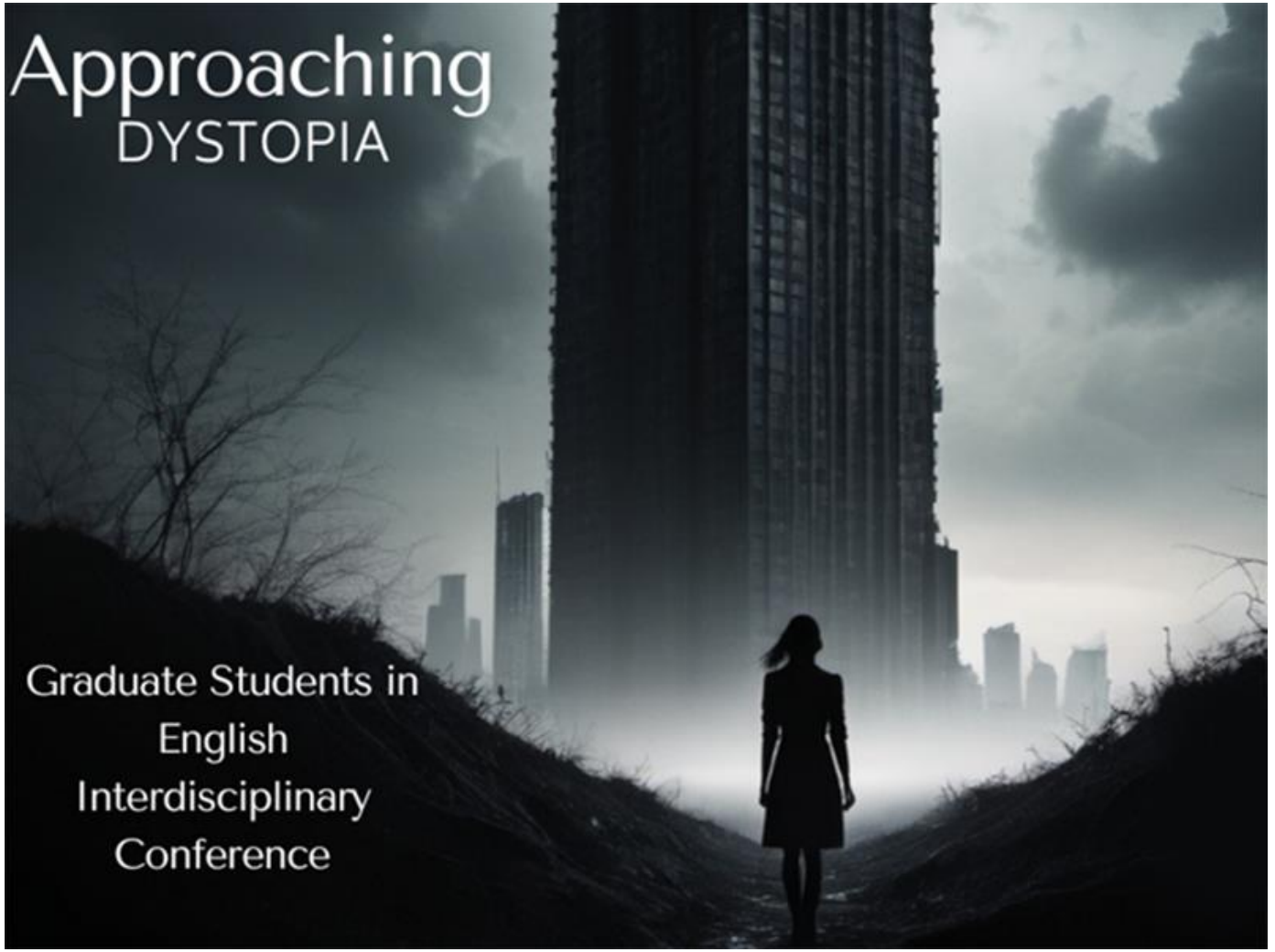
Trustworthy ML Reading Group

31 Jan 2024

Conduct

- **From Ethos of MLC…**
  - https://mlcollective.org/wiki/code-of-conduct/
- **Highlights**
  - Expectation of Confidentiality
  - Reporting -> send me a direct message

