Fiche Projet P21

INTÉRÊT GÉNÉRAL

1. Titre du projet : Evaluation et suivi des performances de systèmes intégrant des briques logicielles à base d'IA - Contexte et enjeux

Aujourd'hui, les innovations technologiques intégrant des algorithmiques à base d'IA, notamment dans le domaine de la numérisation des activités, le déploiement de vecteurs de mobilité autonomes, qu'ils soient terrestres ou aériens, ou même l'intégration de robots/ cobots aux activités humaines, ont pour objet de transformer les expériences de la vie pour l'homme, notamment à travers des services de confort, de facilitation, d'assistance plus ou moins automatisée de tâches complexes ou fastidieuses.

Cela mène souvent au remplacement des actions/manœuvres imitables par des algorithmes / machines, au basculement voire à une délégation complète des capacités humaines, dans un contexte d'optimisation des besoins.

Deux axes sont naturellement adressés par ces innovations technologiques. Le premier concerne l'augmentation de la qualité de vie des usagers, notamment à travers la production de bénéfices individuels et sociétaux, tout en respectant des valeurs éthiques, telles que l'équité, le respect de la diversité et la non-discrimination, mais aussi le respect de la vie privée à travers la protection des données personnelles. Le deuxième axe inclut toutes les dimensions de responsabilisation dans le déploiement de ces innovations : la satisfaction d'impératifs de sécurité et de robustesse, la démonstration du niveau de sécurité/ sûreté, les exigences de transparence, de compréhension et d'explicabilité, et plus généralement la prise en compte du facteur humain.

La question du fonctionnement sûr des logiciels est depuis longtemps au cœur des préoccupations industrielles, qu'il s'agisse du transport, des dispositifs de santé, des systèmes industriels… Les approches actuelles d'Ingénierie Système et d'IVVQ (Intégration, Vérification, Validation, Qualification) sont en cours d'adaptation et d'enrichissement pour prendre en compte les spécificités des systèmes intégrant des briques d'IA.

Ce sujet reste ouvert lorsque les systèmes incorporent de l'IA. Que l'on pense à la sûreté d'une prise de décision « autonome » en temps réel comme dans les domaines évoqués ci-dessus, à des domaines ne tolérant pas l'erreur de décision (décisions de sécurité, de justice, diagnostic de santé, etc.) ou à des attentes d'équité de traitement qui exigent la garantie que ceux-ci ne sont

pas biaisés, la confiance placée dans les systèmes intégrant de l'IA doit impérativement être démontrée et reconnue, comme ce fut le cas précédemment pour les logiciels déterministes «classiques».

La diffusion de résultats issus de l'intelligence artificielle peut se trouver freinée dans bien des secteurs industriels aux motifs qu'on ne sait ni les expliquer, ni les garantir.

L'objectif de ce projet est de :

- dresser un état de l'art des méthodes de conception, développement et validation de systèmes intégrant des briques logicielles à base d'IA ;
- définir un cadre d'évaluation des performances de tels systèmes notamment en termes de « *Safety* » et « métriques de couverture » ;
- comparer différentes techniques sur deux cas exemples ;
- identifier les insuffisances et les inconvénients de ces techniques pour définir des pistes d'amélioration ;
- permettre d'attribuer une certaine confiance aux résultats de sortie.

Dans ce qui suit, on désigne « système à base d'IA », un système complexe intégrant des briques logicielles issues des disciplines de l'Intelligence Artificielle (par exemple : *Machine Learning*, Réseaux de Neurones…).

2. Résultats attendus

- Etat de l'art
- Référentiel de métriques de performance et de couverture pour l'étude et la caractérisation des systèmes à base d'IA
- Construction de deux cas d'application (en collaboration avec les souscripteurs), données nécessaires, structuration.
- Traitement de ces cas d'application à l'aide de différentes méthodes et techniques. Estimation de la confiance dans les résultats.
- Comparaison des résultats *(benchmarking)*. Mise en évidence de l'intérêt et des difficultés.
- Synthèse et perspectives.

Le benchmarking pourra par exemple illustrer les questionnements suivants.

Comment les référentiels normatifs d'ingénierie système intègrent-ils progressivement les spécificités des systèmes à base d'IA ?

