Special Topics in Computer Science:

Fair and Ethical Machine Learning for Social Good

CSCI-GA.3033-112 (8017)
Fall 2024

Course Information

Fair and Ethical Machine Learning for Social Good (“Fair and Ethical ML” for short) is a 3-unit, semester-long course, taught in Fall 2024.

Classes begin Monday September 9th and end Monday December 9th, with a poster session (course project presentations) during final exam week (Monday December 16th at 10am).

Course Schedule

Mondays, 10:15am-12:15pm (all times Eastern/NYC)
238 Thompson St (GCASL), Room 265, Manhattan

Instructor Information

- Professor Daniel Neill
- Email: daniel.neill@nyu.edu
- Office hours: Mondays and Wednesdays, 1:30-2:15pm, 60 Fifth Avenue #304

Course Description

Machine learning (ML) is increasingly used to inform high-stakes decisions that impact individuals' lives and livelihoods. For example, ML algorithms may determine who is given a loan, hired for a job, or sentenced to prison, or how scarce resources are allocated by a hospital or a government agency. It is critical to ensure that these decisions are made in a fair, responsible, and ethical manner, both maximizing the societal benefits and minimizing the harms of ML in practice. The first half of the course will focus primarily on fair ML: defining fairness, building fairer models, and auditing algorithms to identify and correct biases. Fair ML methods, embedded in a broader socio-technical perspective, can mitigate the risks that ML algorithms will create inequity and injustice or that they will exacerbate existing societal disparities. Moreover, fair ML can benefit society by correcting systematic biases in human decisions and by identifying and reducing societal disparities. The second half of the course will consider numerous other aspects of ML ethics (e.g., reliability and trustworthiness; privacy and security; interpretability and explainability; transparency and accountability), and will discuss practical challenges and solutions for measuring and optimizing the societal impacts of deployed ML systems. The course will include both lectures on core content and class discussion of current research papers. Deliverables will include paper presentations, short weekly assignments (including Python coding), and a substantial, semester-long research project.

Prerequisites

- Students are expected to have already taken at least one technical course in machine learning at graduate or advanced undergraduate level, and to be familiar with basic ML concepts. (For example, I expect to be able to teach concepts of fair classification without having to first explain “what is a classifier?”)
- Python programming including numpy, pandas, and scikit-learn. Students should already be able to implement standard ML classifiers (e.g., decision trees or random forests) using the scikit-learn library.

Learning Objectives

Upon completion of this course, the student will be able to:

1. Enumerate the various ethical questions that arise in the practical use of machine learning and artificial intelligence, along with potential challenges and solutions.
2. Articulate the differences between different definitions of algorithmic fairness, and choose and justify appropriate definitions for real-world problem domains (e.g., healthcare decision making).
3. Implement (in Python), apply (to real-world datasets), and evaluate technical approaches for (a) learning fair ML models; (b) auditing ML models for bias; and (c) mitigating bias.

Learning Assessment Table

Graded Assignment	**Course Objective Covered**
Participation	All
Homeworks and in-class exercises	All
Course Project	All

Relationship to Other Courses

Ethical ML has some overlap with the Responsible AI courses taught at NYU’s Center for Data Science (DS-UA 202 and DS-GA 1017) but goes into more technical detail on fairness (definitions, auditing, and learning fair models) and less detail on aspects of data management and legal/policy aspects of regulating the use of AI.

Course Materials

The course has one **required textbook**: *Fairness and Machine Learning: Limitations and Opportunities*, by Solon Barocas, Moritz Hardt, and Arvind Narayanan, aka the "Fair ML book". It is available freely online (as a PDF) at https://fairmlbook.org/pdf/fairmlbook.pdf.

I will also provide readings of current research papers (some required, some optional) corresponding to each class meeting. Finally, I recommend the following two **optional textbooks** for a broader (non-technical) perspective on the many ethical considerations surrounding machine learning and AI:

- *Moral AI: And How We Get There*, by Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong.
- *AI Ethics* by Mark Coeckelbergh.

