Paper Summary

Fairness and Abstraction in Sociotechnical Systems

ANDREW D. SELBST, Data & Society Research Institute
DANAH BOYD, Microsoft Research and
Data & Society Research Institute
SORELLE A. FRIEDLER, Haverford College, PA
SURESH VENKATASUBRAMANIAN, University of Utah
JANET VERTESI, Princeton University

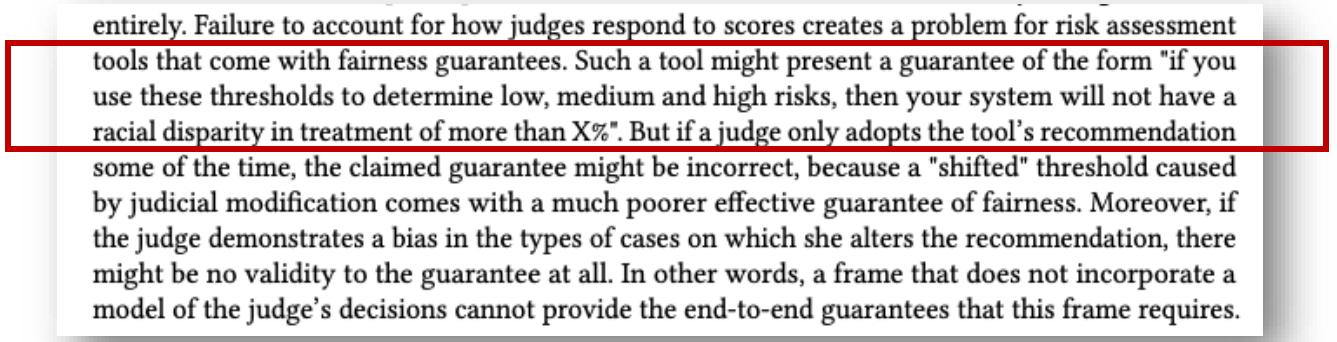
A key goal of the fair-ML community is to develop machine-learning based systems that, once introduced into a social context, can achieve social and legal outcomes such as fairness, justice, and due process. Bedrock concepts in computer science—such as abstraction and modular design—are used to define notions of fairness and discrimination, to produce fairness-aware learning algorithms, and to intervene at different stages of a decision-making pipeline to produce "fair" outcomes. In this paper, however, we contend that these concepts render technical interventions ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems. We outline this mismatch with five "traps" that fair-ML work can fall into even as it attempts to be more context-aware in comparison to traditional data science. We draw on studies of sociotechnical systems in Science and Technology Studies to explain why such traps occur and how to avoid them. Finally, we suggest ways in which technical designers can mitigate the traps through a refocusing of design in terms of process rather than solutions, and by drawing abstraction boundaries to include social actors rather than purely technical ones.

Key points

- To achieve social and legal outcomes such as fairness, justice; ML systems should consider their social context.
- Five traps that fair-ML researchers could fall into:
  The Framing Trap, The Portability Trap, The Formalism Trap, The Ripple Effect Trap, The Solutionism Trap.
- Focus on process not the solution! Draw abstraction boundaries carefully.

The Framing Trap

"Failure to model the entire system over which a social criterion, such as fairness, will be enforced"



entirely. Failure to account for how judges respond to scores creates a problem for risk assessment tools that come with fairness guarantees. Such a tool might present a guarantee of the form "if you use these thresholds to determine low, medium and high risks, then your system will not have a racial disparity in treatment of more than X%". But if a judge only adopts the tool's recommendation some of the time, the claimed guarantee might be incorrect, because a "shifted" threshold caused by judicial modification comes with a much poorer effective guarantee of fairness. Moreover, if the judge demonstrates a bias in the types of cases on which she alters the recommendation, there might be no validity to the guarantee at all. In other words, a frame that does not incorporate a model of the judge's decisions cannot provide the end-to-end guarantees that this frame requires.

The Portability Trap

"Failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context"

The Formalism Trap

"Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms"

The Formalism Trap

- Procedurality: The biggest difference between law and the fair-ML definitions is that the law is primarily procedural, and the fair-ML definitions are primarily outcome-based.
- Contextuality: Wrongful discrimination is better defined in their cultural context. Law falls short in incorporating ideas about discrimination.
- Contestability: Discrimination and fairness are politically contested and shifting (e.g., legal definitions, social norms).

The Ripple Effect Trap

"Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system"

The Solutionism Trap

"Failure to recognize the possibility that the best solution to a problem may not involve technology"

We need to reflect on the potential of technology to improve the current situation.

The Solutionism Trap

When is wrong to start with technology?

