, etc).

Ces technologies façonnent notre sensibilité et le paysage moral de nos sociétés – ce qui nous touche, nous indigne, nous relie aux autres.

Quelle « IAffectivité » est alors en gestation dans cette nouvelle matrice culturelle ?

Agentic AI: Ethical Issues

Raja Chatila

Sorbonne Université

The rapid development and deployment of Agentic AI systems raises several concerns that go unaddressed by current ethics or legal frameworks, unless their outcome becomes apparent. The OpenAI-Hugging Face incident is only one example of an agentic AI systems breaking the law by intruding into the infrastructure of an organization. It raised plenty of discussions and concerns mainly because of its cybersecurity character. But there are several ethics principles that Agentic AI systems don't align with, and therefore such incidents and others of diffent nature are bound to happen. This talk explores this mismatch between Agentic AI and ethics guidelines, specifically those issued by the EU's HLEG-AI and that are foundational of the AI Act. It also questions the completeness of these principles.

The use of SOM or Neuron gaz in spatial linear models

Cécile Hardouin

Université Paris Nanterre

The aim of spatial econometrics is to analyze and/or predict the relationship between one dependent variable Y with other variables, building a model that takes into account the spatial dependence. Usual spatial econometric models are based on a neighbourhood matrix whose elements are linked to geographical distances. We propose to use dis tances between prototypes resulting from a neural classification instead. The results are at least as well as the ones obtained from the geograph ical distances based design. In some cases, we need two neighbourhood matrices and the difficulty rising then is to find a second matrix; then this issue is simply solved by using one matrix based on geographical distances and the other based on neural distances. Finally, the use of neural distances opens the door to use spatial econometrics methods for the analysis and the modelling of any data sets, non necessary geograph ically referenced. We illustrate our approach with two studies of real datasets.

Handling uncertainty in humanitarian logistic

Sira M. Allende Alonso

Université de La Havane

The increasing frequency and unpredictability of hurricanes pose significant challenges to humanitarian logistics, where critical decisions on facility location, inventory pre-positioning, and relief distribution must be made under deep uncertainty. This paper examines the application of robust optimization models to address these challenges, with particular emphasis on distributionally robust optimization (DRO) as an advanced framework for handling uncertainty in hurricane relief operations. Unlike traditional stochastic programming, which requires precise probabilistic information often unavailable in disaster contexts, or classical robust optimization, which tends to yield overly conservative solutions, DRO offers a flexible middle ground by optimizing against the worst-case distribution within an ambiguity set constructed from limited data. Recent developments employ Wasserstein metric-based ambiguity sets to capture uncertainties in post-disaster demand, supply survivability, and road network disruptions, enabling data-driven decision-making with probabilistic guarantees. Risk-averse measures such as worst-case mean-Conditional Value-at-Risk (WMCVaR) are integrated to guard against extreme shortage scenarios while balancing cost-efficiency and service reliability. Case studies focusing on hurricane threats in the Gulf of Mexico region demonstrate that these robust optimization approaches consistently outperform conventional two-stage stochastic programming and sample average approximation benchmarks, achieving favorable trade-offs between out-of-sample reliability, operational costs, and humanitarian service quality. The findings provide actionable insights for humanitarian decision-makers seeking resilient pre-positioning and response strategies under the inherent unpredictability of hurricane disasters.

Comparative Analysis of Statistical, Machine Learning, and Deep Learning Time Series Models for Forecasting Dengue, Zika, and Chikungunya Incidence in Cuban Provinces

Yanet Garcia Serrano, Jennifer de la Caridad Sanchez Santana

Université de La Havane

Arboviral diseases, specifically dengue, Zika, and chikungunya, represent a critical public health challenge in Cuba due to their high morbidity and the direct influence of climatic variables on their spread. Despite entomological surveillance efforts, the non-linear nature of these phenomena demands advanced predictive methods. This research aims to determine the most accurate model for forecasting the incidence of these pathologies across Cuban provinces through a comparative study of time series algorithms. The methodology integrates epidemiological and meteorological data (processed using Thiessen Polygons), applying statistical models (SARIMAX), machine learning (SVR, Random Forest, XGBoost), and deep learning (LSTM). The effect of time lags is examined using Pearson correlation hypothesis tests and the cross-correlation function (CCF). The findings identify the optimal architecture for each territory, providing a scientific methodological basis to strengthen early warning systems and decision-making within the national health system.

