The *Disparate Equilibria* of Algorithmic Decision Making when Individuals Invest Rationally

Lydia T. Liu[*◇], Ashia Wilson[†], Nika Haghtalab[*⊙], Adam Tauman Kalai[†], Christian Borgs[*◇], Jennifer Chayes[*◇]

[*]Work done at Microsoft Research [◇]University of California, Berkeley [†]Microsoft Research [⊙]Cornell University


Machine learning models are being trained and used to make decisions about *people*, allocating *resources* and *opportunities*.

People tend to *change* their behavior in response to how these decisions are made.

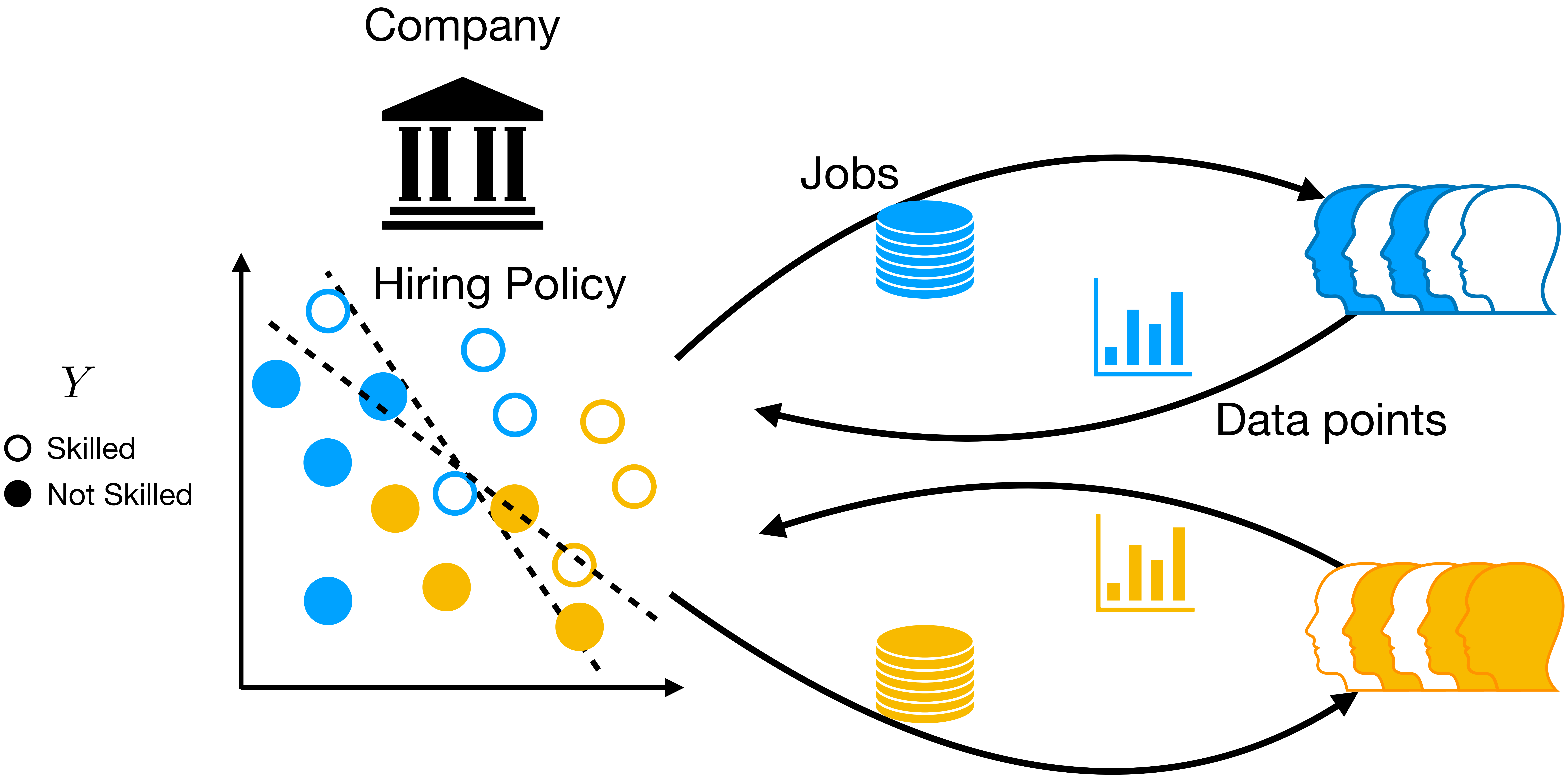
Humans responding to algorithms

Pros

- Algorithms can incentivize humans to take "improving" actions over "gaming" actions [KR19]
- Algorithm rewards people appropriately, encouraging them to pursue beneficial investments, e.g. acquiring job skills, preparing for college [CL93, this work]

