

- Faculté des sciences
- www.unine.ch/sciences

Seminar Trustworthy AI Models and Systems (3IN2090)

Filières concernées	Nombre d'heures	Validation	Crédits ECTS
Master en informatique	Cours: 2 ph	Voir ci-dessous	5

ph=période hebdomadaire, pg=période globale, j=jour, dj=demi-jour, h=heure, min=minute

Période d'enseignement:

- Semestre Automne

Equipe enseignante

Enseignante responsable : Prof. Lydia Y. Chen

Enseignant : Prof. Christos Dimitrakakis

L'équipe enseignante comprend également plusieurs doctorant-es qui interviennent dans le cours et accompagnent les groupes dans leurs projets de recherche

Contenu

Ce séminaire est consacré aux modèles et systèmes d'intelligence artificielle de confiance (Trustworthy AI), avec un accent particulier sur les modèles génératifs et l'apprentissage par renforcement.

Le cours aborde les principaux enjeux liés au développement et au déploiement de systèmes d'IA fiables, notamment la robustesse, la transparence, l'équité, la confidentialité et l'alignement. Une attention particulière est accordée aux risques associés aux modèles génératifs, tels que la génération de contenus synthétiques trompeurs, les deepfakes, les problèmes de propriété intellectuelle et les risques pour la vie privée.

La partie consacrée à la Generative AI couvre notamment la détection et la vérification de contenus synthétiques, la sélection et l'évaluation de modèles génératifs, ainsi que l'analyse des risques liés à la confidentialité. Le cours s'appuie principalement sur des articles scientifiques récents et demande aux étudiant-es de reproduire, analyser et étendre des travaux existants.

Forme de l'évaluation

L'évaluation est principalement basée sur un projet de recherche en groupe et sur la participation aux activités du séminaire.

Projet de groupe (80 %) : les étudiant-es sélectionnent un article scientifique, reproduisent les résultats principaux, puis développent et évaluent une extension du travail. L'évaluation prend en compte la qualité de la reproduction, la pertinence et l'efficacité de l'extension proposée, l'étendue de l'évaluation expérimentale, ainsi que la qualité du rapport, du code et de la présentation.

Contribution individuelle (20 %) : évaluation de l'engagement et de la contribution individuelle au projet.

Modalités de rattrapage

En cas d'échec, les modalités de rattrapage sont définies conformément au règlement de la formation et de l'Université de Neuchâtel.

Documentation

Le cours s'appuie principalement sur des articles scientifiques récents et classiques dans le domaine du Trustworthy AI, notamment sur la détection des contenus générés, l'évaluation et la sélection des modèles génératifs, la confidentialité et l'apprentissage par renforcement.

La liste des articles à lire et à choisir pour le projet est fournie dans le dépôt du cours

Pré-requis

Une bonne connaissance des concepts fondamentaux de machine learning et d'intelligence artificielle est recommandée.

Une expérience de programmation en Python et une familiarité avec la lecture d'articles scientifiques sont également fortement recommandées. Les étudiant-es doivent être capables de comprendre des méthodes d'apprentissage automatique,

Forme de l'enseignement

Le cours prend la forme d'un séminaire de recherche combinant présentations de concepts fondamentaux, lecture d'articles scientifiques, travail individuel et en groupe, et séances de suivi avec les superviseur-es.

Les premières semaines introduisent les concepts fondamentaux du Trustworthy AI, de la Generative AI et du reinforcement learning. Les

- Faculté des sciences
- www.unine.ch/sciences

Seminar Trustworthy AI Models and Systems (3IN2090)

semaines suivantes sont principalement consacrées au travail autonome sur le projet, avec des réunions régulières avec le ou la superviseur-e pour discuter de la sélection de l'article, de la reproduction des résultats, de l'extension proposée et des résultats finaux. Les dernières semaines sont consacrées aux présentations des projets.