239 - REFLEX

PhD Research and Training proposal

From Fact-Checking to Reality-Checking: Operationalizing Discursive Modalization to Predict and Prevent Cognitive Vulnerability

1. EXCELLENCE (4 pages max)

1.1. Pre-proposal's context, positioning and objective(s)

Context and Vision: From Verification to Cognitive Adherence. Disinformation has evolved from simple factual errors into a socio-cognitive phenomenon exploiting human vulnerabilities, posing a significant risk to contemporary society. In an ecosystem where AI-driven persuasive content (e.g., Sora **[Zho+24]**) blurs the line between truth and credibility and changes the citizens' relationship with information, the challenge for decision-makers shifts from binary fact-checking ("*Is it true ?*") to understanding the "microscopic" mechanisms of adherence ("*Why is it believed ?*"). This shift prioritizes critical thinking training (French Digital Strategy for Education 2023-2027) **[MEN26]** over mere technical detection **[Noë24; KH24]**.

Current research remains centered on content detection and lacks models of how individuals interpret, credit, or reject information based on their context **[Nan+25]**. To address this, the LIS and the University of Windsor (Canada) have established since 2024 a long-term strategy structured in two axes:

- Phase 1 (Macroscopic; EIFFEL grant): Systemic and formal modeling of the "how" and "what" of information propagation using formalisms like DEVS.
- Phase 2 (Microscopic): Modeling and prediction of the "why" of the receiver's perceived credibility, moving from passive flow simulations to adaptive agents possessing authentic cognitive adherence (cognitive biases, context, and mental state).

However, we posit that a critical transition mechanism between these phases is necessary, and that *discursive modalization* serves as the operational bridge between message dynamics and cognitive reception. Transforming these linguistic markers into computational variables enables the conversion of static simulation nodes into context-sensitive agents, requiring a transdisciplinary synergy between linguistics, discourse analysis, and AI computing.

This context poses three fundamental questions: Why is information, regardless of veracity, perceived as credible, adopted and eventually relayed ? How does content bypass critical thinking safeguards, especially through linguistic forms or cognitive biases ? How artificial systems can detect and respond to that ?

General Problem Statement and Scientific Questions. The current state of the art in computer science primarily assesses content consistency, treating truth as an absolute, static label while viewing users as identical nodes incapable of cognitive evolution **[Srb+25]** and without perception. This overlooks that truth is a situated, dynamic notion, depending on information available at a specific moment and which may be credible, factually false yet "contextually coherent," accelerating its propagation **[SWL19; Ham22]**. Furthermore, failing to distinguish between false and forged information at an epistemic level prevents systems from explaining why "fake news" succeeds, leading to unrealistic simulations and an inability to anticipate persuasion dynamics **[KT20]**.

From this observation, we can extract scientific hurdles that need to be investigated and understood:

1. Cognitive Blindness & Illusion of Competence – What does the recipient actually perceive ? Current approaches treat truth-seeking as a simple sorting of content and ignore internal psychological patterns and predispositions that dictate trust. Verification methods fail to account for the "illusion of competence" introduced by LLMs and the biasing influence of authoritative or status-based arguments.

2. Absence of a Predictive Credulity Model: No computational framework currently estimates the probability of information acceptance based on a socio-cultural profile. We lack the means to anticipate "cognitive reception", whether content will be accepted without verification and relayed regardless of veracity.
3. Absence of a Likelihood Measure – What makes an information likely “true” ? Current systems fail to explain why forged information, as information created intentionally to harm, "seems genuine". This hurdle involves identifying the linguistic and rhetorical patterns favoring adherence, as well as the impact of affective states, cognitive fatigue, or prior trust.
4. Absence of Subjectivity and Contextual Truth Metrics – How contextual truth can be formalized ? Truth is treated as a global label rather than a context-dependent one. There is no unified mathematical model to quantify the "cognitive distance" between information *I* and an individual's belief state *C*, moving beyond ternary classification.
5. Lack of Simulation Plasticity – How does news acceptance impact propagation ? Rigid models assume uniform behaviors and ignore belief dynamics. The absence of cognitive plasticity mechanisms (learning/adaptation) prevents the simulation of "truth bubbles" or backfire effects.
6. Deficit of Pedagogical Explainability – How should justification be presented ? Systems classify but do not explain persuasion. Training critical thinking requires justifications that make explicit the cognitive levers activated by disinformation, while measuring how these justifications are themselves perceived.

Central Research Question and Synthesis: Beyond factual accuracy, how can the feeling of trust and the situational validity of information be mathematically defined to provide modeling frameworks with the capacity to anticipate the adaptive patterns of human behavior ?

We analyze this cognitive dimension as a dual process bridging the issuer’s strategic persuasive intentionality, encoded in specific linguistic markers, and the recipient’s dynamic reception. The central scientific challenge is to define a congruence metric between the message's persuasive profile and the receiver's mental state. This requires mathematically modeling hidden cognitive variables and situational truth to quantify persuasion potential before propagation.

This thesis focuses specifically on the first dimension only, investigating how intent is woven into the language itself, while the comprehensive modeling of cognitive reception remains the logical next stage for this research program.

Positioning regarding the State of the Art. The project identifies two pillars in the current landscape:

1. **Psycho-Cognitive Sciences and Linguistics**: While psychology has identified qualitative credibility factors (illusion of truth, fluency) **[PR21]**, this project operationalizes discursive modalization as the tangible trace of persuasive intentionality **[Hyl05; RDJ13]**. By constructing a formal typology of persuasion markers (e.g., authoritative framing, intensity), we aim to transform abstract cognitive biases into calculable vectors, effectively bridging the gap between qualitative linguistic theory and quantitative simulation **[Ras+17]**.
2. **Cognitive AI, System Modelling and Simulation:** This axis integrates NLP with Multi-Agent Simulation to overcome isolated limitations. On the NLP side, while current models achieve benchmark performance **[Tho+18; GSV22; Quy+25]**, they suffer from instability, hallucinations **[Kau+25]**, cognitive blindness and are unable to model perceived likelihood or explain why false information appears credible **[Ji+23; ATN25]**. To our knowledge, existing architectures lack a dedicated computational layer for human perception, while the specific linguistic and logical mechanisms driving their classification remain largely understudied. On the Simulation side, while DEVS formalism effectively models macroscopic propagation **[Far+25; Jak+25]**, current models remain constrained by homogeneous agents and static probabilities that fail to capture dynamic cognitive phenomena like polarization or the "backfire effect" **[ZZ20; RCR25; Che+25]**.

Project Objectives and Research Hypotheses. The primary objective of this thesis proposal is to design a *Computational Persuasion Estimator*. This model focuses on qualifying and quantifying the persuasive intentionality of a statement by analyzing specific linguistic markers and discursive modalization.

Ultimately, this metric is intended to be integrated as a foundational dimension within the broader *Human Cognitive Credulity Estimator* (HCSE), a comprehensive framework representing the perceived credibility and contextual validity of information.

