Thrust 1: Individual decision making under computational and cognitive constraints

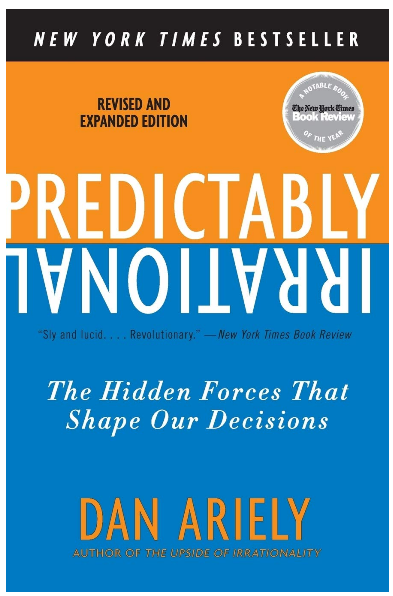
PIs involved: Joe Halpern, Jon Kleinberg, and Josh Tenenbaum

October 21, 2020

The big picture

Many studies have shown that humans are "predictably irrational"

- they do not act in a fully rational way, as assumed by standard economic theory
- but their deviations from rational behavior are quite systematic



Can we explain "predictably irrational" human behavior as the outcome of computational and cognitive constraints?

We focus on four well-studied problems:

- multi-armed bandits:
  - You want to choose the best slot machine, given m choices
  - Classic exploitation/exploration problem
  - Models managing research investments, portfolio selection, ...

We focus on four well-studied problems:

- multi-armed bandits:
  - You want to choose the best slot machine, given m choices
  - Classic exploitation/exploration problem
  - Models managing research investments, portfolio selection, ...
- A security game:
  - poachers are trying to catch rhinos at one of n sites; rangers are trying to stop them
  - Similar issues arise when trying to stop smugglers on one of a small number of routes, checking for terrorists at airports, deploying randomized patrol schedules for the Coast Guard, ...

We focus on four well-studied problems:

- multi-armed bandits:
  - You want to choose the best slot machine, given m choices
  - Classic exploitation/exploration problem
  - Models managing research investments, portfolio selection, ...
- A security game:
  - poachers are trying to catch rhinos at one of n sites; rangers are trying to stop them
  - Similar issues arise when trying to stop smugglers on one of a small number of routes, checking for terrorists at airports, deploying randomized patrol schedules for the Coast Guard, ...
- "Moral" decisions
  - Is it reasonable to use a fuel-inefficient car? Butt in line?

We focus on four well-studied problems:

- multi-armed bandits:
  - You want to choose the best slot machine, given m choices
  - Classic exploitation/exploration problem
  - Models managing research investments, portfolio selection, ...
- A security game:
  - poachers are trying to catch rhinos at one of n sites; rangers are trying to stop them
  - Similar issues arise when trying to stop smugglers on one of a small number of routes, checking for terrorists at airports, deploying randomized patrol schedules for the Coast Guard, ...
- "Moral" decisions
  - Is it reasonable to use a fuel-inefficient car? Butt in line?
- Screening decisions
  - How do we make hiring/admissions decisions?

We focus on four well-studied problems:

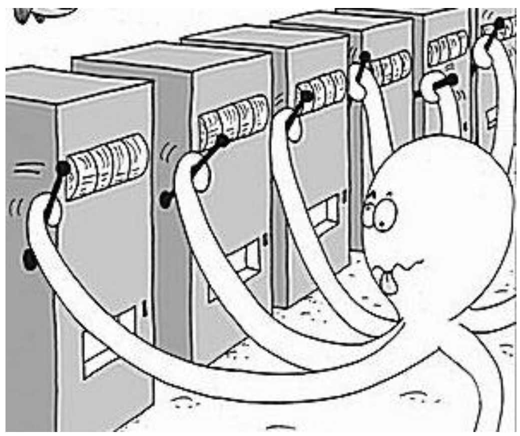
- multi-armed bandits:
  - You want to choose the best slot machine, given m choices
  - Classic exploitation/exploration problem
  - Models managing research investments, portfolio selection, ...
- A security game:
  - poachers are trying to catch rhinos at one of n sites; rangers are trying to stop them
  - Similar issues arise when trying to stop smugglers on one of a small number of routes, checking for terrorists at airports, deploying randomized patrol schedules for the Coast Guard, ...
- "Moral" decisions
  - Is it reasonable to use a fuel-inefficient car? Butt in line?
- Screening decisions
  - How do we make hiring/admissions decisions?