On multiple levels models to design geoparks

Gemayqzel Bouza Allende

Université de La Havane

The UNESCO has conceived geoparks as a way to conserve geological interesting sites. The interest of this work is that Cuba has important geosites and the proposal of at least one geopark. The goal of this contribution is to present a model based on data to evaluate the feasibility of a geopark and the economical efficiency of it based on data collected to declare the geosites that would integrate it. To this aim we present a game problem and analyze the use of multiobjective and bilevel problems.

Vectorial Habitus: Embeddings, Class Language, and the Limits of the Computational Proxy in the Social Sciences

Carlos León González

Université de La Havane

Text embeddings are increasingly used as proxies for social phenomena, but their interpretability in social stratification contexts still lacks systematic examination. This research seeks to clarify whether pretrained language models capture structural social position through discursive style, or whether vector separation responds solely to differences in thematic interest. The distinction is epistemologically crucial for the responsible use of artificial intelligence in the social sciences. The theoretical framework draws on the Bourdieuian notion of habitus: language not merely as content, but as a practice endowed with registers, valuations, and lexical choices that reflect acquired social dispositions. Community embeddings will be constructed from corpora of social media groups with documented socioeconomic profiles, using pretrained models without fine- tuning. The central axis is a lexical ablation experiment under four progressive conditions: complete corpus, removal of hyper-specific terms, removal of explicit consumption markers, and exclusive retention of evaluative and argumentative terms. Under each condition, separability metrics will be recalculated to diagnose which linguistic dimension sustains group differentiation. The main methodological contribution is the ablation delta: a diagnostic metric that quantifies the relative drop in separability between conditions. If separability collapses when thematic nouns are removed but persists when only evaluative adjectives and argumentative connectors are retained, this would constitute evidence of discursive habitus capture. The study will also explore asymmetries between community pairs, providing direct sociological interpretation of the model's selectivity as a proxy for social class.

Extremal properties of heavy-tailed shot noise processes with applications to food risk assessment

Patrice Bertail

Université Paris Nanterre

Joint work with C. Tiller (UVSQ) and A. Medina-Garay (UFPN, Recife).

Risk analysis in food safety has gained increasing attention in recent years, particularly regarding nutritional aspects and exposure to foodborne contaminants, with cadmium and methylmercury among the most extensively studied cases. This paper develops a dynamic framework for dietary exposure assessment based on shot noise processes, extending the Kinetic Dietary Exposure Model (KDEM) introduced by Bertail Clemencon and Tressou (2008)(2010). The proposed approach provides a flexible way to describe the accumulation, assimilation, and elimination of contaminant intakes over time. Under the assumption of heavy-tailed intake distributions, we derive asymptotic equivalents for the tail distribution of the stationary exposure process and obtain an explicit expression for the extremal index of the associated embedded sequence. These results provide theoretical tools to quantify both the long-run tail behavior of the exposure process and the clustering of high-exposure episodes. We then use this framework to analyze several model specifications, including random elimination rates, delayed assimilation, and an immigration-emigration mechanism with instantaneous release. For these models, we derive explicit or asymptotically tractable expressions for risk indicators such as tail probabilities, the extremal index, the Expected Severity, and the Integrated Expected Severity. We conclude with an application to methylmercury exposure in the French population, where the resulting risk indices are compared across men, women, and children.

Business Intelligence and Business Analytics for supporting decision strategies: facing extreme events

Vivian del Rosario Sistachs Vega

Université de La Havane

Joint work with P. Roura (Universidad de La Habana) and M.A. Diaz (Instituto de Meteorología).

