Western Engineering

Western University
Faculty of Engineering
Artificial Intelligence Systems Engineering Program

AISE 3020A – AI: Ethics, Bias, and Privacy

Course Outline Fall 2025

COURSE DESCRIPTION: This course explores fairness, bias, and privacy in machine learning. Students will examine how human bias and data quality affect model outcomes, and learn techniques for building ethical models that balance trade-offs between fairness, accuracy, and privacy. The course covers sources of bias, group fairness in supervised and unsupervised learning, and notions of individually fair predictions, as well as how context (such as decisions made based on predictions) determines. It also addresses privacy risks, differential privacy, federated learning, and the fairness and privacy of generative AI models, equipping students to design machine learning systems that are both effective and responsible.

ACADEMIC CALENDAR: AISE 3020A

https://www.westerncalendar.uwo.ca/Courses.cfm?CourseAcadCalendarID=MAIN_030388_1&SelectedCalendar=Live&ArchiveID=

This course explores the fundamental issues of fairness and bias in machine learning. In addition, the course explores many aspects of building ethical models while considering human bias and dataset awareness. Furthermore, the course will explore fundamental concepts involved in privacy and security of machine learning projects. Topics such as how to protect users from privacy violations while building useful transparent predictive models will be explored.

PREREQUISITE: DS 3000A/B

Unless you have either the requisites for this course or written special permission from your Dean to enroll in it, you will be removed from this course and it will be deleted from your record.

ANTIREQUISITES: n/a

CEAB ACADEMIC UNITS: n/a

CONTACT HOURS:

Lectures occur weekly.

LECTURE:	3 hours/week
LAB:	2hrs / 5 times during the term

RECOMMENDED/ REQUIRED SOFTWARE: Python 3 installation, Jupyter Notebooks, Visual Studio Code

RECOMMENDED RESOURCES/REFERENCES:

Fairness and Machine Learning: Resources and Opportunities

Additional materials relevant to specific sections will be posted on OWL with that week's materials.

GENERAL LEARNING OBJECTIVES (CEAB GRADUATE ATTRIBUTES)

Knowledge Base	D	Engineering Tools		Impact on Society	**I**
Problem Analysis		Individual & Teamwork		Ethics and Equity	**I**
Investigation	I	Communication		Economics and Project Mgmt.	
Design		Professionalism		Life-Long Learning	

Notation: x represents the content level code as defined by the CEAB. blank = not applicable; I = introduced (introductory); D = developed (intermediate) and A = applied (advanced).

Rating: I – The instructor will introduce the topic at the level required. It is not necessary for the student to have seen the material before. D – There may be a reminder or review, but the student is expected to have seen and been tested on the material before taking the course. A – It is expected that the student can apply the knowledge without prompting (e. g. no review).

COURSE MATERIALS: Weekly content will be available on the course OWL site.

UNITS: SI

COURSE TOPICS AND SPECIFIC LEARNING OUTCOMES:

An important aspect of designing artificial intelligence systems is ensuring that bias is accounted for, and that systems are designed in a way that accounts for privacy, fairness, and ethics. This course will cover various topics related to bias, fairness, and privacy, as well as examine trade-offs where desirable objectives are incompatible. Students will develop skills that enable them to translate goals for privacy and fairness into technically-achievable objectives while understanding the limitations of this approach, as well as implement machine learning models that ensure desired objectives are met.

The following table summarizes the course learning outcomes along with CEAB GAIs where the GAIs in bold indicate ones to be measured and reported annually.

Course Topics and Specific Learning Outcomes	CEAB Graduate Attribute Indicators
1. Introduction to Fairness and Machine Learning By the end of this section students should be able to: a. Understand the different stages of machine learning modelling where bias can occur (data collection, pre-processing, model development) and where it can be mitigated (pre-processing, in-processing, post-processing). b. Describe the difference between allocative and representational fairness. c. Describe, in mathematical terms and in language, different concepts of group fairness and individual fairness and assess whether each conception of fairness is met by a given model. d. Identify incompatible notions of fairness and describe trade-offs between different definitions.	EE2, IESE1
2. Sources of Bias By the end of this section students should be able to: a. Identify sources of bias in data used to train machine learning models such as labelling bias, representation bias, and measurement bias. b. Identify how model choice affects fairness, and explain trade-offs of different models given particular classification tasks and data. c. Identify and describe feedback loops resulting from model outputs.	EE2, I3
3. Fair Classification By the end of this section, students should be able to: a. Recognize when data quality or balance results in biased models, and be able to apply pre-processing techniques such as resampling. b. Train classifiers that include group or individual fairness constraints and recognize any trade-offs with accuracy as well as computation time. c. Apply and recognize the advantages and disadvantages of post-processing techniques (such as group-specific thresholds) for group fairness.	KB4, I1