Conclusions:

- Computational limitations help explain human behavior
- Building human-like behavior into our algorithms can significantly improve performance!
- "Irrational" behavior is not always so irrational

Multi-armed bandits (MABs)

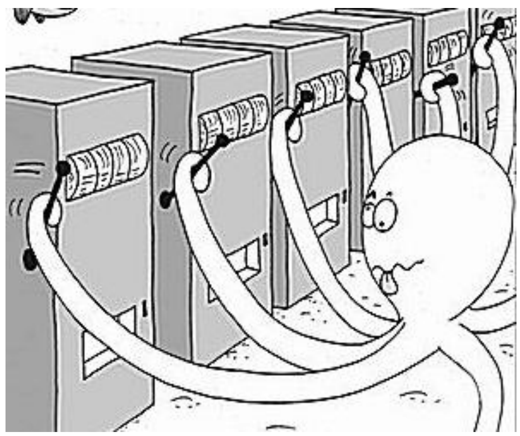
[Liu and Halpern]



Exploration vs. *Exploitation*

Multi-armed bandits (MABs)

[Liu and Halpern]



Exploration vs. *Exploitation*

A K-armed MAB can be identified with a tuple $(\mu_1, \ldots, \mu_K)$

- each component represents an arm
- μ_i is a distribution over the possible rewards of arm i
    - We take the possible rewards to be 1 ("success") or 0 ("failure").
    - We assume the distributions do not vary over time.

MAB protocols

There are many protocols for finding the best arm. Two of the best-known are:

- *Thompson sampling:* uses Bayesian methods
- *Epsilon-greedy: explores* (play a random arm) with probability ϵ and otherwise *exploits* (play the current best arm)

MAB protocols

There are many protocols for finding the best arm. Two of the best-known are:

- *Thompson sampling:* uses Bayesian methods
- *Epsilon-greedy: explores* (play a random arm) with probability ϵ and otherwise *exploits* (play the current best arm)

Both use infinitely many states to keep track of history

- not very "human-like"!

MAB protocols

There are many protocols for finding the best arm. Two of the best-known are:

- *Thompson sampling:* uses Bayesian methods
- *Epsilon-greedy: explores* (play a random arm) with probability ϵ and otherwise *exploits* (play the current best arm)

Both use infinitely many states to keep track of history

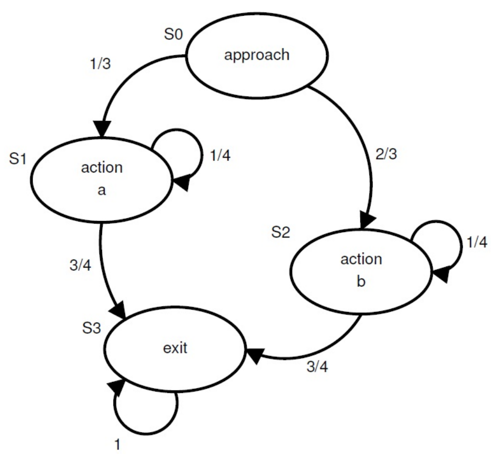
- not very "human-like"!

Our goal is not to find the best protocol, but to explain human behaviors.

Probabilistic finite automata (PFA)

To capture resource-boundedness, we model people as probabilistic finite automata (PFA).

- Just like deterministic finite automata, except that we allow probabilistic state transitions.

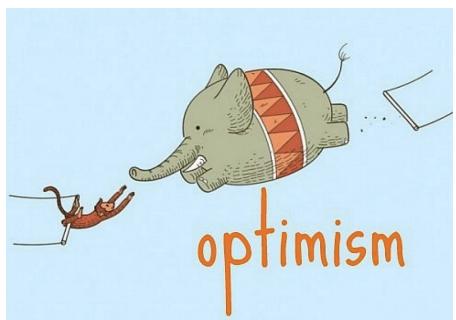


Key Ideas of PFA

- We play each arm sequentially against a “virtual arm”
- We build in *satisficing* [Simon 1955]: quit as soon as you find an arm that's better than the virtual arm

Key Ideas of PFA

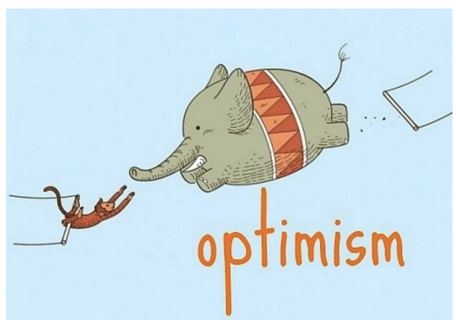
- We play each arm sequentially against a "virtual arm"
- We build in *satisficing* [Simon 1955]: quit as soon as you find an arm that's better than the virtual arm
- We build in an *optimism bias*: we start by assuming that the virtual arm has a high success probability.



