Digital Systems for Humans Graduate School

2024-2025 PhD Subject Proposition

Proposition de Sujet de Thèse 2024-2025

AI AS AN ECONOMIC AGENT: AN EXPERIMENTAL APPROACH TO THE STUDY OF ECONOMIC PREFERENCES AND INTERRACTIONS BETWEEN HUMAN AND ARTIFICIAL INTELLIGENCE

Doctoral School: Doctoral School in Law, Political Science, Economics and Management (ED DESPEG)

Thesis supervisor: Agnès Festré et Guilhem Lecouteux (co-encadrant)

Host laboratory: GREDEG

Subject description:

L'Intelligence Artificielle (IA) est une révolution technologique qui bouleverse notre compréhension et notre interaction avec le monde qui nous entoure. Au fil du temps, les IA ont évolué et sont devenus progressivement plus efficaces et capables d'effectuer de plus en plus de tâches. Ces avancées ont conduit à une intégration croissante des IA dans notre société. Il existe deux principaux types d'IA, les IA performatives qui prennent des décisions de manière autonome, et les IA conseillers qui influencent l'humain dans la prise de décision (Candrian & Scherer, 2022).

L'empreinte des IA dans la prise de décision, impactant les activités économiques, est en adéquation avec la définition même de ce qu'est un agent économique. En effet, l'intégration croissante des intelligences artificielles dans la prise de décision économique reflète leur statut d'agent économique. En outre, comme les agents humains, les IA sont des optimisateurs, cherchant à maximiser les performances dans diverses tâches. Étudier les IA en tant qu'agents économiques permet de comprendre leur impact sur l'économie. Cette thèse défend l'idée qu'il devient légitime d'utiliser des méthodes traditionnellement utilisées pour l'étude des comportements humains (en substance, pour ce projet, l'économie expérimentale) afin d'étudier les propriétés émergentes des IA.

Ce projet de thèse se concentre sur les IA Large Language Models (IA-LLM), qui grâce à des progrès récents dans le traitement du langage naturel, ont donné naissance à des modèles de langage autorégressifs, à l'instar du Generative Pre-trained Transformers (GPT) de l'entreprise OpenAI. Ces modèles connaissent une adoption massive, surtout les IA-LLM ChatGPT-3.5 et ChatGPT-4. Les IA-LLM, sont également capables de remplacer l'homme dans des tâches de production intellectuelle (rédaction de textes créatifs, traduction, génération de codes de programmation, etc.). De plus, bien qu'elles puissent agir en tant que conseillers en répondant aux requêtes posées par les individus, elles peuvent également être intégrées dans des systèmes autonomes, leur permettant, dès lors, d'agir en tant que preneurs de décisions à part entière. Le très grand nombre de paramètres de ces IA-LLM et l'utilisation du deep learning, en font des IA très complexes dont les propriétés émergentes commencent à être étudiées en sciences sociales. En effet, une littérature émergente fait état des biais cognitifs retrouvés dans les IA-LLM (Chen and al. 2023 ; Talboy and *al*. 2023) et cherche à déterminer si les réponses AI-LLM sont plus intuitives ou réflexives, en référence au systèmes 1 et 2 de Kahneman (Palminteri et *al*. 2023 ; Hagendorff et *al*. 2023).

L’objectif de ce projet de thèse réside en une étude sous un angle nouveau des IA-LLM, en les abordant comme des agents économiques à part entière. Cette approche sera guidée par l'utilisation de la méthode de l’économie expérimentale.

Quatre temps sont envisagés.

Dans un premier temps, il est prévu d’étudier les préférences économiques que les conseils et décisions de ces agents peuvent refléter. La science économique suppose que les décisions des agents peuvent être guidées par leurs préférences économiques, notamment les préférences sociales et à l’égard du risque.

Dans un deuxième et troisième temps, il est envisagé d’étudier comment les conseils des IA-LLM influencent l’expression des préférences humaines, tant dans les décisions individuelles que dans les interactions stratégiques entre humains. Une littérature récente met en lumière un phénomène d'excès de confiance des individus dans les conseils des IA, désigné comme un biais d'automatisation. Ce biais se caractérise par une surestimation des performances et de la précision constante de l'IA, conduisant à une idéalisation de ses capacités. (Cummings, 2017 ; Lee, 2018). Ces études soulignent le risque que les individus soient influencés par les IA.

Enfin, dans un quatrième temps, il s’agit d’analyser les interactions stratégiques entre humains et IA-LLM afin de comprendre leur dynamique coopérative. La collaboration IA-humain est inévitable, et se produit déjà dans de nombreux domaines. Il est donc important d’étudier la dynamique de coopération de ces agents économiques. D’autant plus qu’une littérature sur l’aversion aux algorithmes (Dietvorst et al., 2015) sous-tend que les individus préfèrent les humains aux IA même quand elles sont plus performantes (Alvarado-Valencia & Barrero, 2014 ; Bucher, 2017 ; Dietvorst et *al*., 2015).

Ce projet de thèse marque une avancée méthodologique majeure en ouvrant la voie à une approche interdisciplinaire pour étudier les IA. En mobilisant les méthodes expérimentales traditionnellement utilisées pour étudier le comportement humain en économie, ce projet a pour ambition d’enrichir la compréhension des IA et ouvre de nouvelles perspectives de recherche. Bien que certains puissent considérer cette approche prématurée étant donné l'état actuel des IA, leur développement rapide suggère que de telles méthodes seront inévitablement nécessaires à l'avenir. Ainsi, anticiper ces besoins dès aujourd'hui est essentiel. En contribuant à l'établissement de normes méthodologiques et de bonnes pratiques, ce projet prépare le terrain pour une étude plus approfondie des IA en économie. En adoptant une perspective à long terme, ce projet offre non seulement des réponses immédiates, mais jette également les bases d'une approche méthodologique innovante dans ce domaine en pleine expansion.

References:

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. Computers in Human Behavior, 36, 102–113. https://doi.org/ 10.1016/j.chb.2014.03.047

Bucher, T. (2019). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. In The social power of algorithms (pp. 30-44). Routledge.

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, *134*, 107308.

Chen, Y., Andiappan, M., Jenkin, T., & Ovchinnikov, A. (2023). "A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?". *Working Paper. Available at SSRN 4380365*

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In Decision making in aviation (pp. 289-294). Routledge.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 1-6

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society. 5(1). https://doi.org/10.1177/2053951718756684

Palminteri, S., Yax, N., & Anllo, H. (2023). Studying and improving reasoning in humans and machines.

Talboy, A. N., & Fuller, E. (2023). "Challenging the appearance of machine intelligence: Cognitive bias in LLMs". *arXiv preprint arXiv:2304.01358*.