Quelles sont les métriques de performances liées à la « *Safety* » et aux enjeux de « couverture » dans les démarches de validation de systèmes à base d'IA ?

Quelles sont les différentes activités techniques et scientifiques qui peuvent contribuer à produire un argumentaire justificatif de performances « *Safety* » liées à l'exploitation d'un système à base d'IA ?

3. Programme des travaux

Ce programme est donné à titre indicatif. Il sera bien évidemment précisé dans un cahier des charges, si le projet a un nombre suffisant de souscripteurs, en fonction de leurs besoins. Le programme proposé dans cette fiche peut donc se trouver modifié en fonction des besoins des souscripteurs.

Tâche 0 : Analyse des besoins des souscripteurs

Tâche 1 : Etat de l'art des méthodes et outils de validation et évaluation des performances des systèmes à base d'IA, notamment en ce qui concerne la « *Safety* » et la confiance dans les résultats

Tâche 2 : Choix de deux cas d'application de systèmes à base d'IA

Ces cas, issus de besoins en maîtrise des risques ou sûreté de fonctionnement (par exemple démonstration d'un niveau de sécurité/ sûreté, maintenance prévisionnelle, …) nécessiteront vraisemblablement une structure, des liens d'influence, des données d'entrée, une expertise… Ces informations seront fournies par les souscripteurs ou, à défaut, un exemple fictif pourra être créé ou extrait de la documentation.

Remarque : la validation - qualification des résultats de l'IA dépend aussi beaucoup de la qualité et de la pertinence des données en entrée, lesquelles ne peuvent plus être validées « manuellement » ; curieusement ce point est rarement évoqué dans la documentation, c'est pourtant important. Rien ne garantit que les données initiales soient bonnes, d'autant plus qu'elles sont nombreuses et acceptées telles quelles. D'autres facteurs comme un contexte et un environnement en évolution, la non explicabilité des résultats des modèles… peuvent également porter atteinte à la confiance d'un utilisateur dans les résultats.

Tâche 3 : Traitement de ces deux cas

On s'attachera à appliquer des méthodes, techniques ou outils susceptibles d'évaluer certaines performances de systèmes à base d'IA, et à mettre en avant l'apport de ces méthodes et outils par rapport aux méthodes applicables aux algorithmiques classiques, leurs conditions d'utilisation, l'interprétation des résultats et les éventuelles difficultés rencontrées.

Tâche 4 : Analyse critique de l'utilisation de la méthode ou des méthodes choisie(s), exploration des variantes, et des contraintes présentées par les

modalités d'implantation, les hypothèses prises en compte et la maîtrise des informations à prendre en compte à l'entrée de la méthode.

Tâche 5 : Synthèse et identification des insuffisances et incomplétudes pour définir des pistes de développement de méthodes innovantes pour la validation des systèmes à base d'IA et leur utilisation.

4. Références bibliographiques

[1] Livre blanc de l'AFNOR : L'impact et les attentes pour la normalisation dans l'intelligence artificielle. Avril 2018.
[2] Christoph Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2018.
[3] W. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu. Interpretable machine learning:definitions, methods, and applications. Janvier 2019.
[4] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry. On Evaluating Adversarial Robustness. https://arxiv.org/abs/1902.06705 Février 2019.
[5] Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. https://www.declarationmontreal-iaresponsable.com/, Décembre 2020
[6] Concepts of Design Assurance for Neural Networks, rapport, 31 mars 2020, (CoDann)

- [7] A. De Galizia,
- A. Bracquemond,
- E. Arbaretier,

- A scenario-based risk analysis oriented to manage safety critical situations in autonomous driving, In book: Safety and Reliability – Safe Societies in a Changing World, , 2018.

[8] Autonomous Driving System: Model Based Safety Analysis (Tlig et al., 2018, https://hal.archives-ouvertes.fr/hal-01906465).
[9] New Methodological and Technical Fields for Safe Design Engineering of Autonomous Systems (Zhao et al., 2018).
[10] Societal acceptance and regulation facing innovations and associated risks: development from autonomous systems (Arbaretier and Zhao, 2021).
[11] IMdR (2020), Projet *Big Data in Reliability*, Janvier 2020.

5. Durée

12 mois

6. Montant de la souscription

9800 € HT