Architecture modulaire pour l'alignement des systèmes d'intelligence artificielle : découplage fonctionnel du langage, de la décision et de la modélisation d'autrui afin d'assister la collaboration entre humains

Nom(s) du/des encadrant(s) :	Catherine Pelachaud, Mehdi Khamassi
Laboratoire d'accueil :	Institut des Systèmes Intelligents et de Robotique (ISIR), Paris
Court résumé :	Ce projet de thèse propose de concevoir une architecture modulaire pour l'intelligence artificielle visant à améliorer l'alignement et l'explicabilité des systèmes dans des contextes de collaboration humaine. Contrairement aux approches monolithiques actuelles, l'hypothèse centrale est qu'un découplage fonctionnel entre traitement du langage, prise de décision et modélisation d'autrui (Théorie de l'Esprit) permet de mieux reproduire la structure des jugements humains et d'expliciter les sources de désaccords. L'objectif est de développer un agent cognitif capable d'agir comme médiateur explicite des interactions humaines, en clarifiant les intentions, les rôles et les arbitrages normatifs au sein de dyades ou de groupes. Le projet s'appuie sur des travaux préliminaires combinant une évaluation normative model-based (COMETH) et la modélisation des croyances multiples dans des contextes sociaux incertains (jeu du Loup-Garou), afin de construire un agent dédié à la facilitation de la collaboration humaine.
Brève description du groupe de recherche / laboratoire d'accueil :	L'Institut des Systèmes Intelligents et de Robotique (ISIR), situé sur le campus de Jussieu à Paris, est un laboratoire pluridisciplinaire spécialisé en robotique, intelligence artificielle et sciences cognitives. Ses recherches couvrent notamment l'interaction humain-machine, la robotique cognitive et sociale, ainsi que la modélisation computationnelle du comportement. L'ISIR offre un environnement de recherche propice à l'étude de systèmes intelligents capables d'interagir avec des humains dans des contextes complexes et dynamiques.

1 Contexte

L'émergence des grands modèles de langage a marqué un tournant majeur en intelligence artificielle. Leur capacité à traiter le langage à un niveau quasi-humain repose toutefois sur des architectures monolithiques de type end-to-end, dans lesquelles les fonctions de représentation, de raisonnement et de décision sont fusionnées. Si cette intégration s'avère efficace sur le plan computationnel, elle engendre une opacité structurelle qui complique l'alignement des systèmes, entendu comme la convergence entre leur comportement et les attentes humaines. **Ceci limite l'utilité de ces modèles dans des contextes de collaboration où l'explicabilité est cruciale** [Gab20] [Key25].

Cette difficulté est particulièrement manifeste dans le domaine des jugements moraux. Les systèmes actuels rencontrent des limites lorsqu'ils doivent traiter des situations ambiguës impliquant simultanément intentions, conséquences et normes concurrentes [Sch+23][KNC24]. Or les attentes humaines en matière de morale ne se réduisent pas à l'optimisation d'une fonction de récompense unique : elles reposent sur des régularités cognitives et comportementales complexes, issues de l'interaction entre notre compréhension du langage, nos mécanismes de décision et notre cognition sociale.

Dans un contexte d'autonomie croissante des systèmes artificiels, il devient nécessaire de **concevoir des architectures permettant non seulement de produire des décisions, mais aussi d'en expliciter les fondements dans des situations de collaboration humaine**.

2 Problématique et objectif

Problématique : Les architectures actuelles d'intelligence artificielle, en raison de leur caractère monolithique, rendent difficile l'explicitation des processus internes de décision, ce qui limite leur capacité à s'intégrer dans des contextes de collaboration humaine impliquant des désaccords, des normes multiples et des rôles différenciés.

Hypothèse centrale : Une architecture modulaire opérant une séparation fonctionnelle entre le langage, la prise de décision et la modélisation d'autrui (ToM) permettrait de mieux reproduire la structure des jugements humains et d'expliciter les intentions et rôles des agents impliqués.

Objectif : Concevoir un agent intelligent capable d'agir comme un **médiateur explicite de la collaboration humaine**, en clarifiant les états mentaux et les arbitrages normatifs au sein d'un groupe.

3 Bref état de l'art

Les architectures actuelles de type end-to-end dominent le paysage de l'IA, mais présentent des limites importantes en termes d'explicabilité et d'alignement. Des approches récentes proposent de réintroduire des formes de modularité inspirées des sciences cognitives [LeC22][Mah+24].

Par ailleurs, les approches d'alignement reposent majoritairement sur l'optimisation de fonctions de récompense, ce qui reste insuffisant pour capturer la complexité des jugements humains, notamment dans les situations impliquant des normes concurrentes [Hay+22].

Enfin, les travaux en Théorie de l'Esprit computationnelle ont permis de modéliser les croyances d'autrui dans des contextes contrôlés [Rab+18], mais leur intégration dans des systèmes visant à faciliter la collaboration humaine reste encore limitée.

