JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Syllabus
Computer Science EN.601.787
Advanced Machine Learning: Machine Learning for Trustworthy AI
Spring 2022 (3 credits, CSCI-RSNG, in person)

*(*The instructors reserve the right to make adjustments to this syllabus as deemed necessary with notice.)

Course Description

- This course teaches advanced machine learning methods for the design, implementation, and deployment of trustworthy AI systems. The topics we will cover include but are not limit to different types of robust learning methods, fair learning methods, safe learning methods, and research frontiers in transparency, interpretability, privacy, sustainability, AI safety and ethics. Students will learn the state-of-the-art methods in lectures, understand the recent advances by critiquing research articles, and apply/innovate new machine learning methods in an application. There will be homework assignments and a course project.
- **Prerequisites**
  601.475/675 or equivalent required, 601.476/676 and 601.482/682 or equivalent preferred, undergraduate could contact the instructor for approval, experience of machine learning research preferred
- **Limit: 15**

Instructors

Prof. Anqi Liu, Assistant Professor
aliu@cs.jhu.edu, https://anqiliu-ai.github.io/,
Office hours: After the class, 4:15-5:00 MW

Meetings

Hodson 303

Monday, Wednesday 03:00–04:15pm

Textbooks

There is no required textbooks, but the following books are recommended for background reading.

- Recommended: Michael Kearns and Aaron Roth. 2019. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA.
- Recommended: Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and Machine Learning. fairmlbook. org, 2019.
- Additionally, you will be expected to read various materials posted on the Blackboard.

Online Resources

The following sites will be used heavily during the course:

- Piazza will serve as our primary communication channel. Find our class signup link at: this link
- Blackboard (blackboard.jhu.edu) will be used for video distribution, assignment submission, grades and feedback.

Course Objectives

Upon successful completion of this course, you should be able to:

- Recognize various challenges on the machine learning methods imposed by trustworthy AI systems;
- Understand the state-of-the-art methods in machine learning for trustworthy AI;
- Investigate and critique the recent methods in a chosen sub-field (from robustness, fairness, safety, and so on);
- Apply recent advanced techniques in a chosen sub-field;
- Design and implement your own solutions to problems in trustworthy AI applications.

This course will address the following ABET Outcomes:

(1) (SO1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.
(2) (SO2) Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.
(3) (SO3) Communicate effectively in a variety of professional contexts.
(4) (SO4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.
(5) (SO5) Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.
(6) (SO6) Apply computer science theory and software development fundamentals to produce computing-based solutions.

Tentative Course Schedule

Week1: Class policy. Motivation. Introduction of various topics under the umbrella in trustworthy AI.
Week2: Guest Lecture on AI ethics, full paper list released and explained.
Week3: Statement of Interest due. ML Robustness (I) and (II). HW1 released.
Week4: Statement of Interest feedback. Uncertainty Quantification, Distribution Shift.
Week5: HW1 due. Paper choice due. ML interpretability (I) and (II). HW2 released.
Week6: Paper choice feedback. Fair ML (I) and (II).
Week7: HW2 due. Differential Privacy. Discussion class (I).
Week8: Paper critique presentation (I) and (II).
Week9: Spring Break. No Class.
Week10: Paper critique due. Safety considerations in sequential decision making (I) and (II)
Week11: Project proposal due. Human-in-the-loop Learning (I) and (II).
Week12: Proposal feedback. AI for social good. Discussion Class (II).
Week13: AI Safety, Additional topics that are of interest.
Week14 - Final exam weeks: Project Presentation. Project report due.

ACTUAL Course Schedule

Week1: Class policy. Motivation. Introduction of various topics under the umbrella in trustworthy AI.
Week2: Full paper list released and explained. Adversarial robustness
Week3: Statement of interest due. Distribution Shift. Distributional Robustness. HW1 released.
Week4: Statement of interest feedback. More Distributional Robustness. AI ethics guest lecture.
Week5: HW1 due. Paper choice due. Uncertainty quantification.
Week6: HW1 feedback. HW2 released. ML interpretability. Review start for mini-conference.
Week7: ML fairness. Presentation I for mini-conference. Paper review due.
Week8: Presentation II and III for mini-conference. Paper review feedback. HW2 due.
Week9: Spring Break. No Class.
Week10: Midterm review. ML fairness. Paper critique due.
Week11: Project proposal due. Human-in-the-loop Learning (I) and (II).
Week12: Proposal feedback. Safety considerations in sequential decision making, AI safety. Privacy
Week13: AI responsibility discussion. Project presentations.
Week14 - Final exam weeks: Project Presentation. Project report due.