Cons

- People strategically change their features to game the algorithm [HMPW16, HIV19, MMDM19]
- Algorithms fail to reward certain groups, discouraging them from making beneficial investments [CL93, this work]
  - There is **heterogeneity across groups** leading to different responses [this work]



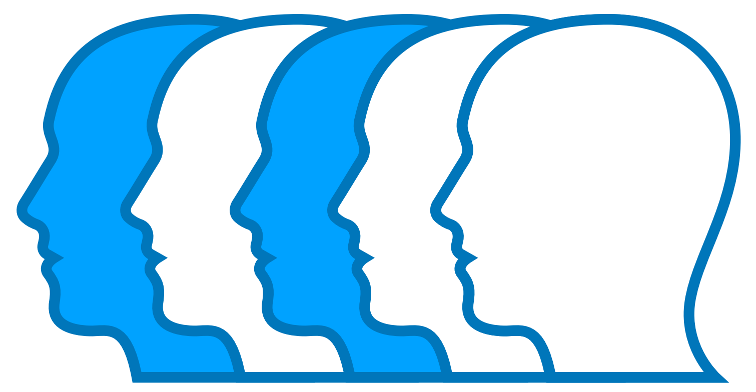
Company
Hiring Policy
Y
Skilled
Not Skilled
Jobs
Data points

Under these dynamics…

1. What kind of **long-term outcomes (equilibria)** are produced?
2. What kind of **interventions** produce desirable equilibria?