4. Fair Predictions vs Fair Decisions By the end of this section, students should be able to: a. Explain the difference between "learning to predict" and "learning to decide" and understand when ignoring the final use of a prediction can result in suboptimal outcomes (e.g., problems with missing labels). b. Explain how different types of decisions (such as ranking tasks) amplify errors in predictions.	KB4, I1
5. **Fair Unsupervised Learning** By the end of this section, students should be able to: a. Explain and implement a model for fair clustering. b. Describe and implement fairlets. c. Compare approaches to resolving biased representations (such as word embeddings) learned from large datasets.	EE2, KB4, I1
6. Causal Models and Counterfactual Fairness By the end of this section, students should be able to: a. Explain the limitations of observational fairness metrics in capturing individual-level discrimination. b. Construct simple causal models and identify paths that potentially represent unfair influence. c. Understand and apply the definition of counterfactual fairness and discuss limitations (such as contexts where it requires Pareto-dominated outcomes).	KB4
7. Privacy Implications of Machine Learning Models By the end of this section, students should be able to: a. Explain different types of privacy threats in machine learning models, including membership inference attacks, model inversion, and training data extraction. b. Evaluate the conditions under which ML models leak private information, considering factors such as model overfitting, black-box vs white-box models, and data distribution.	KB4, I1
8. Differential Privacy By the end of this section, students should be able to: a. Understand the (ε, δ)-differential privacy model. b. Be able to implement models with differentially private training methods (starting from appropriate libraries), and analyze the impact of privacy parameters on model performance. **c.** Understand trade-offs between privacy budgets and other desired outcomes (accuracy, fairness, transparency) and justify trade-offs for	KB4

particular settings.	
9. Federated Learning By the end of this section, students should be able to: a. Explain the how federated learning works, including the roles of clients, the central server, and the algorithm to combine models. b. Evaluate the advantages and limitations of federated learning in terms of privacy, model performance, and efficiency. **c.** Be able to run a basic federated learning setup (e.g., PySyft), and assess how the distribution of data between clients affects model training.	KB3, I1
10. Fairness and Privacy in Generative AI Models By the end of this section, students should be able to: a. Identify privacy risks such as training data memorization and prompt-based data leakage. b. Explain how generative AI models can amplify social biases in the training data. c. Critically evaluate common techniques for reducing bias and protecting privacy**.**	EE2, IESE1, KB4

EVALUATION:

Name	**% Worth**	**CEAB GAs ASSESSED**
Labs (Total = 5)	30%	KB3, KB4, I1, I3
Mid-Term Examination	30%	KB3, KB4, I3
Final Examination	40%	KB3, KB4, I3

Note that the dates listed above are **tentative** and may be adjusted if needed. Marks will be assigned on the basis of method of analysis and presentation, correctness of solution, clarity and neatness.

COURSE POLICIES:

All work submitted must be of professional quality in the requested format. Material that is handed in illegible, disorganized, or in an unapproved format will be returned to the student for

resubmission and the late submission penalty will take effect. Any penalty of 10% may be deducted for poor grammar, incoherence, or lack of flow in any written questions.

GENERATIVE AI POLICY: Generative AI may be used for the coding portions of assignments, but how it was used and any changes made must be documented (in the notebooks for labs). Written questions associated with lab should be answered in your own words.

LABORATORIES: There will be 5 lab assignments. Attendance at all laboratory sessions is mandatory. Absence from any session, or a portion of a session, without permission will result in a zero assigned to the corresponding assignment. Students who arrive 20 min after the scheduled lab time without a legitimate reason, leave the lab early without permission from the teaching assistant, or miss the lab without a legitimate reason will receive a zero for the corresponding laboratory assignment. Students who miss a lab with academic consideration are required to contact the course instructor within 3 days for further instructions. Failure to do so will result in a zero mark for that lab.

FINAL EXAMINATION: The final exam will take place during the regular examination period. The final exam will be three hours long, closed book. Only a basic, non-programmable calculator is allowed.

To obtain a passing grade in the course, a mark of 60% or more must be achieved on the final examination. A final examination mark < 60% will result in a final course grade of 48% or less. If the above conditions are not met, your final grade cannot be greater than 48%. Students who have failed this course (i.e., final average < 50%) must repeat all components of the course.

LATE SUBMISSION POLICY:

All lab assignments will be penalized by 20% of the available marks per day for late submission,. Assignments submitted more than 5 days late will not be accepted.

This course has 4 labs with only the best 4/5 labs counted towards your final grade. If students miss one lab, the remaining 4 labs will be used in the calculation of the final grade. If students miss greater than 1 lab, they will receive a grade of zero on each missed lab. Because of this, academic consideration will be granted in exceptional circumstances only. See https://www.eng.uwo.ca/undergraduate/academic-consideration-for-absences.html if more than one lab must be missed.

ATTENDANCE: Attendance is mandatory for all labs.

FACULTY OF ENGINEERING POLICIES:
Students must familiarize themselves with the policies of the Faculty of Engineering https://www.eng.uwo.ca/electrical/pdf/2024-UG-BOILERPLATE-OUTLINES.pdf