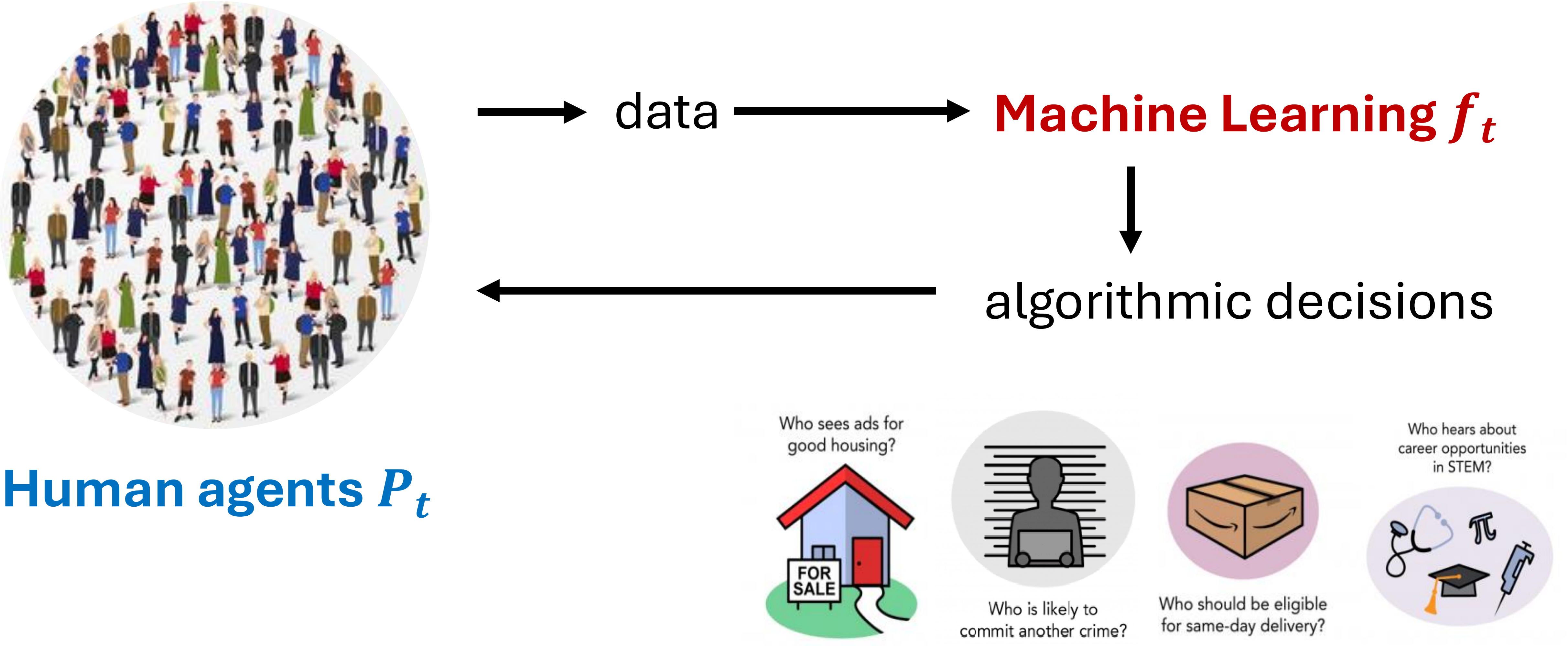
Long-Term Safety for Human-AI Ecosystem

Xueru Zhang, Assistant Professor
zhang.12807@osu.edu
Computer Science & Engineering
The Ohio State University

