Private Sector AI: Ethics and Incentives

Tom Slee

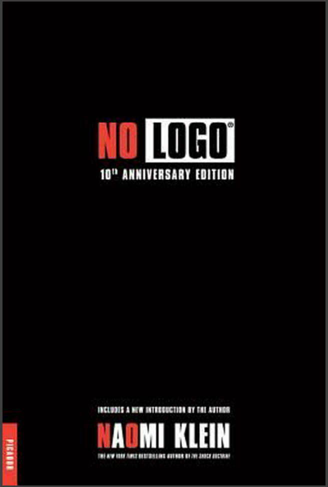
March 13, 2019

Private sector AI: ethics and incentives

1. Limits of the ethical algorithm
2. Elasticity and accuracy
3. Incompatible incentives
4. Algorithms demand rules
5. Rules create temptations
6. Governing algorithmic governance

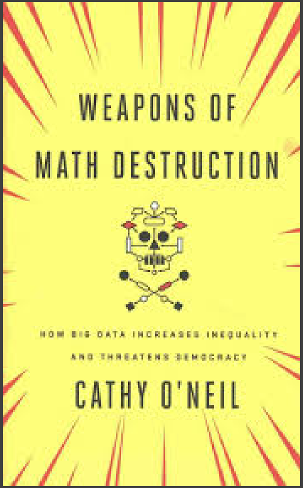
The values gap (1999)

- Brand values: personal empowerment
- Supply-chain management values: sweatshops



The values gap (2019)

- Brand values: Don't be evil
- Algorithmic values: black-boxes and bias



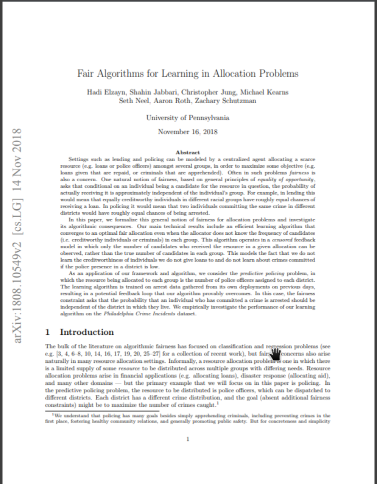
Private sector responses

Themes

- Assert responsible stewardship
- Build values into software

Actions

- Statements of principle
- Industry bodies ()Partnership on AI) and advisory councils
- Investments in FAT-ML



arXiv:1808.10549v2 [cs.LG] 14 Nov 2018

Fair Algorithms for Learning in Allocation Problems

Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns
Seth Neel, Aaron Roth, Zachary Schutzman

University of Pennsylvania

November 16, 2018

Abstract

Settings such as lending and policing can be modeled by a centralized agent allocating a scarce resource (e.g. loans or police officers) amongst several groups, in order to maximize some objective (e.g. loans given that are repaid, or criminals that are apprehended). Often in such problems *fairness* is also a concern. One natural notion of fairness, based on general principles of *equality of opportunity*, asks that conditional on an individual being a candidate for the resource in question, the probability of actually receiving it is approximately independent of the individual's group. For example, in lending this would mean that equally creditworthy individuals in different racial groups have roughly equal chances of receiving a loan. In policing it would mean that two individuals committing the same crime in different districts would have roughly equal chances of being arrested.

In this paper, we formalize this general notion of fairness for allocation problems and investigate its algorithmic consequences. Our main technical results include an efficient learning algorithm that converges to an optimal fair allocation even when the allocator does not know the frequency of candidates (i.e. creditworthy individuals or criminals) in each group. This algorithm operates in a *censored* feedback model in which only the number of candidates who received the resource in a given allocation can be observed, rather than the true number of candidates in each group. This models the fact that we do not learn the creditworthiness of individuals we do not give loans to and do not learn about crimes committed if the police presence in a district is low.