However, to preclude any ambiguity regarding the operational scope: This PhD focuses exclusively on the design, formalization, and validation of the CPE. The large-scale integration of the CPE into the broader HCSE simulation framework constitutes a separate research program beyond the strict scope of this doctoral work. Nevertheless, the development of a functional Proof of Concept (PoC) of this integration remains a prospective extension, envisioned as a high-value bonus contingent upon the candidate's progress and the efficiency of the validation phase.

Application. The HCSE framework enriches media assessment by weighting factual truth against a personalized trust indicator, serving as the foundation for adaptive cognitive simulations and holistic fact-checking tools that contextualize information within the recipient's unique worldview, facilitating a deeper study of how ideas spread and the generation of explanations tailored to specific human leanings.

Originality and Innovation. The project's novelty lies in its systemic approach to address a fundamental challenge: cognitive grounding. It models truth as a tripartite relation between an utterance, its surrounding circumstances, and the recipient's mental landscape (Statement-Context-Receiver) rather than a binary property of text. It moves beyond text classification to model the interaction between text and a dynamic mental state, aiming to quantify "persuasion" before diffusion occurs. It proposes first step to a new systemic approach combining cognitive sciences, computer science, and system sciences to treat not only the truth of information but its perceived credibility by taking into account the reception context: passing from "true or false information" to "credible information for whom, why, and how".

Hypotheses. We posit two key hypotheses to verify during the investigation in this thesis proposal: **(H1)** Perceived credibility is predictable using psycho-linguistic markers and underlying cognitive models, independent of factual truth. This perception is closely tied to the issuer's persuasive intentionality, which is manifested within the discourse through specific narrative and linguistic structures. **(H2)** Integrating a credulity score based on a persuasion estimator improves the detection of "persuasive" fake news where LLMs currently fail.

These two hypotheses are the conditions for three broader hypotheses to be tested:

(H3) The divergence between "contextual truth" (receiver-dependent) and "factual truth" is a predictor of disinformation potential. **(H4)** Introducing cognitive plasticity (adaptation) in simulation agents reproduces emergent phenomena (polarization, resistance) absent in static models. **(H5)** Explaining the "cognitive levers" of adherence offers better pedagogical prevention than simple refutation.

Methodology. The research methodology in this thesis proposal is structured around three incremental layers, designed to move from theoretical formalization to practical application within the Computational Persuasion Estimator (CPE).

Layer 1: Formalization of Persuasive Intentionality. The first step is to deconstruct the "mechanics of influence" by identifying the specific discursive modalizations used to manipulate perception (e.g., markers of authority, emotional intensity, pseudo-logical connectors). The goal is to operationalize these qualitative linguistic features into measurable variables. This layer involves constituting a specialized corpus where content is annotated not by its truth value, but by its persuasive density, creating a taxonomy of "cognitive hooks".

Layer 2: Design of the Computational Persuasion Estimator (CPE). This step involves developing the hybrid architecture (combining Expert Rules and Machine Learning) using different techniques that powers the CPE. This model takes a message as input and calculates a *Persuasion Score*, a quantitative index reflecting the intensity of

the issuer's intent to convince, independent of factual veracity. This engine transforms the abstract concept of "seductiveness" into a tangible, graded metric.

Layer 3: Validation of the CPE. The relevance of the CPE index will be empirically validated. Comparisons will be performed with non-enriched pipelines to demonstrate that the *Persuasion Score* provides a distinct and necessary signal compared to simple factual verification, effectively isolating the "persuasive charge" of a message.

Opening: Integration into the HCSE Framework. The CPE is designed to be the foundational module of the broader HCSE. Its integration opens two critical horizons:

1. **Augmented Fact-Checking:** Producing enriched verdicts (e.g., *Factual Verdict: FALSE; Contextual Veracity: TRUE; Persuasion Score: HIGH; Bias: Confirmation*) to generate specific justifications adapted to a target audience.
2. **Adaptive Cognitive Simulation:** Synergizing with the LIS existing research (Phase 1) to move beyond static parameterization. The CPE will drive dynamic adaptation mechanisms, such as repeated exposure effects or confrontation with inconsistencies, to simulate the resilience of a network facing disinformation, information bubbles, or polarization phenomena.

Gender Dimension in Research Content. As this research models human cognition and susceptibility to persuasion, the Gender Dimension is intrinsic to the project's scientific integrity. Recognizing that persuasion strategies and linguistic markers may impact genders differently due to socio-cultural conditioning, the project implements a rigorous protocol to avoid algorithmic bias. Specifically, profiling models will be designed to account for diverse cognitive styles across genders, while training corpora and experimental cohorts will be strictly balanced to prevent the system from overfitting to specific demographics, such as erroneously flagging female speech patterns as less credible. Furthermore, the analysis of contextual truth will explicitly investigate whether specific disinformation narratives disproportionately target or exploit gender-specific biases.

1.2. Interdisciplinary dimension of the project

Contribution of Disciplines. This project bridges Computer Science (AI/Simulation) and Linguistics/Cognitive Science to address the "cognitive blindness" of current technological solutions. This collaboration is not merely additive but mandatory for the project's success: without linguistics, the AI remains blind to meaning; without Computer Science, linguistic theory cannot model propagation at scale. Consequently, the Computer Science axis (LIS / UWindsor) provides the formalism for systemic and agent-based modeling, the architecture for the CPE/HCSE, and the computational power required to simulate complex social dynamics. Complementarily, the Linguistics axis (LPL) moves beyond simple sentiment analysis to focus on the micro-mechanisms of language that trigger belief, ensuring the model captures the semantic nuance necessary for accurate prediction.

Specific Interdisciplinary Methodology: The Linguistic Marker Analysis. A key originality of our approach is the operational translation of qualitative concepts into computational variables. This approach will be particularly situated within the framework of French discourse analysis, which places special emphasis on enunciation and discourse regimes **[MAI25]**. The collaboration is structured around a "Translation Loop": linguists identify specific discursive modalizations (e.g., epistemic modals, evidentiary markers, rhetorical intensity) **[VIO14]** through experimental study. These qualitative features are then converted into quantitative weights (vectors) for the computational model. This methodology ensures that the AI's "learning" is constrained and guided by established linguistic theory rather than random statistical correlation. Our qualitative methodology will effectively integrate the conceptual frameworks for analyzing digital discourse **[DEI25]**.

Originality of the Approach. Unlike standard NLP approaches that rely on opaque embeddings, this interdisciplinary method ensures explicability by design. By grounding the AI's weights in linguistic theory, we can explain *why* the system flags content as persuasive (e.g., "This text uses high-intensity epistemic markers to mask a lack of evidence"). This creates a Hybrid AI that leverages the predictive power of machine learning while retaining the semantic transparency of human sciences, directly serving the project's pedagogical goals.