Course Approach

One of the aims of this class is to inspire research interest and ideas in machine learning for trustworthy AI. We will combine lectures, paper discussions, and project presentations in our class. Students will be required to attend all the paper discussions and finish course projects in groups.

Trustworthy AI is generally a large area. At the beginning of the class, a statement of interest will be collected from the students. Students are expected to choose one or multiple topics that they have concerns about based on their own observations and expectations on trustworthy AI systems. The course materials would be slightly tailored to fit the students' interests better.

Expectations and Grading

Students will be expected to complete a variety of assignments, including reading suggested books and papers, finishing written homework, presenting paper critique, and implementing machine learning algorithms to solve real-world problems. Some assignments may be done in pairs or groups as specified; others must be completed alone.

Grading Breakdown

- Statement of interest: state what and why you think the proposed aspect is important to trustworthy AI, list at least one application that is associated with this aspect. (10%)
- Paper critique: choose one paper under the proposed topic and discuss the pros and cons of one current method developed, as well as difficulties and challenges that still exist. (20%)
- Written homework. (15%)
- Class participation and discussion. (10%)
- Project proposal (each group): propose a solution, plan out timeline and task assignment. Students will be encouraged to apply learned techniques in class, but are also welcomed to explore novel methods. (15%)
- Project report (each group), including the actual solution and its implementation details, result analysis, as well as the lessons learned in the project. (15%)
- Project presentation (each group): present the motivation, the design consideration, the solution, and the conclusions in front of the class and receive the feedback from the class. (15%)

All scores and grader commentary on your homework and project submissions, as well as exams, will be available via Blackboard. Please keep your own record of your grades so that you will know your standing in the course. At the end of the term, letter grades are generally assigned according to the following scale.

You should not expect a curve in this course.

$[97, 100]$: A+,	$[93, 97)$: A,	$[90, 93)$: A-
$[87, 90)$: B+,	$[83, 87)$: B,	$[80, 83)$: B-
$[77, 80)$: C+,	$[73, 77)$: C,	$[70, 73)$: C-
$[67, 70)$: D+,	$[60, 67)$: D,	$[0, 60)$: F

Classroom Climate

As your instructors, we are committed to creating a classroom environment that values the diversity of experiences and perspectives that all students bring. Everyone here has the right to be treated with dignity and respect. We believe that fostering an inclusive climate is important because research and my experience show that students who interact with peers who are different from themselves learn new things and experience tangible educational outcomes. Please join us in creating a welcoming and vibrant classroom climate. Note that you should expect to be challenged intellectually by instructors, the course assistants (CAs), and your peers, and at times this may feel uncomfortable. Indeed, it can be helpful to be pushed sometimes in order to learn and grow. But at no time in this learning process should someone be singled out or treated unequally on the basis of any seen or unseen part of their identity.

If you ever have concerns in this course about harassment, discrimination, or any unequal treatment, or if you seek accommodations or resources, we invite you to share directly with the instructor or CAs. We promise that we will take your communication seriously and to seek mutually acceptable resolutions and accommodations. Reporting will never impact your course grade. You may also share concerns with the Department Head (Randal Burns, randal@cs.jhu.edu), the Director of Undergraduate Studies (Joanne Selinski, joanne@cs.jhu.edu), the Assistant Dean for Diversity and Inclusion (Darlene Saporu, dsaporu1@jhu.edu), or the Office of Institutional Equity (oie@jhu.edu). In handling reports, people will protect your privacy as much as possible, but faculty and staff are required to officially report information for some cases (e.g. sexual harassment).