ML interacting with humans



data

Machine Learning f_t

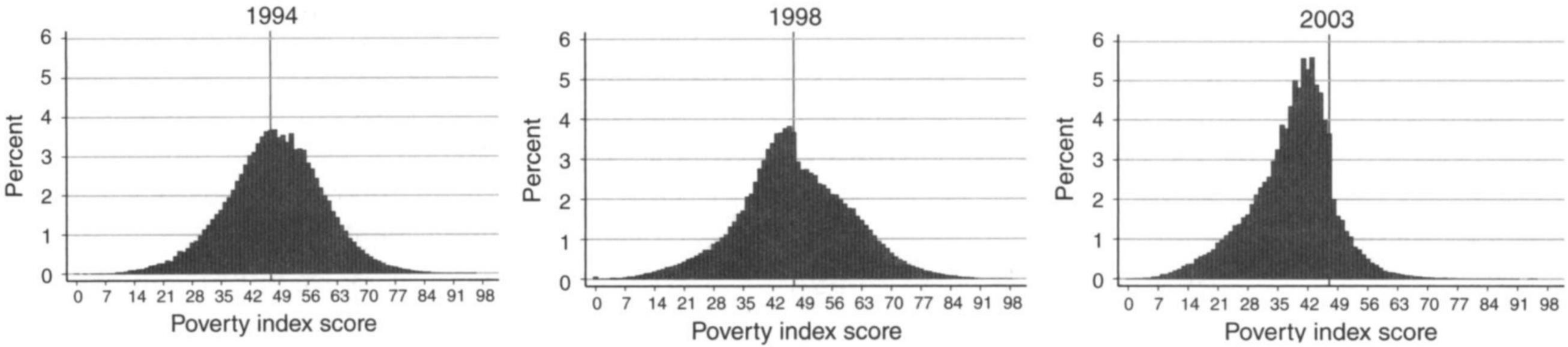
algorithmic decisions

Human agents P_t

Who sees ads for good housing?
FOR SALE
Who is likely to commit another crime?
Who should be eligible for same-day delivery?
Who hears about career opportunities in STEM?

ML reshapes the human agents

- Humans strategically manipulate data to receive favorable decisions.
  - E.g., Eligibility for social welfare program in Columbia



Source: Camacho, Adriana, and Emily Conover. "Manipulation of social program eligibility." American Economic Journal, 2011.

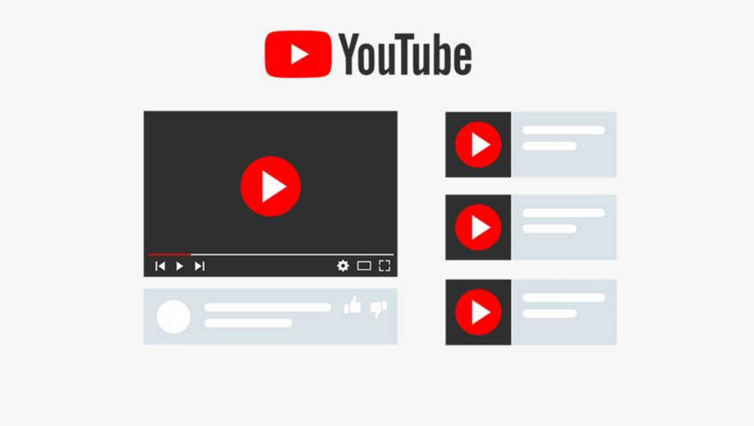
ML reshapes the human agents

- Financial models drive the market

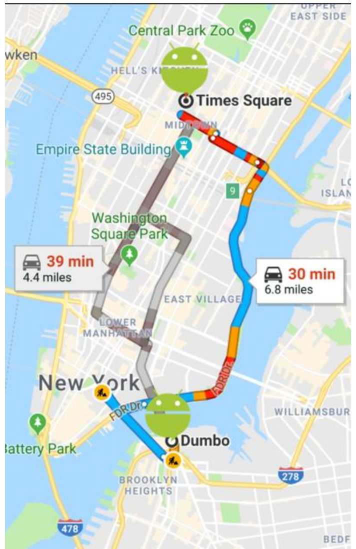
> *“Option pricing theory—a “crown jewel” of neoclassical economics—succeeded empirically not because it discovered preexisting price patterns but because once it was adopted, it pushed the market to conform to its predictions […].”*
>
> *MacKenzie & Millo, American Journal of Sociology, 2003*

