**HUCAI: Human Collaboration enhanced by AI: How to ensure AI alignment with the human goals, preferences and values?**

**Noms des conseillers doctoraux :** Mehdi Khamassi (1), Cédric Paternotte (2), Raja Chatila (1)
**Laboratoires d'accueil :** (1) Équipe ACIDE, Institut des Systèmes Intelligents et de Robotique (ISIR), UMR 7222 ; (2) Sciences, Normes, Démocratie (SND), UMR 8011, CNRS/Sorbonne U.
**Nom du/de la candidat/e :** Marceau Nahon
**Genre :** M
**Institution qui sera l'employeur du doctorant (partenaire PC5 et PC3) :** Sorbonne U.

# Description du projet de doctorat (3 pages max)

Contexte (et scénarios éventuels). Le contexte envisagé est celui de l'évaluation et de l'amélioration de l'alignement entre des systèmes à base d'intelligence artificielle (IA) et les objectifs, préférences et valeurs d'humains effectuant des tâches collaboratives.

Deux scénarios particuliers sont envisagés ici :

- Une situation de co-manipulation d'objets dans un espace collaboratif impliquant plusieurs humains et un robot (e.g., rangement d'objets dans un atelier), et la nécessité pour le robot de chercher à la fois l'efficacité mais aussi un alignement avec les humains. Le robot doit ainsi conseiller les humains pour éviter tout comportement qui pourrait s'avérer dangereux (e.g., bousculades, empilements instables d'objets), proposer une répartition équilibrée des tâches entre les humains, un juste dosage dans ses interventions (ne pas nuire à l'agentivité des humains), éviter les biais (e.g., biais de genre ou autre dans la répartition des tâches).
- Un scénario encore ouvert à ce stade, et idéalement à faire coller à un scénario déjà existant dans le PEPR, pour pouvoir favoriser les synergies avec d'autres partenaires. Cela pourrait être une situation d'assistant à base d'IA en milieu hospitalier. L'agent doit collaborer avec le personnel médical pour optimiser la gestion des soins, tout en s'assurant d'une bonne répartition des tâches entre le personnel, gestion des données médicales, transparence dans la justification et l'explication des recommandations et actions proposées. L'agent à base d'IA doit faire attention à ne pas réduire l'agentivité des agents humains, et doit veiller à favoriser plutôt qu'entraver la tâche collaborative.

Problématique et objectif. La problématique générale abordée dans cette thèse est celle de l'évaluation et l'amélioration par apprentissage de l'alignement entre un système d'IA et des humains en situation de collaboration. L'alignement doit se faire avec les objectifs (l'objectif du système est-il le même que celui des humains ? L'objectif de la tâche collaborative et les intentions des humains qui collaborent sont-ils bien identifiés par le système d'IA pour qu'il les aide plutôt que les entraver ?), les préférences, les valeurs (le système respecte-t-il les préférences et valeurs des humains dans le contexte de la tâche ?). L'objectif est de mettre au point une méthode permettant d'évaluer et d'améliorer cet alignement.
Le travail de thèse a pour but de :

- développer des modèles et métriques permettant de mesurer :
    - l'alignement entre les décisions prises par un système à base d'IA et les objectifs définis par les humains dans des contextes collaboratifs.
    - le respect des préférences et valeurs individuelles et collectives (e.g., conventions, valeurs éthiques/morales) des humains impliqués dans la collaboration.
- développer des modèles à base d'apprentissage par renforcement visant à améliorer l'alignement des systèmes d'IA intégrés dans la collaboration.
- Faire une analyse philosophique du problème pour vérifier que les termes employés (collaboration, alignement, valeurs humaines, etc.) sont bien définis et formalisés dans le système de manière cohérente avec les théories philosophiques.

- définir des recommandations pour l'intégration responsable et éthique de l'IA en contexte collaboratif.