4 Questions de recherche

- Dans quelle mesure une architecture modulaire permet-elle d'améliorer l'explicabilité et l'alignement des systèmes d'IA dans des contextes sociaux ?
- Comment modéliser et expliciter des points de vue multiples dans des situations de désaccord ?
- Comment un système peut-il identifier et **expliciter les sources de désaccord** au sein d'une interaction ou d'un groupe ?
- Quel est l'impact d'un tel système sur la qualité de la collaboration humaine ?

5 Fondements théoriques

Ce projet s'inscrit dans une approche interdisciplinaire à l'interface de la science cognitive, de la psychologie morale et de la cognition sociale.

La psychologie morale a mis en évidence des régularités robustes dans la manière dont les individus arbitrent entre intentions, conséquences et normes. Les neurosciences sociales ont montré que ces jugements mobilisent conjointement des mécanismes décisionnels et des processus de cognition sociale [Gre+04].

Par ailleurs, les architectures cognitives [RTO18] [Ala+98] [Cha+18] fournissent un cadre théorique pour modéliser des systèmes reposant sur des modules spécialisés.

L'hypothèse sous-jacente est que les jugements humains émergent de l'interaction de processus spécialisés, et que leur explicitation nécessite de préserver cette structure fonctionnelle.

6 Approche et méthodes

Travaux préliminaires Ce projet s'inscrit dans la continuité directe de deux lignes de recherche préalablement initiées.

D'une part, le projet COMETH, développé lors d'un stage de master, a permis de mettre en œuvre un prototype computationnel modulaire comprenant une abstraction sémantique, une décision *model-based* et une évaluation normative explicite. Testé sur un corpus de dilemmes moraux, ce système a fourni des résultats exploratoires encourageants [Mor+25] et a démontré la viabilité du découplage fonctionnel proposé.

D'autre part, un projet de recherche actuellement mené à l'ISIR porte sur la modélisation des croyances de joueurs humains dans le jeu de déduction sociale du Loup-Garou. Ce travail

vise à formaliser des mécanismes avancés de Théorie de l'Esprit pour inférer les croyances multiples sous incertitude d'un groupe d'humains collaborant à une tâche commune (ici jouer à un jeu de société).

La thèse vise à unifier ces deux approches en combinant robustesse de l'évaluation normative et modélisation dynamique des croyances collectives, afin de construire un agent cognitif complet dédié à la facilitation de la collaboration humaine.

Le projet propose une architecture modulaire séparant explicitement trois fonctions :

- **Langage :** extraction de représentations sémantiques du contexte
- **Décision :** arbitrage entre dimensions multiples (sociales, normatives, instrumentales)
- **Modélisation d'autrui :** inférence des croyances et des points de vue

6.1 Médiation dyadique

Dans un premier temps, l'étude se concentrera sur des scénarios conflictuels exprimés sous forme textuelle, en étendant les bases du projet COMETH. L'objectif sera d'élargir l'analyse aux normes sociales et professionnelles régissant la collaboration.

Un module de Théorie de l'Esprit permettra de faire varier le point de vue et d'expliciter les représentations propres à chaque acteur. Le système pourra ainsi fournir une explication croisée visant à restaurer la compréhension mutuelle.

6.2 Dynamique de groupe

Dans un second temps, le projet étendra cette modélisation à des configurations multi-agents. L'agent devra inférer les croyances croisées et les asymétries d'information au sein du groupe.

Cette étape s'appuiera sur les travaux menés sur le jeu du Loup-Garou, en tant que paradigme de cognition sociale en contexte d'incertitude [Lai+23].

7 Évaluation des contributions

L'évaluation reposera sur deux niveaux complémentaires.

Évaluation computationnelle :

- comparaison avec des données issues de la psychologie morale et sociale
- analyse de la cohérence des inférences produites

Évaluation expérimentale :

- interactions entre humains assistés par le système
- mesure de métriques collaboratives :
  - temps de résolution des conflits
  - confiance interpersonnelle
  - sentiment d'agentivité

8 Nature de la collaboration numérique

Le projet s'inscrit dans un cadre de collaboration humaine médiée par un système d'intelligence artificielle jouant un rôle actif dans la coordination des interactions.

- **Fonction :** communication, coordination, médiation
- **Type :** synchrone et asynchrone
- **Échelle de temps :** de l'interaction en temps réel (secondes à minutes) à la collaboration étendue (heures à semaines, voire projets de plus long terme)
- **Taille du groupe :** dyades à petits groupes
- **Espace :** contextes hybrides (présentiel et distant), dans lesquels le système agit comme médiateur des interactions

Le système agit comme un médiateur explicite capable de clarifier les intentions, expliciter les désaccords et faciliter la compréhension mutuelle.

