**Topic:** The Impact of AI Ethics on AI-Powered Education

**Algorithmic Fairness/Explainability**

1. Friedman, B., & College, C. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3).
   1. Early framework specifying how bias can show up in computer systems, describing the categories of preexisting, technical, and emergent bias (Table 1). Counts freedom from bias as one important criterion for technical systems.
   2. Bounds of bias that we care about. Namely, systematic computer errors do not in essence involve unfairness (p. 333)
   3. The 7th grade gender bias section (p. 345).
2. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. <https://doi.org/10.1145/2090236.2090255>.
   1. Early framework for achieving statistical parity in a computational algorithm. Early fairness in classification work with an educational application (university acceptance).
   2. Works to achieve individual fairness.
   3. The Lipschitz Condition. Talk about how we are mapping to distributions. It is distributions all the way down
   4. Quantitative trade-off between fairness and utility
   5. Solving the local issue always. Not particularly solving the global problem.
   6. Also related to the videos I watched by Dwork
3. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. <https://doi.org/10.1007/s10115-011-0463-8>
   1. An early educational piece that outlines some techniques to transform data for less discrimination.
   2. Four methods of preprocessing data for non-discrimination: suppression, massaging the dataset, reweighing, sampling
   3. Theoretical analysis of the accuracy: discrimination trade-off (p. 8). DA-optimal (discrimination and accuracy) classifiers can be found from the perfect classifier with a linear trade-off. Linear trade-off is also found with an imperfect classifier.
   4. Massaging the dataset means changing labels. Use a ranker to determine probabilities of class membership and swap labels of the ones that have the lowest probabilities.
   5. Conclude that if minimal discrimination is the first priority, an unstable classifier, i.e., a classifier more sensitive to noise as a base learner, is the better option (p. 26).
4. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *I*nherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*. arXiv. <http://arxiv.org/abs/1609.05807>
   1. A formalization of the three fairness ideals that are most commonly found in algorithmic fairness discussion. Also focuses on the unsatisfiability of all three ideals. All three are only satisfiable with a perfect prediction, or equal base rates.
   2. (p. 7) related work section has many good examples of early fairness literature. (p. 8) describes incompatibility proof
   3. Unless the identity assignment happens to be fair, every fair assignment must have larger loss than that of *I*, forcing a tradeoff between performance and fairness. (p. 13) Where *I* is the assignment where each score has its own bin.
   4. Section 4.1 could be useful for ICMI fairness analysis (paragraph 1 in particular)
   5. Well-calibrated: if the algorithm predicts probability for a group, then a *z* fraction of that group should indeed be positive instances (p. 2). Given some outcome, the probability that a member in the outcome is a member of a group is independent of the outcome.
   6. Balance for the positive class/balance for the negative class
5. Green, B., & Hu, L. (2018). The myth in the methodology: towards a recontextualization of fairness in machine learning. *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*.
   1. Discusses the current state of how we measure fairness with statistical metrics, and it concludes that true social deliberation must exist for fair systems.
   2. Defines an inequality between the fairness metrics and true normative fairness that we value. Unfairness is diagnosable in context…but fairness is hard to pin down - it required values, principles, and commitments that only have substance in specificity.
   3. Two primary technical approaches to achieving fairness in ML. Procedural and statistical.
   4. Section 2.2 the terms with their statistical definitions is helpful and important to memorize.
   5. Changes in society cannot be reflected in ML systems. Conclusion is powerful.
6. Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, 22–27. <https://doi.org/10.1257/pandp.20181018>
   1. A piece that recontextualizes classical algorithmic fairness work with a focus towards more humanistic fairness insights. Also describes how fairness through unawareness does not lead to more fair algorithms.
   2. More equity could come from including protected attributes in the prediction task. (Section II)
   3. Three predictors, race-blind, unfairness mitigated, and race-aware.
   4. Also a fairness vs accuracy type of piece.
7. Barocas, S., Hardt, M., & Narayana, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. Retrieved from <https://fairmlbook.org/>
   1. A textbook that describes the fundamental technical aspects of fair machine learning.
   2. Chapter 1: Should automated decision making even be applied in this context? Two types of arbitrary decisions. Three types of automation.
   3. Classification is often attacked by first solving a regression problem to summarize the data in a single real-values risk score. We then turn the risk score into a classifier by thresholding. (Chapter 3)
8. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1, 206-215.* <http://arxiv.org/abs/1811.10154>
   1. Describes the difference between explaining black box models and creating/using more interpretable models. Illuminates the importance of interpretability in machine learning.
   2. Explainable vs interpretable. Explanations don’t actually explain anything. They just give summary statistics of the predictions. Thus the explanation could be wrong.
   3. There is not a necessary tradeoff between accuracy and interpretability
   4. Simple model more accurate than COMPAS
   5. Goodhart’s law
   6. <https://arxiv.org/pdf/1801.01489.pdf>
9. Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review (arXiv:2006.00093). arXiv. <http://arxiv.org/abs/2006.00093>
   1. A review of popular explainable AI methods.
10. Khan, F. A., Manis, E., & Stoyanovich, J. (2021). Fairness as equality of opportunity: normative guidance from political philosophy(arXiv:2106.08259). arXiv. <http://arxiv.org/abs/2106.08259>
    1. A piece that moves algorithmic fairness outside of a technical ideal, and describes how fair machine learning can benefit from more interdisciplinary contextualization.
    2. Formal EOP is equivalent to fairness through blindness.
    3. The SAT lacks test validity as it is not equally effective of predicting college success across all groups.
    4. Substantive EOP: Rawl’s fair EOP vs luck-egalitarian EOP
       1. Rawl: Equally talented babies should have equal life prospects
          1. Statistical parity/equality of odds
       2. Roemer: luck egalitarian
          1. Comparing apples to oranges is hard
    5. Rawl’s veil of ignorance
    6. This whole piece is super important
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. <https://doi.org/10.1145/3457607>
    1. A survey of fair machine learning techniques and applications.
12. Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: revisiting the impossibility theorem in practice. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 400–422. <https://doi.org/10.1145/3593013.3594007>
    1. There can be relaxations of the impossibility theorem in fairness where multiple metrics can be optimized for. This is done through deciding the level in which the prevalence of the data is and how large a difference is okay to imply the size of a “fairness region”, and determining how inaccurate a model is allowed to be to still imply a large “fairness region”.
13. Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). Humans, AI, and Context: understanding end-users’ trust in a real-world computer vision application. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 77–88. <https://doi.org/10.1145/3593013.3593978>
    1. A recent example of researchers analyzing trust in AI systems. This piece helps represent opportunities for future research in how AI systems and people interact.
    2. Bird paper
14. Baron, S. (2023). Explainable AI and causal understanding: counterfactual approaches considered. *Minds and Machines*, *33*(2), 347–377. <https://doi.org/10.1007/s11023-023-09637-x>
    1. A more modern approach to explainable AI which argues for causal understanding in some counterfactual explainable AI approaches.