Model for individual investment

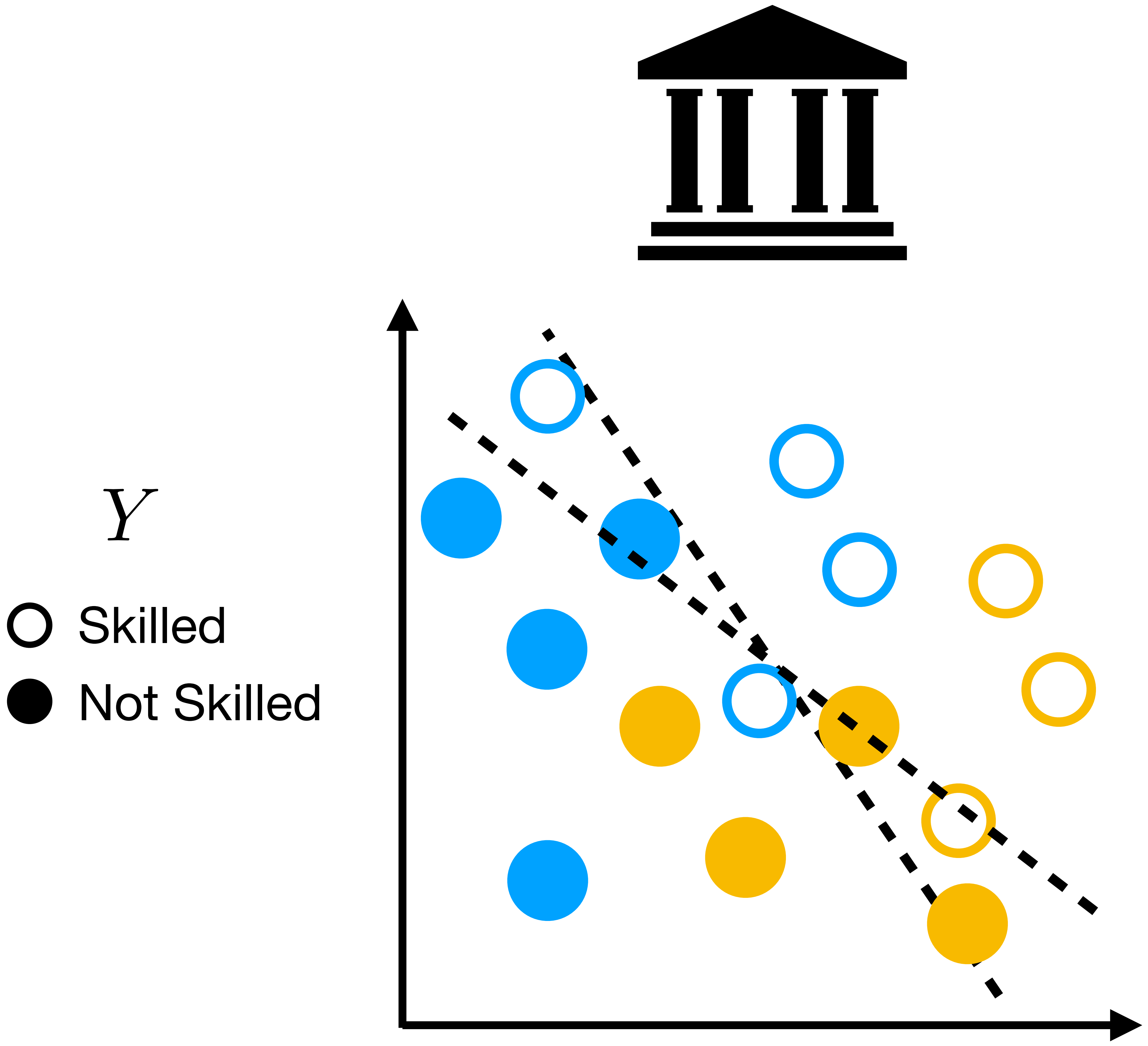
- Given the **current hiring policy**, should I invest in acquiring job skills (become $Y = 1$) if
  - It costs me $\boldsymbol{C}$ to do that
  - I will develop features (e.g. resume, scores) that depend on my group $\boldsymbol{A}$ and this boosts my chances of being hired by $\boldsymbol{\beta(A)}$



- I will invest in job skills if and only if my expected gain > 0.

- Individual-level decisions determine the overall **qualification rate** in each group.

Model for institution's response



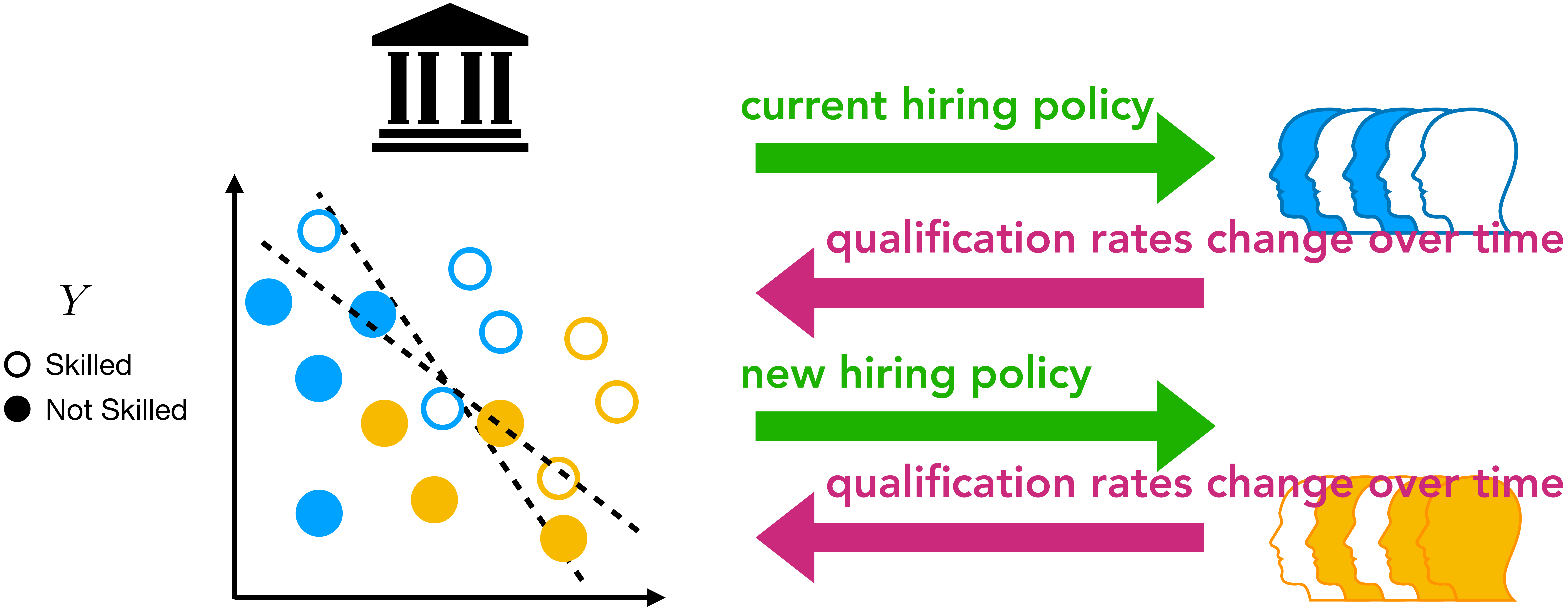
- Accepting skilled individuals is a gain, accepting unskilled individuals is a loss.