- if no arm can beat the virtual arm, we slowly lower its success probability

Key Ideas of PFA

- We play each arm sequentially against a "virtual arm"
- We build in *satisficing* [Simon 1955]: quit as soon as you find an arm that's better than the virtual arm
- We build in an *optimism bias*: we start by assuming that the virtual arm has a high success probability.



- if no arm can beat the virtual arm, we slowly lower its success probability
- We build in a *negativity bias*:
  - we quickly eliminate arms that do not beat the virtual arm, but are reluctant to declare an arm a winner

- Experiments show that our aspiration-level protocol does extremely well in practice
  - It does better that ϵ-greedy
  - It does not do as well as Thompson sampling (which is known to be optimal) but that is an inherent problem
    - For all finite-state protocols P, there exists an $\epsilon > 0$ such that after N steps, the regret is $> \epsilon N$.
    - Thompson sampling has logarithmic regret
- The protocol's performance degrades gracefully as we decrease the number of states, with human-like biases becoming more emphasized.

The Ranger-Poacher Game

[Liu and Halpern]

- There are two player, a ranger and a poacher, and a fixed number n of sites that rhinos might go to.
- The distribution of rhinos at each site is commonly known.
  - e.g., may have $(.8, .9, .3, .2)$: with probability .8, there is a rhino at site 1, with probability .9, there is a rhino at site 2, . . .
- It's a zero-sum game:
  - If the poacher catches a rhino, then he gets $+1$; the ranger gets -1
  - if the ranger catches the poacher, then she gets $+1$, the poacher gets -1.
- The game has a unique Nash equilibrium (NE).

Fictitious Play

There are many protocols that converge to NE. We consider perhaps the best studied: *fictitious play* (FP):

- At each step, players best respond to the mixed strategy where the probability that an action is played is the frequency with which that strategy has been played thus far.
  - This requires an unbounded number of states to implement, since again, we must keep track of history.
  - It is easy to approximate this with an FP
- Although the convergence time of FP is slow, it has been very well-studied in large part because it is so natural.

PFAs play FP

With only finitely many states, a PFA must approximate frequency that the opponent has gone to each site:

- Assume that the memory has the form $[q_1, \ldots, q_n]$, where $q_1 + \cdots + q_n \leq M$
  - q_i is (an approximation to) the number of times that the ranger has been to site i in the last M steps
  - If the poacher observes the ranger go to site i then
    - q_i is increased by 1
    - q_j is chosen at random according to its frequency ($\frac{q_j}{q_1+\cdots+q_n}$) and is decreased by 1
  - Experiments show that this gives quite a good approximation to the actual frequency for $M = 100$.

PFAs play FP

With only finitely many states, a PFA must approximate frequency that the opponent has gone to each site:

- Assume that the memory has the form $[q_1, \ldots, q_n]$, where $q_1 + \cdots + q_n \leq M$
  - q_i is (an approximation to) the number of times that the ranger has been to site i in the last M steps
  - If the poacher observes the ranger go to site i then
    - q_i is increased by 1
    - q_j is chosen at random according to its frequency $(\frac{q_j}{q_1+\cdots+q_n})$ and is decreased by 1
  - Experiments show that this gives quite a good approximation to the actual frequency for $M = 100$.
  - For smaller M, the approximation fluctuates more around the actual frequency, which leads to *probability matching*
    - The smaller M is, the more the poacher goes to sites proportional to the probabilities of rhinos being there.
    - The ranger continues to play NE, because probability matching is essentially the NE strategy for the ranger
    - Key observation: Best responding + variations in estimates due to small memory lead to probability matching!

The Significance of Significance

From a human perspective, some events are more significant than others.

- Observing a potentially poisonous snake is far more significant than observing a beetle.

Significant events are typically ones with very bad outcomes:

- Kahneman and Tversky observe that we tend to overweight the negative

In the ranger-poacher game, we take the significant events to be

- Getting caught, for the poacher
- The poacher catching a rhino, for the ranger

We overweighted these events in the PFA, so as to be able to reproduce what we observed in our Amazon Turk experiments:

- Some poachers go to sites with high rhino likelihood less often that would be predicted by NE
    - Intuitively, they are trying to avoid being caught

The Significance of Significance

From a human perspective, some events are more significant than others.