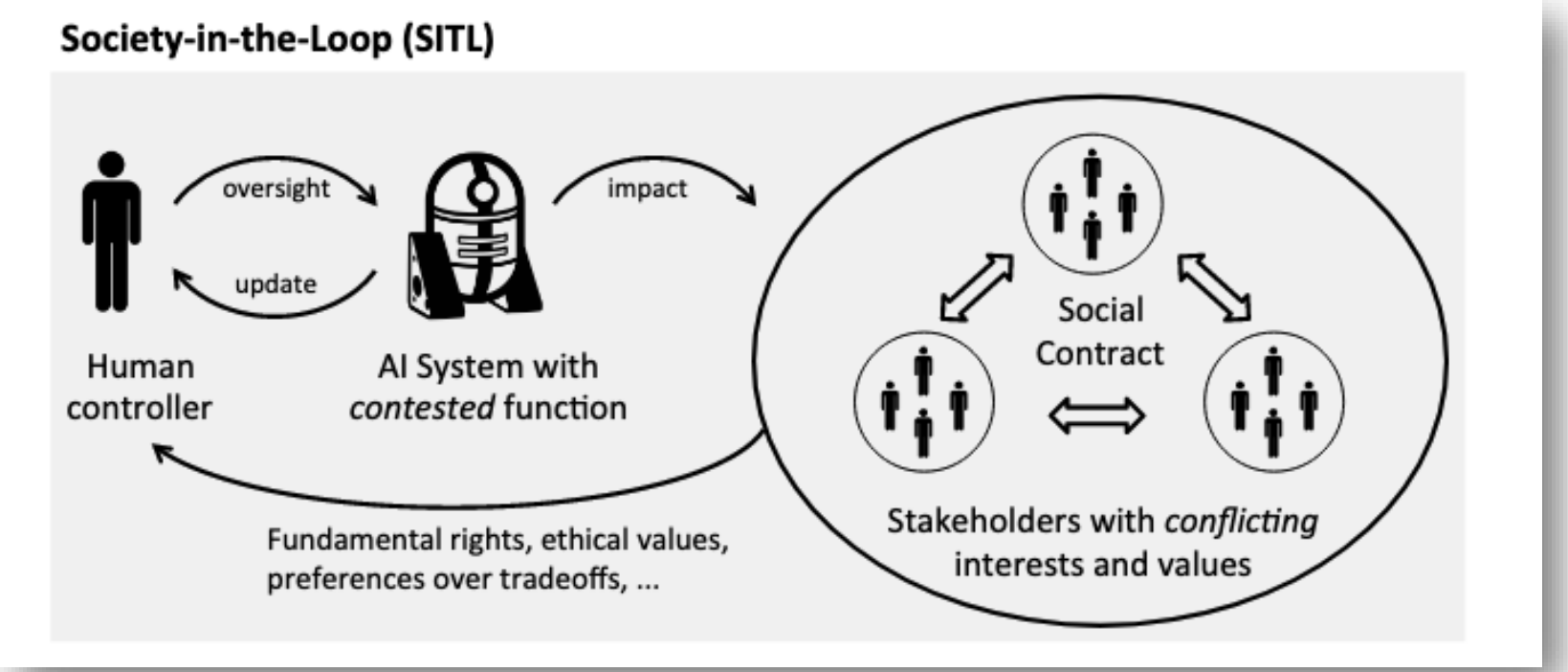
- Formalism Trap: Modelling requires pinning down definitions, that are changing and can be politically contested.
- Complexity leads to computational intractability.

A Science and Technology Studies (STS) Lens on all traps

What Fair-ML Researchers Can and Should Do

(1) is appropriate to the situation in the first place, which requires a nuanced understanding of the relevant social context and its politics (Solutionism);
(2) affects the social context in a predictable way such that the problem that the technology solves remains unchanged after its introduction (Ripple Effect);
(3) can appropriately handle robust understandings of social requirements such as fairness, including the need for procedurality, contextuality, and contestability (Formalism);
(4) has appropriately modeled the social and technical requirements of the actual context in which it will be deployed (Portability); and
(5) is heterogeneously framed so as to include the data and social actors relevant to the localized question of fairness (Framing).

Recommended Reading



Rahwan, Iyad. "Society-in-the-loop: programming the algorithmic social contract." Ethics and Information Technology 20.1 (2018): 5-14.

Discussion Questions

1. Consider the Portability trap. How could a fair-ML researcher ensure that an approach works properly in their social context?
2. What do you think about the feasibility of the five-trap approach recommended for Fair-ML researchers?

Feel free to discuss the questions by using AI examples.