- Picks **current hiring policy**

  - out of a chosen model class (e.g. linear models on observable features)

  - to maximize its *expected profit*, which depends on the **qualification rates** in each group.

Dynamics of qualification rates



eventually **qualification rates** stabilize —
reached **equilibrium**!

What ensures "good" equilibria?

Result: If there exists a **zero-error** hiring policy in the model class, there is a unique (non-trivial) equilibrium.

- All groups have the **same qualification rate** at equilibrium. This is also the *optimal* qualification rate.
- This also holds approximately if there exists a low-error hiring policy.

Challenge: Heterogeneity across groups

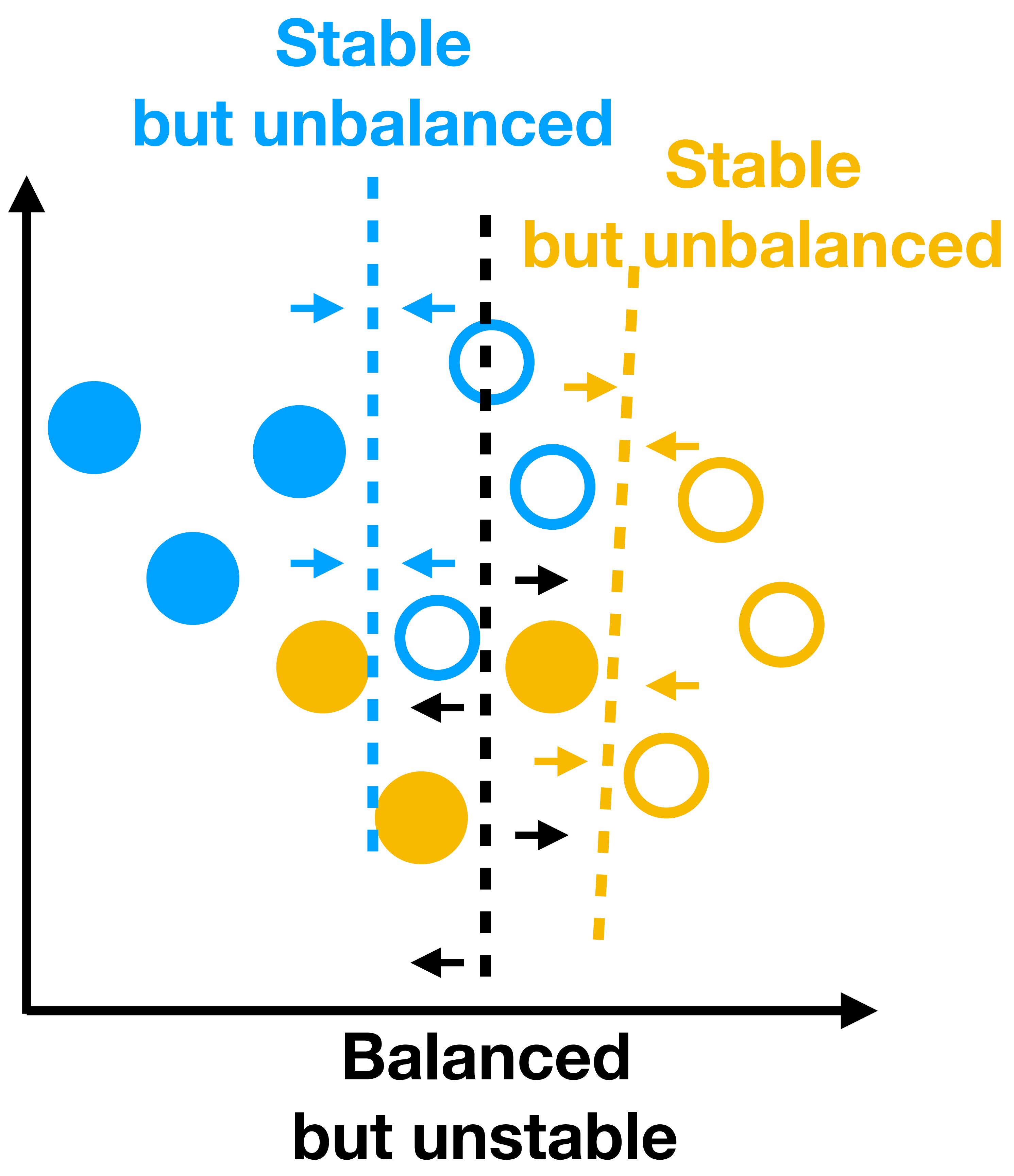
- There exists a **zero-error** hiring policy for each group separately but not together.