**Bref état de l'art.** Dans les débats philosophiques actuels concernant les impacts de l'IA, on trouve la discussion de ce qui signifierait une *collaboration* entre humains et agents artificiels (Evans et al., 2023). On trouve aussi tout un débat autour d'influences et de biais sur la psychologie humaine que peuvent introduire les systèmes d'IA, comme le biais d'anthropomorphisation et d'attribution de personnalité (Araujo, 2018 ; Haring et al., 2018 ; Korteling et al., 2021), le biais d'automaticité qui fait que les humains auront tendance à davantage faire confiance aux algorithmes et à moins vérifier leurs recommandations (Cummings, 2017), les questions autour de la valeur épistémique et de la véracité des informations produites par les systèmes d'IA (Huneman, 2024), et le problème de l'alignement avec les valeurs humaines (Scherrer et al., 2023), soulignant la difficulté pour des systèmes à base d'IA d'identifier certaines conséquences éthiques négatives de leurs actions/recommandations pour les humains, telles que les impacts sur la dignité (**Khamassi, Nahon, Chatila, soumis**). Les débats portent aussi sur la possibilité ou non de ces systèmes d'IA de contribuer eux-mêmes à estimer les impacts des actions sur la tâche (Khamassi, Nahon, Chatila, soumis), et en particulier les actions conjointes lors d'une situation de collaboration, qui sont plus difficiles à caractériser car elles impliquent une identification du but conjoint fixé à la tâche, et donc des intentions des humains impliqués dans la collaboration (Khamassi et al., 2016).

Il existe un ensemble de travaux computationnels dans le domaine du *machine ethics* pour tenter de formaliser et d'estimer les impacts éthiques des actions dans différents contextes. Les méthodes actuelles dans ce domaine se focalisent sur les réactions humaines et les conséquences immédiates des actions (e.g., Bonnefon. Shariff & Rahwan 2016), en outre avec des modèles prédéfinis sans estimation des incertitudes (Berreby et al., 2015 ; Han et al., 2021), ne permettant donc pas l'apprentissage d'un modèle statistique des conséquences à long-terme de séquences d'actions.

D'autres méthodes provenant du domaine de l'ingénierie existent pour la planification de séquences d'actions lors de tâches de co-manipulation nécessitant une interaction humain-robot avec un but commun et un plan d'actions conjoint (donc des situations de collaboration au sens du psychologue Michael Tomasello). Mais ces méthodes n'ont été que très récemment étendues pour estimer l'incertitude sur les effets des actions (You et al., 2023). De plus, elles requièrent toujours un modèle prédéfini et ne font pas de lien explicite avec des principes éthiques ni avec des normes.

La question de l'alignement entre un agent à base d'IA et un humain dans une tâche collaborative n'a que récemment commencé à être étudiée en s'intéressant à la capacité d'un système à aider des humains dans des activités ménagères (Puig et al, 2021). Ce travail a été approfondi (Ying et al, 2024) avec la proposition d'un cadre d'alignement mental orienté vers les objectifs (Goal-oriented Mental Alignment framework). Cependant, le système n'infère que le but de l'humain et ne s'intéresse pas à ses valeurs ni préférences, ni aux implications quant aux manières de l'aider à accomplir son but.

Questions de recherche. Comment mesurer et évaluer l'alignement entre les actions d'un système à base d'IA et les objectifs, préférences et valeurs individuelles et collectives des humains dans une situation de collaboration ? Comment capturer les préférences individuelles et collectives des humains dans un

contexte collaboratif ? Quelles sont les implications d'un mauvais alignement entre un système à base d'IA et des humains ? Comment aligner les actions d'un système à base d'IA avec les objectifs, préférences et valeurs individuelles et collectives des humains dans un situation de collaboration ?

**Fondements théoriques. Collaboration entre humains au sens du psychologue Michael Tomasello (op. cit.) et des théories philosophiques sur l'action conjointe (Paternotte, 2020). Problématique de l'alignement de l'IA avec les valeurs humaines (Gabriel, 2020, Minds and Machines). Théorie de l'apprentissage par renforcement direct (Sutton & Barto, 2018) et inverse (Ng & Russell, 2000).**

**Approche et méthodes.** Le travail de thèse proposé s'ancre dans les méthodes d'alignement forward (par apprentissage pendant la tâche), par opposition à l'alignement backward, i.e., a posteriori (Ji et al., 2023, arxiv). L'approche consiste à élaborer un modèle computationnel à base d'IA ancré dans la théorie de l'apprentissage par renforcement (RL) fondé sur un modèle (Sutton et Barto, 2018) pour estimer de manière probabiliste les impacts de séquences d'actions effectuées lors de tâches collaboratives. De plus, nous adopterons les méthodes d'inverse RL (Ng & Russell, 2020) permettant d'estimer la "fonction de récompense" latente qui explique le mieux une séquence d'actions effectuée par un agent (ici par les humains pendant la tâche collaborative, afin de vérifier que cette fonction correspond à l'objectif prédéfini de la tâche). Enfin, nous utiliserons les extensions récentes du formalisme RL auxquelles l'équipe ACIDE acontribué pour que les impacts des actions puissent être également évalués au regard du respect de normes, principes, conventions (Baldassarre et al. 2024 arxiv 2403.02514). Appliquées à l'inverse RL, nous évaluerons leur capacité à détecter les objectifs, préférences et valeurs sous-jacentes aux comportements des humains qui collaborent.

