INSTITUT DES SYSTEMES INTELLIGENTS ET DE ROBOTIQUE
OFFRE D'EMPLOI

Fiche de poste

Intitulé du poste : Post-doctorat en « Éthique de l'IA neuro-computationnelle »

Type de poste : ☒ Post–Doc ☐ Ingénieur·e ☐ Autre : …

Date de début de contrat : 01/08/2024

Durée du contrat : 26 mois (jusqu'au 30/09/2026)

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

Expérience souhaitée :
☒ Débutant
☒ 1 - 4
☐ 4 - 10
☐ + de 10

Niveau d'études souhaité : Doctorat

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 [CAVAA 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 : 07/05/2024

Description du poste (en français)

Titre du poste : Post-doctorat en « Éthique de l'IA neuro-computationnelle »

Contexte :
Le projet européen CAVAA (https://cavaa.eu/) propose de réaliser une théorie de la conscience instanciée sous la forme d'une architecture informatique intégrée et de ses composants afin d'expliquer la conscience dans les systèmes biologiques et de l'intégrer dans les systèmes technologiques. Dans un monde régi par des états cachés, la conscience permet de traiter l'"invisible", depuis les environnements inexplorés (passés et futurs contrefactuels) jusqu'aux interactions sociales qui dépendent des états internes des agents et des normes morales. En particulier, nous étudierons la capacité et la propension des agents dotés d'une telle architecture cognitive à raisonner, à prendre des décisions ou à revenir sur des expériences passées, à réfléchir sur ce qui était bien ou mal selon certaines normes morales, et sur les états futurs

possibles qui pourraient être bien ou mal. L'ingénierie de la conscience de CAVAA s'accompagne d'un cadre éthique à l'égard des utilisateurs humains et des artefacts conscients dans le spectre plus large de l'IA digne de confiance, en tenant compte des objectifs partagés, des contrefactuels et des projections vers de nouveaux scénarios futurs, ainsi que de la prédiction de l'impact des choix. CAVAA vise à offrir une meilleure expérience à l'utilisateur grâce à sa capacité d'explication, d'adaptation et de lisibilité.

Lieu et environnement :
Le poste de post-doctorant 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 Salpétrière), 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 CAVAA.

Missions :
Le travail post-doctoral se concentrera sur le raisonnement éthique à travers la virtualisation, la délibération et l'alignement sur les valeurs humaines. Le cadre théorique sera ancré dans l'apprentissage par renforcement probabiliste fondé sur un modèle (*model-based*), étendu pour inclure les valeurs homéostatiques, épistémiques et sociales, y compris les conventions sociales et les normes morales comme point de départ. Les travaux étudieront l'apprentissage par l'interaction avec l'environnement et avec d'autres agents, la prise de décision sociale, la simulation mentale et le raisonnement contrefactuel pour informer les humains des conséquences potentielles à long terme de leurs actions. Le modèle sera confronté à des données expérimentales sur la prise de décision humaine face à divers dilemmes sociaux et moraux. Le modèle sera également intégré dans l'architecture cognitive CAVAA et appliqué à des agents artificiels et à des robots dans des scénarios virtuels et réels impliquant la navigation spatiale et l'interaction sociale.

Profil recherché :
Nous recherchons des candidats très motivés ayant un solide dossier académique. Une excellente expérience est attendue à l'interface entre les neurosciences computationnelles et l'apprentissage automatique. Une expérience significative dans les architectures cognitives et la modélisation computationnelle pour les neurosciences, la psychologie, l'IA ou la robotique cognitive sera appréciée. Un intérêt marqué pour la philosophie de l'esprit et la philosophie morale est attendu. Admissibilité : 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).

INSTITUT DES SYSTEMES INTELLIGENTS ET DE ROBOTIQUE

OFFRE D'EMPLOI

Description du poste (en anglais)

Job title: Post-doc position in neuro-computational AI ethics

Context:
The CAVAA European project (https://cavaa.eu/) proposes to realize a theory of awareness instantiated as an integrated computational architecture and its components to explain awareness in biological systems and engineer it in technological ones. In a world governed by hidden states, awareness enables to deal with the "invisible", from unexplored environments (counterfactual pasts and futures) to social interaction that depends on the internal states of agents and moral norms. In particular, we will study such cognitive architecture agents' ability and propensity of reasoning, decision-making, or revisiting past experiences, but also reflecting upon what was right or wrong given some moral norms, and which possible future states could be right or wrong. CAVAA's awareness engineering is accompanied by an ethics framework towards human users and aware artefacts in the broader spectrum of trustworthy AI, considering shared goals, counterfactuals and projections towards new future scenarios, and prediction of the impact of choices. CAVAA aims to deliver a better user experience because of its explainability, adaptability, and legibility.

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 the Seine river, from other academic institutions (La Sorbonne, Collège de France, Muséum d'Histoire Naturelle, Ecole Normale Supérieure, Université Paris Cité, Hôpital la Pitié Salpétrière), 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 CAVAA consortium.

Missions:
The post-doc work will focus on ethical reasoning through virtualization, deliberation and alignment with human values. The theoretical framework will be anchored on probabilistic model-based reinforcement learning, extended to include homeostatic, epistemic and social values, including social conventions and moral norms as starting point. Research will investigate learning through interaction with the environment and with other agents, social decision-making, mental simulation and counterfactual reasoning to inform humans about potential long-term consequences of actions. The model will be confronted to experimental data about human decision-making when confronted to various social and moral dilemmas. The model will be integrated into the CAVAA cognitive architecture and applied in artificial agents and robots in virtual and real-world scenarios involving spatial navigation and social interaction.

Required profile:
We are looking for highly motivated candidates with a strong academic record. An excellent background is expected at the interface between computational neuroscience and machine learning. Significant experience in cognitive architectures and computational modeling for neuroscience, psychology, AI or cognitive robotics will be appreciated. A strong interest in philosophy of mind and moral philosophy is expected. 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).