2. IMPACT (2 pages max)

2.1. Expected impact of the project on the candidate's career

Acquisition of Transverse "Human-Centric" AI Profile. This doctoral project addresses a critical societal need: the training of a new generation of transdisciplinary experts operating at the frontier of Artificial Intelligence (Computer Science), Cognitive Science, and Systems Science. It is specifically designed to cultivate a rare, hybrid profile capable of acting as a strategic bridge between academic rigor and industrial application. As society moves toward tighter structural coupling between human cognition and algorithmic systems, it is imperative to develop professionals capable of anticipating the reciprocal impacts of this integration. By mastering the complex translation of qualitative concepts (linguistics) into rigorous computational models (hybrid AI), the fellow acquires a dual competence that dissolves the traditional boundary between theoretical research and technological deployment. Methodological hybridization is vital for bridging the "interpretative gap" currently limiting AI systems, while cognitive modeling and impact anticipation enable the researcher to go beyond standard statistical approaches to formalize dynamic human behaviors (intentionality, belief evolution). Ultimately, this expertise allows for the integration of the human dimension not merely as a user, but as a core parameter of conception and evaluation.

Unified Career Prospects: Leading the "Responsible AI" Transition. This versatility ensures high employability by positioning the candidate at the exact intersection of public strategy and private necessity. The acquired skills are equally critical for leading transdisciplinary research initiatives aligned with sovereign ambitions (e.g., *France 2030*) and for designing compliant, explainable, and ethically grounded systems required by tightening regulations like the *EU AI Act*. Consequently, the fellow will be prepared to occupy high-stakes pivoting roles, *such as Socio-Technical Systems Architect or AI Safety Strategist*, where the ability to align algorithmic performance with societal and ethical standards is the primary driver of value, regardless of the sector. These emerging professions are dedicated to ensuring that the next generation of algorithms, whether in education, media, or governance, serves human cognitive resilience rather than exploiting its vulnerabilities. With this training program, the candidate positions themselves at the forefront of "Societally Aware AI". This PhD will enable them to lead research initiatives that require understanding both the algorithmic engine and the human driver, or to design and engineer reliable products integrating ethical systems.

Development of Cross-Sectoral Agility. The international cooperation (UWindsor/AMU) serves as an incubator for this transverse agility. Navigating between the high-performance computing requirements of the Canadian partner and the theoretical exigencies of the French laboratory fosters advanced project management skills suited for complex, multi-stakeholder, large-scale collaborative environments. Furthermore, by designing tools that directly serve civil society (education, media), the researcher learns to translate complex scientific innovation into tangible public value, a critical asset for strategic leadership in any organization bridging the gap between science and society. By addressing education and democratic integrity, the project breaks the "ivory tower", grounding scientific innovation in civic reality through direct dialogue with societal actors by responding to an urgent demand from citizens. Finally, leveraging Canada's status as a "strategic trusted partner" of the EU, this initiative fosters an ethical transatlantic synergy that reinforces the laboratory's global standing.

2.2. Expected impact for the thematic axis

Alignment with the Axis. This project perfectly embodies the SCHADOC AI thematic axis's originality: *combining historical symbolic AI with recent numerical learning techniques.* Current Large Language Models (Numerical AI) are powerful but opaque "black boxes" that generate text without understanding truth or intent. By integrating a symbolic layer (Formalization of cognitive states, rules of discursive modalization) into the numerical architecture (Machine Learning classifiers), this project creates a Hybrid AI. It directly answers the specific call of the axis to apply AI to "human behavior studies" by modeling the most complex of interactions: the granting of trust. Furthermore, this project pioneers a "Cognitive NLP" framework that treats society as a complex adaptive system. By leveraging recent advances in linguistics, we aim to use language analysis not merely to classify text, but to

predict human behavior (adherence, rejection, polarization). This synergy allows us to operationalize linguistic markers as behavioral drivers within computational models, enabling the rigorous reproduction of complex socio-cognitive dynamics. Consequently, this represents a reciprocal breakthrough for both disciplines: while AI gains semantic depth, linguistics acquires a novel computational lens to better understand and anticipate the causal mechanisms linking discursive strategies to observable human behavior. This research transforms static information processing into a dynamic tool capable of simulating how information flows, mutates, and impacts the collective mental state of a population.

Advancement of the Research Field. The project addresses a critical stagnation point in the field of Disinformation Studies. By shifting the focus to perceived credibility and persuasive intentionality, this research offers a new paradigm:

- From Detection to Prediction: Moving from classifying static text to anticipating dynamic reception.
- From Binary to Situated Truth: Operationalizing the concept that truth is context-dependent, a major leap for computational epistemics.

Societal Impact: Strengthening Democratic Resilience. Beyond academia, the project addresses an urgent societal threat. By modeling the "cognitive flaws" that disinformation exploits, we aim to provide tools that do not just censor content, but help citizens understand *why* they are being targeted. This aligns with the need to protect democratic deliberation from automated manipulation and "cognitive hacking."

2.3. Dissemination, exploitation and communication activities planned

Scientific Dissemination (Peers and Academic Community). The results will be disseminated through high-impact channels prioritizing both AI, Social Computational Sciences and Linguistics.

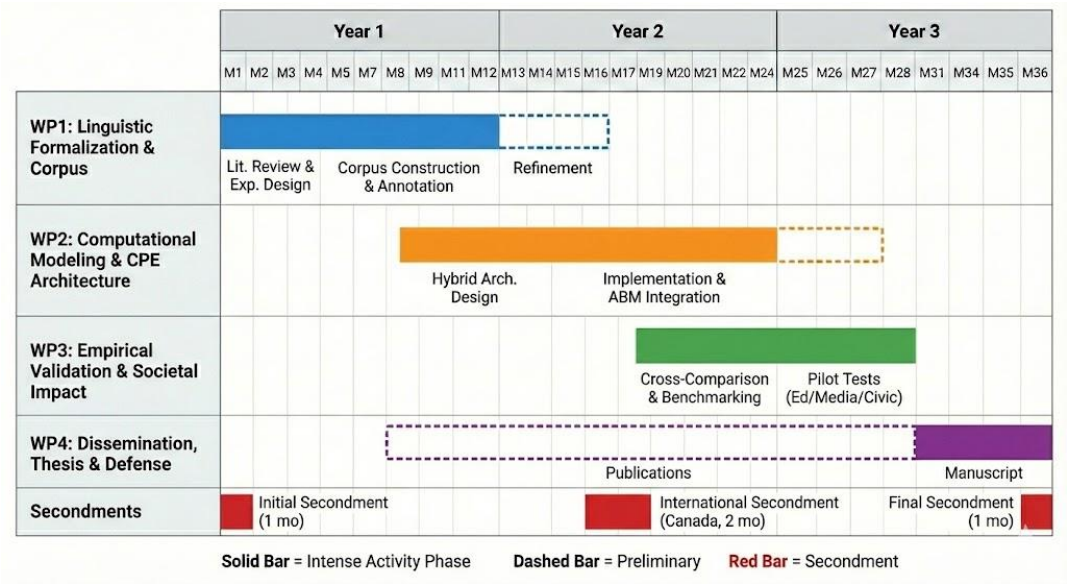
- **Publications:** Targeting Rank A conferences in NLP (ACL, EMNLP), AI/Simulation (EGC, CIKM, AAMAS, IEEE Transactions on Computational Social Systems), Cognitive Computing (Cognitive Computation, IEEE Transactions on Cognitive Systems).
- **Open Science:** Adhering to the guiding European principle "As Open as Possible, as Closed as Necessary", the project adopts a tiered dissemination strategy that rigorously balances scientific reproducibility with ethical security. While the annotated datasets and the underlying architectural code will be made fully available as open-source resources to allow the community to validate findings, the release of the fully trained CPE weights will be governed by a *Responsible Release* protocol. To specifically mitigate the "Dual Use" risks identified in the ethical assessment, access to the operational detector may be subject to a Controlled Access license restricted to verified academic and civil society actors, thereby preventing malicious actors from reverse-engineering the system for propaganda optimization while still fostering legitimate research.