- Observing a potentially poisonous snake is far more significant than observing a beetle.

Significant events are typically ones with very bad outcomes:

- Kahneman and Tversky observe that we tend to overweight the negative

In the ranger-poacher game, we take the significant events to be

- Getting caught, for the poacher
- The poacher catching a rhino, for the ranger

We overweighted these events in the PFA, so as to be able to reproduce what we observed in our Amazon Turk experiments:

- Some poachers go to sites with high rhino likelihood less often that would be predicted by NE
  - Intuitively, they are trying to avoid being caught
- Big surprise: taking significance into account not only replicated observed results, but significantly improved performance of the PFA, especially with little memory.

MTurk Experiments

We ran experiments on Amazon Turk (MTurk), using a number of different rhino distributions, with people playing the role of poacher.

People largely cluster into three groups:

- *Level-0*: nonstrategic, visit all sites with equal probability
  - We suspect that these are often people trying to finish the game as quickly as possible, just to get the base payment

MTurk Experiments

We ran experiments on Amazon Turk (MTurk), using a number of different rhino distributions, with people playing the role of poacher.

People largely cluster into three groups:

- *Level-0*: nonstrategic, visit all sites with equal probability
  - We suspect that these are often people trying to finish the game as quickly as possible, just to get the base payment
- *Level-1*: probability match with rhino distribution
  - As we saw, this can be explained by limited memory
  - people could still be best responding!

MTurk Experiments

We ran experiments on Amazon Turk (MTurk), using a number of different rhino distributions, with people playing the role of poacher.

People largely cluster into three groups:

- *Level-0*: nonstrategic, visit all sites with equal probability
  - We suspect that these are often people trying to finish the game as quickly as possible, just to get the base payment
- *Level-1*: probability match with rhino distribution
  - As we saw, this can be explained by limited memory
  - people could still be best responding!
- *Level-2:* Best responding to level-1 rangers, so go to sites with higher rhino distribution less often
  - As we saw, this can be explained by overweighting negative outcomes
  - Key point: this behavior is quite rational!

Moral Decision Making

[S. Levine, M. Kleiman-Weiner, L. Schulz, J. Tenenbaum, F. Cushman]

Current theories of moral decision-making use two main approaches to choosing the “moral” action:

- Rules
  - precompiled answers that apply to a wide range of cases
- Expected utility calculations
  - Perform the action that maximizes expected utility (EU)

Moral Decision Making

[S. Levine, M. Kleiman-Weiner, L. Schulz, J. Tenenbaum, F. Cushman]

Current theories of moral decision-making use two main approaches to choosing the "moral" action:

- Rules
  - precompiled answers that apply to a wide range of cases
- Expected utility calculations
  - Perform the action that maximizes expected utility (EU)

Perhaps the function of moral decision-making is to maximize EU

- Whose utility?
  - The agent's? The agent's community?
- Even ignoring that, doing EU calculations is hard!
- Rules may provide a way for agents to get desired results in a computationally-efficient way.

Universalization

On the other hand, rules are limited:

- Sometimes, there is no obvious rule that exists to guide us
  - How do we create novel rules for novel circumstances?
- Sometimes a simple heuristic gives the wrong answer
  - When is it is acceptable to override/modify a heuristic rule?
    - E.g., when is it acceptable to cut in line?

Universalization

On the other hand, rules are limited:

- Sometimes, there is no obvious rule that exists to guide us
  - How do we create novel rules for novel circumstances?
- Sometimes a simple heuristic gives the wrong answer
  - When is it is acceptable to override/modify a heuristic rule?
    - E.g., when is it acceptable to cut in line?

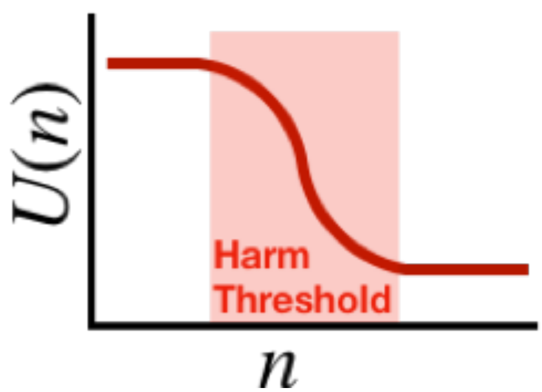
Universalization is a tool that resource-bounded agents can use to create new rules and refine their current rules

- It asks "what if everyone felt at liberty to do that?"
- Helps us figure out when and how to override rules.