Business intelligence is the field of knowledge that encompasses a wide range of methodologies, processes, and technologies that generate information and knowledge to support decision-making in any organization. Within this field is business analytics, which specifically refers to the use of analytical techniques that make it possible to achieve the purpose of business intelligence. In line with this idea, the Climate Center of the Institute of Meteorology has focused its attention on the intense cyclonic activity that has occurred in Cuba recently, which is considered extreme weather events. Therefore, a research project was conducted on the climatology of cyclonic activity, its variability, and its long-term trend, since the greatest natural disasters recorded in the history of our country have been associated with tropical cyclones. This motivated the Climate Center, as part of its project, to update the TkHURS software (statistical software used to calculate hurricane return periods) in order to obtain return periods and calculate estimated frequencies by fitting a Poisson model to the variable that counts the number of hurricanes per year that have struck Cuba during the period from 1791 to 2023, based on the Chronology of Tropical Cyclones and the General Weather Reports, following a working methodology for the software that enables the Cuban government to make decisions aimed at protecting the population and economic resources from this type of extreme event.

Sessions posters

Fondements de l'IA explicable dans les systèmes d'apprentissage adaptatif : vers une compréhension des mécanismes de décision des plateformes éducatives intelligentes (Annulé)

Douha El-Otmani * [1]

[1] Université Hassan II de Casablanca (UH2C) – Maroc

Les systèmes d'apprentissage adaptatif basés sur l'intelligence artificielle occupent une place croissante dans les environnements éducatifs numériques. Ils ajustent automatiquement les contenus, les activités et les parcours en fonction des données issues des interactions des apprenants. Cependant, ces systèmes reposent sur des modèles complexes dont les mécanismes de décision restent souvent opaques pour les utilisateurs finaux. Cette proposition s'inscrit dans le champ de l'IA explicable appliquée à l'éducation. Elle interroge les fondements théoriques et méthodologiques permettant de rendre compréhensibles les décisions produites par les plateformes éducatives intelligentes. L'objectif est d'analyser les conditions dans lesquelles une recommandation pédagogique peut être interprétée par les enseignants et les apprenants, sans réduction excessive de la complexité des modèles. L'étude porte également sur la notion de transparence algorithmique dans les dispositifs d'apprentissage adaptatif. Elle examine la manière dont les systèmes traduisent des données comportementales en décisions pédagogiques, ainsi que les formes d'explicitation proposées aux utilisateurs. Une attention particulière est accordée à la traçabilité des résultats, à la lisibilité des critères de recommandation et à la capacité des acteurs éducatifs à se réapproprier ces informations dans leurs pratiques. L'enjeu central réside dans l'articulation entre performance algorithmique et intelligibilité des systèmes. Il s'agit de comprendre comment l'explicabilité influence la confiance, l'acceptation et l'usage des technologies éducatives, tout en questionnant les limites des modèles actuels dans la médiation des apprentissages.

Mots-Clés: IA explicable, apprentissage adaptatif, transparence algorithmique, éducation numérique, prise de décision

*Intervenant

Transcrire des doléances au XXIe siècle : le cas des Cahiers Citoyens issus du Grand Débat National

Julie Lenoir *† [1], Léa Maronet‡ [1,2], Alyx Taounza-Jeminet§ [1]

[1] Établissement public Campus Condorcet (EPCC) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France
[2] Huma-Num : l'IR* des humanités numériques (Huma-Num IR*) – Aix Marseille Université, Centre National de la Recherche Scientifique, Établissement public Campus Condorcet – France