Exploitation: Pedagogical Tools for Critical Thinking. Unique in its application, the CPE model powers a Critical Thinking Training Simulator targeting education, media, and governance. Collaborating with INSPE, it equips teachers for proactive manipulation analysis while offering media a forensic tool to detect influence campaigns. Ultimately, it provides citizens with essential "intellectual self-defense", enabling them to visualize the critical distinction between objective facts and emotional persuasion strategies.

Public Engagement and Outreach. To engage society, the candidate will organize interactive workshops (e.g., "Fête de la Science") where citizens test their "credulity" against the AI model to visualize biases in real-time. This outreach extends to media contributions (e.g., *The Conversation*) and targets youth education: digital literacy campaigns will equip secondary students with "Explainable AI" tools to detect manipulation, while university seminars will showcase the Linguistics-Computer Science synergy to inspire future transdisciplinary research careers among undergraduate and Master's students.

3. IMPLEMENTATION (2 pages max)

Work Packages. The project is structured into four interconnecting Work Packages (WPs) designed to achieve the objectives within the standard 36-month doctoral timeframe. The schedule ensures a progressive transition from theoretical formalization to computational validation. Expected deliverables are in bold and underlined.



WP1: Linguistic Formalization & Experimental Corpus Construction (Months 1-12)

- **Objective:** To identify and categorize markers of "persuasive intentionality" through an experimental approach, aiming to construct a qualitatively **annotated corpus** capable of quantitative operationalization.
- **Tasks:** Experimental study of discursive markers; Definition of the annotation protocol (transitioning from qualitative analysis to calculable vectors); Construction of the reference dataset.
- **Supervision:** Led by Dr. Damien Deias, in close collaboration with the intersectoral partner.

WP2: Computational Modeling & CPE Architecture (Months 10-24)

- **Objective:** To formalize and develop the learning model for the Computational Persuasion Estimator (CPE).
- **Tasks:** Design of a Hybrid Architecture combining quantitative metrics (ML) and Agent-Based Modeling (ABM) logic; **Implementation of the estimator** to bridge linguistic inputs with cognitive state predictions.
- **Supervision:** Led by Dr. Amine Hamri, with the support of Dr. Aznam Yacoub, Dr. Ismail Badache and Dr. Erwan Tranvouez.

WP3: Empirical Validation & Societal Impact Assessment (Months 20-30)

- **Objective:** To validate the model through cross-comparison with real-world data and assess its utility across targeted societal sectors.
- **Tasks:** **Benchmarking** the model's outputs against reality (ground truth); Deploying pilot tests with **samples** from educational, journalistic, and civic environments to measure the tool's effectiveness in "intellectual self-defense".
- **Supervision:** Led by Dr. Damien Deias, with the support of Dr. Ismail Badache and Dr. Erwan Tranvouez, in close collaboration with the intersectoral partner.

WP4: Dissemination, Thesis Writing & Defense (Months 6-36)

- **Tasks:** **Continuous publication** in top-tier conferences (Ranking A) across NLP, AI, and Social Simulation; Public outreach (critical thinking workshops using the tool); Manuscript writing and final defense.

Feasibility. The feasibility of the schedule is secured by a structural risk-mitigation strategy designed to ensure timely delivery regardless of specific experimental hurdles. The project relies on a "Reusability Strategy," building directly upon simulation modules previously developed in the laboratory to bypass the initial prototyping phase. This is reinforced by a "Modular Contingency" approach: the development of the Estimator constitutes a standalone scientific contribution, ensuring that even in the event of delays in the final systemic integration, the core results remain sufficient for a successful defense. Furthermore, the timeline employs deliberate parallelization, where the overlap between modeling and validation allows for iterative testing without sequential bottlenecks. Operational feasibility is further guaranteed by secured access to High-Performance Computing resources through the international partnership, while the research scope is strictly contained to a specific set of discursive markers to prevent experimental drift. Finally, the mobility plan is optimized for efficiency: the international secondment to UWindsor is strategically scheduled in Year 2 to coincide with the computationally intensive training phase (WP2), granting the candidate direct access to the required HPC infrastructure and supervision exactly when needed without disrupting the workflow.

Management Structure and Procedures. The project relies on a robust international co-supervision structure ensuring complementary expertise:

- **Main Supervisor (Modelling and Simulation/AI):** Dr. Maamar El Amine Hamri (AMU/LIS) – Supervision of the Computational and Systems Modeling axis.
- **Co-Supervision (Linguistics)** : Dr. Damien Deias (AMU/LPL) – Supervision of the Linguistics axis.
- **International Scientific Advisor (Cognitive AI/Cyber-Physical Systems Modelling):** Dr. Aznam Yacoub (University of Windsor) – Supervision of the "Human-Centric" AI architecture and systemic formalization.
- **Technical/Scientific Advisors (AI and Agent-based Modelling):** Dr. Ismail Badache and Dr. Erwan Tranvouez – Expertise in Information Retrieval and Multi-Agent Systems / Cognitive Modelling.
- **Intersectoral Primary Partner:** Anonymal – Marie-Julier PELTIER.
- **Partnerships under Study:** To expand the validation spectrum and societal reach, strategic collaborations are currently being explored with La Provence, the EFSCN (European Fact-Checking Standards Network), and EDMO (European Digital Media Observatory). Furthermore, additional collaborations are envisioned with key AMU structures, namely InCIAM, Ampiric, IMSIC, and ILCB, which constitute a robust interdisciplinary ecosystem directly relevant to the project's cognitive and societal dimensions.

Monitoring Procedures.