As an application of our framework and algorithm, we consider the *predictive policing* problem, in which the resource being allocated to each group is the number of police officers assigned to each district. The learning algorithm is trained on arrest data gathered from its own deployments on previous days, resulting in a potential feedback loop that our algorithm provably overcomes. In this case, the fairness constraint asks that the probability that an individual who has committed a crime is arrested should be independent of the district in which they live. We empirically investigate the performance of our learning algorithm on the *Philadelphia Crime Incidents* dataset.

1 Introduction

The bulk of the literature on algorithmic fairness has focused on classification and regression problems (see e.g. [3, 4, 6–8, 10, 14, 16, 17, 19, 20, 25–27] for a collection of recent work), but fairness concerns also arise naturally in many resource allocation settings. Informally, a resource allocation problem is one in which there is a limited supply of some *resource* to be distributed across multiple groups with differing needs. Resource allocation problems arise in financial applications (e.g. allocating loans), disaster response (allocating aid), and many other domains — but the primary example that we will focus on in this paper is policing. In the predictive policing problem, the resource to be distributed is police officers, which can be dispatched to different districts. Each district has a different crime distribution, and the goal (absent additional fairness constraints) might be to maximize the number of crimes caught.[1]

[1] We understand that policing has many goals besides simply apprehending criminals, including preventing crimes in the first place, fostering healthy community relations, and generally promoting public safety. But for concreteness and simplicity

1

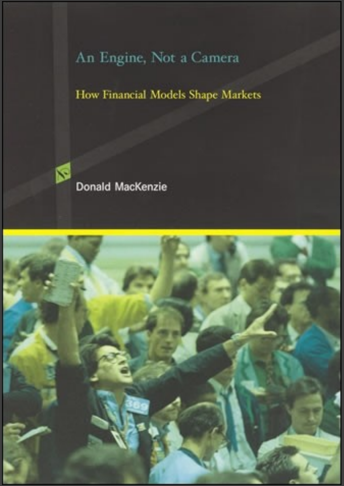
An engine not a camera

Camera

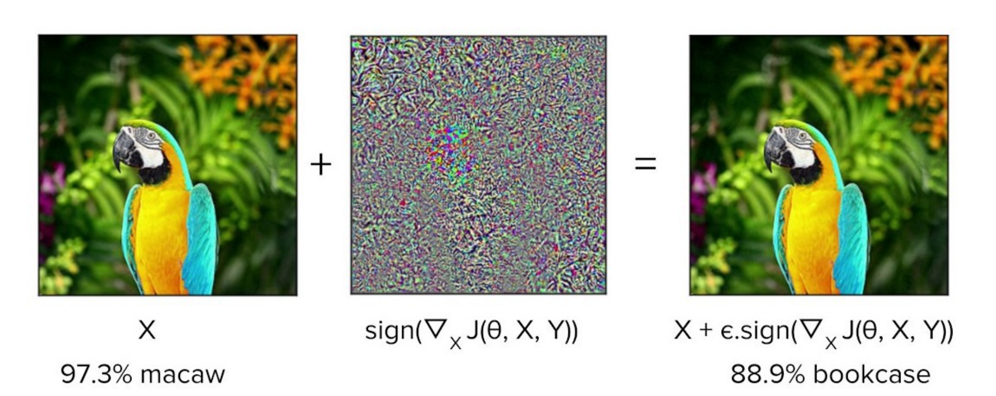
- Fairness as a statistical concept
- A problem of inaccuracy

Engine

- People respond to being sorted
- A problem of elasticity



Elasticity and accuracy



A slight perturbation of this picture of a macaw causes it to be classified as a bookcase

Factors affecting elasticity

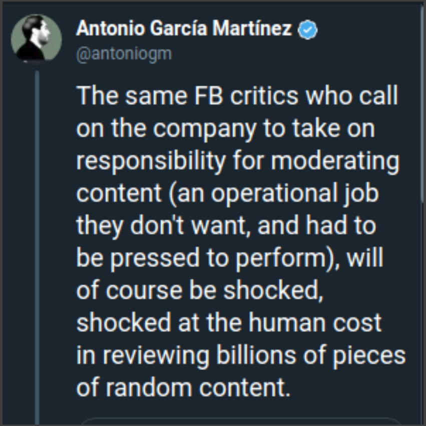
Elasticity increases with:

- Affordability (cost of change required)
- Sensitivity (magnitude of change required)
- Impact (benefit of change required)

Malicious actors attacking the commons?