Dans la suite de la résolution de mars 2025 concernant la publicisation des doléances du Grand Débat National, un projet est né autour du corpus des Cahiers Citoyens, coordonné par l'Établissement Public Campus Condorcet avec l'appui de l'IR* Huma-Num. Il a pour objectif de développer une chaîne de traitement en vue de la mise en ligne de leurs contenus. Le corpus en question est constitué de numérisations de Cahiers alimentés par des citoyens et citoyennes dans 17 000 mairies françaises entre 2018 et 2019, où ils et elles ont pu écrire, notamment à la main, leurs " doléances ". Dans la perspective d'une diffusion large et ouverte du corpus, il est aujourd'hui nécessaire d'établir une chaîne de traitement " *transcription – anonymisation – publication* " qui permette une reconnaissance fiable des caractères manuscrits et une anonymisation des données personnelles mentionnées dans les Cahiers. La principale difficulté que présente ce corpus est sa variabilité, tant du point de vue du format des Cahiers (parfois très structurés, avec des encarts définis, parfois rédigés sur pages blanches, parfois sur des cahiers d'écoliers) que du point de vue de l'anonymat (des données directement identifiantes, telles que les noms et prénoms, côtoient des expressions indirectement identifiantes). Ces difficultés liées au corpus lui-même sont amplifiées par l'absence actuelle d'un modèle de reconnaissance des écrits manuscrits (HTR) spécialisé sur les écritures françaises contemporaines, ce qui rend nécessaire l'entraînement d'un tel modèle sur le corpus. Nous présenterons donc les premiers travaux de constitution des " vérités terrain " et de calibration du modèle de HTR, en donnant à voir la variabilité interne aux corpus et les stratégies mises en place pour contrôler cette variabilité.

Mots-Clés: Humanités numériques, HTR, chaîne de traitement, parole citoyenne, histoire contemporaine

*Intervenant
†Auteur correspondant: julie.lenoir@campus-condorcet.fr
‡Auteur correspondant: lea.maronet@huma-num.fr
§Auteur correspondant: alyx.taounza-jeminet@campus-condorcet.fr

Reducing Inequalities in Access to Medical Information through AI: Which Data, Which Interpretable Models, and for Which Health Equity? (Annulé)

Mouheb Mehdoui * [1], Amel Fraisse [1], Widad Mustafa El Hadi [1], Mounir Zrigui [2]

[1] Université de Lille – Université Lille III - Sciences humaines et sociales – France
[2] Université de Monastir – Tunisie

Access to understandable medical information is a major determinant of reducing health inequalities. Yet, specialized documents often remain inaccessible to populations with low health literacy. Artificial intelligence, particularly Natural Language Processing (NLP), offers solutions for text simplification and personalization. However, the effectiveness and fairness of these systems depend heavily on the quality of training data and the actual performance of the deployed models. We address three fundamental questions:
- Which data? We mobilize heterogeneous corpora: anonymized electronic health records, online patient forums (capturing lay language), and biomedical ontologies (for semantic enrichment). We critically examine representativeness biases inherent to these sources.

- Which interpretable models? We comparatively evaluate several language model architectures: general-purpose models (Mistral, Llama 3) and biomedical specialized models (BioBERT, BioGPT). Our evaluation goes beyond automatic metrics (BLEU, ROUGE); we integrate readability measures (Flesch scores) and qualitative semantic error analyses to assess interpretability and reliability.

- For which health equity? We analyze how model performances vary across medical topics and baseline complexity levels, aiming to identify scenarios where AI might inadvertently widen disparities.
Our preliminary comparative results reveal significant gaps between general and specialized models, underscoring the need for rigorous, context-aware selection. This poster will discuss the ethical and methodological implications for responsible AI in health.

Mots-Clés: Health Literacy, Natural Language Processing, Large Language Models, Interpretable AI, Health Equity, Text Simplification

*Intervenant

CRiDiT : Calibrer la confiance humain-IA dans les systèmes à haute responsabilité

Yuntian Ding *† [1], Nicolas Herbaut‡ [2], Camille Salinesi§ [1]

[1] Centre de Recherche en Informatique de Paris 1 (CRI) – Université Paris 1 Panthéon-Sorbonne – France

[2] Laboratoire Bordelais de Recherche en Informatique (LaBRI) – Université de Bordeaux, École Nationale Supérieure d'Électronique, Informatique et Radiocommunications de Bordeaux (ENSEIRB), Centre National de la Recherche Scientifique – France