- **Weekly Technical Meetings:** With the local co-supervisors to address immediate hurdles.
- **Monthly Steering Committees:** With the full direction team (including Windsor via video-conference) to review WP progress against the Gantt chart.
- **Semester Reviews:** To validate milestones (e.g., Corpus validation, Prototype V1) and adjust the roadmap if necessary.

Risk Management. Feasibility is secured by three strategies. Data scarcity, specifically the difficulty of annotating "credibility", is addressed by enrichment: rather than creating massive datasets from scratch, we will overlay a specific "persuasion" layer onto existing standards like FEVER. Interdisciplinary complexity is bridged by collaboration with linguists (Damien Deias) to yield operational markers for direct algorithmic implementation. Finally, integration with Phase 1 is secured by a shared formalism (DEVS) and common laboratory infrastructure (LIS/UWindsor), ensuring native compatibility between "Container" and "Content."

4. ETHICS SELF-ASSEMENT

1) Studies involving humans: Details on recruitment, inclusion and exclusion criteria and informed consent procedures.

Recruitment & Inclusion/Exclusion Criteria: Human participants may be involved during two specific phases: (1) the validation of linguistic markers (WP1) and (2) the testing of the pedagogical tool (WP3).

- **Target Population:** Recruitment will be strictly limited to adult volunteers (aged 18+). No minors or vulnerable populations will be recruited.
- **Recruitment Method:** Participants will be primarily recruited within the university environment (students, staff) via internal mailing lists and campus announcements. Secondary recruitment may be facilitated through the intersectoral partner's network to broaden the demographic scope (civil society). In such cases, the partner will act solely as a communication relay; the formal inclusion process and collection of consent will remain the exclusive responsibility of the research team to ensure identical ethical standards.
- **Diversity & Equity:** Recruitment will be conducted with a strict commitment to non-discrimination and gender equality. As the study design does not aim to isolate specific demographic groups for correlation analysis, we will apply standard unbiased selection procedures to ensure a diverse and representative cohort. Special attention will be paid to monitoring gender balance throughout the inclusion process.
- **Exclusion Criteria:** Individuals under 18 years of age or those unable to provide legal consent.

Informed Consent Procedures: All participants will be required to sign a formal Informed Consent Form prior to any interaction. This document, drafted in collaboration with the Ethics Committee of Aix-Marseille Université (AMU) and in compliance with GDPR (and Canadian TCPS 2 where applicable), will clearly state:

- The purpose of the study (analyzing perception of text, not assessing the individual's intelligence).
- The voluntary nature of participation and the right to withdraw at any time without penalty.
- Data anonymity and protection measures.
- Contact information for the researchers and the ethics board.

Unexpected Findings Policy: Although this research is non-clinical and non-invasive (focusing on linguistic perception and critical thinking), protocols are in place for incidental findings:

- **Scope:** The research team does not have the mandate or qualification to provide medical or psychological diagnoses. Therefore, no medical feedback will be provided to participants regarding their cognitive performance.
- **Distress Protocol:** In the unlikely event that a participant exhibits signs of severe psychological distress or discloses sensitive information indicating a risk of harm to themselves or others during the study (e.g., in open-ended responses or interviews), the protocol requires the researcher to stop the session. The participant will be gently directed toward appropriate university support services (e.g., Student Health Services or Counseling Center), without the researcher attempting to manage the situation personally.
- **Data Anomalies:** If data reveals a significant anomaly in a participant's interaction with the tool (e.g., systematic incoherence), this data will be excluded from the dataset for scientific rigor, but no notification will be sent to the participant to avoid causing undue alarm, unless explicitly required by the Ethics Committee.

2) Personal Data.

Details of the technical and organisational measures to safeguard rights and freedom of the participants. The project strictly adheres to the GDPR and the Canadian PIPEDA.

- **Governance**: The project will be registered with the Data Protection Officer (DPO) of Aix-Marseille Université. A specific Data Management Plan (DMP) will be drafted at M6.
- **Security**: All collected data will be stored on secured, encrypted servers provided by the CNRS (Huma-Num infrastructure), with access restricted to authorized consortium members via strong authentication.
- **Pseudonymization**: A strict separation will be maintained between direct identifiers (names of annotators/participants) and research data (cognitive scores). Identity keys will be stored locally on an offline-encrypted drive, separate from the cloud processing servers.

Details of the informed consent procedures.

- **For Experimental Participants (Annotators/Testers):** Prior to any interaction with the CPE tool, participants will sign a granular informed consent form detailing: the research purpose, the type of data collected (e.g., reaction time, credibility assessment), their right to withdraw at any time without penalty, and the data retention period (5 years post-publication).
- **For Public Social Media Data:** Data collection from platforms (e.g., X, Comments) will rely on the "Legitimate Interest" legal basis for research, ensuring strict data minimization and excluding private conversations.

Data minimisation. The project applies a "Privacy by Design" approach. We collect only the metrics strictly necessary to model the "Credibility Perceived" vector (e.g., linguistic markers, user interaction logs, hesitation time). No extraneous Personal Identifiable Information (PII) such as political views, religious beliefs, or geolocation will be stored unless it constitutes the specific object of the linguistic bias analysis, in which case it will be immediately tokenized and dissociated from the user's identity.

Justification of why personal data will not be anonymized. Data will be pseudonymized during the data construction phase to ensure the quality control of the linguistic annotation process. Retaining a temporary link between the annotator ID and the labeled data is strictly necessary to calculate Inter-Annotator Agreement metrics (e.g., Cohen's Kappa) and to detect/correct potential systematic biases in the labeling of "credibility markers." Full anonymization (destruction of the correspondence keys) will occur immediately upon the validation of the final dataset.

Data transfer. Encrypted transfer of fully anonymized datasets (irreversible stripping of all direct and indirect identifiers).

- **Destination**: Canada (University of Windsor).
- **Justification & Framework**: The Machine Learning training phase hosted at UWindsor relies exclusively on abstract cognitive vectors and linguistic patterns, requiring no user re-identification capability. Consequently, all data is anonymized prior to export. Although anonymized data technically falls outside the scope of personal data restrictions, the transfer and storage infrastructure at UWindsor will still adhere to the security standards aligned with the European Commission's Adequacy Decision regarding Canada (PIPEDA).

3) Non-EU Countries.

Countries involved: Canada**.** (Specific partner: University of Windsor, School of Computer Science).

Benefits. The involvement of Canada provides the project with critical resources that significantly outweigh the logistical risks:

- **High-Performance Computing (HPC):** Access to the *Digital Research Alliance of Canada* (formerly Compute Canada) infrastructure is essential for training the complex hybrid models (CPE) and running large-scale multi-agent simulations (WP2 & WP3), offering computational power that complements the French laboratory's resources.

- **Scientific Excellence:** The University of Windsor provides unique expertise in socio-cognitive systems modeling, essential for the "Society as a System" simulation component.

