Cornell
SC Johnson College of Business

NRE5220: Doctoral Seminar in Quant. Models

(a.k.a. Human-Algorithm Interactions)

Fall 2024, Wed. 11:40AM — 2:20PM, Sage 135 (link)

Tutorial Topic	Readings & Assignments
October 16: Improving Human Decisions	**Paper I**: Human decisions and machine predictions (link) **Paper II**: Diagnosing physician error (link)
October 23: Improving Human Decisions (contd.)	**Paper I**: Identifying prediction mistakes in observational data (link) **Paper II**: Machine learning as a tool for hypothesis generation (link)
October 30: Human-Algorithm Interactions	**Paper I**: Strategic classification (link) **Paper II**: Generalized strategic classification (link) **Paper III**: Performative prediction (link) **Paper IV**: Performative power (link)
November 6: Large Language Models	**Large language models primer**: Tutorial by Emaad, covers loss functions and pre-training, fine-tuning, and alignment (PPO and DPO) **Homework 1**: GPT (link) — due November 20, 11:59PM
November 13: Large Language Models as Simulated Humans	**Paper I**: What can we learn from *homo silicus?* (link) **Paper II**: Using LLMs for market research (link) **Paper III**: Can language models capture human preferences? (link) **Homework 2**: Attend talks by Sendhil Mullainathan this week (link)
November 20: Artificial Intelligence Alignment	**Paper I**: Consequences of misaligned artificial intelligence (link) **Paper II**: Incentive compatibility for AI alignment (link) **Paper III**: Model alignment as prospect-theoretic optimization (link) **Paper IV**: Model-agnostic alignment via Bayesian persuasion (link)
Thanksgiving Break	
December 4: Course Project Presentations	Course project writeup due before class on December 4 Time per presentation TBD based on the number of students enrolled

Last updated: October 2024

Course Description

This course is structured as a series of paper readings broadly focused on the interactions between humans and machine learning algorithms. In each session, students will lead in-depth discussions on recent research papers, with the rest of the attending group participating in discussions.

Learning outcomes: Upon completion of this course, you will be able to: (i) describe approaches to compare human decisions and algorithmic predictions; (ii) describe the key results of the strategic classification and performative prediction frameworks; (iii) describe the key results of the algorithmic triage and learning to defer frameworks; (iv) describe how large language models work; and (v) describe approaches to and current issues with aligning large language models.

Paper Presentations

The paper presentations are the most important and challenging component of this course. As a presenter, you need to understand not just the paper you are presenting, but also acquire the background required to deliver a *self-contained* presentation in class (*tip:* ask me for the most important background papers before you present). Prioritize depth over breadth: it is alright to skip some theorems, but *not* alright to present material that the audience does not learn from. Presentations do not require slides if you prefer using chalk or a whiteboard.

Date	Presenters	Time
October 16	**Paper I**: Mohammed Mosaffa **Paper II**: Sepehr Khadem Masjedi	75 minutes 75 minutes
October 23	**Paper I**: Lijuan Luo **Paper II**: Gonzalo Diaz	75 minutes 75 minutes
October 30	**Paper I, II**: Jongho Kim **Paper III, IV**: Meng Yang	75 minutes 75 minutes
November 6	**Large language models primer**: Emaad	150 minutes
November 13	**Paper I**: Via Wenya Guo **Paper II**: Suyeon Cho **Paper III**: Lena Kim	50 minutes 50 minutes 50 minutes
November 20	**Paper I, II**: Si Zuo **Paper III, IV**: Ziyi Lin	50 minutes 50 minutes 50 minutes

Course Project — An Aspirational Abstract & Introduction[1]

Write the abstract and introduction for a paper your *aspire* to write — feel free to make up actual results, within reason. The goal of the aspirational abstract and introduction is to test: if you *actually* did the work, would *you* be excited to read about it? If your imaginary results do not excite you, chances are that real ones will not either.

Here is a template I like to use for introductions (minus the budget and timeline): link. Your aspirational introduction should argue for why your research is important, challenging, and urgent. It should also include a selective literature review woven into the introduction (i.e., the most closely related work *in your research field* contrasted with what your paper aspires to do).

On December 4 during class, you will present your aspirational introduction as a set of "pitch" slides (similar to the introductory slides of an academic seminar talk) and respond to questions. The course project writeup is due before class on December 4.