Testing Universalization

- We are interested in seeing to what extent people use universalization and feel it appropriate
- We tested whether subjects use universalization in a class of collective action problems called "threshold problems", when there are no pre-agreed rules
  - There are two possible actions
  - If few people choose one of the actions, the are better off, and there are no negative consequences for the group
  - If many people choose that action things go badly for everyone
  - E.g., climate change

Universalization captures (some of) our moral judgment process.



Modeling Universalization

The critical components of the model are

- The number n_i of "interested parties"
  - the people who would act if they felt at liberty to do so
  - If everyone felt at liberty, then only the people who were interested in doing the action would do it.
- The utility consequences of those people acting

Assumption: the probability that the action would be found acceptable given that n_i people performed it has the form

$$P_{Univ}(Acceptable) = \frac{1}{1 + e^{\tau(U(0)-U(n_i))+\beta}}$$

- This is a standard "soft max"
- Note the dependence on $U(0) - U(n_i)$
  - This measures the harm done by n_i agents performing the action.

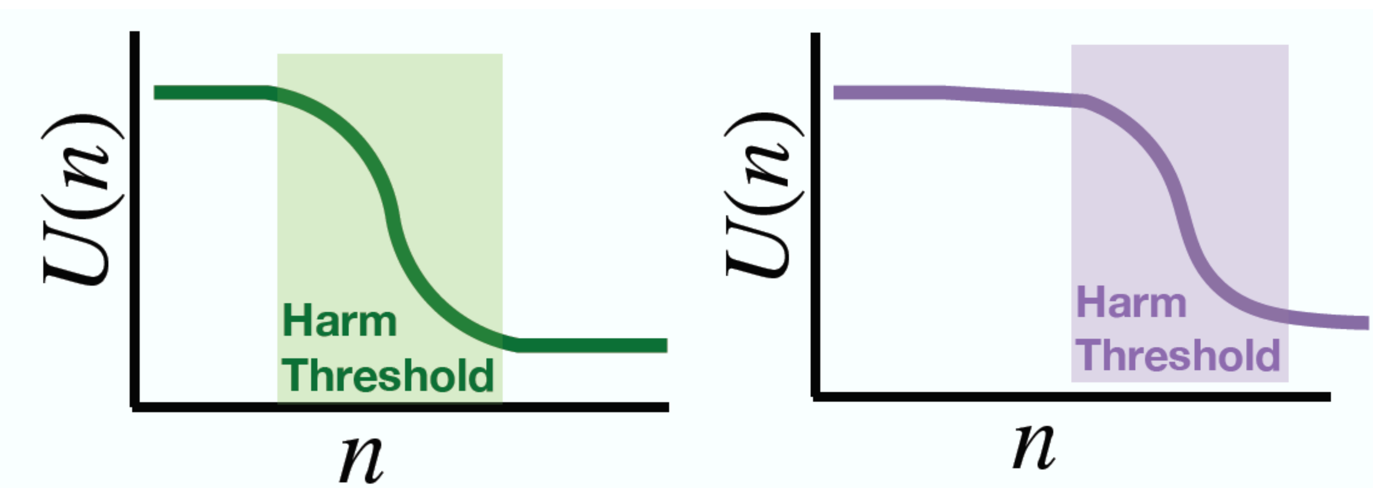
Experiment

Over-fishing scenario

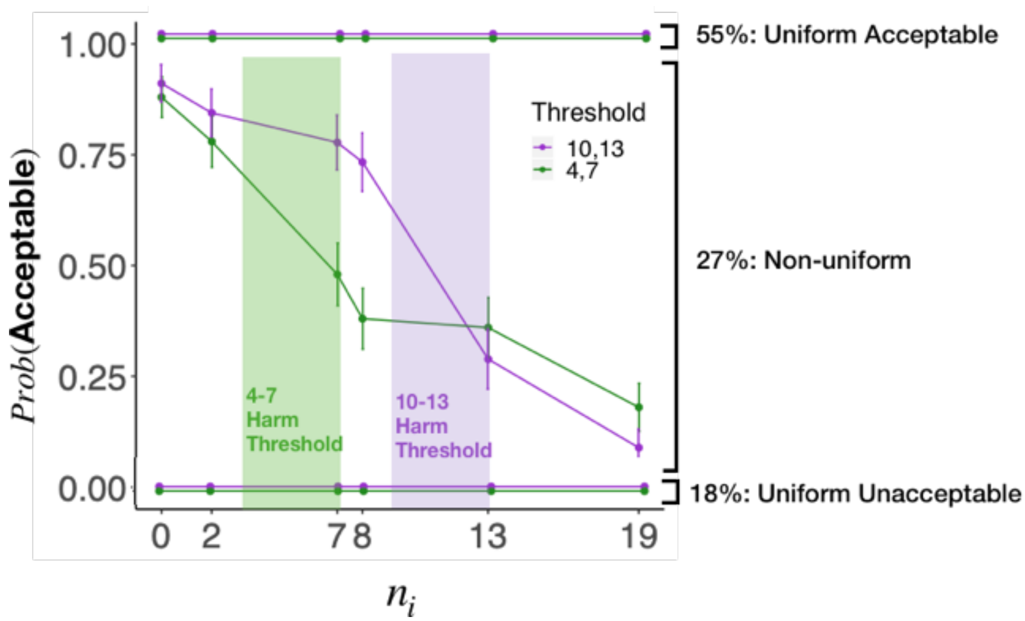
- Everyone in the village can fish sustainably with the traditional fishing method
- A new method allows them to catch more fish, but would lead to the fish going extinct if many people use the hook

Two parameters of interest:

- the number of interested parties (0-20 people)
- the utility consequences of multiple people using the new hook (which affects the harm threshold)



Moral Judgments: Experimental Results



- The model fits the data really well for the 27% whose judgments affected by n_i
- The others seem to care only about the outcome

Universalization: Creating new rules

- Given a stable number of interested parties and a stable harm threshold, a new simple rule can be created:
  - it is OK/not OK to use the new fishing hook
- The output of universalization reasoning can then be codified/amortized and reused in future cases that have the same structure

Current work

[S. Levine, J. Halpern, M. Kleiman-Weiner, J. Tenenbaum]

