

Fiche de poste

Intitulé du poste : Post-doctorat en « Éthique de l'apprentissage ouvert en robotique et en IA »

Type de poste : ☒ Post-Doc ☐ Ingénieur-e ☐ Autre : ...

Date de début de contrat : 01/10/2025

Durée du contrat : 18 mois (jusqu'au 30/03/2027)

Quotité de travail : ☒ 100% ☐ autre précisez (50 % minimum) :

Expérience souhaitée :

- ☒ Débutant
☐ 1 - 4
☐ 4 - 10
☐ + de 10

Niveau d'études souhaité : Doctorat

Montant rémunération : Rémunération habituelle pour les post-doc

Source du financement : Projet européen PILLAR-Robots

Laboratoire d'accueil : ISIR (*Institut des Systèmes Intelligents et de Robotique*), Campus Pierre et Marie Curie, 4 place Jussieu, 75005 Paris.

Personne à contacter

Prénom Nom : Mehdi Khamassi

Tel : +33 6 50 76 44 92

Email : mehdi.khamassi@sorbonne-universite.fr

Candidature :

- ☐ En ligne. Lien vers le portail emploi :
☒ Par mail. Envoyer votre candidature par mail, avec [PILLAR post-doc application] en objet, un CV, une lettre de motivation (max 2 pages) et une liste de deux références.

Date limite de dépôt de la candidature : 07/05/2025

Description du poste (en français)

Titre du poste : Post-doctorat de 1,5 ans en éthique de l'apprentissage ouvert en robotique et en IA

Contexte :

Le projet européen PILLAR-Robots (<https://pillar-robots.eu>) vise à développer une nouvelle génération de robots dotés d'un niveau d'autonomie plus élevé, capables de déterminer leurs propres objectifs et d'établir leurs propres stratégies, en s'appuyant de manière créative sur

Sous la co-tutelle de :

l'expérience acquise au cours de leur vie pour satisfaire les demandes de leurs concepteurs/utilisateurs humains dans des cas d'utilisation d'applications réelles. Pour ce faire, il vise à mettre en œuvre une nouvelle extension théorique du formalisme de l'apprentissage par renforcement, appelée *the purpose framework**, qui constitue une abstraction indépendante du domaine d'un ensemble d'objectifs exprimant ce que le concepteur/utilisateur attend du robot. Cela permet de contraindre l'exploration et l'apprentissage autonomes des robots de façon à (1) se concentrer sur les objets et les parties de l'environnement pertinents pour la tâche, et à (2) éviter d'atteindre des états particuliers de l'environnement ou d'agir d'une manière qui ne correspond pas aux objectifs, aux préférences et aux valeurs humaines.

* Baldassarre, G., Duro, R. J., Cartoni, E., Khamassi, M., Romero, A., & Santucci, V. G. (2024). *Purpose for Open-Ended Learning Robots: A Computational Taxonomy, Definition, and Operationalisation*. arXiv preprint arXiv:2403.02514.

Lieu et environnement :

Le poste de postdoctorant sera situé à l'Institut des Systèmes Intelligents et de Robotique (ISIR, <http://www.isir.upmc.fr>), Paris, France. L'ISIR appartient à Sorbonne Université, au CNRS et à l'INSERM, et est situé dans le centre de Paris, à quelques pas de la Seine, d'autres institutions académiques (La Sorbonne, le Collège de France, le Muséum d'Histoire Naturelle, l'École Normale Supérieure, l'Université Paris Cité, l'Hôpital la Pitié Salpêtrière), et de monuments célèbres (Notre Dame, la Conciergerie, le Panthéon, le Théâtre du Châtelet, l'Institut du Monde Arabe). Il n'est pas nécessaire de parler ou de comprendre le français. Ce travail se fera en étroite collaboration avec les philosophes, les ingénieurs et les neuroscientifiques computationnels du consortium européen PILLAR-Robots.

Missions :

Le travail postdoctoral portera sur les possibilités d'alignement et de désalignement de tels robots reposant sur l'apprentissage ouvert. En particulier, des recherches seront menées sur la manière dont la combinaison du moteur de motivation du robot PILLAR, qui s'appuie sur *the purpose framework*, avec de grands modèles de langage (LLM) peut faciliter l'alignement, et dans quelles conditions. L'explicabilité, la confiance et l'efficacité de la coordination humain-robot seront évaluées dans un ensemble de scénarios réels, correspondant aux trois cas d'usages du projet : un scénario d'entrepôt, un scénario de robotique agricole et un scénario ludo-éducatif. Une analyse éthique sera réalisée en consultant les membres du consortium sur l'ensemble de ces trois scénarios.