Risks & Mitigation (GDPR Compliance). The primary risk concerns the transfer of personal data (e.g., annotated corpora or participant feedback) outside the European Economic Area (EEA).

- **Mitigation Strategy:** This risk is considered low and controlled. The European Commission has recognized Canada as providing an adequate level of data protection (Adequacy Decision 2001/519/EC). Consequently, personal data can flow from the EU to Canada without requiring additional specific authorization. All transfers will be governed by a strict Data Management Plan (DMP). Data will be stored on secure University of Windsor servers, while heavy processing will occur on the Digital Research Alliance of Canada infrastructure. The Alliance operates under rigorous federal security protocols and privacy standards (PIPEDA), ensuring that data processed on national clusters remains protected at a level equivalent to EU requirements.

Details on activities are carried out in non-EU countries. Activities conducted in Canada will be strictly scientific and technical, focusing on the heavy computational aspects of the project:

- **Computational Modeling & Hybrid Architecture (WP2):** The heavy lifting of algorithmic training for the Computational Persuasion Estimator (CPE) will utilize Canadian HPC resources. This includes the development of the Agent-Based Modeling (ABM) components which require significant processing power provided by the *Digital Research Alliance of Canada*.
- **Scientific Co-supervision:** Regular technical workshops with the Canadian supervisor to refine the hybrid quantification methods used in the estimator.
- **International Dissemination:** Presentation of the CPE model and validation results at North American conferences to ensure global visibility.

4) Artificial Intelligence.

Explanation as to how the participants and/or end-users will be informed / Transparency and Explainability (XAI). Participants in the validation phase and future end-users (teachers, journalists) will be systematically informed that they are interacting with an Artificial Intelligence system. A clear "transparency notice" will precede any usage, specifying that the generated "credibility scores" are probabilistic estimates derived from linguistic analysis, not absolute truths. To ensure users understand the logic behind the decisions, the tool acts as a "Glass Box" rather than a "Black Box": it will not only output a score but also highlight the specific linguistic markers (e.g., emotional intensity, logical fallacies) that triggered the verdict. This "pedagogical feedback loop" ensures that users understand the system's reasoning, adhering to the principle of explicability. Finally, the system's limitations will be explicitly stated, warning users that the tool is a decision-support aid and must not replace critical human judgment.

2) Details on the measures taken to avoid bias in input data and algorithm design / Data Balance and Hybrid Architecture. To prevent the replication of societal biases, the training corpora will undergo a rigorous "bias audit" before use. We will apply strict balancing protocols to ensure equal representation of diverse demographic sources, specifically monitoring for gender or cultural bias (e.g., ensuring female speech patterns are not algorithmically penalized as "less credible"). Regarding algorithm design, the project utilizes a Hybrid Architecture combining Machine Learning with symbolic Expert Rules. This structural choice is a key mitigation strategy: unlike pure Deep Learning models which can behave unpredictably, our Expert Rules impose logical constraints that prevent the model from learning or amplifying discriminatory correlations found in raw data. Continuous monitoring during the WP3 validation phase will specifically test for disparate impact across different user groups.

Explanation as to how the respect to fundamental human rights and freedoms will be ensured / Cognitive Sovereignty and Human-in-the-Loop. The project is fundamentally designed to enhance, not replace, human autonomy. The tool serves as an instrument of "intellectual self-defense," empowering the user to make informed

decisions rather than automating the belief process. We strictly adhere to a "Human-in-the-Loop" design philosophy: the AI provides forensic evidence, but the final judgment on a text's credibility remains with the human operator. Privacy and data protection are ensured through strict GDPR compliance; no personal user data is required for the core functioning of the linguistic analysis models. Any data collected for validation (WP3) will be anonymized, and the system is designed to run locally or on secure servers to prevent intrusive profiling of the user's reading habits.

Detailed explanation on the potential ethics risks and the risk mitigation measures / Dual Use and Automation Bias. We have identified two primary ethical risks. First, the "Dual Use" risk: the theoretical knowledge gained on persuasion mechanics could theoretically be repurposed to generate more effective disinformation. To mitigate this, the generative capabilities of the model will be restricted. The tool is engineered as a *detector* (classifier), not a *generator*. Detailed weights and "persuasion recipes" will be kept confidential within the secure research environment. Second, the risk of "Automation Bias" (users blindly trusting the AI's score). To mitigate this, the user interface will avoid binary "True/False" labels, instead using nuanced "Confidence Scores" and visual visualizations of the analysis complexity. This design forces the user to engage cognitively with the results rather than passively accepting a machine verdict. Additionally, training modules for teachers and journalists will emphasize the system's fallibility to maintain critical vigilance.

Continuous Ethical Oversight & Expert Consultation / Institutional Supervision. Beyond these initial design measures, the project commits to a structured and ongoing dialogue with the Ethics Committee of Aix-Marseille Université (AMU). A dedicated ethics advisor or data protection officer (DPO) from the university will be consulted at critical milestones, specifically before the deployment of the validation phase (WP3), to audit the specific protocols regarding human-AI interaction. This external expert review ensures that the project's evolving methodologies remain strictly aligned with the latest regulatory updates, including the emerging standards of the EU AI Act, and provides an additional layer of independent oversight to validate the effectiveness of our bias mitigation strategies throughout the project lifecycle.

Detailed explanation of the measures set in place to avoid potential bias, discrimination and stigmatisation. To prevent the CPE (Computational Persuasion Estimator) from stigmatizing individuals based on their linguistic background, gender, or social origin, we implement a multi-layered "Fairness-by-Design" strategy:

1. Representative Data Curation & Stratification: Stigmatization in NLP often stems from training data dominated by "standard" or academic language (often white, male, Western), leading the model to flag minority dialects or sociolects as "less credible."

- **Measure:** The training corpora (WP1) will be strictly stratified. We will ensure the inclusion of diverse linguistic registers, separating "grammatical correctness" from "rhetorical credibility."
- **Specific Protocol:** We will actively monitor the dataset to prevent "Standard Language Ideology" bias. The model will be trained to recognize that the use of non-standard English/French (e.g., specific sociolects) or non-native phrasing does not equate to a lack of trustworthiness.

2. Hybrid Architecture as a "Safety Lock": Pure Deep Learning models can latently learn to correlate specific topics (e.g., minority rights) or identity markers with "bias" or "emotion."

- **Measure:** By using a Hybrid Architecture (Machine Learning + Symbolic Expert Rules), we impose logical constraints. The Expert Rules (defined by linguists) explicitly define what constitutes a "persuasion marker" (e.g., a logical fallacy).
- **Effect:** The system is hard-coded to penalize specific *rhetorical structures* (manipulation techniques), preventing it from penalizing *identity markers*. The AI is forced to justify its score based on linguistic evidence, not statistical correlations with protected attributes.

3. Counterfactual Testing (Perturbation Analysis): During the validation phase (WP3), we will employ "Counterfactual Fairness" testing.