L'intégration croissante de l'IA générative dans des contextes décisionnels à fort enjeu rend critique l'alignement entre la confiance de l'utilisateur et la fiabilité réelle du système. Une confiance mal calibrée, qu'il s'agisse de sur-confiance ou de sous-confiance, engendre des risques opérationnels et éthiques. Nous présentons CRiDiT (Computational Risk-Sensitive biDirectional Trust Management for AI-Infused Systems), un modèle de haut niveau de calibration de la confiance structuré en une architecture à quatre couches, reformulé comme une boucle fermée reliant les signaux de fiabilité machine et les signaux de confiance humaine, afin de détecter les situations de mauvaise calibration et de déclencher des *trust cues* ou interventions. Nous illustrons son application dans plusieurs domaines à forte responsabilité : santé (aide au diagnostic et interfaces d'IA explicable), cybersécurité (détection d'attaques et délégation à un second détecteur), éducation (adaptation du niveau d'explication à l'apprenant) et détection de deepfakes. Dans chaque cas, les données système et les informations utilisateur alimentent la mesure de la confiance, tandis que les interventions côté machine et les trust cues côté humain soutiennent une calibration continue. Nous identifions ensuite trois directions de recherche ouvertes. D'abord, la relation entre *trust cues* et confiance humaine : comment un *trust cue* (par exemple : message verbal, indicateur visuel) est perçu à travers le processus cognitif de l'utilisateur (perception, compréhension, projection, décision) et comment cette chaîne, qui va du trust cue à la cognition puis à la confiance, conditionne une calibration effective. Ensuite, l'orchestration de la confiance côté machine : quand et comment intervenir, notamment en déléguant une tâche à un outil ou un système jugé plus fiable. Enfin, la validation du modèle par simulation, articulant données côté humain et exigences issues de l'étude des interventions. Ce travail vise à ouvrir la discussion sur l'opérationnalisation interdisciplinaire de la confiance et de la responsabilité dans les systèmes à base d'IA en sciences humaines et sociales.

Mots-Clés: calibration de la confiance, IA responsable, explicabilité, IHM, systèmes à base d'IA

*Intervenant

†Auteur correspondant: yuntian.ding@univ-paris1.fr

‡Auteur correspondant: nicolas.herbaut@u-bordeaux.fr

§Auteur correspondant: camille.salinesi@univ-paris1.fr

Active Tigger

Adrien Rougier * [1], Emilien Schultz [1], Julien Boelaert [2], Axel Morin [1], Annina Claesson [1,3], Emma Bonutti D'agostini [1,3], Arnault Chatelain [1], Étienne Ollion† [1], Francesco Colonna [1]

[1] Centre de Recherche en Économie et STatistique / Institut Polytechnique de Paris (CREST / IPP) – Institut Polytechnique de Paris, Centre de Recherche en Économie et STatistique (CREST) – France
[2] Centre d'Etudes et de Recherches Administratives, Politiques et Sociales - UMR 8026 (CERAPS) – Université de Lille, Centre National de la Recherche Scientifique, Sciences Po Lille - Institut d'études politiques de Lille – France
[3] médialab (Sciences Po) (médialab) – Sciences Po – France

Across social sciences, researchers are increasingly interested in skyrocketing amounts of textual data. Even as computer scientists have developed useful tools for analysing textual data, these often place high demand on technical mastery. Widening access to these research methods requires evidence-based, inclusive and transparent tools. Active Tigger is a software that aims to : - Facilitate collaborative human annotation and exploration of textual datasets - Use AI to carry out text classification and information extraction - Democratise computational social sciences practices for other communities - Create a lightweight, frugal and open source application.