We hope that all students will be able to actively participate in all course meetings. Out of respect for your fellow students, be sure to attend on time each day. If you miss a classroom meeting for whatever reason, you are responsible for the material presented. Also, when attending zoom office hours please use your full name in your zoom account settings, and add your picture to your Zoom profile.

Personal Wellbeing

- If you are sick, please notify us by email so that we can make appropriate accommodations should this affect your ability to attend class, complete assignments, or participate in assessments. The https://studentaffairs.jhu.edu/student-health/Student Health and Wellness Center is open and operational for primary care needs. If you would like to speak with a medical provider, please call 410-516-8270, and staff will determine an appropriate course of action based on your geographic location, presenting symptoms, and insurance needs. Telemedicine visits are available only to people currently in Maryland. See also https://studentaffairs.jhu.edu/student-life/student-outreach-support/absences-from-class/illness-note-policy/
- The Johns Hopkins COVID-19 Call Center (JHCCC), which can be reached at 833-546-7546 seven days a week from 7 a.m. to 7 p.m., supports all JHU students, faculty, and staff experiencing COVID-19 symptoms. Primarily intended for those currently within driving distance of Baltimore, the JHCCC will evaluate your symptoms, order testing if needed, and conduct contact investigation for those affiliates who test positive. More information on the JHCCC and testing is on the https://covidinfo.jhu.edu/health-safety/johns-hopkins-covid-19-call-center/coronavirus information website.

- All students with disabilities who require accommodations for this course should contact us at their earliest convenience to discuss their specific needs. If you have a documented disability, you must be registered with the JHU Office for Student Disability Services (Shaffer 101; 410-516-4720; http://web.jhu.edu/disabilities/) to receive accommodations.
- Students who are struggling with anxiety, stress, depression or other mental health related concerns, please consider connecting with resources through the JHU Counseling Center. The Counseling Center will be providing services remotely to protect the health of students, staff, and communities. Please reach out to get connected and learn about service options based on where you are living this fall at 410-516-8278 and online at http://studentaffairs.jhu.edu/counselingcenter/.
- Student Outreach & Support will be fully operational (virtually) to help support students. Students can self-refer or refer a friend who may need extra support or help getting connected to resources. To connect with SOS, please email deanofstudents@jhu.edu, call 410-516-7857, or students can schedule to meet with a Case Manager by visiting the Student Outreach & Support website and follow "Schedule an Appointment".

Family Accommodations Policy

You are welcome to bring a family member to class on occasional days when your responsibilities require it (for example, if emergency childcare is unavailable, or for health needs of a relative). In fact, you may see children in class on days when their school is closed. Please be sensitive to the classroom environment, and if your family member becomes uncomfortably disruptive, you may leave the classroom and return as needed.

Deadlines for Adding, Dropping and Withdrawing from Courses

Students may add a course up to the end of the second week of class. For more information on these and other academic policies, see https://e-catalogue.jhu.edu/engineering/full-time-residential-programs/undergraduate-policies/academic-policies/grading-policies/

The Office of Academic Support at JHU

All programs are free to students. Please see below for specifics:

- PILOT Learning—Peer-Led Team Learning
  - Students are organized into small study teams who meet weekly to collaborate on faculty-developed problems-sets. Students work together as a team to solve problems.
  - A trained student leader acts as captain and facilitates the weekly meetings using various strategies to foster a collaborative learning environment.
  - Contact: Ariane Kelly <ariane.kelly@jhu.edu>
  - Instagram: @jhupilot
- Learning Den Tutoring Program - Small Group Tutoring
  - Small group, tailored tutoring of 4 students or less which is headed by one tutor. Visit the website (above) to access zoom links for drop-in sessions
  - Tutors can assist with but are not limited to:
    - Review and strengthening of subject-specific material knowledge
    - Assist with homework-like problems
    - Course-specific study skills and exam preparation
    - Contact: Kaitlin Quigley <quigley@jhu.edu>
    - Instagram: @jhulearningden
- The Study Consulting Program
  - Students work one-on-one with a study consultant to set academic goals and develop customized strategies for success. Areas addressed include but are not limited to:

    - ∗ Time management
    - ∗ Note taking and test preparation
    - ∗ Mastering large amounts of information
  - Contact: Dr. Sharleen Argamaso <sharleen.argamaso@jhu.edu>
  - Instagram: @jhustudyconsulting
- The Writing Center
  - Undergraduate and graduate students in KSAS/Whiting School of Engineering can schedule 50-min sessions with a Writing Center tutor to look over a draft of written work (up to 10 pages) or a personal statement for graduate study
  - Contact: Robert Tinkle <rtinkle1@jhu.edu>
  - Web Address: https://krieger.jhu.edu/writingcenter/

Ethics

The strength of the university depends on academic and personal integrity. In this course, you must be honest and truthful, abiding by the *Computer Science Academic Integrity Policy*:

> Cheating is wrong. Cheating hurts our community by undermining academic integrity, creating mistrust, and fostering unfair competition. The university will punish cheaters with failure on an assignment, failure in a course, permanent transcript notation, suspension, and/or expulsion. Offenses may be reported to medical, law or other professional or graduate schools when a cheater applies.
>
> Violations can include cheating on exams, plagiarism, reuse of assignments without permission, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition. Ignorance of these rules is not an excuse.
>
> Academic honesty is required in all work you submit to be graded. Except where the instructor specifies group work, you must solve all homework and programming assignments without the help of others. For example, you must not look at anyone else's solutions (including program code) to your homework problems. However, you may discuss assignment specifications (not solutions) with others to be sure you understand what is required by the assignment.
>
> If your instructor permits using fragments of source code from outside sources, such as your textbook or on-line resources, you must properly cite the source. Not citing it constitutes plagiarism. Similarly, your group projects must list everyone who participated.
>
> Falsifying program output or results is prohibited.
>
> Your instructor is free to override parts of this policy for particular assignments. To protect yourself: (1) Ask the instructor if you are not sure what is permissible. (2) Seek help from the instructor, TA or CAs, as you are always encouraged to do, rather than from other students. (3) Cite any questionable sources of help you may have received.
>
> On every exam, you will sign the following pledge: ”I agree to complete this exam without unauthorized assistance from any person, materials or device. [Signed and dated]”. Your course instructors will let you know where to find copies of old exams, if they are available.

In addition, the specific ethics guidelines for this course are as follows:

(1) This course is about learning, and encourages collaboration toward that end. In general, if a collaborative act helps you to learn, it is probably permitted. If, on the other hand, it helps you avoid learning, it is not permitted. For example, helping your friend learn how to use the debugger is great. Helping your friend by debugging their code for them is bad, because your friend will never learn how to do it by watching you. A main focus of this course is learning skills, and you can't acquire

skills without practice. Therefore, ”helping” other students by allowing them to skip the practice endangers the learning outcomes of the course, and is prohibited. Helping other students practice more efficiently and effectively (e.g. not waste 3 hours trying to fix one bug), on the other hand, actively supports the learning goals of the course, and is not only allowed, but encouraged.

(2) In general, when helping another student, never do something for them; instead, try to think like a teacher and ”teach” them how to do it themselves. This will help you both learn, since teaching something is a great way of learning more about it as well.
   - Asking a friend to let you look at their working code is not allowed, nor is offering to let someone else look at your working code.
   - If a friend is helping you to debug, you should only share the minimal amount of buggy code necessary.
   - In general, when helping others, think ”teach, not ”do”.
   - Always thank anyone who helped you on a given assignment in your README file.
   - The two coding projects for this course will allow for and encourage collaboration in teams. There are no limits to communication within a team, but work should be done ”as a team” with all members present; don’t just split the assignment into pieces and work on separate parts. No working source code should be shared outside your team.

Report any violations you witness to the instructor.

You can find more information about university misconduct policies on the web at these sites:

- For undergraduates:
  https://studentaffairs.jhu.edu/policies-guidelines/undergrad-ethics/
- For graduate students:
  http://e-catalog.jhu.edu/grad-students/graduate-specific-policies/