- **Measure:** We will run automated tests where sensitive attributes in a text are swapped while keeping the semantic content identical (e.g., changing names from male to female, or changing location markers).
- **Success Metric:** The system passes the test only if the generated Credibility Score remains invariant despite these changes. If the system rates a text differently simply because the author is female or from a specific region, the model will be retrained.

4. Focus on Intent via Contextualization: The conceptual framework distinguishes between "Epistemic Uncertainty" (not knowing the truth) and "Sociolinguistic Politeness" (cultural or gendered norms).

- **Measure:** Instead of blindly excluding "hedging" markers (which would degrade the model's accuracy), we implement a Sociolinguistic Normalization Layer. The algorithm is trained to evaluate modalization relative to the speaker's baseline context. This ensures that "politeness strategies" (often used by specific demographic groups) are mathematically dissociated from "deceptive markers," preventing the misclassification of a polite style as a lack of trustworthiness.

5. Intrinsic Nature of Bias: Finally, we openly acknowledge that total neutrality is neither possible nor desirable in the specific context of modeling persuasion. Since the project aims to detect manipulative intent, which is inherently a deviation from neutral objectivity, sanitizing the training data of all "subjectivity" or "rhetorical bias" would strip the model of its ability to detect the very phenomena it aims to study. In this research, cognitive bias is not the "noise" to be eliminated, but the "signal" to be analyzed. The ethical safeguard relies on the output's framing: the system is designed to identify and label these biases as persuasion strategies, never to validate or reproduce them as objective truths.

Detailed explanation on how humans will maintain meaningful control over the most important aspects of the decision-making process / Principle of "Cognitive Sovereignty" (Human-in-the-Loop). The system is strictly engineered as a forensic instrument for Decision Support, not for automated decision-making or censorship. Meaningful human control is maintained through three specific design choices:

- **Non-Binary Output:** The CPE (Computational Persuasion Estimator) will strictly avoid delivering binary verdicts (e.g., "True/False" or "Safe/Unsafe"). Instead, it provides a graded "Persuasion Complexity Score" accompanied by a dashboard of linguistic evidence. This forces the user to interpret the data rather than passively accepting a machine judgment.
- **Friction by Design:** To prevent "Automation Bias" (the tendency of humans to blindly trust AI suggestions), the user interface will incorporate "cognitive friction." Before displaying a final score on a controversial text, the system will prompt the user to review the highlighted "persuasion markers" (e.g., specific emotional triggers or logical fallacies). This workflow ensures that the human cognition remains the final arbiter of credibility.
- **Override Capability:** In the educational and journalistic tools, the human operator always retains the ability to dismiss or override the AI's annotation if they detect a false positive (e.g., satire or irony), ensuring that nuanced human understanding prevails over algorithmic rigidity.

Explanation on how the presence/role of the AI will be made clear and explicit to the affected individuals / Transparency and Explainability (XAI). Users will never be subjected to "invisible" profiling. The presence of the AI will be made explicit through Transparency by Design:

- **Visual Signaling:** Every interface (pedagogical module or browser plugin) will feature a permanent visual indicator (e.g., "AI-Assisted Analysis") clearly stating that the content is being processed by an algorithm.

- **Probabilistic Framing:** The system's output will be linguistically framed as an estimation, not a fact. Visualizations will use confidence intervals (e.g., "The model detects high *likelihood* of persuasive intent") rather than definitive assertions, reminding the user of the system's probabilistic nature.
- **"Glass Box" Explanations:** When the system flags a text, it will provide an "Explainability Layer" accessible via a click (e.g., "Why was this highlighted?"). This feature will display the specific linguistic rules triggered, demystifying the black box and educating the user on how the AI reached its conclusion.

Here is a comprehensive and strategically framed response. It positions your project not just as "safe," but as a necessary "Counter-Measure" to existing threats, while acknowledging the environmental cost of AI.

Justification of the need for developing/using this particular technology. The development of the CPE (Computational Persuasion Estimator) is justified by the critical "Asymmetry of Defense" currently weakening democratic institutions. Traditional fact-checking is a reactive, labor-intensive process that cannot scale to match the velocity and volume of automated disinformation. By the time a falsehood is debunked, the cognitive damage is often irreversible. This technology is necessary to shift the paradigm from post-hoc correction to proactive resilience. By automating the detection of "manipulative intent" rather than just "factual inaccuracy," the tool fills a crucial gap in our cognitive security infrastructure. It empowers citizens and journalists to identify influence campaigns in real-time, restoring trust in the information ecosystem by revealing the hidden mechanics of persuasion rather than arbitrarily censoring content.

Assessment of the ethics risks and detailed description of the measures set in place to mitigate the potential negative impacts.

Social Risk: Dual Use and "Surveillance" Concerns. A plausible negative impact is the "Dual Use" dilemma: the same tools used to detect persuasion could theoretically be repurposed by authoritarian actors to optimize their own propaganda (using the detector as a feedback loop) or to automate censorship of dissenting voices.

- **Mitigation Strategy:** The project adopts a "Defensive-Only" Architecture. The system is engineered as a classifier (detector), not a generator. The "weights" and specific configurations that constitute the detection logic will be protected under a restricted license, preventing malicious actors from easily reverse-engineering the system to train "evasive" propaganda bots. Furthermore, to prevent mass surveillance risks, the deployment strategy prioritizes "Client-Side" or "Edge" processing. The tool is designed to empower the individual user (e.g., via a browser plugin running locally) rather than establishing a centralized server that monitors all reading habits, thereby preserving user privacy and preventing the creation of a surveillance apparatus.

Environmental Impact: Carbon Footprint of AI Training. The training of complex AI models poses a risk of significant energy consumption and carbon emissions.

- **Mitigation Strategy:** We strictly adhere to "Green AI" and "Frugal Learning" principles. Instead of training massive Large Language Models (LLMs) from scratch, which is environmentally costly, we utilize a Transfer Learning approach, fine-tuning existing open-source models (like CamemBERT or RoBERTa) which requires a fraction of the energy. Additionally, the Hybrid Architecture (integrating symbolic rules) reduces the reliance on pure brute-force computation. Finally, the heavy computational workloads (WP2) will be executed on the *Digital Research Alliance of Canada* infrastructure, which is largely powered by hydroelectricity, significantly lowering the carbon intensity compared to standard fossil-fuel-reliant data centers.

5. REFERENCES

[ATN25] Dang Anh-Hoang, Vu Tran et Le-Minh Nguyen. “Survey and analysis of hallucinations in large language models : attribution to prompting strategies or model behavior”. In : *Frontiers in Artificial Intelligence* Volume 8 - 2025 (2025). doi : 10.3389/ frai.2025.1622292.