All announcements, resources (including lecture slides, datasets, and supplemental readings), and assignments will be delivered through the NYU Brightspace site. I may modify assignments, due dates, and other aspects of the course as we go through the term, with advance notice provided as soon as possible through the course website.

Evaluation Method

Grades will be based on the following:

Class participation (including in-class exercises): 10%
Weekly homework assignments (including paper presentations): 30%
Project checkpoints (including project pitches): 10%
Final project posters/presentations: 15%
Final project reports: 35%

Much of your course grade will be based on a substantial, semester-long **course project** on a topic of your choice related to Fair and Ethical ML, to be done in teams of 2 students. Projects should attempt to extend the current state of the art in Fair and Ethical ML, and will be graded on four criteria: significance, novelty, correctness, and clarity.

Key project dates:

- Oct 21: Each team presents a 5-minute "project pitch" for feedback from the class.
- Oct 28: Each team submits a 1 to 2-page project proposal for instructor approval. The proposal should incorporate and address feedback from your project pitch.
- Dec 13: Final project reports due (ACM FAccT format and length restrictions)
- During final exam week: Class poster session (December 16th from 10:00-11:50am)

You can also expect **each week** to complete in-class responses (5 min), assigned readings, and short homework assignments (typically no more than 1-2 problems, plus completing the response if not submitted in class). Homework problems may involve Python coding and assume basic ML knowledge not covered in class. Each homework assignment is due at the **start** of class the week after it is assigned.

Grading Scale

Grade in Course	Points Earned
A	94 – 100
A-	90 – 93
B+	87 – 89
B	84 – 86
B-	80 – 83
C+	76 – 79
C	70 – 75
C-	65 – 69
F	Less than 65

Depending on the overall distribution of final numeric grades at the end of the semester, I may revise the grading scale to be more lenient than this distribution, but you are guaranteed at least the letter grade above for a given final average. This will be at my sole discretion (do not ask me for a higher grade) and all decisions are final.

Cheating and Plagiarism Notice

Projects will be done in teams of two students; we encourage discussion among teams, but any work that is submitted for grading must be the work of your team alone. Problem sets must be done individually, i.e., any work that is submitted for grading must be the work of that student alone. Sanctions for cheating include lowering your grade including failing the course. In egregious instances, the instructor may recommend the termination of your enrollment at NYU.

Please note that use of large language models (LLMs) such as ChatGPT to complete your assignments/projects is **not** allowed, and is considered to be academic dishonesty, unless you obtain permission **in advance** from the course instructor. Permission for certain uses (e.g., grammatical editing for non-native speakers, or course projects related to LLMs) may be granted, in which case you will be required to document these uses, including providing your prompts and the LLM outputs as supplementary material.

Late Work Policy

You are expected to turn in all work on time (at the start of class on the due date). Because we understand that exceptional circumstances may arise, each student will be permitted to turn in one assignment up to 48 hours late, or two assignments up to 24 hours late, with no penalty. Any other late assignments will not be accepted.

NOTE: Assignments turned in more than five (5) minutes after class starts will be counted as "late" and treated according to the Late Work Policy above.

Additional (School-Wide) Course Policies

NYU's Calendar Policy on Religious Holidays states that members of any religious group may, without penalty, absent themselves from classes when required in compliance with their religious obligations. Please notify me in advance of religious holidays that might coincide with exams or other major course events (e.g., course project poster session) to schedule mutually acceptable alternatives.

Academic accommodations are available for students with disabilities. Please visit the Moses Center for Student Accessibility (CSA) website and click on the Reasonable Accommodations and How to Register tab or call or email CSA at (212-998-4980 or mosescsa@nyu.edu) for information. Students who are requesting academic accommodations are strongly advised to reach out to the Moses Center as early as possible in the semester for assistance.

Course Outline – topics are tentative and subject to change!

