University of Tehran
School of Electrical and Computer Engineering

Course	**Interpretable AI**		
Course type, level, credit	Optional	Graduate	3 units
Field, Major	Electrical and Computer Engineering	All majors	
Co-requisite(s)	-		
Prerequisite(s)			
Prerequisite by topic	-		
Goals	The use of machine learning in practical settings, is different than in a lab setting. In practical settings, several issues appear, that do not generally appear in lab settings. Some of these practical issues include: Robustness, Domain Generalization, Interpretability, Explainability, Safety, Security, Fairness, Transparency, privacy and regulatory. The goal of this course is to equip students with the knowledge and experience of these practical problems of machine learning and allow them to make more informed decision in their design of machine learning systems.		
Outcome	Upon successful completion of the course, the students will gain a knowledge and understanding about the practical issues of using AI in practice.		
Topics	1- Robustness and Generalization: - General AI - Bias-Variance trade-off and regularization - Inductive biases vs. generalization - Thought process to choose inductive biases - How to use inductive biases in a model - Self-supervision and its useful inductive biases - Multi-task learning and its emerging properties - The emergence of deduction (AI) from induction (learning) - Learning with loss function vs learning with rules - Generalization vs robustness - Generalization to different domains - Robustness to input noise in a given domain - trade-offs between robustness and generalization - Learning with noisy inputs - Effects of noise on generalization		

    - data augmentation and associated inductive biases

- Learning with low-quality supervision
    - Robustness to noisy labels
    - Evaluation given noisy labels
    - Sampling, rare events and black swan problem
    - Sparse and low-entropy labels
    - Anomaly detection

2- Interpretability and Explainability:

- Prescriptive models (models that make decisions for us)
    - Why predictive models need some causal understanding
    - How prescriptive models make decisions affecting humans
    - Tragedies in prescriptive machine learning
    - Why interpretability matters in prescriptive models

- Interpretable models
    - Linear models, decision trees, etc. and how to interpret them.
    - The relation between model interpretability vs causal analysis
    - Trade-off between interpretability and prediction accuracy
    - Practical cases where interpretability becomes more important than accuracy

- Explaining the workings of machine learning models
    - Explainability vs interpretability
    - Importance of explaination in advancing machine learning
    - Model-dependent vs model-agnostic methods
    - Example-based methods
    - Global and local explaination techniques
    - Understanding the workings of deep models

3- Safety and Security:

- Security attacks in machine learning:
    - poisoning of a dataset
    - backdoor attacks and trojan attacks
    - open-access models vs closed-access models
    - Security attacks in generative models and language models
    - Watermarking

- Defense against attacks:
    - Trojan detection and characterization
    - Input/output monitoring

- Safety and reliability:
    - Supervised/unsupervised Anomaly detection
    - Out of distribution detection
    - Accumulation of error in autoregressive models
    - Safety in prescriptive models
    - How to quantify reliability (domain shift, augmentation, adversarial)

4- Fairness, Transparency, ethics and Privacy

- Fairness and justice
    - How to define fairness in mathematical terms
    - Controversies in expressing defining justice mathematically
    - How to measure fairness/bias
    - Use of interpretable models to analyze fairness
    - Ways to enforce fairness in machine learning models

	- Transparency and Opqueness in decision making - In what contexts decision making process should be transparent - Importance of interpretability in transparency - Transparency regulations - Ethics of AI - How AI affects the job market, democracy, environment - Who is legally liable if an AI causes an injury - The use of chat bots and their impact. - Fiar use of data - Regulatory bodies of AI
Required software	- Python, PyTorch
Assignments	This course will cover four projects. (Details below)
Projects	Students are expected to work on four course projects throughout the semester. Projects must be completed individually. Projects cover: - Robustness - Interpretability - Machine Learning Security - Fairness
Grading	Projects: 50 % Midterm exam: 20 % Final exam: 30 %
Textbook(s)	https://christophm.github.io/interpretable-ml-book/ https://fairmlbook.org/ https://fairmlclass.github.io/ http://www.math.ku.dk/~peters/elements.html
Further readings	https://www.fatml.org/ https://facctconference.org/
Prepared by	Mohammad Amin Sadeghi, Mostafa Tavassolipour

Prepared by: Mohammad Amin Sadghi, Last revision 20.10.1401