9 Contribution à la collaboration numérique : résultats attendus et impact

Le projet vise plusieurs types de contributions :

- **Théoriques :** formalisation d'une architecture modulaire pour la collaboration
- **Méthodologiques :** protocoles d'évaluation de la collaboration assistée par IA
- **Techniques :** développement d'un agent cognitif modulaire
- **Empiriques :** validation expérimentale sur des interactions humaines

Les impacts attendus concernent l'amélioration de la collaboration homme-machine, le développement de systèmes explicables et la préservation de l'agentivité humaine.

10 Positionnement dans le programme eNSEMBLE

Le projet contribue directement à deux axes du programme :

- **PC3 – Collaboration humaine :** facilitation de la coordination et de la compréhension mutuelle
- **PC5 – Alignement et agentivité :** explicitation des processus décisionnels et soutien à la délibération humaine

References

[Ala+98] R. Alami et al. "An Architecture for Autonomy". In: *The International Journal of Robotics Research* 17.4 (Apr. 1998), pp. 315–337. ISSN: 0278-3649, 1741-3176. DOI: 10.1177/027836499801700402. URL: https://journals.sagepub.com/doi/10.1177/027836499801700402 (visited on 04/23/2026).

[Cha+18] Raja Chatila et al. "Toward Self-Aware Robots". In: *Frontiers in Robotics and AI* Volume 5 - 2018 (2018). ISSN: 2296-9144. DOI: 10.3389/frobt.2018.00088. URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2018.00088.

[Gab20] Iason Gabriel. "Artificial Intelligence, Values and Alignment". In: *Minds and Machines* 30.3 (Sept. 2020), pp. 411–437. ISSN: 0924-6495, 1572-8641. DOI: 10.1007/s11023-020-09539-2. arXiv: 2001.09768 [cs]. URL: http://arxiv.org/abs/2001.09768.

[Gre+04] Joshua D. Greene et al. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment". In: *Neuron* 44.2 (Oct. 2004), pp. 389–400. ISSN: 08966273. DOI: 10.1016/j.neuron.2004.09.027. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627304006348.

[Hay+22] Conor F. Hayes et al. "A Practical Guide to Multi-Objective Reinforcement Learning and Planning". In: *Autonomous Agents and Multi-Agent Systems* 36.1 (Apr. 2022), p. 26. ISSN: 1387-2532, 1573-7454. DOI: 10.1007/s10458-022-09552-y. arXiv: 2103.09568 [cs]. URL: http://arxiv.org/abs/2103.09568.

[Key25] Hu Keyi. *Beyond the Black Box: A Cognitive Architecture for Explainable and Aligned AI*. Dec. 8, 2025. DOI: 10.48550/arXiv.2512.03072. arXiv: 2512.03072 [cs]. URL: http://arxiv.org/abs/2512.03072. Pre-published.

[KNC24] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. "Strong and Weak Alignment of Large Language Models with Human Values". In: *Scientific Reports* 14.1 (Aug. 21, 2024), p. 19399. ISSN: 2045-2322. DOI: 10.1038/s41598-024-70031-3. URL: https://www.nature.com/articles/s41598-024-70031-3.

[Lai+23] Bolin Lai et al. "Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Findings 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 6570–6588. DOI: 10.18653/v1/2023.findings-acl.411. URL: https://aclanthology.org/2023.findings-acl.411/ (visited on 01/27/2026).

[LeC22] Yann LeCun. "A Path Towards Autonomous Machine Intelligence Version 0.9.2," in: (June 27, 2022).

[Mah+24] Kyle Mahowald et al. *Dissociating Language and Thought in Large Language Models*. Mar. 23, 2024. DOI: 10.48550/arXiv.2301.06627. arXiv: 2301.06627 [cs]. URL: http://arxiv.org/abs/2301.06627. Pre-published.

[Mor+25] Geoffroy Morlat et al. *Morality Is Contextual: Learning Interpretable Moral Contexts from Human Data with Probabilistic Clustering and Large Language Models*. Dec. 24, 2025. DOI: 10.48550/arXiv.2512.21439. arXiv: 2512.21439 [cs]. URL: http://arxiv.org/abs/2512.21439. Pre-published.

[Rab+18] Neil C. Rabinowitz et al. *Machine Theory of Mind*. Mar. 12, 2018. DOI: 10.48550/arXiv.1802.07740. arXiv: 1802.07740 [cs]. URL: http://arxiv.org/abs/1802.07740. Pre-published.

[RTO18] Frank Ritter, Farnaz Tehranchi, and Jacob Oury. "ACT-R: A Cognitive Architecture for Modeling Cognition". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 10 (Dec. 7, 2018), e1488. DOI: 10.1002/wcs.1488.

[Sch+23] Nino Scherrer et al. *Evaluating the Moral Beliefs Encoded in LLMs*. July 26, 2023. DOI: 10.48550/arXiv.2307.14324. arXiv: 2307.14324 [cs]. URL: http://arxiv.org/abs/2307.14324. Pre-published.