Pour le travail philosophique de la thèse, l'approche consiste à partir des théories de la collaboration comme cas particulier de la coopération impliquant des actions conjointes orientées vers un but conjoint (e.g., Paternotte et al., 2014). Du côté de l'alignement, nous partirons des théorisations classiques des valeurs humaines (Schwartz, 1994) et comparerons les situations collaboratives aux catégorisations philosophiques entre différents types de valeurs personnelles et sociales, les dernières incluant les conventions et les normes morales (Habermas, 1984; Turiel, 1991).

Évaluation des contributions. Les contributions méthodologiques seront évaluées par leur capacité à (1) mesurer l'alignement entre les actions d'un système à base d'IA et les objectifs définis par des humains dans des tâches collaboratives, ainsi que leurs préférences et valeurs, (2) pallier au désalignement (misalignment) en proposant de meilleures actions.
Les contributions philosophiques seront évaluées par la cohérence théorique et pratique entre d'une part la clarification et formalisation de concepts (collaboration, alignement, objectifs, valeurs, préférences), et d'autre part les modèles et métriques développés durant la thèse.

# Nature de la collaboration numérique (1 page max)

Le terme "collaboration" doit être compris dans un sens large, couvrant les **activités humaines (médiées par la technologie) qui impliquent un groupe d'au moins deux personnes**.
Types d'activités de groupe ciblés par le doctorat en termes de :

- **Fonction (communication, partage, coordination, autre) : co-réalisation de tâches, i.e., coordination, communication, partage d'informations, répartition et partage des actions (manipulation d'objets (scénario 1) +** scénario ouvert (hospitalier ?) (s2))
- **Type (synchrone, asynchrone) :** synchrone (scénario 1), asynchrone (scénario 2)
- **Échelle de temps (seconde, heures, mois, années, ...) : minut**es (s1), mois (s2)
- **Taille du groupe (deux, une douzaine, des centaines, des milliers, ...) :** deux (s1), une douzaine (s2)
- **Espace (co-localisé, distant, hybride) :** espace de co-manipulation (ex : atelier/workshop industriel) + hybride présentiel/distant.
- **Autre :** N.A.

# Contribution à la collaboration numérique : Résultats attendus et impact (1 page max)

Contributions visées par la thèse :

- **Empirique :** simulations numériques et analyses des résultats
- **Méthodologique** : nouvelle méthode d'estimation des impacts.
- **Autre (philosophique, éthique) :** évaluation des impacts philosophiques et éthiques de l'intervention de systèmes d'IA dans la collaboration entre humains.

Résultats attendus

Ces travaux auront d'abord pour vocation de contribuer à l'élaboration de nouvelles méthodes à base d'intelligence artificielle pour (1) évaluer l'alignement entre un agent artificiel et des humains en collaboration, (2) s'adapter et agir de manière plus alignée avec les objectifs, préférences et valeurs des humains en collaboration. Ceci implique donc le développement de modèles et métriques d'évaluation d'alignement, et des algorithmes d'apprentissage par renforcement visant à optimiser les actions des agents artificiels, et s'adapter en temps réel.

Le travail de thèse a de plus pour vocation de contribuer à l'analyse philosophique des impacts philosophiques et éthiques de l'intervention de systèmes d'IA dans des situations de collaboration entre humains. Ceci pourrait contribuer à renforcer la réflexivité et l'agentivité des êtres humains en les aidant à être mieux informés des conséquences possibles des actions dans des situations de collaboration numérique entre humains impliquant des systèmes d'IA.

Enfin, un des résultats attendus du travail thèse est la mise en place de recommandations pour une intégration éthique des systèmes à base d'IA en contexte collaboratif.

Impact

Les travaux auront pour contribution des interactions plus éthiques et productives, une meilleure compréhension des implications éthiques de l'utilisation de technologies de ce type, une meilleure intégration de ce type de technologie.

