UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Seminar 1

Friday November 24, 2023 at 3PM online via Zoom (remote)

Sonal Allana

Privacy Risks and Preservation Methods in Explainable Artificial Intelligence (XAI)

Advisor: Dr. Rozita Dara
Advisory: Dr. Xiaodong Lin
Advisory: Dr. Luiza Antonie

Abstract:

Transparency is a fundamental principle and a pillar of Trustworthy Artificial Intelligence (AI). Its main goal is to bring interpretability in complex AI models, such as deep learning networks, that are essentially black-boxes. In recent years, interpretability techniques based on different approaches have been proposed in the field of Explainable Artificial Intelligence (XAI). However, releasing transparency reports using XAI techniques, is found to impact privacy of AI models adversely. Privacy attacks exploiting explanations have targeted the sensitive data of individuals used in training the model and/or the confidentiality of the model. To counter this, various preservation methods using Privacy Enhancing Techniques (PETs) are proposed in literature.

In this seminar, we present the main approaches of generating explanations and the popular methods within each approach. We elaborate on the interplay of privacy and explainability by describing the privacy risks introduced in AI models by transparency techniques. An overview of privacy preserving methods applied to explanations is presented future direction towards achieving a balance of privacy and transparency.