- <u>Result</u>: Then 2 types of equilibria exist

  1. Only one group has the optimal **qualification rate** (*unbalanced*) — **Stable**

  2. Both groups have the **same qualification rate** — **Unstable**

- Almost never converge to a "balanced" long term outcome, even if you started close to one!



Takeaways

- Long-term effectiveness of **interventions** depends on the dynamics

    1. **Decoupling** the hiring policy by group: *helps in the static setting, but not necessarily in the dynamic setting*

    2. **Subsidizing** the cost of investment in a disadvantaged group

(More details in paper!)

- Algorithms and re-training **impact human decisions** beyond their intended scope
    - Principled view of how feedback loops arise and implications for system design - more work is needed!

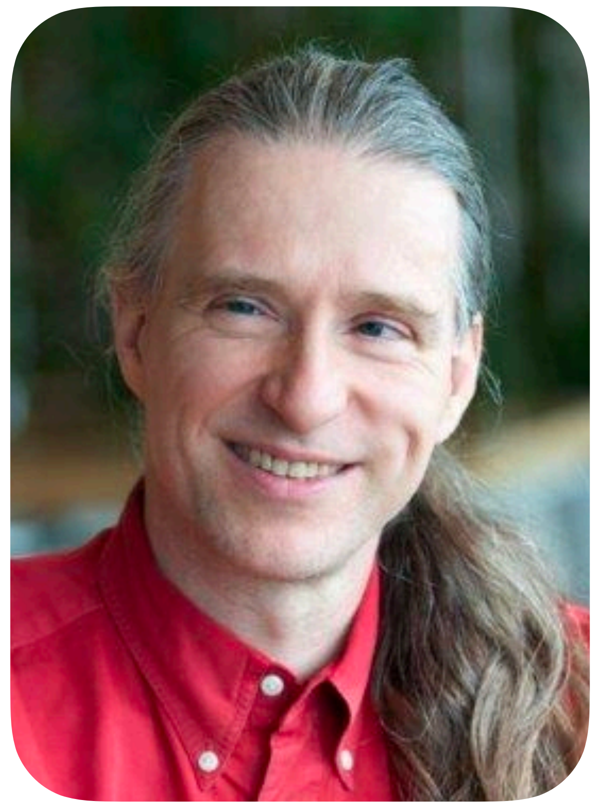
The *Disparate Equilibria* of Algorithmic Decision Making when Individuals Invest Rationally

Thank you!

Ashia Wilson

Nika Haghtalab

Adam Kalai



Christian Borgs

Jennifer Chayes