Course Number: CSC 859
Course Title: AI Explainability and Ethics
Number of Credits: 3
Schedule: Three hours of lecture/discussion per week
Prerequisite: CSC 869 or CSC 872 or CSC 876**;** or consent of instructor

03/20/19

Catalog Description: Impact of Artificial Intelligence on society and business, Need for explainable and ethical AI systems and related problems in current AI applications. Analysis and evaluation of technologies and methods for the design, development, evaluation and deployment of explainable and ethical AI systems

Expanded Description

Topics will include but not be limited to the following:

- Ethics and explainability in AI from Computer Science perspective
- 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 AI algorithms and applications in terms of their explainability
- Techniques and best practices for improving transparency and explainability of selected AI algorithms and methods
- Guest lectures on selected topics

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) who have basic knowledge of AI and will serve the purpose to educate and sensitize CS students (many who will become developers and implementers of AI technologies) in: a) Issues and societal implications of non-ethical and hard to explain AI systems; b) Technical issues leading toward more explainable AI systems; and c) Best ways how to design, develop, evaluate and deploy explainable AI applications

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 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 and definitions relevant to explainable AI systems
- Algorithms and technologies to make selected AI methods more explainable
- Development, evaluation and deployment of explainable AI technologies and applications
- Practical ways to develop and implement small-scale explainable AI applications

Method of Evaluation

Each student will have to develop one small scale explainable AI project, one written assignments and one class presentations, as follows:

- *Mid term paper* of min. 10 pages addressing some specific problem or application related to AI ethics and/or explainability which demonstrate student's understanding of these issues and their implications and related technical challenges (30% of the grade)
- *In class presentation* on selected topic upon approval of the instructor (optional, for extra credit of up to 10% of the grade)
- *Final small scale project in Explainable AI* where student would evaluate and compare chosen AI algorithms/application for explainability and explore ways to improve their explainability (70 % of the grade). Use of standard open source AI implementations/tools and open source databases is encouraged

Reading Material:

To address the latest works in this fast moving area, instructor (and students) 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 will 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