In communities of resource-bounded agents, some universalizable actions might not be morally permissible because of cognitive constraints.

- An agent might not be able to understand the intended universalizable action or figure out what it is

All universalizable actions should also be ...

- legible?
  - Can others figure out what you're doing
- tamperproof?
  - Not easily gameable
  - A policy like "use the hook when your relatives visit" might result in me inviting my relatives frequently
- robust?
  - Nothing disastrous happens if people deviate
- communicable?
  - Is it easy to explain what you're doing

Stereotype Formation

Cases where people evaluate each other.

- Formal settings like hiring or admissions.
- Informal evaluation in everyday interaction.

Where do negative stereotypes come from? [Greenwald-Banaji 95]

- In the presence of low information or limited available time, we are more prone to fall back on stereotypes.
- These stereotypes often work to the detriment of groups that are already disadvantaged.

Can we find a formal basis for these properties?

- Can such a model suggest useful interventions?

Screening Decisions and Disadvantage

A stylized scenario:

Applicants and feature vectors:

- Applicants are described by (Boolean) variables $x = (x^{\langle 1 \rangle}, x^{\langle 2 \rangle}, \ldots, x^{\langle k \rangle})$.
- Function f describes productivity $f(x)$ of applicant with features x.
- Plan: Sort by f-value, admit top r fraction.

Screening Decisions and Disadvantage

A stylized scenario:

Applicants and feature vectors:

- Applicants are described by (Boolean) variables $x = (x^{\langle 1 \rangle}, x^{\langle 2 \rangle}, \ldots, x^{\langle k \rangle})$.
- Function f describes productivity $f(x)$ of applicant with features x.
- Plan: Sort by f-value, admit top r fraction.

Group membership:

- Applicants can belong to *advantaged* group A or *disadvantaged* group D. Extended feature vector (x, A) or (x, D).
- Function f is independent of group: $f(x, A) = f(x, D) = f(x)$.

Screening Decisions and Disadvantage

A stylized scenario:

Applicants and feature vectors:

- Applicants are described by (Boolean) variables $x = (x^{\langle 1 \rangle}, x^{\langle 2 \rangle}, \ldots, x^{\langle k \rangle})$.
- Function f describes productivity $f(x)$ of applicant with features x.
- Plan: Sort by f-value, admit top r fraction.

