Proposition de sujet de thèse :

**Méthodes mathématiquement rigoureuses d’explicabilité et d’interprétabilité de l'Intelligence Artificielle**

**Mathematically rigorous methods of explicability and interpretability of Artificial Intelligence**

Intelligence Artificielle, explicabilité de l'IA, xAI, acceptabilité de l'IA, IA de confiance, Equations aux Dérivées Partielles, Analyse Numérique

Artificial Intelligence, AI explicability, xAI, AI acceptability, Confidence AI,Partial Differential Equations, Numerical Analysis

Directeur : Emmanuel Frénod

Co-directeur : François Septier

Les questions d'explicabilité et d'interprétabilité de l'Intelligence Artificielle (IA) sont aujourd'hui très présentes dans les débats autour de l'émergence de cette technologie. En particulier l'irruption de « ChatGPT » amène une grande partie de la société à prendre conscience de l'urgence d'adresser ces questions. Ce sont en effet les deux éléments centraux pour arriver à bâtir ce que l’on appelle une « IA de confiance ».

Parmi les concepts qui ont émergé récemment autour de ces questions, il y a l’xAI (pour explainable Artificial Intelligence). L'approche décrite dans les références qui abordent l’xAI (c.f <https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/>, <https://www.alteryx.com/fr/resources/whitepaper/essential-guide-to-explainable-ai>) semblent très descriptive est relativement peu mathématisée.

Le but de cette thèse est d’utiliser une approche mathématiquement rigoureuse pour aborder ces questions.

L’approche se base sur la très grande proximité entre un « réseau de neurones de convolution » et la discrétisation par des méthodes d’analyse numérique d’équations aux dérivées partielles (EDP) de diffusion non stationnaires et non linéaires. Il s’agit donc d’utiliser cette proximité afin de mesurer, apostériori, en quoi un réseau de neurones est le résultat d’un processus cognitif sous-jacent modélisé par une EDP.

L’explicabilité et d'interprétabilité seront alors une conséquence de la description du processus cognitif sous-jacent : il y aura ainsi un raisonnement à expliquer.

Au-delà de cette question d'explicabilité et d'interprétabilité, le travail de cette thèse permettra aussi de mettre en œuvre des méthodes de pré-entraînement de réseaux de neurones - en particulier dans le cas du Deep Learning, où ces entraînements peuvent être très coûteux en temps de calcul.

Étapes de la thèse

1. Revue de littérature sur
   1. L’explicabilité et l'interprétabilité de l’IA ;
   2. Les liens entre IA et EDP.
2. Construction d'une distance sur les coefficients de réseaux de neurones entraînés.
3. Interprétation des coefficients de réseaux de neurones entraînées comme des coefficients résultats d’une discrétisation d’une EDP de diffusion non stationnaire et non linéaire.
4. Identification précise des EDP en jeu.
5. Interprétation de ces EDP, via de la psychologie cognitive, comme des descriptions de processus cognitifs.
6. Étude mathématique de ces EDP.
7. Développement d’un outil d'explicabilité et d'interprétabilité d’un réseau de neurones.
8. Développement d’outils pour pré-entraîner un réseau de neurones.