2-years 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 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 CAVAA consortium.

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. Mastery of reinforcement learning and game theory, and advanced programming skills in modern C++ and python are required. A strong interest in philosophy of
mind and moral philosophy is expected.

Mission
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.

To apply
Applicants should send a CV, letter of motivation (max 2 pages), and a list of two references via e-mail to mehdi.khamassi@sorbonne-universite.fr and raja.chatila@sorbonne-universite.fr. Please put [CAVAA post-doc application] in the subject of the mail. Review of applicants will begin immediately, and will continue until the position is filled.