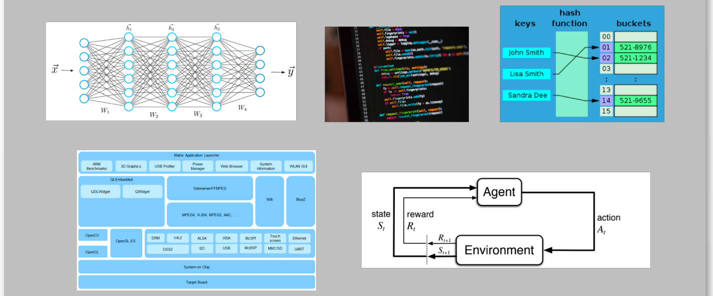
Reflexive Design for Fairness and Other Human Values in Formal Models

Benjamin Fish & Luke Stark

Problem

The social impacts of automated unfairness and other forms of discrimination in AI systems are of increasingly urgent public concern



- **"Fair" computational models often fail to satisfy even their own limited criteria for fairness when deployed**
- **There are few specific methods for ensuring human values are built adequately into models**

Prior Approaches

DEFINITION

A binary classifier $\hat{Y}$ satisfies (M, m)-individual fairness if for every $x, y \in X$,

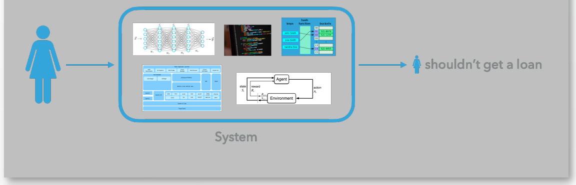
$$M(\hat{Y}(x), \hat{Y}(y)) \leq m(x, y),$$

where M is a statistical distance and m is a metric.

- **Machine learning models are often generic and domain-independent (e.g. binary classification)**
- **"Abstraction traps" (Selbst et al. 2019) a major problem: how to get around them?**

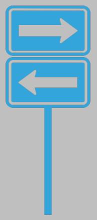
Reflexive Values

- **Our contribution: highlighting four reflexive values to guide model design, to help clarify:**
- **a) does model bear a reasonable relation to the human values it schematizes?**
- **b) is model used and useful for a purpose which in turn supports those same values?**



Value Fidelity & Accuracy

- **Value Fidelity: A reflexive assessment of the context/domain for your formal model. Do they align?**



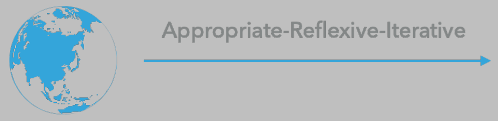
- **Appropriate Accuracy: Do your data proxies and model mechanics actually represent the value to be modelled?**

Value Legibility & Contestation

- **Value Legibility: are broader consequences of a formal model's design and deployment modeled or considered?**
- **Value Contestation: are you aware/flexible to conflicts around the normative valence of particular models?**

Reflexive Values in Design Practice

What guidance for the incorporation of human values into formal models do we provide modelers?



- **pre-design stage: assess whether it is appropriate to design or deploy a formal model in the first place**
- **design stage: determine what and how to model based on reflexive values (value fidelity, accuracy, legibility, and contestation)**
- **post-design stage: work iteratively on evaluation, and maintenance, and potential modifications with reflexive values in mind**

Mila Western