1. **Course.** CS 181: Machine Learning

2. **Course Level.** Upper-level Undergraduate

3. **Course Description.** "Introduction to machine learning, providing a probabilistic view on artificial intelligence and reasoning under uncertainty. Topics include: supervised learning, ensemble methods and boosting, neural networks, support vector machines, kernel methods, clustering and unsupervised learning, maximum likelihood, graphical models, hidden Markov models, inference methods, and computational learning theory. Students should feel comfortable with multivariate calculus, linear algebra, probability theory, and complexity theory. Students will be required to produce non-trivial programs in Python."(Course Description)

4. **Module Topic.** Machine Learning, Policing, and Discrimination

5. **Module Author.** Diana Acosta-Navas

6. **Semesters Taught.** Spring 2017-2018, Spring 2018-2019, Spring 2019-2020

7. **Tags.** machine learning [CS], discrimination [phil], disparate treatment [phil], disparate impact [phil]

8. **Module Overview.** The module analyzes the potential consequences of using location-based predictions to channel law enforcement efforts and resources. The module focuses on the case of *PredPol* to reveal the discriminatory potential of machine learning algorithms when employed in contexts that meet specific conditions. Based on this case analysis, students learn to discern between discriminatory intent (or disparate treatment) and discriminatory impact (or disparate impact). Through this discussion, students are prompted to reflect on how algorithms operate in specific social, political and economic contexts, and how their potential for wrongful discrimination changes as a result of the context in which they are implemented.

9. **Connection to Course Material.** This topic connects to course content about bias and discrimination as basic operation principles of machine learning algorithms. The module addresses the question of how technical bias can give rise to different forms of discrimination.

10. **Module Goals**

1. Distinguishing between disparate treatment and disparate impact.
2. Understanding how software that does not engage in disparate treatment can wrongfully discriminate individuals on the basis of protected attributes.
3. Identifying relevant features (e.g. social, political, economic) by virtue of which an algorithm may wrongfully discriminate.
4. Recognizing the tension that arises when algorithms offer the possibility to correct for human biases and thus employ critical resources more effectively, on the one hand; and, on the other, pose the risk of reinforcing existing patterns of discrimination.

11. **Key Philosophical Questions**

1. What constitutes wrongful discrimination?

2. Can decisions be wrongfully discriminatory even if they fail to satisfy the *disparate treatment* definition of discrimination?
3. Why is disparate impact ethically problematic?
4. Is indirect discrimination equally problematic when it impacts vulnerable and non-vulnerable populations?
5. Are there ethical reasons to use these algorithms *despite* their discriminatory potential?

12. **Key Philosophical Concepts.**

- Disparate treatment
- Disparate impact
- Direct discrimination
- Indirect discrimination

13. Assigned Readings.

● Lum, K., & Isaac, W. (2016). “To predict and serve?” *Significance*, 13(5), 14-19.	This paper examines the results of applying *Predpol’s* algorithm to a biased police data set. The authors feed the algorithm with the historical data of drug arrests by the Oakland Police Department. The results are contrasted against a high-resolution prediction of drug-use based on the 2011 National Survey on Drug Use and Health. Though the latter predict drug use to be evenly spread across the city, *Predpol*’s algorithm consistently directs policing efforts towards low-income, majority black and Latino neighborhoods.

14. Class Agenda

1. Wrongful discrimination
2. Case Study1: *Predpol*
   Direct Discrimination vs. Indirect Discrimination
   Disparate Treatment vs. Disparate Impact
3. Case Study2: *White Collar Crime Early Warning System* (a parody)
   The role of context in enhancing an algorithm’s discriminatory potential

15. Sample Class Activity

During the course of the module students are polled on the potential for wrongful discrimination of two algorithms: (1) *Predpol*; and (2) *White-Collar Crime Early Warning System.*

The latter is a parody of the former and is meant to draw attention to some problematic consequences of using machine learning algorithms to allocate law enforcement resources.

After the second poll, students are asked to pair and discuss the reasons why someone may have different intuitions in these cases. They are then asked to share their ideas with the class and prompted to reflect on the importance of context, in particular the existence of historical disadvantage, to enhance the discriminatory potential of machine learning algorithms applied in the context of policing practices.

16. Module Assignment.

Prompt: Imagine that you are a product manager in a technology company and the board of directors requests a detailed report of the potential social impact of the product you're building. You are currently working on an algorithm designed to allocate economic resources for various health providers. You are instructed to address various kinds of possible impacts, giving special attention to the algorithm's potential to wrongfully discriminate against vulnerable populations.

Having followed the coding process closely, you can confidently assume that there is no discriminatory intent animating the design. In fact, the product's design is based on the notion that resources should be distributed to health centers in a manner that is proportional to the needs of the populations they attend to. However, you worry that there may still be wrongful discrimination due to disparate negative impact on members of low-income populations, migrants, and racial minorities.

What further questions must you address in order to confidently determine whether the algorithm wrongfully discriminates against members of these groups? Write two questions and, for each, write a short paragraph explaining how addressing this question can help you assess the algorithm's potential for wrongful discrimination.

17. Lessons Learned.

Student response to this module was overall positive. A few lessons stand out.

- Student engagement could be further motivated with the following strategies: (1) introducing active learning exercises during the early stages of the module; (2) reducing the philosophical content, thus leaving more time for in-depth discussion and student questions.
- Philosophical distinctions could be further illuminated through the use of examples.
- Greater attention and time could be directed towards the ethical tensions that arise from the use of predictive policing algorithms.
- Students found it difficult to understand the second case-study and its role as a parody of existing predictive policing algorithms. The framing of the case and the associated discussion prompt could incorporate more scaffolding. Otherwise, it could be replaced by a more realistic case study that requires less framing.