Part I: Fairness in Machine Learning

Class 1 (9/9): Introduction to Fair and Ethical ML

- Course overview
- Benefits and Harms of AI/ML
- Ethical ML Principles
- Introduction to Algorithmic Fairness
- Philosophical Basis of Fair ML

Class 2 (9/16): Group Fairness in Classification

*** Homework due each week at the start of class; HW 1 due today! ***

- Defining Fairness: The COMPAS/ProPublica Debate
- Balanced Error Metrics
- Independence, Separation, and Sufficiency-Based Definitions
- Incompatibility Results (and How to Resolve Them)

Class 3 (9/23): Beyond Group Fairness

- Individual fairness
- Fair representations
- Pipelined fairness
- Broader conceptions of fairness

Class 4 (9/30): Learning Fair(er) Models

- Learning Fair Classifiers: pre-processing, in-processing, and post-processing methods

Class 5 (10/7): Auditing ML Models; Intersectionality and Subgroup Fairness

- Simple Auditing Approaches
- Intersectionality
- Auditing Approaches for Subgroup Fairness: fairness gerrymandering, multicalibration and multiaccuracy, bias scan and extensions.

Class 6 (***TUESDAY 10/15***): Causality and Fairness; Algorithmic Recourse

- Introduction to Causal Inference Frameworks
- Counterfactual Fairness
- Algorithmic Recourse

Class 7 (10/21): 5-minute Project Pitches

*** Students present their project pitches in class ***

Part II: Other Aspects of Ethical Machine Learning

Class 8 (10/28): Transparency, Accountability, Interpretability, and Explainability

*** Students submit their project proposals at the start of class ***

- Broad overview of these four closely related criteria, benefits and risks/drawbacks.
- Transparency: the degree to which a human can understand an AI/ML system.
- Accountability: the ability to determine whether a decision was made in accordance with standards and hold someone responsible if standards are not met.
- Interpretability: the degree to which a human can look at the internals of a model and readily understand how the model makes decisions.
- Explainability: the degree to which we can accurately and understandably answer questions about the way a complex model works.

Class 9 (11/4): Methods for Interpretable and Explainable ML

- In-depth look at various methods for learning interpretable models and explaining complex models (including class paper discussion).

Class 10 (11/11): ML Safety and Security

- Goal of ML safety: preventing harms from both present-day and future AI/ML systems.
- Overview of research areas: robustness, monitoring, alignment, systemic safety.
- Security: preventing attacks and manipulation by an adversary.
- Attacks and defenses: backdoors, data poisoning, adversarial examples, model inversion, and membership inference.
- Safety threats from rare, extreme events (“Black Swans”).
- Anomaly/OOD detection and uncertainty calibration.

Class 11 (11/18): AI Alignment

- Risks of misalignment, including existential risks to humanity
- Causes of misalignment (e.g., reward misspecification and goal misgeneralization)
- Approaches to alignment (e.g., learning from human feedback, scalable oversight, handling distribution shift, assurance, and governance)

Class 12 (11/25): Stability and Model Multiplicity

- 1st half of class: **guest lecture** on model multiplicity by Prof, Emily Black.
- 2nd half of class: additional lecture and discussion on stability and fairness, and finding fair and interpretable models within the Rashomon set.

Class 13 (12/2): Risks and Ethical Considerations for Large Language Models

- Brief overview of LLMs
- Taxonomy of LLM risks: discrimination, exclusion, and toxicity; information hazards; misinformation harms; malicious uses; human-computer interaction harms; automation, access, and environmental harms.
- Mitigating dishonesty and hallucinations in LLMs

Class 14 (12/9): Putting it All Together: Challenges and Solutions When Deploying ML Systems for Social Good

- Fair and Ethical ML course recap
- ML for Social Good: reflections and lessons learned
- What can we do to make deployed AI/ML systems fairer and more ethical?

*** Students submit their project reports by Friday December 13th at 11:59pm ***

Monday December 16th, 10:00-11:50am: Poster session for course project presentations!