System Dynamics Society

Online Poster Presentation

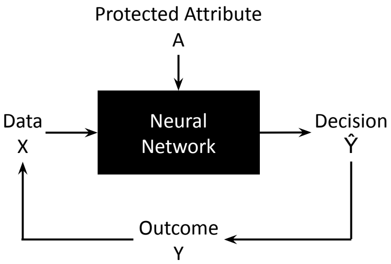
A Systems Thinking Approach to Algorithmic Fairness

Chris Lam, Epistamai

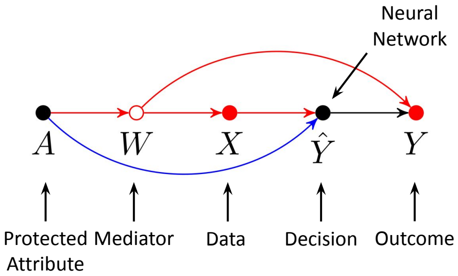
Introduction

To build fair machine learning systems in highly regulated domains, we need to translate fairness into a ***complex systems*** problem by “thinking outside of the black box.”

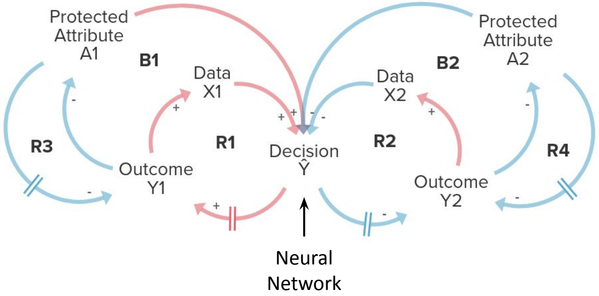
Machine Learning
Neural Network



Causal Inference
Causal Bayesian Network



System Dynamics
Causal Loop Diagram

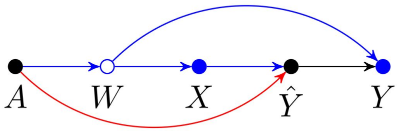


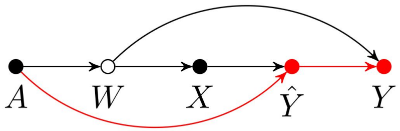
Application	Protected Attribute A	Mediator W	Data X	Decision Ŷ	Outcome Y
Credit scoring	Race, gender	Creditworthiness	Income, credit history	Deny loan?	Loan default?
Resume screening	Race, gender	Qualifications	Experience, education	Screen out resume?	Employee turnover?
College admissions	Race, gender	Merit	Grades, test scores	Reject applicant?	Student failure?

Modeling Fairness as a Linear System

We can use ***causal Bayesian networks*** to visualize bias in a machine learning model, understand how the model causes discrimination, and perform interventions to make fair decisions.

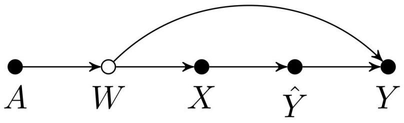
Fairness through supremacism
(Far-right politics)

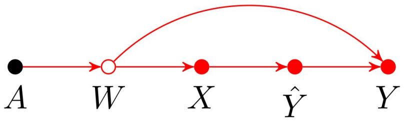




Overt discrimination
(e.g. Disparate treatment)

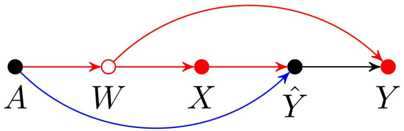
Fairness through unawareness
(Right-wing politics)

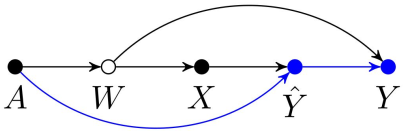




Covert discrimination
(e.g. Disparate impact)

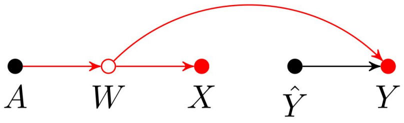
Fairness through affirmative action
(Left-wing politics)

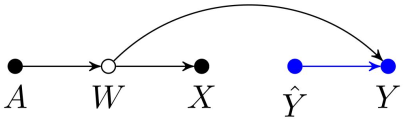




Reverse discrimination

Fairness through lottery
(Far-left politics)



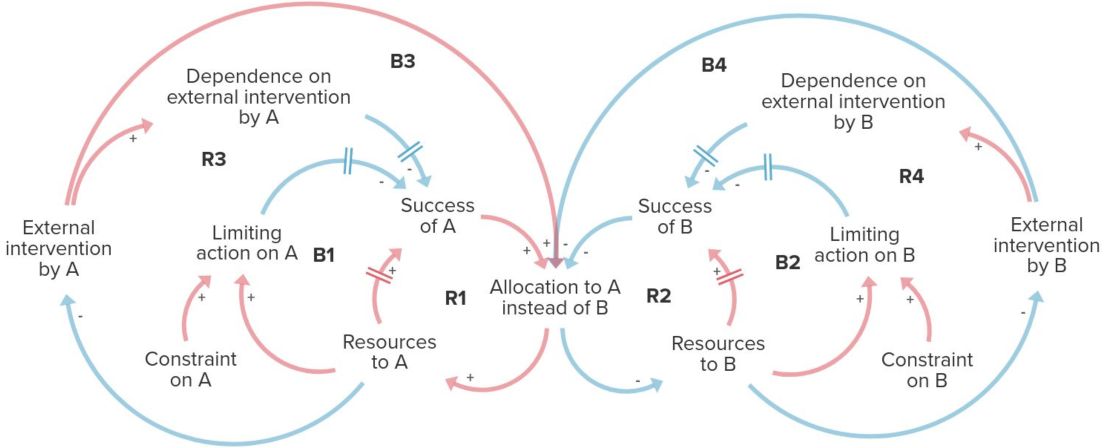


“No” discrimination

Modeling Fairness as a Nonlinear System

We can use ***causal loop diagrams*** to model bias from the data generating process and to identify counterintuitive behavior using system archetypes.

