Proposal for a New Course

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Course number: IS301
Course title: Ethics in AI
Course prerequisites: NA
Course credits: [3] (1-0-0-0)
Course duration: Full semester
Course type: DC
Proposing instructor: Sayak Ray Chowdhury
Other faculty members interested in teaching the course:
Other departments interested in the proposed course: None

Course description:
a. *Objectives*: The increasing level of autonomy enjoyed by AI agents and their pervasive use in critical decision-making processes with life-altering consequences makes it imperative for AI development to be aware of ethical considerations. This course will introduce students to the foundations of ethical principles of AI design and introduce techniques to identify and remove the biases from AI systems.
b. *Logistics*: The course will serve as a DC for IS UG students (BT, double major, digital IS minor, cyberphysical IS minor), and as a DE for PG students (MT, MS, PhD) of the CSE and IS departments and CSE UG students (BT, double major, ML minor). The course may be offered one or more times every academic year depending on demand and availability of resources.
c. *Content*: There will be an equivalent of 14 lectures of 50 minutes each. A weekly breakup of lecture content is given below.
d. *Evaluation*: Evaluation will use a combination of graded lab exercises, lab exams, take- home assignments and projects, and traditional sit-down quizzes and exams.

Weekly breakup of content: Numbers in square brackets [] against each topic indicate the number of lectures/labs for that topic.

Lecture content (14 lectures):
1. Basics of AI [2]: Recap of Deep learning and Vision Models, Introduction to Transformers, Hands on with Pytorch and Hugging Face
2. AI Alignment [3]: Supervised Fine-tuning, Reinforcement learning with Human Feedback, Direct Preference Optimization, Hands-on with Transformer Reinforcement Learning Library
3. Risks of AI models [2]: Toxicity, Bias, Goal misspecification, Adversarial attacks, jailbreaking
4. Principles of AI Ethics [2]: Fairness, Robustness, Privacy, Safety
5. Transparency and accountability in AI [2]: Explainable AI, Mechanistic Interpretability
6. AI Policy and Regulations [3]: EU Act, DDPA Act, US Presidential Elections

Short summary for inclusion in the Courses of Study booklet: The course aims to train future professionals and engineers with techniques to identify and remove the biases from AI systems.

Textbook: There will be no textbook for this course.

Course proposer: Sayak Ray Chowdhury
Date:

Convener DPGC:
Date:

The course is approved/not approved

Chairperson, SPGC
Date: