Applied Ethical and Governance Challenges in AI
Harvard Law School and MIT Media Lab, Spring Semester 2019

Course Heads: Jonathan Zittrain & Joi Ito
TAs: John Bowers (Harvard), Natalie Saltiel (MIT), and Samantha Bates (MIT)

Tuesdays 5:00-7:00 PM
WCC 4063
Office hours to be announced

COURSE OVERVIEW

This course will pursue a cross-disciplinary investigation of the development and deployment of the opaque complex adaptive systems that are increasingly in public and private use. We will explore the proliferation of algorithmic decision-making, autonomous systems, and machine learning and explanation; the search for balance between regulation and innovation; and the effects of AI on the dissemination of information, along with questions related to individual rights, discrimination, and architectures of control.

The structure of the course is somewhat nontraditional (see the "Schedule Overview" section) –This process will be bound together by a central theme – tradeoffs in the design of AI systems. Fixing AI's problems will mean imposing burdens and constraints on the performance of critical systems, but the costs of sticking with the status quo might be much worse.

We will apply this analytical process to 3 distinct problem areas, committing one diagnosis session, one prognosis session, and one intervention session to each. Our three problem areas are *fairness, interpretability,* and *adversarial example attacks*. The first two are well developed within the AI literature and have been rearing their ugly heads in the real world for years, but are still host to numerous unresolved questions. The last is still largely *terra incognita* – adversarial examples haven't yet made it out of the lab, but there are strong indications that they may do so soon. The literature around adversarial examples is highly technical, and needs much more attention from lawyers and social scientists. As such, it could very well provide fertile ground for meaningful scholarly contributions on the part of students.

It is important to note that many of the readings assigned for this course are quite technical in nature. We assume no computer science or mathematics background whatsoever, and simply ask that you give all of the material a shot. Feel free to skip math and figures that you can't understand, but do try to dip a toe in where possible!

COURSE MATERIALS

We have worked to ensure that the readings for this course are very manageable. As discussion will be tightly focused around assigned readings, we require that students prepare carefully and completely for each session. Given our discussion-based seminar format, failure to complete the readings will be both obvious and an impediment to the flow of class.

Most course readings – except for those distributed in print – are available on this course's H2O playlist: https://bit.ly/2SJU8Lq. Links to readings are also provided in the reading list that follows. Students who would prefer a physical copy of the readings can pick up a free printed packet at the Harvard Law School Copy Center in the basement of the Wasserstein Community Center. Enrolled students will be informed when these packets become available.

ASSESSMENT

In addition to class participation, students will be assessed on the basis of final projects structured around problems and challenges encountered throughout the course. Many students will choose to write a 12-15 page paper to meet this requirement, but other types of projects (technical, artistic, etc.) are also welcome. The TA staff will meet with students individually to define and scope these projects partway through the semester.

Please feel free to contact John Bowers (Harvard) <jobowers@law.harvard.edu> or Natalie Saltiel (MIT) <nsaltiel@media.mit.edu> with any questions or concerns relating to the course. Our first class session will be held on January 29th, 2019.

SCHEDULE OVERVIEW

1. Introduction (no guest) [1/29]

Part I: Diagnosis

2. Fairness (Cathy O'Neil) [2/5]
3. Interpretability (with Zach Lipton) [2/12]
4. Adversarial Examples (with LabSix) [2/19] **(Course Dinner 7:00-8:00)**

Part II: Prognosis

5. Fairness (with Solon Barocas) [2/26]
6. Interpretability (with Sandra Wachter) [3/5]
7. Adversarial Examples (no guest) [3/12] **(Course Dinner 7:00-8:00)**

Part III: Intervention

8. Fairness (with Sendhil Mullainathan and Rodrigo Ochigame) [4/2]
9. Interpretability (Guest TBD) [4/9]
10. Adversarial Examples (no guest) [4/16] **(Course Dinner 7:00-8:00)**

READINGS AND SESSION INFORMATION

Class Session 1: Introduction [1/29]
Our first class session will be used primarily to describe and discuss the structure and motivations behind the course, and to offer a high-level perspective on the bodies of work we will be examining throughout the term.

- "Artificial Intelligence—The Revolution Hasn't Happened Yet" by Michael Jordan, Medium (April 2018)
- "Troubling Trends in Machine Learning Scholarship" by Zachary C. Lipton & Jacob Steinhardt (July 2018)

Part I: Diagnosis
Class sessions 2-4 will constitute our effort to consider pressing problems in AI through a *diagnostic* lens, identifying their social, technical, and philosophical roots. We'll apply this analysis to our three major topics – Fairness, Interpretability, and Adversarial Examples – in succession. As you complete the readings and participate in the class sessions under this heading, consider the ways in which the predicates of power in AI systems can be responsible for their most harmful behaviors. To what extent are the problems we are confronting separable from the fundamental mechanics of the systems they affect? And what implications do these tight connections have for the sorts of tradeoffs we might need to take to get AI technology under control?

Class Session 2: Diagnosing problems of fairness [1/29]
For our second class session, we will be joined by Cathy O'Neil, a data scientist and activist who has become one of the leading voices on fairness in machine learning. As you read a selection from Cathy's book *Weapons of Math Destruction*, think about what fairness means in the context of AI/ML, and why we are so bad at embedding it within our systems.

- *Weapons of Math Destruction* by Cathy O'Neil, Broadway Books (2016). Read Introduction and Chapter 1: "Bomb Parts: What Is a Model?"
- [OPTIONAL] "The scored society: due process for automated predictions" by Danielle Keats Citron and Frank Pasquale, Washington Law Review (2014)

Class Session 3: Diagnosing problems of interpretability [2/12]
This session will be joined by Zachary Lipton, an Assistant Professor at Carnegie Mellon University who is working intensively on defining and addressing problems of interpretability in machine learning. We ask that you carefully consider what it means for a model to be interpretable, and how should we think about interpretability in relation to other desirable attributes.

