

Proposition de sujet de thèse – AAP ENACT

Titre

L'IA exégète : effets des systèmes d'IA générative sur les pratiques et les discours religieux contemporains

Mots-clés

IA générative, autorité, religion, interprétation, normativité, sociologie du numérique

Problématique

La thèse analyse les effets des systèmes d'IA générative sur les pratiques et les discours religieux contemporains, en s'appuyant principalement sur l'analyse de systèmes de type grands modèles de langage (LLM), au cœur de l'axe « Traitement automatique des langues et LLM multimodaux » privilégié par le programme ENACT. Elle s'inscrit dans une sociologie des médiations numériques et des autorités symboliques, en dialogue avec les *Science and Technology Studies* (STS) et la sociologie du religieux.

Les systèmes d'IA générative produisent aujourd'hui des discours interprétatifs et normatifs (explications, conseils, jugements), traditionnellement réservés à des autorités humaines. Lorsqu'ils sont mobilisés dans des contextes religieux (interprétation de textes, réponses à des questions morales ou spirituelles, médiation pédagogique, prédication), ces systèmes peuvent reconfigurer les régimes d'autorité, de légitimité et de confiance.

À partir d'un ensemble de cas exemplaires à définir mais représentatifs des principales traditions religieuses (christianismes, judaïsme, islam) présentes en France (des ouvertures comparatives sont envisageables à moyen terme), la thèse analysera comment, à quelles conditions et avec quels effets les systèmes d'IA générative sont perçus, utilisés ou rejetés comme acteurs interprétatifs dans des pratiques religieuses contemporaines (Plaza-del-Arco, et al, 2024). Plutôt que de poser la question normative de la légitimité de l'IA à interpréter le religieux, l'approche attendue conduira à interroger les processus concrets par lesquels, selon les dispositifs, les situations et les usages, l'IA devient de facto légitime ou illégitime. On s'interroge sur les biais liés aux représentations des religions sur le web qui sont présents dans les données utilisées par les LLMs et qui peuvent déformer les discours (Abid et al, 2021, Shankar et al, 2025, Seth et al, 2025) et designer des prompts pour éviter ces biais (Gillespie, 2024). Cette perspective permet d'analyser les conditions d'acceptabilité, les résistances, les controverses et les ajustements pratiques associés à ces usages.

Objectifs scientifiques

- Identifier les usages effectifs de l'IA générative dans différents contextes religieux.
- Analyser les modes de légitimation, de délégation ou de rejet de l'IA comme source d'interprétation, tant du point de vue des autorités religieuses que des usagers.
- Analyser les représentations biaisées des religions dans les textes générés par l'IA
- Étudier les transformations discursives et pratiques induites par ces usages.
- Clarifier les limites symboliques, institutionnelles et éthiques assignées à l'IA.

Corpus et terrain

- Constitution expérimentale d'un double corpus de productions d'IA générative et de productions humaines (réponses à des questions religieuses ou morales), suivie de focus groups portant sur les différences identifiées auprès de publics variés.
- Analyse de discours institutionnels et publics d'acteurs religieux portant sur l'IA.
- Enquêtes qualitatives ciblées (entretiens, observations) auprès d'acteurs religieux, de médiateurs numériques et d'utilisateurs sur les usages des IA génératives en contexte religieux.

Le périmètre empirique sera volontairement circonscrit afin de garantir une analyse approfondie et comparative.

Méthodologie

- Dialogue avec la linguistique et la sociolinguistique pour une analyse comparative de discours (productions humaines / productions d'IA).
- Dialogue méthodologique avec les sciences de l'IA, portant notamment sur le fonctionnement des modèles génératifs, leurs biais (hallucinations, alignement normatif) et leurs limites épistémiques et opérationnelles.
- Sociologie des usages et de la réception.
- Analyse des controverses et des cadres normatifs.

Apports scientifiques attendus

- Conceptualisation de l'IA comme forme d'autorité interprétative non humaine, prise dans des chaînes d'usages humains.
- Contribution empirique originale aux études contemporaines sur le religieux, notamment sur les régimes exégétiques en jeu dans l'interaction humains/IA.
- Apport aux débats sur la gouvernance éthique et sociale de l'IA.
- Production de résultats transférables aux réflexions institutionnelles au-delà des seules institutions religieuses.

Positionnement ENACT et financement

- **Thématique** : IA & Sciences humaines et sociales – sujet émergent à fort enjeu sociétal.
- **Interdisciplinarité** : SHS avec interaction explicite avec les sciences de l'IA et l'éthique.
- **Financement** : co-financement ENACT / ReligiS envisagé ; également compatible avec un financement complet, conformément aux priorités ENACT pour les SHS.
- **Assises institutionnelles** : Le projet s'inscrit en cohérence avec les axes scientifiques du Work Package 4.2 (axe1) du programme ReligiS, sur les recompositions contemporaines du religieux à l'ère numérique. Le projet pourra s'adosser à des infrastructures de recherche labellisées, notamment le Bnu_Lab (labellisé CoRTecS), ainsi qu'à des plateformes de calcul et d'ingénierie en IA telles que la plateforme GAIA, disposant d'ingénieurs spécialisés en LLM, afin de soutenir la constitution des corpus, les expérimentations et l'analyse des productions d'IA générative.

Directeurs de thèse : Amalia Todirascu (LiLPa, Université de Strasbourg, todiras@unistra.fr), Anne Jeannin-Girardon (iCube, Université de Strasbourg, jeanningirardon@unistra.fr), Josselin Tricou (Université de Lausanne, josselin.tricou@unil.ch)

Profil du candidat

- être diplômé d'un master dans un des domaines de sciences du langage, sciences sociales, sciences de données ou Traitement Automatique du Langage
- excellents résultats académiques
- maîtrise des méthodes statistiques et des outils d'analyse de discours
- maîtriser des méthodes de collecte de données textuelles et des notions de corpus
- avoir des connaissances/compétences dans le domaine du TAL
- connaître le fonctionnement d'un LLM et de l'ingénierie de prompts

Bibliographie

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. <https://doi.org/10.1145/3461702.3462624>

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in largelanguage models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2): 1–21

Flor Miriam Plaza-del-Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Gillespie, T. 2024. Generative AI and the politics of visibil-ity. *Big Data & Society*, 11(2): 20539517241252131.

Hari Shankar, Vedanta S P, Tejas Cavale, Ponnurangam Kumaraguru, Abhijnan Chakraborty (2025) Sometimes the Model doth preach: Quantifying Religious Bias in Open LLMs through Demographic Analysis in Asian Nations. arXiv:2503.07510v1 [cs.CY] 10 Mar 2025

Seth, A., Choudhury, M., Sitaram, S., Toyama, K., Vashistha, A., & Bali, K. (2025). How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3), 2319-2330. <https://doi.org/10.1609/aies.v8i3.36718>