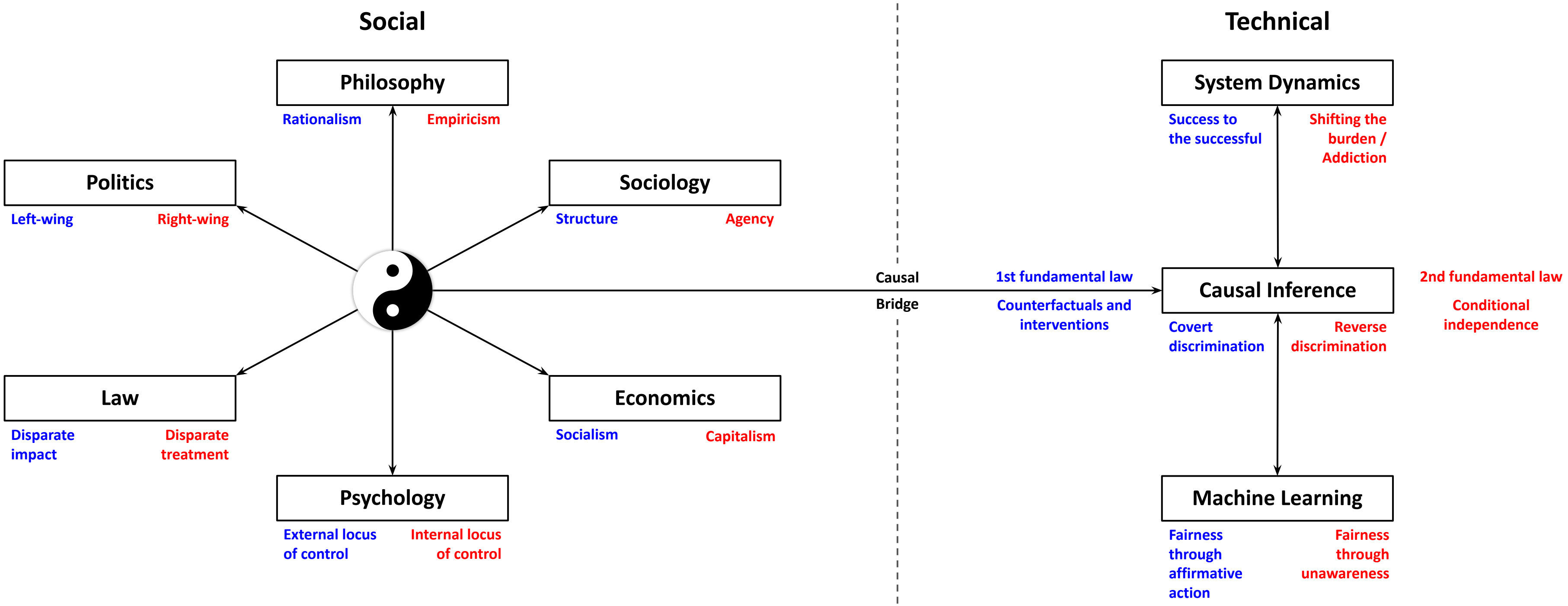
Let's say that there are two groups. Group A has historically had more resources and success than Group B, which leads to the following:


B3
Dependence on external intervention by A
R3
External intervention by A
Limiting action on A
B1
Success of A
Constraint on A
Resources to A
R1
Allocation to A instead of B
R2
Resources to B
Success of B
B4
Dependence on external intervention by B
R4
Limiting action on B
B2
External intervention by B
Constraint on B
+
-

Fairness through unawareness		Fairness through affirmative action	
Success to the successful (R1/R2)	**Limits to success (R1/B1 and R2/B2)**	**Shifting the burden / Addiction (R1/B3/R3 and R2/B4/R4)**	
Group A's historical success means more resources are allocated towards Group A (R1) over Group B (R2), thus reinforcing Group A's success over Group B.	As Group A receives more resources (R1), it may face a limiting action due to some constraint. This reduces their success which results in less resources being allocated to Group A (B1).	As Group B receives fewer resources (R2), there is greater demand for an external intervention to allocate more resources towards Group B (B4).	But the external intervention may create a dependence that harms the success of Group B. This decreased success causes less resources to be allocated to Group B (R4).

Modeling Fairness as a Complex System

We can model the sociotechnical nature of the fairness problem using a ***systems map***. On the left, we can model the social aspect of fairness as a ***connection circle***. On the right, we can model the technical aspect of fairness as an ***hierarchy***.



Social
Philosophy
Rationalism
Empiricism
Politics
Left-wing
Right-wing
Sociology
Structure
Agency
Law
Disparate impact
Disparate treatment
Economics
Socialism
Capitalism
Psychology
External locus of control
Internal locus of control
Causal
Bridge
1st fundamental law
Counterfactuals and interventions
Technical
System Dynamics
Success to the successful
Shifting the burden / Addiction
Causal Inference
2nd fundamental law
Conditional independence
Covert discrimination
Reverse discrimination
Machine Learning
Fairness through affirmative action
Fairness through unawareness