[Che+25] Mengyang Chen, Lingwei Wei, Wei Zhou et Songlin Hu. “Structure-aware Propagation Generation with Large Language Models for Fake News Detection”. In : *Findings of the Association for Computational Linguistics : EMNLP 2025*. Suzhou, China : Association for Computational Linguistics, nov. 2025, p. 13258-13272. doi : 10.18653/v1/2025.findings-emnlp.714.

[DEI25] **Damien Deias**. « Perspectives épistémologiques de l’analyse du discours numérique : l’exemple de TikTok. » Langue française. 2025. 91-108. https://doi.org/10.3917/lf.226.0091.

[Far+25] David Farr, Lynnette Hui Xian Ng, Stephen Prochaska, Iain J. Cruickshank et Jevin West. *Simulating Misinformation Vulnerabilities With Agent Personas*. 2025. arXiv : 2511.04697[cs.SI]. url : https://arxiv.org/abs/2511.04697.

[GSV22] Zhijiang Guo, Michael Schlichtkrull et Andreas Vlachos. “A Survey on Automated Fact-Checking”. In : *Transactions of the Association for Computational Linguistics* 10 (2022). Sous la dir. de Brian Roark et Ani Nenkova, p. 178-206. doi :10.1162/tacl_a_00454.

[Ham22] Michael Hameleers. “Disinformation as a context-bound phenomenon : toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination”. In : *Communication Theory* 33.1 (oct. 2022), p. 1-10. doi : 10.1093/ ct/qtac021.

[Hyl05] Ken Hyland. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173-192. 2005. https://doi.org/10.1177/1461445605050365

[Jak+25] Lise Jakobsen, Anna Johanne Holden, Önder Gürcan et Özlem Özgöbek. *Agent-Based Exploration of Recommendation Systems in Misinformation Propagation*. 2025. arXiv : 2507.21724[cs.MA]. url : https://arxiv.org/abs/2507.21724.

[Ji+23] Ziwei Ji et al. “Survey of Hallucination in Natural Language Generation”. In : *ACM Comput. Surv.* 55.12 (mars 2023). doi :10.1145/3571730.

[Kau+25] Gagandeep Kaur, **Aznam Yacoub, Ismail Badache et Maamar Hamri**. *DeepSeek R1 à l’épreuve, au-delà des scores*. Unpublished. 2025. (Originally accepted at EGC26)

[KH24] Amandine Kervella et Nicolas Hubé. *(Ré)éduquer aux médias. L’action publique contre les "désordres informationales"*. Séminaire Fait didactique et éducation (Podcast). 2024. url : https://inspe.univ-lorraine.fr/entendu-episode-40.

[KT20] Neema Kotonya et Francesca Toni. “Explainable Automated Fact-Checking : A Survey”. In : *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online) : International Committee on Computational Linguistics, déc. 2020, p. 5430-5443. doi : 10.18653/v1/2020.coling-main.474.

[MAI25] Dominique Maingueneau. Les régimes discursifs. In : Langage et société, 186(3), 153-172. 2025. https://shs-cairn-info.lama.univ-amu.fr/revue-langage-et-societe-2025-3-page-153?lang=fr .

[MEN26] Ministère de l’Éducation nationale. *Feuilles de route de l’Éducation nationale*. 2026. URL : https://www.education.gouv.fr/feuilles-de-route-450426 (visité le 05/02/2026).

[Nan+25] Arghodeep Nandi, Megha Sundriyal, Euna Mehnaz Khan, Jikai Sun, Emily K. Vraga, Jaideep Srivastava et Tanmoy Chakraborty. “The Psychology of Falsehood : A Human-Centric Survey of Misinformation Detection”. In : *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China : Association for Computational Linguistics, nov. 2025, p. 8519-8536. doi : 10.18653/v1/2025.emnlp-main.428.

[Noë24] Elisabeth Noël. “Désinformation : Former les professionnels!” In : *Balisages* 7 (fév. 2024). doi : 10.35562/balisages.1244.

[PR21] Gordon Pennycook et David G. Rand. “The Psychology of Fake News”. In : *Trends in Cognitive Sciences* 25.5 (2021), p. 388-402. doi : 10.1016/j.tics.2021.02.007.

[Quy+25] Thanh Quy Le, **Ismail Badache, Aznam Yacoub et Maamar el-amine Hamri**. “LIS at CheckThat! 2025: multi-stage open-source large language models for fact-checking numerical claims”. In : *Notebook for the CheckThat! Lab at the 16th Conference and Labs of the Evaluation Forum*. Vol. 4038. CLEF 2025. 2025. hal-id: HAL Id : hal-05296845.

[Ras+17] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova et Yejin Choi. “Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking”. In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark. 2017. doi: 10.18653/v1/D17-1317.

[RCR25] Raquel Rodríguez-García, Roberto Centeno et Álvaro Rodrigo. “Simulating misinformation diffusion on social media through CoNVaI : a textual- and agent-based diffusion model”. In : *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. IJCAI ’25. Montreal, Canada, 2025. isbn : 978-1-956792-06-5. doi : 10.24963/ijcai.2025/29.

[RDJ13] Marta Recasens, Cristian Danescu-Niculescu-Mizil et Dan Jurafsky. "Linguistic Models for Analyzing and Detecting Biased Language". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. 2013.

[SWL19] Kai Shu, Suhang Wang et Huan Liu. “Beyond News Contents : The Role of Social Context for Fake News Detection”. In : *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM ’19. Melbourne VIC, Australia : Association for Computing Machinery, 2019, p. 312-320. doi : 10.1145/3289600.3290994.

[Srb+25] Ivan Srba et al. “A Survey on Automatic Credibility Assessment Using Textual Credibility Signals in the Era of Large Language Models”. In : *ACM Transactions on Intelligent Systems and Technology* (sept. 2025). doi : 10.1145/3770077.

[Tho+18] James Thorne, Andreas Vlachos, Christos Christodoulopoulos et Arpit Mittal. “FEVER : a Large-scale Dataset for Fact Extraction and VERification”. In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, juin 2018, p. 809-819. doi : 10.18653/v1/N18-1074.

[VIO11] Robert Vion. « La modalisation. Un mode paradoxal de prise en charge. » *in* P. Dendale et D. Coltier La prise en charge énonciative : Études théoriques et empiriques. 2011. 75-91. De Boeck Supérieur. https://doi.org/10.3917/dbu.denda.2011.01.0075.

[Zho+24] Kyrie Zhixuan Zhou, Abhinav Choudhry, Ece Gumusel et Madelyn Rose Sanfilippo. *"Sora is Incredible and Scary" : Emerging Governance Challenges of Text-to-Video Generative AI Models*. 2024. arXiv : 2406.11859 [cs.CY]. url : https://arxiv.org/abs/2406.11859.

[ZZ20] Xinyi Zhou et Reza Zafarani. “A Survey of Fake News : Fundamental Theories, Detection Methods, and Opportunities”. In : Acm. Comput. Srv. 53, 5, Article 109 (sept. 2020). doi : 10.1145/3395046.