Mots-Clés: Text Annotation, Fine, tuning, Open science, Reproducibility

*Intervenant
†Auteur correspondant:

Writing and Evaluating Under an Identity: Gender and Affiliation Bias in Scientific Review Generated by Large Language Models

Adam Faci * [1], Alice Breton† [1]

[1] Huma-Num : l'IR* des humanités numériques (Huma-Num IR*) – Aix Marseille Université, Centre National de la Recherche Scientifique, Établissement public Campus Condorcet – France

Research on bias in generative models has focused largely on decisions or short texts. Yet large language models are increasingly used to review scholarly articles, and what they produce in that setting is of a different nature: a full expert discourse, complete with criticisms and requests for revision. We examine whether, with the scientific content held constant, models evaluate a text differently depending on the identity attributed to its author.
We start from an English-language corpus of roughly three million articles in the social sciences and humanities. From the abstracts we extract controlled materials (research question, methods, results, arguments) and derive discursive profiles for each discipline. Several open models, run locally, produce a review report from these materials. Attributed gender, geographical affiliation and discipline vary independently, while the source content remains identical.

The protocol analyses the review report jointly with the reasoning trace that precedes it in reasoning models. This is where justification is articulated, which makes it possible to separate cases where the author's identity enters the argumentation explicitly from those where it affects only the final wording.

We measure the nature of the criticisms, the qualifiers used, the explanations, the suggestions, the recognition of novelty and the scientific legitimacy granted to the text. We then compare disciplinary, gender and origin profiles on the basis of rhetorical indicators and conditional perplexity. The latter is taken as an index of conformity to a discursive norm rather than as a measure of discrimination.
Two questions frame the study and guide our discussion of the results: who can be recognised as innovative, and judged fully legitimate, within a given discipline? And what effect might the widespread use of generative models have on the future of research writing?

Mots-Clés: generative large language models, discriminative bias, evaluation, social sciences and humanities

*Intervenant
†Auteur correspondant: alice.breton@huma-num.fr

A Review of Fine-Tuning for QA and a Proposal for a New Annotated Corpus

Alice Breton *† [1], Adam Faci‡ [1]

[1] Huma-Num : l'IR* des humanités numériques (Huma-Num IR*) – Aix Marseille Université, Centre National de la Recherche Scientifique, Établissement public Campus Condorcet – France

Developing reliable Retrieval-Augmented Generation (RAG) systems for specific domains like social sciences and humanities (SSH) requires question-answering (QA) datasets tailored to the linguistic and epistemological specificities of these fields. This necessity motivates a comprehensive review of existing QA datasets and corpora, ranging from general benchmarks (e.g., SQuAD, Rajpurkar et al., 2016; Natural Questions, Kwiatkowski et al., 2019; HotpotQA, Yang et al., 2018) to domain-specific corpora. However, these resources remain inadequate for SSH, due to a lack of domain-specific linguistic nuances, interdisciplinary complexity, and epistemological diversity, which limits the development of accurate and contextually grounded RAG systems. This review therefore highlights their limited coverage of SSH and the unique challenges posed by this linguistic context. To address this gap, we introduce a new synthetically generated QA corpus derived from real user queries in a documentary search engine, supplemented by human annotation to ensure the accuracy and relevance of questions and answers. This approach aligns with recent work on automatic QA data generation using large language models (e.g., HistoriQA-ThirdRepublic, Pellet et al., 2026). Beyond proposing this new corpus, we review state-of-the-art fine-tuning methods for adapting language models to QA tasks, such as supervised fine-tuning, instruction tuning, and lightweight adaptation techniques like LoRA, discussing their strengths and limitations for interpretable and responsible use in SSH. This work contributes to the broader discussion on models, methods, and applications of interpretable AI in SSH, while examining trust, explainability, and biases in generation systems when relying on domain-specific corpora and annotations.