OpenAI Blog on GPT-2 language model:

> *[M]alicious actors... have already begun to target the shared online commons... We should consider how research into the generation of synthetic images, videos, audio, and text may further combine to unlock new as-yet-unanticipated capabilities for these actors, and should seek to create better technical and non-technical countermeasures.*



Incompatible incentives

Description	ML goal (task)	Values goal (intent)
Signaling, screening	$+$	$+$
Pooling, gaming	$+$	$-$
Coordination, performative	0	$+$
Workaround	$-$	$+$
Protest	$-$	$-$

Gaming or not?

Credit Scoring: The LenddoScore

Lenddo's patented score is a powerful predictor of an individual's character or 'willingness to pay'. The LenddoScore ranges from 1 to 1000, with higher scores representing a lower propensity to default.

The LenddoScore can be deployed at the wide end of the funnel to prioritize applications or within an existing underwriting scorecard to reduce risk or approve more applications. The LenddoScore complements traditional underwriting tools, like credit scores, because it relies exclusively on non-traditional data derived from a customer's social data and online behavior. When the LenddoScore is added to a traditional underwriting scorecard, it has been proven to better discriminate between good and bad borrowers.

Alternative credit scoring with social media data

Workarounds

"The problem is that peoples' lives are not a drop-down menu... And that's where we run into problems.... And we have to manipulate the system to make the decisions that we want"

From: "Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works", Jennifer Raso

Canadian Journal of Law and Society / La Revue Canadienne Droit et Société

Search Canadian Journa

Article

Volume 32, Issue 1 April 2017 , pp. 75-95

Get access

Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works

Jennifer Raso (a1)

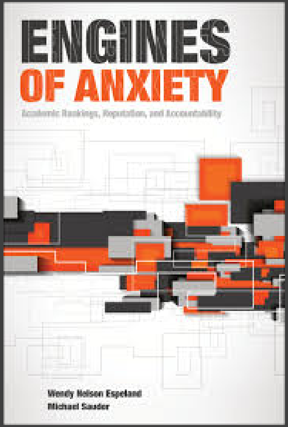
https://doi.org/10.1017/cls.2017.6 Published online: 27 June 2017

Abstract

This paper explores how new regulatory technologies and front-line decision-makers reshape one another. Drawing on a recent qualitative study of caseworker decision-making in the Ontario Works program, it demonstrates the dialectical relationship between new case management software and caseworkers. While new technologies may attempt to deskill and decentre front-line decision-makers, transforming them into data entry clerks, caseworkers learn how to expertly translate and input client data to produce decisions that more closely match their interpretation of clients' needs and welfare laws. The ways in which workers "manipulate the system" to produce a particular decision, though common knowledge among their colleagues, are black boxed to program managers, auditors, and benefits recipients.

Performativity

"The academic legal establishment did not so much fall into this trap as become entangled in it. Like a fly touched by the thread of a spider's web, they were at first only lightly caught up, but then found that each move they made in response only drew them in more tightly." – K.J. Healy



Incompatible incentives again

Description	ML goal (task)	Values goal (intent)
Signaling, screening	+	+
Pooling, gaming	+	−
Coordination, performative	0	+
Workaround	−	+
Protest	−	−

Algorithms demand rules

- Most algorithmic systems are incentive-incompatible
- Algorithms demand rules to keep them functioning (Code is law until it isn't)
- Rules often scale less well than their algorithms
- Some highly-elastic systems may be ungovernable