Group membership:

- Applicants can belong to *advantaged* group A or *disadvantaged* group D. Extended feature vector (x, A) or (x, D).
- Function f is independent of group: $f(x, A) = f(x, D) = f(x)$.

Disadvantage:

- $\mu(x, \gamma) =$ fraction of population with features x and group γ.
- Likelihood-ratio condition: if $f(x) > f(x')$, then

$$\frac{\mu(x, A)}{\mu(x, D)} > \frac{\mu(x', A)}{\mu(x', D)}.$$

Simplification

- True criterion is conjunction of $x^{\langle 1\rangle}$ and $x^{\langle 2\rangle}$.
- Applicants from A have $x^{\langle i\rangle} = 1$ with prob. $2/3$.
- Applicants from D have $x^{\langle i\rangle} = 1$ with prob. $1/3$.

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	f	μ
1	1	D	1	$1/18$
1	1	A	1	$4/18$
1	0	D	0	$2/18$
1	0	A	0	$2/18$
0	1	D	0	$2/18$
0	1	A	0	$2/18$
0	0	D	0	$4/18$
0	0	A	0	$1/18$

Simplification

- True criterion is conjunction of $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$.
- Applicants from A have $x^{\langle i \rangle} = 1$ with prob. $2/3$.
- Applicants from D have $x^{\langle i \rangle} = 1$ with prob. $1/3$.

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	1	$1/18$
1	1	A	1	$4/18$
1	0	D	0	$2/18$
1	0	A	0	$2/18$
0	1	D	0	$2/18$
0	1	A	0	$2/18$
0	0	D	0	$4/18$
0	0	A	0	$1/18$

At all admission rates $r \leq 5/18$, all admitted have f-value 1, with a $1/5$ fraction from group D.

Now suppose we simplify f by using only $x^{\langle 1 \rangle}$, not both features.

- Perhaps collecting $x^{\langle 2 \rangle}$ is too expensive.
- (For larger instances) Interpretability or cognitive complexity.
- Out-of-sample generalization.
- Removing a variable that confers some of the disadvantage.

Simplification

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	any	any	$5/9$	1/2
0	any	any	0	1/2

$\longleftarrow$

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	f	μ
1	1	D	1	$1/18$
1	1	A	1	$4/18$
1	0	D	0	$2/18$
1	0	A	0	$2/18$
0	1	D	0	$2/18$
0	1	A	0	$2/18$
0	0	D	0	$4/18$
0	0	A	0	$1/18$

- An f-approximator: collapse rows; assign each applicant their expected f-value conditional on what we know about them; admit in this order.
- Now at all admission rates $r \leq 5/18$:
  average f-value is $5/9$ (not 1)
  fraction from group D is $1/3$ (not $1/5$).
- Relative to true f, gains in equity, losses in efficiency.

Simplifying may confer many of the aforementioned benefits.
But it also causes two potential difficulties.

First Problem: Incentived Bias

Simplification transforms disadvantage into bias:

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	any	A	$2/3$	$1/3$
1	any	D	$1/3$	$1/6$
0	any	A	0	$1/6$
0	any	D	0	$1/3$

$\longleftarrow$

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	any	any	$5/9$	$1/2$
0	any	any	0	$1/2$

Dropping $x^{\langle 2\rangle}$ creates an (f-maximizing) incentive to use the group membership variable γ in a way that hurts group D.

- Incentive only arises because $x^{\langle 2\rangle}$ is invisible; with true f, no incentive to consult value of γ.
- A basic mechanism for stereotype formation in everyday life [Leyens et al 1994, Greenwald-Banaji 1995].
- Connected to models of statistical discrimination [Arrow 1973, Coate-Loury 1993, Hu-Chen 2018].
- Empirical analogues: "ban the box" policies and effect discrimination [Agan-Starr 2016, Doleac-Hansen 2016, Shoag-Veuger 2016.]
- Related: drug tests [Wozniak 2015], credit history [Bartik-Nelson 2016].

Second Problem: Pareto-Improvement

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	1	D	1	$1/18$
1	any	any	$1/2$	$8/18$
0	any	any	0	$1/2$

$\longleftarrow$

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	any	$5/9$	$1/2$
0	any	any	0	$1/2$

Let g and h be two f-approximators.

- h strictly improves on g in efficiency if for every admission rate r, average f-value of admitted set under h is at least average f-value of admitted set under g; and strictly greater for at least one value of r.
- h strictly improves on g in equity if analogous condition holds for the fraction of members of group D who are admitted.