**Education**

1. Siemens, G. (2013). Learning Analytics: the emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
   1. Describes how learning analytics was first created and what the specific goals and unique challenges are for learning analytics.
   2. The slighted move in the virtual landscape is paid for in lines of code
   3. Learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs.
      1. I care about optimizing learning for marginalized peoples
   4. Techniques vs applications
   5. Very unique problems with temporal and multimodal data.
   6. The dark side section for some philosophy
2. Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics*, *3*(1). <https://doi.org/10.18608/jla.2016.31.3>
   1. Develops an ethical ideal for learning analytics with the goal of upholding the morality and legality of learning analytics.
   2. Goal was to remove barriers for the adoption of learning analytics.
   3. Table 1 that splits the responsibilities of different ethical questions in learning analytics.
   4. I wish to add a section in the validity portion about fairness and explainability.
3. Doroudi, S., & Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335–339. <https://doi.org/10.1145/3303772.3303838>
   1. Analysis of the fairness of using knowledge tracing, a method that I have used extensively in my PhD research.
   2. Emphasizing making sure that learning analytics algorithms act fairly with different populations of students. This paper specifically deals with high-performing versus low-performing students. Slow vs fast learners.
   3. Misspecified models versus models that do not individualize.
4. Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)fairness: lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. <https://doi.org/10.1145/3287560.3287600>
   1. Describes how the education community has had definitions of unfairness for much longer than machine learning. Focuses on how the definitions do or do not overlap.
   2. Mid-1970s section (2.3) references the values that underpin what people determine as fair.
   3. “How do these groups differ under existing cultural conditions? (3.1)
   4. Separation, sufficiency, and independence mentioned
   5. GAPS IN ML: Unfairness versus fairness in machine learning. What different types of unfairness could be present. DIF analogues in ML
   6. In the discussion, I am particularly focused on point 3. Point 4 is discussed in Baker’s “should demographics be used” paper…
5. Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology*, *50*(6), 2855–2870. <https://doi.org/10.1111/bjet.12868>
   1. A modern piece that proposes edge cases that ethical guidelines in learning analytics must grapple with.
   2. Consequentialism, deontological, **virtue ethics**
   3. Who can define “the good”
   4. Three basic ethical principles
   5. *This tension describes the trade-offs—common to all human research—inherent in application of consequentialist principles; how are we to balance each principle, and in particular how are we to balance the minimisation of harm and the maximisation of benefits? In our view, learning analytics has tended to emphasise the former with relatively little discussion of the latter*
6. Paquette, L., Li, Z., Baker, R., Ocumpaugh, J., & Andres, A. (2020). Who’s learning? Using demographics in EDM research. *Journal of Educational Data Mining* *12*(3).
   1. Surveys the usage of sensitive features in educational data mining research.
7. Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining*, *15*(2), 22–52. https://doi.org/10.5281/zenodo.7702628
   1. A thought piece on the philosophy behind using demographic data in educational data mining.
8. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
   1. A review of algorithmic bias methods and applications in the field of education.
9. Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *4*, 100130. <https://doi.org/10.1016/j.caeai.2023.100130>
   1. A modern piece discussing how students ought to adapt to new AI-based technologies that are emerging.
10. Yanagiura, T., Kihira, M., Yano, S., & Okada, Y. (2023). Examining algorithmic fairness for first-term college grade prediction models relying on pre-matriculation data. *Journal of Educational Data Mining 15*(3).
    1. Recent piece that uses a real-world educational dataset that analyzes definitions of fairness on the data and prediction methods. Synonymous to how educational data can be used in the real-world.

**AI Ethics**

1. So, R. J. (2017). All Models Are Wrong. *PMLA/Publications of the Modern Language Association of America*, *132*(3), 668–673. <https://doi.org/10.1632/pmla.2017.132.3.668>
   1. A look at statistical modeling from a researcher in the humanities. Describes where models are useful and where they are not.
2. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
   1. A framework of how AI systems ought to interact with society that supports ethical AI design.
   2. Strong correlation to principles in bioethics. But we willfully cede some of our decision making power to AI (autonomy). Explicability is the unique one to AI.
   3. Justice ideal… To eliminate discrimination from the Asilomar Principles
   4. Are we the patients receiving the “treatment” of AI, or the doctor prescribing it? (p. 699)
3. Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536. <https://doi.org/10.1007/s11948-017-9975-2>
   1. A look at how computational bias impacts different systems, and the piece describes general ways to mitigate biases in such technologies.
   2. Almost a survey of some of the ways bias has found its way into AI systems and provides some real problems/possible opportunities for dealing with the bias
4. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
   1. Describes the nuance of the AI ethics discussion from a holistic perspective outside of only academia. This piece has a main focus on industries and how they act within AI ethical principles.
   2. Five principles: transparency, justice and fairness, non-maleficence, responsibility and privacy
   3. Five years ago - more conversations have happened
5. Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
   1. States the nuance of the term AI, and how the general public perceives and does not perceive AI.
6. Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
   1. Describes that future XAI must involve interpretive models from the ground up rather than using the most optimized models.
7. Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
   1. Focuses on how explainable AI has a true challenge when it comes to modern deep learning models.
   2. DNN can find significant features, but it is not clear how to describe what was found without an understanding of the relationships. DNN cannot communicate linguistically, yet we must.
   3. Specifically, information discovery can be a very real problem.
   4. Predictions vs explanations. What matters more?
8. Kerr, A. D., & Scharp, K. (2022). The end of vagueness: technological epistemicism, surveillance capitalism, and explainable artificial intelligence. *Minds and Machines*, *32*(3), 585–611. <https://doi.org/10.1007/s11023-022-09609-7>
   1. Describes an epistemic outcome of the growth of big data and methods to synthesize that data. Showing that ideals within philosophy ought to grow alongside technology becoming more powerful.