Graded Components & Grading

1. *Presentations [50%]:* Presentations will be graded for depth and clarity, and *not* for breadth (i.e., non-key results in papers can be skipped), with the goal of ensuring collective learning.
2. *Course project writeup and presentation [20%]:* The project proposal writeup should follow the Overleaf "Science" LaTeX template and include (i) the title and abstract on the first page, (ii) an "aspirational introduction" at most 4 pages long.
3. *Paper summaries [10%]:* For every presented paper, all non-presenting students are required to submit a short summary of the paper. Optionally, summarize (i) one *previous* paper that the presented paper builds on, and (ii) one *subsequent* paper that builds on the presented paper.
4. *Homework [10%]:* One graded homework accounting for 10% of the total score. The homework will involve Python programming and be released as a Google Colab notebook.
5. *In-class participation [10%]:* Meaningful participation as non-presenters.

Letter grades will likely be assigned based on the following scale: (i) top 20% = A; (ii) next 20% = A-; (iii) next 50% = B+; (iv) bottom 10% = B or below. Letter grades will heavily depend on the quality of the paper presentation. Since this is a PhD-level class with typically low enrollment and no mandated GPA requirement, actual assigned grades tend to be on the higher end.

[1] This was inspired by the guide "4 steps to an applied micro paper" by Jesse M. Shapiro (link).

Other Administrivia

How Credit Hours are Met by the Course: This class meets for one, 160-minute class period each week for 7 weeks. Students are expected to work on course learning activities for about 3 hours out of the classroom in each of these 7 weeks.

Course Accommodations: Your access in this course is important to me. Please request your accommodation letter early in the semester, or as soon as you become registered with SDS (*https://sds.cornell.edu/*), so that we have adequate time to arrange your accommodations.

- Once SDS approves your accommodation letter, it will be emailed to both you and me. Please follow up with me to discuss the necessary logistics of your accommodations.
- If you are approved for exam accommodations, please consult with me at least two weeks before the scheduled exam date to confirm the testing arrangements.
- If you experience any access barriers in this course, such as with printed content, graphics, online materials, or any communication barriers; reach out to me or SDS right away.
- If you need an immediate accommodation, please speak with me after class or send an email message to me and SDS at *sds_cu@cornell.edu*.
- If you have, or think you may have a disability, please contact Student Disability Services for a confidential discussion: *sds_cu@cornell.edu*, 607-254-4545, *sds.cornell.edu*.

Inclusivity Statement: This course aims to collectively create a welcoming, supportive and tolerant environment for all students and respects the various forms of diversity that they bring, including differences related to race, gender, sexuality, class, nationality, geography, age, size, ability, etc. Towards this purpose, I ask that we be respectful of each other, actively listen, participate, ask relevant questions, and give balanced, specific, and constructive verbal & written feedback to each other. Please let me know if there are circumstances affecting your ability to participate.

Mental Health and Stress Management Resources: If you are feeling overwhelmed, or are worried about a friend, please reach out to one of your instructors or your academic advisor. We can try to help or we can put you in touch with someone who can help. Cornell has trained counselors available to listen and help: Empathy, Assistance, and Referral Service (213 Willard Straight Hall, 607-255-3277), Cornell Health's Counseling and Psychological Services (CAPS,

607-255-5155), and Let's Talk. The Learning Strategies Center offers a range of academic resources. Notably, Cornell has recently invested a great deal of time and effort to offer more comprehensive support for our campus community. See mentalhealth.cornell.edu to learn more.

Academic Integrity: Each student in this course is expected to abide by the Cornell University Code of Academic Integrity and the Johnson Honor Code. Any work submitted by a student in this course for academic credit will be the student's own work. The Johnson Honor Code is available here: https://www.johnson.cornell.edu/wp-content/uploads/sites/3/2021/10/Johnson-Honor-Code-VD.pdf

Intellectual Property: As a graduate school of business education, Johnson places a tremendous value on intellectual property, defined as "any product of the human intellect that the law protects from unauthorized use by others". As a future business leader, you should respect and protect intellectual property at Johnson and the University, as well as within the community of business scholars. This is the same behavior that will be expected of you in your future organizations. It is a theft of intellectual property to photocopy, scan, or otherwise unlawfully obtain course packets, course textbooks, solutions to assignments, etc., for this or any other Cornell class. It is also theft to distribute intellectual property without authorization (e.g. uploading assignments, answer keys, and/or exams to external websites, sharing exams or other materials with future students in the class or students in other programs). Stealing intellectual property or distributing intellectual property without authorization are considered violations of the Johnson Honor Code and of our community's ethical standards.