Profil recherché :

Nous recherchons des candidats très motivés ayant un solide dossier académique. Une excellente formation est attendue en apprentissage automatique, en IA, en robotique cognitive ou en neurosciences computationnelles. Une expérience significative dans les architectures cognitives des robots, l'alignement de l'IA, les LLM ou la modélisation computationnelle pour les neurosciences ou la psychologie sera appréciée. Des connaissances préalables en philosophie de l'esprit et en philosophie morale seront un atout. Critères d'éligibilité : Doctorat dans une discipline quantitative. Il n'y a pas de critère de nationalité ou d'âge.

Compétences requises :

La maîtrise de l'apprentissage par renforcement et de la théorie des jeux, un très bon niveau en mathématiques appliquées et des compétences de programmation avancées en C++ moderne et en python sont nécessaires. Très bon niveau d'anglais (écrit et oral).

Sous la co-tutelle de :

Description du poste (en anglais)

Job title: 1.5-year post-doctorate in the ethics of open-ended learning in robotics and AI

Context: The PILLAR-Robots European project (<https://pillar-robots.eu>) aims at developing a new generation of robots endowed with a higher level of autonomy, that are able to determine their own goals and establish their own strategies, creatively building on the experience acquired during their lifetime to fulfil the requests of their human designers/users in real-life application use-cases. To do so, it aims to implement a new theoretical extension of the reinforcement learning formalism, called *the purpose framework*^{*}, which constitutes a domain-independent abstraction of a set of goals that expresses what the designer/user wants from the robot. This enables us to constraint robots' autonomous exploration and learning to (1) focus on task-relevant objects and parts of the environment, and to (2) avoid reaching particular states of the environment or acting in ways that misalign with humans' objectives, preferences and values.

^{*} Baldassarre, G., Duro, R. J., Cartoni, E., Khamassi, M., Romero, A., & Santucci, V. G. (2024). *Purpose for Open-Ended Learning Robots: A Computational Taxonomy, Definition, and Operationalisation*. arXiv preprint arXiv:2403.02514.

Location and environment: The post-doc position will be located in the Institute of Intelligent Systems and Robotics (ISIR, <http://www.isir.upmc.fr>), Paris, France. ISIR belongs to Sorbonne Université, CNRS and INSERM, and is located in the center of Paris, thus at walking distance from Seine River, from other academic institutions (La Sorbonne, Collège de France, Muséum d'Histoire Naturelle, Ecole Normale Supérieure, Université Paris Cité, Hôpital la Pitié Salpêtrière), and from famous monuments (Notre Dame, Conciergerie, Panthéon, Théâtre du Châtelet, Institut du Monde Arabe). Speaking or understanding French is not required. This work will be done in close collaborations with philosophers, engineers and computational neuroscientists of the PILLAR-Robots European consortium.

Missions: The post-doc work will investigate possibilities of alignment and misalignment in such purposeful open-ended learning robots. In particular, research shall be done on how the combination of PILLAR's robot motivational engine relying on the purpose framework with large language models (LLMs) can facilitate alignment, and under which conditions. Explainability, trust and efficient human-robot coordination will be evaluated in a set of real-world scenarios, corresponding to the project's three use cases: a warehouse scenario, an agricultural robotics scenario and an edutainment scenario. An ethical analysis will be carried out by consulting consortium members on all three scenarios.

Required profile: We are looking for highly motivated candidates with a strong academic record. An excellent background is expected in machine learning, AI, cognitive robotics or computational neuroscience. Significant experience in robot cognitive architectures, AI alignment, LLMs, or computational modeling for neuroscience or psychology will be appreciated. Prior knowledge in philosophy of mind and moral philosophy will be a plus. Eligibility: PhD degree in a quantitative discipline. There is no nationality or age criteria.

Required skills: Mastery of reinforcement learning and game theory, very good level in applied maths, and advanced programming skills in modern C++ and python are required. Very good level of English (written, spoken).