# Positionnement dans eNSEMBLE (½ page max)

Les deux laboratoires (ISIR et SND) sont partenaires du PC5. Le projet de thèse est directement et fortement lié au **Thème 2 du PC5** sur les « aspects éthiques, juridiques et philosophiques de la collaboration ». En effet, ce thème s'intéresse aux « débats et réflexions philosophiques posés par les Humanités [qui] explorent jusqu'à quel point l'humain peut être aidé, assisté, facilité dans ses tâches et ses activités, tout en restant « humain ». En particulier, il est important d'identifier comment la réflexivité humaine, la pensée critique et sa propre délibération éthique peuvent être promues pour augmenter son agentivité pendant l'interaction avec les outils numériques. » Le thème 2 du PC5 a de plus parmi ses défis principaux « l'identification des responsabilités éthiques dans la collaboration humain-machine ». Ce projet fait également des liens avec le **Thème 1 du PC5**, qui vise à élaborer et comparer des méthodes de mesure des impacts de la collaboration. Les méthodes proposées ici sont basées sur le formalisme de l'apprentissage par renforcement direct et inverse venant du domaine de l'IA.

Ce projet s'inscrit enfin fortement dans les **thématiques du PC3**, en particulier en lien avec : le **Thème 1** qui pose pour défi de s'assurer que l'introduction de systèmes intelligents dans un groupe d'utilisateurs humains sert l'objectif de la collaboration ; le **Thème 2** qui s'intéresse à l'agentivité et la confiance des humains impliqués dans la collaboration lorsque des systèmes d'IA sont introduits, grâce notamment à l'explicabilité de ces systèmes. Or notre projet vise à rendre explicites les impacts estimés des actions par le système d'IA et l'incertitude associée, de façon à garantir l'alignement du système d'IA avec les objectifs de la tâche collaborative, et les préférences et valeurs humaines.

# Présentation du candidat (1 page max)

[redacted]

[redacted]

# Autres documents (à inclure dans le même PDF)

- Brève description du groupe de recherche / laboratoire d'accueil (1 page max)

La thèse implique deux laboratoires d'accueil liés à deux disciplines différentes.

**L'institut des systèmes intelligents et de robotique (ISIR), UMR 7222**, est un laboratoire interdisciplinaire sous la triple tutelle de Sorbonne Université, du CNRS et de l'INSERM. Il rassemble des équipes pluridisciplinaires, des chercheuses et des chercheurs créant des drones, micro-pinces, prothèses bioniques, robots sociaux, bras chirurgicaux et toutes sortes de systèmes intelligents et interactifs, physiques, virtuels ou de réalité mixte. Leurs applications adressent des enjeux sociétaux majeurs : santé, industrie du futur, transports, et service à la personne. Outre des recherches en intelligence artificielle et en robotique, le laboratoire inclut aussi des travaux de modélisation pour les neurosciences ou des sciences du mouvement humain, pour décrire les fonctions cognitives et sensori-motrices des êtres vivants et pour les exploiter dans la commande des robots, ainsi que pour la psychologie, pour la caractérisation des comportements socio-interactifs des personnes humaines, entre elles ou avec des machines.

Au sein de l'ISIR, l'équipe ACIDE s'intéresse à l'action, la cognition, l'interaction et la décision *encorporées* : le comportement des agents (vivants ou artificiels) est considéré dans le cadre d'un environnement dynamique comportant d'autres agents. Elle étudie par l'expérimentation et la modélisation, les capacités cognitives des humains et des animaux. D'autre part, elle s'investit dans la conception de dispositifs d'interaction humain-machine plus adaptés, transparents et personnalisés.

**Sciences Normes Démocratie (SND), UMR 8011 CNRS**, a été créé en 2018, et est issue du rapprochement de la FRE 3593 (Fédération de Recherche) Sciences, Normes, Décision (2013-2018) et de l'EA 3559 (Equipe d'Accueil) Rationalités Contemporaines (créé en 2002). Elle est composée d'un groupe de philosophie des sciences et de la connaissance et d'un groupe de spécialistes de philosophie politique et morale. Les membres de SND se donnent pour tâche d'étudier les conditions pratiques de l'exercice de la rationalité collective, aussi bien dans les communautés scientifiques que dans les contextes politiques, à partir d'une vaste gamme de thèmes liés à la place des sciences et des techniques dans la société, allant de la constitution à la circulation et l'appropriation de l'information scientifique et technique, jusqu'à la prise de décision politique et les conditions de son efficacité. Une telle étude s'appuie sur les compétences des membres de l'équipe en philosophie de la connaissance et des sciences d'une part, et en philosophie politique et éthique de l'autre. L'articulation étroite entre ces deux domaines et son expression dans des projets de recherche résolument interdisciplinaires envisagés dans le cadre des Instituts de Sorbonne Université : transition environnementale, calcul et données, sciences du comportement et robotique, humanités bio-médicales. Équipe pivot de l'université unifiée, SND est ainsi capable de rassembler des chercheurs de disciplines différentes en pratiquant au quotidien le dialogue interdisciplinaire.