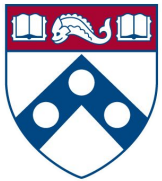
The Next Step in AI Transparency: Taking Standards from Generalities to Reality

Christopher S. Yoo
John H. Chestnut Professor of Law, Communication, and Computer & Information Science

March 15, 2024



Center for Technology, Innovation & Competition
University of Pennsylvania

Introduction

- Proposals call for transparency without providing details (e.g., U.S. Executive Order, NIST Standards, U.S. AI Bill of Rights, OECD Recommendation, Bletchley Declaration, model cards)
- My talk will review three major elements of AI transparency
  - Algorithms
  - Training data
  - Testing and validation
- It will also discuss the necessity of ongoing review

Transparency Regarding Algorithms

- The opaqueness of bare algorithms (inputs, not outputs)
- Limits of bans on including protected classes as criteria
  - The possibility of proxies and omitted variable bias
  - The need to correct biases in training data
  - Foreclosure of focusing on equality of outcome
  - Negative impact on predictive powers of models
- Legal constraints: trade secrets, government privilege
- Compromise of business models

Center for Technology, Innovation & Competition
University of Pennsylvania

Transparency Regarding Training Data

- Source of the training data
- Quality of the training data (limits of adding more data)
  - Potential bias
  - Scope (e.g., LLM hallucination, weather) and other *V*s
  - Rare events and long-tail effects
  - Structural changes (e.g., collapse of LTCM, Zillow iBuying)
- Need to address legal constraints (e.g., de-duplication, personal information)

Transparency Regarding Testing/Validation

- Different approaches to testing
  - IEEE AV: design processes, formal methods, robustness analysis, simulation/closed course/public road testing
  - Limits of approaches (e.g., U.S. seatbelt testing)
- Difficulties in specifying outcomes (Goodhart's Law)
  - Limits to the solution space (e.g., Tetris pause, pancake flip)
  - Specification gaming/reward hacking (e.g., CycleGAN cheating)

The Need for Ongoing Evaluation

- Little guidance in current debate
- Learning nature of AI
- Reality that complex systems have emergent properties
  - Proxy discrimination
  - Adverse environments (e.g., Microsoft's Tay)
  - Multiagent interactions (e.g., flash crashes, Lambrecht & Tucker study, other black swan events)
  - Hallucinations

Conclusion

- Standards need to provide more details regarding ex ante transparency on algorithms, training data, and testing
- AI accountability also needs ex post assessment
- Closing thought: need to frame issues in terms of optimality
  - Compliance is costly, e.g., HIPAA right to know data sharing
  - Predictive analytics are probabilistic/must have permissible error rate

Thank you!

Questions/Discussion?

Center for Technology, Innovation & Competition
University of Pennsylvania