- "The Mythos of Model Interpretability" by Zachary C. Lipton, ArXiv (2016)
- [OPTIONAL] "Towards a rigorous Science of Interpretable Machine Learning" by Finale Doshi-Velez and Been Kim, ArXiv (2017)

Class Session 4: Diagnosing vulnerabilities to adversarial examples [2/19]

Our first session on adversarial examples will include a presentation by LabSix, a student-run AI research group at MIT which is doing cutting-edge work on adversarial techniques. In addition to this technical primer on adversarial examples, we will discuss some of the potential domains in which adversarial attacks might prove particularly destructive.

- ➢ "Motivating the Rules of the Game for Adversarial Example Research" by Justin Gilmer et al., ArXiv (2018).
- ➢ [RECOMMENDED] "Intriguing properties of neural networks" by Christian Szegedy et al., ArXiv (2013)

Part II: Prognosis

In class sessions 5-7, we will build on our diagnostic efforts by moving towards a *prognostic* approach to think about the human implications of the problems we have identified. As before, we will move through each of our three topics in succession. Continue to think about these problems in terms of tradeoffs – in human terms, what do we stand to gain or lose by changing our approach to AI research and deployment? What are – or will be – the costs imposed by the status quo?

Class Session 5: Prognosticating the impacts of unfair AI [2/26]

Solon Barocas, an Assistant Professor at Cornell University, will be joining us for the kickoff session of Part II. The paper assigned for today takes a blended legal and technical approach to identifying key problems of fairness and equity raised by the deployment of autonomous systems in sensitive capacities. How might the disparate impact framing influence how we define – and maybe even address – fairness in AI?

- ➢ "Big Data's Disparate Impact" by Solon Barocas and Andrew D. Selbst, California Law Review (2016)

Class Session 6: Prognosticating the impacts of uninterpretable AI [3/5]

Our sixth session will be joined by Sandra Wachter, a lawyer and research fellow at the Oxford Internet Institute. We'll turn towards governance for the session – the assigned paper explores the power of interpretability through the notion of counterfactual explanations in the context of the GDPR. How does Wachter's framework guide our understanding of how interpretability might actually be useful in motivating our actions?

- ➢ "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" by Sandra Wachter et al., Harvard Journal of Law and Technology (2018)
- ➢ [OPTIONAL] "Algorithmic Transparency for the Smart City" by Robert Brauneis & Ellen P. Goodman, Yale Journal of Law and Technology (2018)

Class Session 7: Prognosticating the impacts of adversarial examples [3/12]

This session will continue our study of adversarial examples by examining some potential scenarios in which it could be leveraged to inflict harm. The readings offers two examples of probable ramifications of adversarial examples in an increasingly ML-driven society – increased opportunities for fraud and complex legal implications.

- ➢ "Adversarial attacks on artificial intelligence systems as a new healthcare policy consideration" by Samuel Finlayson, Joi Ito, Jonathan Zittrain et al., preprint (2019)
- ➢ "Law and Adversarial Machine Learning" by Ram Shankar Siva Kumar et al., ArXiv (2018)

Part III: Intervention

Sessions 8-10 will close out the course by investigating some potential means of intervening against the problems identified in the preceding classes – and against the futures those problems implicate. Here our discussion of tradeoffs is at its most concrete. All of the interventions we consider would require engineers and users to make difficult choices between performance and other criteria – adherence to fairness constraints, say. What is the best approach to making these decisions? Are there any general rules that can be observed across application domains, or must we operate on a case-by-case basis (and, if so, are there any heuristics we might benefit from keeping in mind)?

Class Session 8: Intervening on behalf of fairness [4/2]

The hardest part of enforcing fairness constraints on AI systems is deciding what fairness actually means, and coming up with a suitable mains of representing that definition. This class session will be structured as a conversation involving two guests, University of Chicago Booth School of Business Professor Sendhil Mullainathan and MIT PhD student Rodrigo Ochigame. Both study problems of fairness in machine learning. What implications do Mullainathan's findings have for tradeoff decisions in system design? Are you swayed by Ochigame's argument regarding irreduciblity? If so, how might it modulate our approach?

- ➢ "Inherent Trade-Offs in the Fair Determination of Risk Scores" by Jon Kleinberg, Sendhil Mullainathan, et al., ArXiv (2016)
- ➢ [SUPERSEDING PAPER TO COME] "Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning" by Rodrigo Ochigame et al., ICML (2018)

Class Session 9: Intervening on behalf of interpretability [4/9]

What if we stopped settling for finding predictively powerful correlations in data and insisted on identifying true causal connections? The growing discipline of causal inference, a long-running but increasingly popular statistical approach which aims to identify and make interventions on the basis of causality, aims to do just this. This week's reading comes from one of the field's best and most approachable texts.

- ➢ *The Book of Why* by Judea Pearl and Dana Mackenzie, Basic Books (2018). Read Introduction.
- ➢ [OPTIONAL] "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment" by Joichi Ito, Jonathan Zittrain, et al. *Proceedings of Fat**, (2018)

Class Session 10: Intervening against adversarial examples and course conclusion [4/16]

Our final class session will place us on the frontier of the conversation around resilience against adversarial example attacks. What are some of` the inherent difficulties relating to intervening against a problem that has not yet manifested itself strongly in the wild? Why might it be

difficult to adapt our existing systems to this new threat, particularly given the concepts of technical debt described in the reading? We will also reserve some time to reflect on the broader themes and findings of the course.

- ➢ "Hidden Technical Debt in Machine Learning Systems" by D. Sculley et al., NIPS (2015)