Pareto-improvement:

- Approximator on left strictly improves approximator on right: for any monotone preferences for efficiency and equity, approximator on left is an improvement.
- Empirical analogues in settings like United Steelworkers v. Weber (1974) on programs for selected members of underrepresented groups.

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	f	μ
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

simplification

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

incentivized bias

Pareto-improvement

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	any	A	2/3	1/3
1	any	D	1/3	1/6
0	any	A	0	1/6
0	any	D	0	1/3

$x^{\langle 1\rangle}$	$x^{\langle 2\rangle}$	γ	avg f	μ
1	1	D	1	1/18
1	any	any	1/2	8/18
0	any	any	0	1/2

General Result

Informal version of a general result [Kleinberg-Mullainathan]:

For every Boolean function f with real-valued outputs satisfying the disadvantage condition and a genericity assumption, and for every simplification g of it (partitioning feature vectors into cells by fixing variables):

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

simplification

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

incentivized bias

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	A	2/3	1/3
1	any	D	1/3	1/6
0	any	A	0	1/6
0	any	D	0	1/3

Pareto-improvement

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	1	D	1	1/18
1	any	any	1/2	8/18
0	any	any	0	1/2

(a) There is always an f-approximator that strictly improves g in both efficiency and equity.

(b) If g does not use group membership, then adding group membership as a variable increases efficiency and reduces equity.

The Nature of the Disadvantage Condition

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	.9	.06
1	1	A	.9	.04
1	0	D	.6	.02
1	0	A	.6	.06
0	1	D	.2	.07
0	1	A	.2	.06
0	0	D	.02	.35
0	0	A	.02	.34

$\longrightarrow$

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	D	.825	.08
1	any	A	.72	.10
0	any	D	.05	.42
0	any	A	.047	.40

The Nature of the Disadvantage Condition

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	.9	.06
1	1	A	.9	.04
1	0	D	.6	.02
1	0	A	.6	.06
0	1	D	.2	.07
0	1	A	.2	.06
0	0	D	.02	.35
0	0	A	.02	.34

$\longrightarrow$

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	D	.825	.08
1	any	A	.72	.10
0	any	D	.05	.42
0	any	A	.047	.40

An example where the mean f-value in group A exceeds the mean f-value in group D (a weaker form of disadvantage), but:

- The f-approximator g that uses $x^{\langle 1 \rangle}$ and γ cannot be Pareto-improved.
- The f-approximator h that uses only $x^{\langle 1 \rangle}$ creates an incentive to use γ in a way that favors group D (not A).
- This is a reflection of Simpson's Paradox.
- The technical underpinning of our main theorem can be viewed as proving an "anti-Simpson" result.

Open question: What is the weakest specification of disadvantage where the result holds?

Main Combinatorial Lemma

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	.9	.06
1	1	A	.9	.04
1	0	D	.6	.02
1	0	A	.6	.06
0	1	D	.2	.07
0	1	A	.2	.06
0	0	D	.02	.35
0	0	A	.02	.34

$\longrightarrow$

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	D	.825	.08
1	any	A	.72	.10
0	any	D	.05	.42
0	any	A	.047	.40

Assume the likelihood-ratio condition for disadvantage:
if $f(x) > f(x')$, then $\dfrac{\mu(x,A)}{\mu(x,D)} > \dfrac{\mu(x',A)}{\mu(x',D)}$.

Consider any non-trivial partition of the feature vectors for A into cells and (separately) the feature vectors for D into cells.

- Assign each feature vector (x,γ) a value $g(x,\gamma)$ equal to the (measure-weighted) average of f in its cell.
- Then there exists a feature vector x for which $g(x,A) > g(x,D)$.

Reflections

Simplifying a function f with groups A and D.

- Creates an incentive to use group membership in a way that hurts group D.
- Can always be strictly improved in both efficiency and equity.

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	f	μ
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

simplification

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

incentivized bias

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	any	A	2/3	1/3
1	any	D	1/3	1/6
0	any	A	0	1/6
0	any	D	0	1/3

Pareto-improvement

$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	γ	avg f	μ
1	1	D	1	1/18
1	any	any	1/2	8/18
0	any	any	0	1/2

Ongoing open questions:

- Consider other formulations of simplicity.
  Large alternate category: linear approximations to f.
- Consider other formulations of the disadvantage condition.
  What is the weakest condition for which these results hold?
- Studying the space of all simplifications of f w.r.t. efficiency and equity.
- Further implications for empirical analysis and interventions.