Coming Up



Contact: Emily Birtwistle -> ebb006@uark.edu

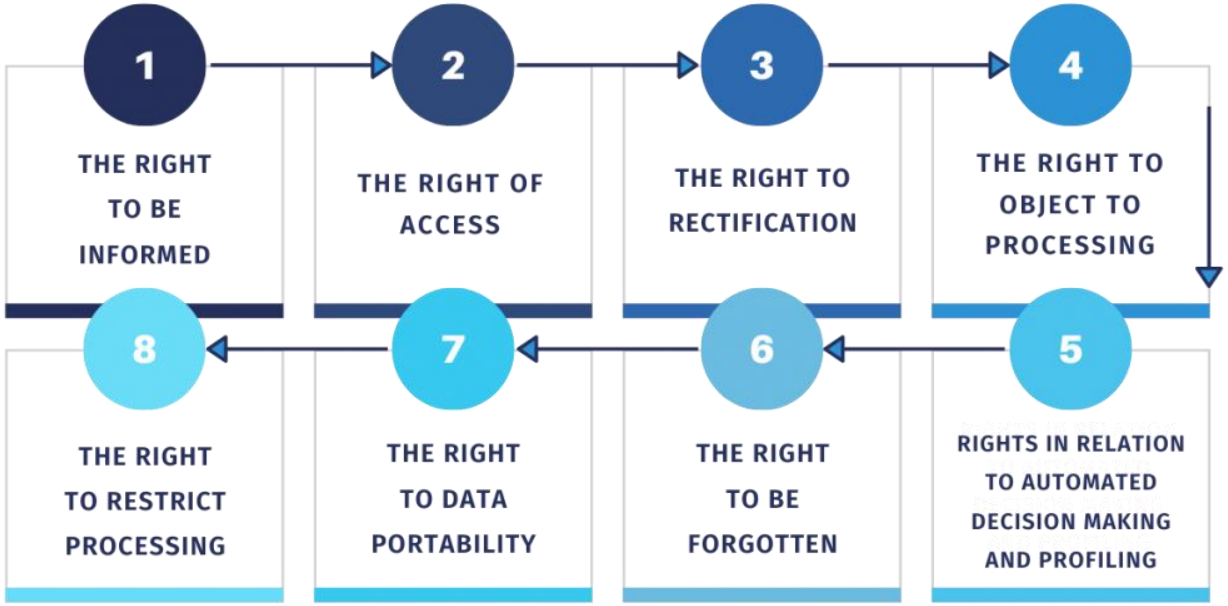
https://news.uark.edu/articles/71087/call-for-papers-approaching-dystopia-graduate-students-in-english-conference-2025

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez and Been Kim. 3/2017. “Towards A Rigorous Science of Interpretable Machine Learning”.

Motivations for Explanation

- The right to explanation (be informed)



Questions the Author Consider

1. Do all applications have the same interpretability needs?
2. What justifies using explanations

What is Interpretability

- Feature-Based Explanations
- Causal Explanations
- Concept based Explanations
- Global vs Local explanations

2. What justifies using explanations?

- Incomplete problem formalization
  - Lack of Scientific Understanding
  - Safety Risk
  - Ethics (e.g., bias)
  - Mismatched objectives
  - Competing objectives

Evaluation Approaches

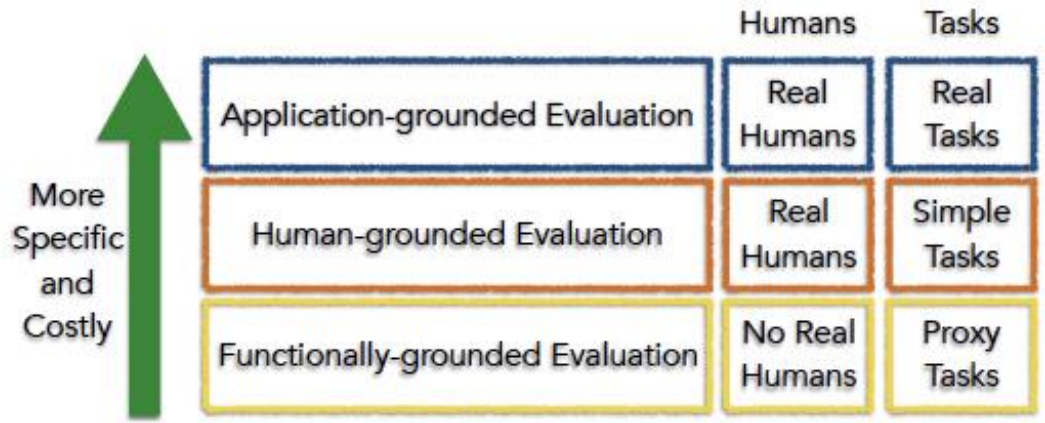


Figure 1: Taxonomy of evaluation approaches for interpretability

Open Interpretability Questions

- What Proxies are best for what real-world applications?
- What are the important factors to consider when designing simpler tasks that maintain the essence of the real end-task?
- What are the important factors to consider when characterizing proxies for explanation quality?

Quantifying Explanations as a Task

- Global vs. Local explanations
- Area, Severity of Incompleteness
- Time Constraints (for end users)
- Nature of User Expertise

Determining Method of Explanation

- What is the basic unit of explanation that can be used?
  - E.g., Features, Prototypes, etc.
- How complex can an explanation be?
  - How much information should be included in an explanation
- How should explanations be structured
  - What metrics may be helpful?
- Should explanations be linear or nonlinear?
  - What explanation methods are most natural to an end user?
- How should Uncertainty be quantified?

Suggested Starting Points

- How is the problem formulation incomplete?
- What level is evaluation being performed at (application, general user study, proxy)?
- What are task-related relevant factors (global vs. local explanation, severity of incompleteness, level of user expertise, time constraints)
- What are method-related relevant factors?
  - How complex does explanation need to be, monotonicity, uncertainty, compositionality