Graduate Research School
University of Bordeaux
LaBRI (C.S. Lab in Bordeaux)

Stage de Master Recherche en Intelligence Artificielle

Apprentissage de systèmes I.A. de confiance, une approche neuro-symbolique

Stage débutant en Février 2023
(avec une bourse de thèse associée, démarrant en Sept. 2023)

Encadrant
Laurent Simon, Professor at the Enseirb Matmeca / LaBRI (Bordeaux)
Head of the Computer Science teaching department of the ENSEIRB-Matmeca engineering school
Head of the industrial chair “Trustworthy AI” of Fondation Bordeaux Université
Co-Head of the A.I. axis of the LaBRI
Chair of the French Association on Constraint Programming
lsimon@labri.fr

Thèmes de recherches associés
Machine Learning, SAT, Logic, Knowledge Compilation, Causality

Localisation
LaBRI, Laboratoire Bordelais de Recherche en Informatique, Talence, France

Le stage et la bourse de doctorat sont entièrement financés par la chaire "IA digne de confiance" dirigée par Laurent Simon et soutenue par la Fondation Bordeaux Université.
Renseignements supplémentaires :
https://www.labri.fr/perso/lsimon/fr/research/internships/

Les progrès impressionnants réalisés ces dernières années autour de l’apprentissage automatique permettent d’envisager de nombreuses applications jusqu’ici hors de portée. Cependant, la précision atteinte dans la prédiction (et/ou la recommandation) de ces outils ne permet pas à elle seule de garantir leur adoption, par exemple dans le cadre d’applications critiques. Il existe en effet de nombreuses barrières à leur industrialisation, dès lors que ces outils imposent de pouvoir

comprendre leurs décisions et/ou de les expliquer / justifier. De manière générale se pose ainsi de manière de plus en plus cruciale **la question de la confiance** que l'on peut avoir dans les recommandations calculées par ces outils.
Ce constat a donné lieu à **un nouveau sous domaine de l'intelligence artificielle**, s'intéressant aux garanties offertes par ces systèmes et à la confiance que l'on pouvait y avoir. Cette confiance peut revêtir différents aspects et a des liens forts avec les problématiques d'explicabilités, également sur le devant de la scène scientifique depuis quelques années.

Dans ce stage de M2 Recherche, nous proposons d'aborder la question de la confiance exclusivement par la capacité du système à admettre des preuves formelles, en lien avec certaines questions précises. Les progrès observés en vérification formelle (grâce notamment à l'amélioration des solveurs SMT / SAT) pourraient laisser espérer que l'on puisse prouver des propriétés de systèmes issus de l'apprentissage automatique, comme les réseaux de neurones profonds. Cependant, de par leur construction, les fonctions apprises par ces systèmes sont d'une complexité telle qu'il demeure impossible d'appliquer les techniques usuelles de vérification de manière frontale. Il convient donc d'étudier le compromis entre possibilité calculatoire de raisonner sur les propriétés de la fonction apprise et précision offerte par cette dernière.

Pour cela, nous proposons **d'étudier une approche hybride de l'apprentissage automatique** permettant d'apprendre directement des fonctions offrant des propriétés structurelles ou sémantiques **qui permettraient d'appliquer, en pratique, des méthodes automatiques de preuves** et/ou de **raisonnements logiques**. Il sera également étudié comment des connaissances préalables peuvent permettre de converger plus rapidement ou de garantir que les fonctions apprises respectent bien les propriétés connues d'avance.

Le but du stage est d'identifier des langages cibles bien adaptés à la recherche d'un compromis performance / confiance. Nous avons ainsi pour ambition **de proposer des méthodes d'apprentissages permettant de prendre en compte les garanties visées dès le début de l'apprentissage**, offrant ainsi la possibilité pour les modèles IA d'être plus ou moins efficaces ou explicables, suivant le contexte. Nous allons par exemple étudier comment la compilation de connaissances (Darwiche, Marquis, 2001) permet d'appréhender la confiance en un système de décision complexe par l'intermédiaire d'une série de questions expliquant les raisons de la décision. L'étude théorique des capacités des différents systèmes sera proposée par l'intermédiaire d'une cartographie claire des garanties possibles offertes par les systèmes d'apprentissage. L'originalité de l'approche repose sur une hybridation progressive, plus ou moins forte, dans le but d'obtenir des systèmes de décision réalistes dans le cadre de déploiement effectifs.

Les travaux que nous proposons se positionnent ainsi au cœur de l'IA dite "hybride" (mêlant apprentissage statistique et raisonnement symbolique), domaine de l'IA en pleine expansion et ayant de plus en plus de workshops et conférences dédiées. Les thèmes de recherche en lien avec ce sujet, en plus de l'Apprentissage Automatique (Réseaux de Neurones, Forêts Aléatoires, …), sont : les Réseaux Binaires de Décision, la Compilation de Bases de connaissances (représentation décomposable, déterministe), mais aussi les problématiques d'apprentissage de modèles causaux.

Le candidat devra avoir de solides connaissances en Apprentissage Automatique et en IA symbolique. Une très bonne connaissance de la programmation est également attendue.