Course Number: CSC 859
Course Title: AI Explainability and Ethics
Number of Credits: 3
Schedule: Three hours of lecture/discussion per week
Prerequisite: Graduate standing or consent of the instructor

Catalog Description: Brief overview of AI. Impact of Artificial Intelligence on society and business. Motivation for explainable and ethical AI systems. Analysis, practicing and evaluation of technologies and methods for the design, development and evaluation of explainable and ethical AI systems. Also recommended for non CS majors.

Expanded Description

Topics will include but not be limited to the following:

- Overview of AI and methods for its evaluation, with more details on one selected AI algorithm (e.g. Random Forest)
- Overview of recent notable concerns and problems with some AI applications (e.g. fake news filtering, autonomous cars, bias in various AI applications, etc.)
- Factors causing bias in current AI systems
- Review of historical and recent regulatory, political and industry initiatives (EU GDPR laws, CA Assembly 23 Asilomar Principles of AI etc., Google AI principles),
- Definitions of Explainable AI and its impact toward more ethical and easier to adopt AI systems
- Review of selected explainable AI algorithms and evaluation methods
- Usage of selected open source tools for AI and explainable AI
- Guest lectures on selected topics (optional)

Course objectives and role in the program

We are witnessing emergence of AI and "AI economy and society" where AI technologies are impacting more and more areas such as biomedical research, health, business (e.g. credit approvals), military (e.g. autonomous robots), self driving cars, management of news (e.g. filtering of fake news), and automation of many business practices (loan approvals, hiring etc.). Much attention emerged recently in political, legal, social and technical communities addressing strong concerns about the impact of these AI systems to society including how to make AI systems better e.g. fair and non-biased, explainable/transparent and legally defendable, privacy-protecting, safe etc. These concern resulted in recent legal, regulatory and political actions such as EU GDPR privacy and data protection laws May 2018; CA Assembly 23 Asilomar AI Principles June 2018) and consequently motivated new research efforts, dedicated conferences and workshops.

This course will be positioned as advanced graduate course for CS majors (and non-matriculated students with relevant background) as well as for non-CS majors who intend to learn basics of applied AI. This course is also positioned as key course in Technology segment of new Ethical AI Graduate Certificate. By working in teams of mixed backgrounds (CS and non-CS majors) students will learn how to understand, explain, evaluate and communicate about AI applications with non-experts.

As such, this course directly supports SFSU mission of social justice by ensuring that CS graduate students attain the knowledge and understanding on how to develop explainable and ethical AI technologies and applications which benefit the society.

Learning Outcomes

Student will be able to learn and understand:

- Basics of AI and machine learning (ML) with more detailed understanding of one ML algorithm e.g. Random Forest
- Basics of ethics as it applies to AI systems and its impact to society and business and social justice

- Specific issues of AI explainability and ethics in various applications such as biomedical research, health, robotics, business
- Issues that can lead to bias in AI systems
- Recent initiatives at business, political and regulatory level related to AI ethics and explainability/transparency
- Requirements, definitions and technologies relevant to explainable AI systems and their evaluation
- Development and implementation of small-scale explainable coding AI experiment using open source/free AI tools like Python with SciKit or R system
- Basic steps in assessment and audit of AI systems for Ethics and Explainability

Method of Evaluation

Evaluation shall be based on combination of several individual student homework assignments and a team project.

Homework: there will be several individual homework assignments carrying total of 40-50/100 points (TBD in each class)

- Analysis of case studies of failures of AI necessitating better AI explainability and ethics
- Review of selected research papers in Explainable AI
- Review of a selected AI algorithms
- Individual AI experiment in analyzing a data set, and presenting and evaluating results
- In class presentation on selected topic upon approval of the instructor

Team project: Student teams of 2-3 students of diverse backgrounds (e.g. CS and Philosophy, CS and Business) will be formed to work an applied ethical and explainable AI project comprising of data selection, AI method selection, AI data analysis on selected data including explainability, and ethical assessment of

developed application. Each student team shall write a report and present it to the class – total of 50-60/100 points (TBD).

Reading Material:

To address the latest works in this fast moving area, instructor will provide recent relevant technical and other kinds of papers (blogs, news), legal/regulatory documents and pointers to open source AI algorithm implementation/tools and databases.

Notes:

This course may include guest lectures by other CS and non-CS faculty

Created: D. Petkovic, November 2018, with input from CS Department faculty and consultation withy Chair of Philosophy department Prof. J. Tiwald. Approved February 2019. Revised 01/13/21