Mots-Clés: RAG, QA, SSH, Fine, tuning, State of the art, Benchmark

*Intervenant

†Auteur correspondant: alice.breton@huma-num.fr

‡Auteur correspondant: adam.faci@huma-num.fr

Identifying EEG signatures of psychological flow state using virtual reality in educational environments

Sadya Martinez * [1,2]

[1] MIRAI Innovation Research Institute – Japon
[2] Facultad de Ingeniería Mecánica y Eléctrica (FIME, UANL) – Mexique

Flow, a state of deep absorption and optimal engagement, is a central construct in the psychology of learning, yet it remains difficult to measure objectively, relying largely on retrospective self-report. This project addresses that gap by combining EEG, virtual reality, and interpretable machine learning to identify neural correlates of flow during learning tasks, and to make the resulting model's reasoning explicit and verifiable.
Building on a virtual reality learning protocol developed in 2023, EEG is recorded from participants engaged in a series of VR-based learning tasks structured around varying levels of challenge, following Csikszentmihalyi's challenge-skill balance framework. Task epochs corresponding to high challenge ("Test") are compared with lower level of challenge epochs ("Memory/Practice") to construct a supervised classification problem grounded in psychological theory rather than self-report alone. A machine learning classifier is trained on EEG features extracted from these epochs, with leave-one-subject-out (LOSO) cross-validation used to assess generalization across individuals. Crucially, feature importance analysis is applied to identify which EEG spectral bands and brain regions drive the model's predictions, discovering the classifier's internal logic. As external validation, flow-relevant subscores from Csikszentmihalyi's questionnaire are compared against each participant's EEG-based classification confidence, testing whether model interpretability aligns with subjective experience.
This approach shows how AI methods can be applied not only to text or social data, but to inherently subjective, psychological states providing a case study on interpretable AI within cognitive and behavioral research, with direct implications for how learning technologies could be evaluated and trusted.

Mots-Clés: Flow State, Machine Learning, Education, EEG, Virtual Reality, Cognitive Neuroscience

*Intervenant

Apprentissages et usages de l'intelligence artificielle en archéologie : un retour d'expériences, une mise en perspective et des propositions

Christophe Tufféry * [1,2]

[1] Ministère de la Culture (MC) – ''Ministère de la culture '' – France
[2] Technologie et Ethnologie des Mondes Préhistoriques (TEMPS) – Université Paris 1 Panthéon-Sorbonne, Université Paris Nanterre, Centre National de la Recherche Scientifique – France

Pendant notre seconde thèse, soutenue en décembre 2022, nous avons utilisé plusieurs solutions d'intelligence artificielle (IA) pour la transcription de carnets de terrain en archéologie. Si nous avons rapidement perçu l'intérêt de ces dispositifs, nous en avons aussi perçu les limites et exigences. Depuis notre thèse, les solutions d'IA, qu'elles soient " grand public " (ChatGPT, Gemini, Claude, Perplexity, Mistral AI, etc.) ou spécialisées, se sont multipliées. En archéologie, les projets d'utilisation d'IA ne cessent de se développer (ex. ArcheologIA, AUTOMATA, SIAMOIS) et de nombreux financements viennent appuyer cet essor (Guyot et al. 2024, Durost et al. 2024, Vidal et al. 2025). Nous continuons à explorer plusieurs solutions d'IA dans divers domaines (transcription, développement d'applications de gestion de données scientifiques, reconnaissance de formes). Depuis 2024, nous assurons une formation à l'histoire de l'informatique et des IA à des élèves ingénieurs à l'EPITA, une des principales écoles supérieures en IA. Dans le cadre de notre habilitation à diriger des recherches en cours, nous mettons en perspective nos pratiques dans le cadre d'une approche réflexive pour interroger ce que peut avoir comme conséquences pour l'archéologie la délégation à de tels dispositifs numériques d'une part croissante des opérations de détection, d'analyse et de partage de l'information sur des sites et des vestiges archéologiques. À la suite de plusieurs auteurs (Guénot 2023, Baillifard et Carbonel 2025, Gefen 2026), nous faisons des propositions pour un renforcement de la formation des étudiant.es mais aussi des professionnels de l'archéologie aux solutions d'IA (Tufféry 2026). Cette communication propose de rendre compte de l'avancée de nos réflexions pour identifier les convergences possibles avec celles d'autres domaines des SHS mais aussi d'isoler les spécificités du domaine de l'archéologie, par exemple en matière épistémologique et d'administration de la preuve.

Mots-Clés: archéologie, IA, humanités numériques, réflexivité, formation

*Intervenant

Does AI Take Her Complaint Seriously? A Counterfactual Audit of Gender Fairness in AI-Generated Hotel Service Recovery

Milena Nikolic * [1], Zarko Radjenovic [2]

[1] The Academy of Applied Technical and Preschool Studies (AATPS) – Serbie
[2] Innovation Center, University of Nis (ICUN) – Serbie

Large language models (LLMs) increasingly draft responses to hotel complaints. Differences in validation, redress, or escalation can affect guests. Hospitality research compares human and machine responses but rarely tests whether an identical complaint receives equivalent treatment when only the guest's gender cue changes. This study introduces a counterfactual audit of gender fairness in AI-generated hotel service recovery and frames unequal credibility assessment as algorithmically mediated testimonial injustice.
Twenty scenarios are derived from 20,491 TripAdvisor hotel reviews and cover five complaint categories, ranging from booking disputes to harassment and safety incidents. Each is presented with a female guest, a male guest, or no stated gender. All other facts, harm, and requested remedy remain fixed. GPT-5.6, Gemini 3.1 Pro Preview, and Claude Sonnet 5 are tested using standard prompts and prompts with an explicit fairness instruction. Each condition is repeated three times, yielding 1,080 responses.

The framework distinguishes testimonial parity, covering credibility, validation, and blame, from remedial parity, covering compensation, procedural escalation, and safety action. Two independent coders rate validation, empathy, blame, remedy adequacy, escalation, autonomy, and action specificity on anchored five-point scales. Categorical decisions are assessed through action reversal rates across matched female and male versions. Ordinal shifts are analyzed with cumulative link mixed models and random intercepts for scenarios. Text embedding similarity separates wording variation from changes in recommended action. Intercoder agreement is assessed with ordinal Krippendorff's alpha.
The audit pipeline checks prompt equivalence and records model versions, generation settings, and raw outputs. It distinguishes empathy from concrete action and does not treat sentiment as evidence of fairness. Differences count as disparities only when they alter credibility, remedy, escalation, or autonomy without justification from the scenario. The analysis compares models and prompt conditions without conflating warranted safety escalation with gender disparity.

Mots-Clés: Large language models, Gender bias, Counterfactual auditing, Hotel service recovery, Testimonial injustice

*Intervenant

Liste des auteurs

Allende Alonso, Sira M., 14
Andler, Daniel, 5

Bertail, Patrice, 18
Boelaert, Julien, 25
Bonutti D'Agostini, Emma, 25
Bouza Allende, Gemayqzel, 16
Breton, Alice, 26, 27

Chatelain, Arnault, 25
Chatila, Raja, 12
Claesson, Annina, 25
Colonna, Francesco, 25

Denis, Christophe, 7
Ding, Yuntian, 24

El-otmani, Douha, 21

Faci, Adam, 26, 27
Fort, Karën, 8
Fraisse, Amel, 23

Garcia Serrano, Yanet, 15
Grinbaum, Alexei, 10
Guerouaou, Nadia, 11

Hammer, Barbara, 9
Hardouin, Cécile, 13
Herbaut, Nicolas, 24

Lenoir, Julie, 22
León González, Carlos, 17

Maronet, Léa, 22
Martinez, Sadya, 28
Mehdoui, Mouheb, 23
Morin, Axel, 25
Musiani, Francesca, 6
Mustafa El Hadi, Widad, 23

Nikolic, Milena, 30
Nourdin, Ivan, 4

Ollion, Étienne, 25

Radjenovic, Zarko, 30
Rougier, Adrien, 25

Salinesi, Camille, 24
Sanchez Santana, Jennifer de la Caridad, 15
Schultz, Emilien, 25
Sistachs Vega, Vivian del Rosario, 19

Taounza-Jeminet, Alyx, 22
Tufféry, Christophe, 29

Zrigui, Mounir, 23