Johnny Tanner (YouTube video producer)

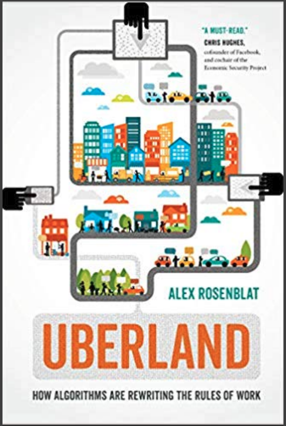
> *"The algorithm is the thing we had a relationship with since the beginning. That's what got us out there and popular... We learned to fuel it and do whatever it took to please the algorithm."*

Rules create temptations

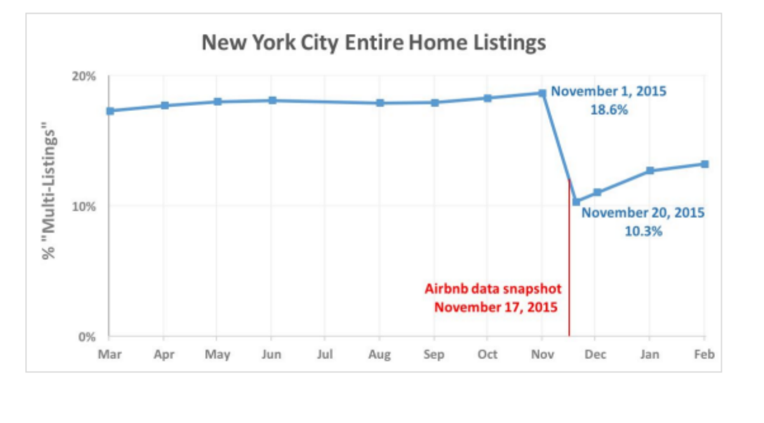
Techniques of regulatory arbitrage:

- Invoke unintended consequences
- Invoke the software process
- Invoke values ad-hoc
- Keep problems hidden
- Use data as leverage

Bug fixing


"A MUST-READ."
CHRIS HUGHES,
cofounder of Facebook,
and cochair of the
Economic Security Project
ALEX ROSENBLAT
UBERLAND
HOW ALGORITHMS ARE REWRITING THE RULES OF WORK

Community values



Rules create temptations, again

Techniques of regulatory arbitrage:

- Invoke unintended consequences
- Invoke the software process
- Invoke values ad-hoc
- Keep problems hidden
- Use data as leverage

Governing algorithmic governance

- Section 230
- Data limitation
- Competition
- Wikipedia

Summary

- Algorithms are getting more accurate, but not more robust

Summary

- Algorithms are getting more accurate, but not more robust
- Sorting creates incentives, so algorithms demand supplementary rules to manage people's behaviour.

Summary

- Algorithms are getting more accurate, but not more robust
- Sorting creates incentives, so algorithms demand supplementary rules to manage people's behaviour.
- The rules become a form of governance for which the platform owner has no expertise. In cases of high elasticity, effective governance may not be possible.

Summary

- Algorithms are getting more accurate, but not more robust
- Sorting creates incentives, so algorithms demand supplementary rules to manage people's behaviour.
- The rules become a form of governance for which the platform owner has no expertise. In cases of high elasticity, effective governance may not be possible.
- Algorithm owners have a temptation to engage in regulatory arbitrage. They also have an incentive to keep the brand/practice gap wide.

Summary

- Algorithms are getting more accurate, but not more robust
- Sorting creates incentives, so algorithms demand supplementary rules to manage people's behaviour.
- The rules become a form of governance for which the platform owner has no expertise. In cases of high elasticity, effective governance may not be possible.
- Algorithm owners have a temptation to engage in regulatory arbitrage. They also have an incentive to keep the brand/practice gap wide.
- There are rationales for external action, whether through competition rules, constraints on the algorithms themselves, or limitations to the data that can be used.

Thank you

Tom Slee
tom@tomslee.net