ML reshapes the human agents

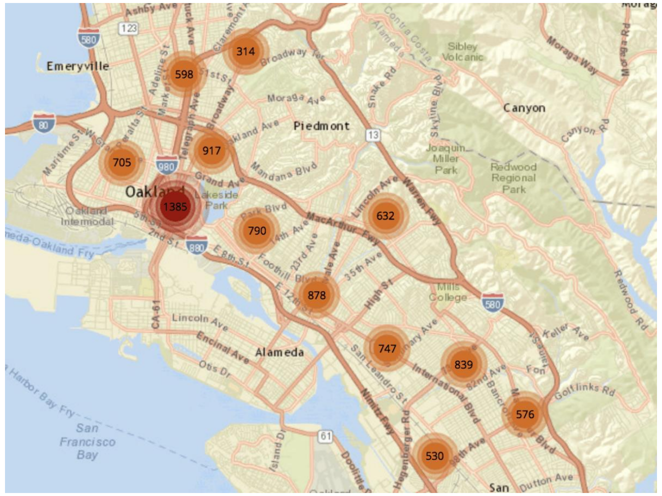
- More examples



Recommendations steer consumer preference & behavior



Route predictions affect traffic condition



Predictive policing tools affect crime patterns

ML interacting with humans

Human agents P_t

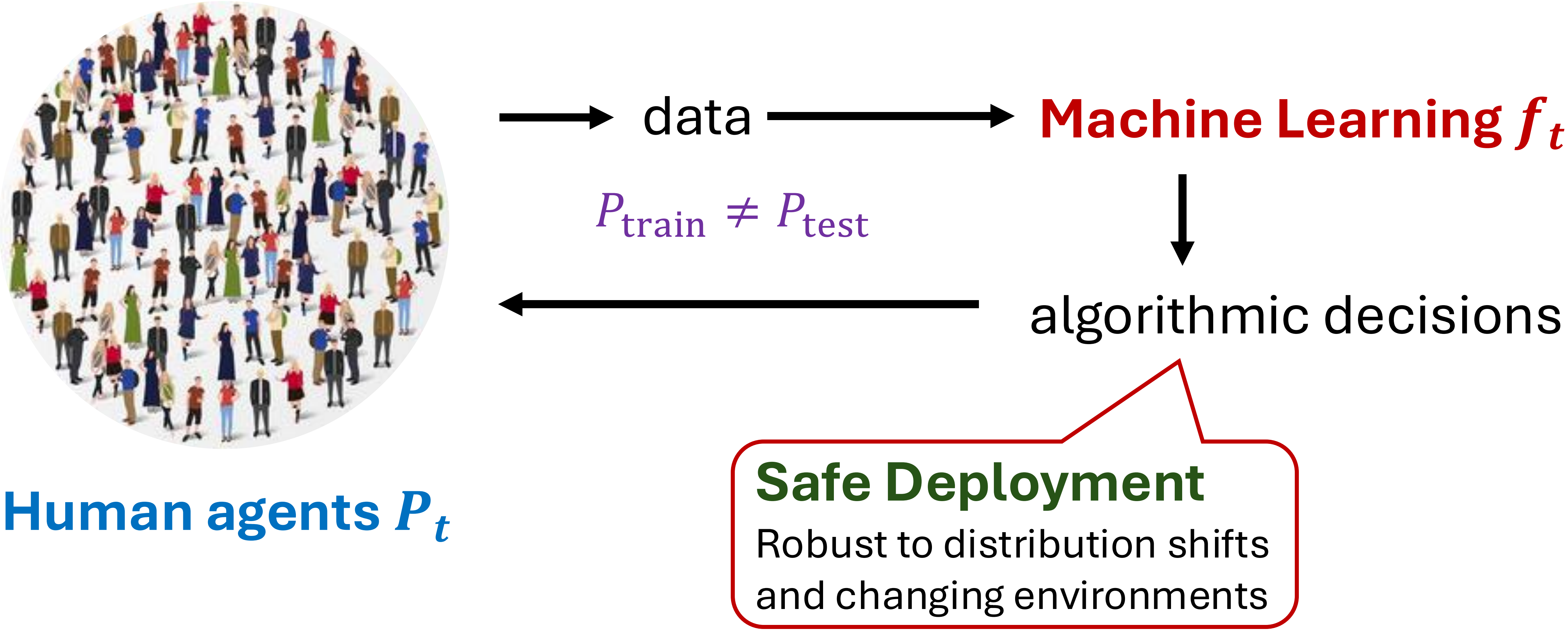
→ data → **Machine Learning f_t**

FAKE

Safe Training

Robust to noisy, erroneous, manipulated inputs

ML interacting with humans



data

Machine Learning f_t

$P_{\text{train}} \neq P_{\text{test}}$

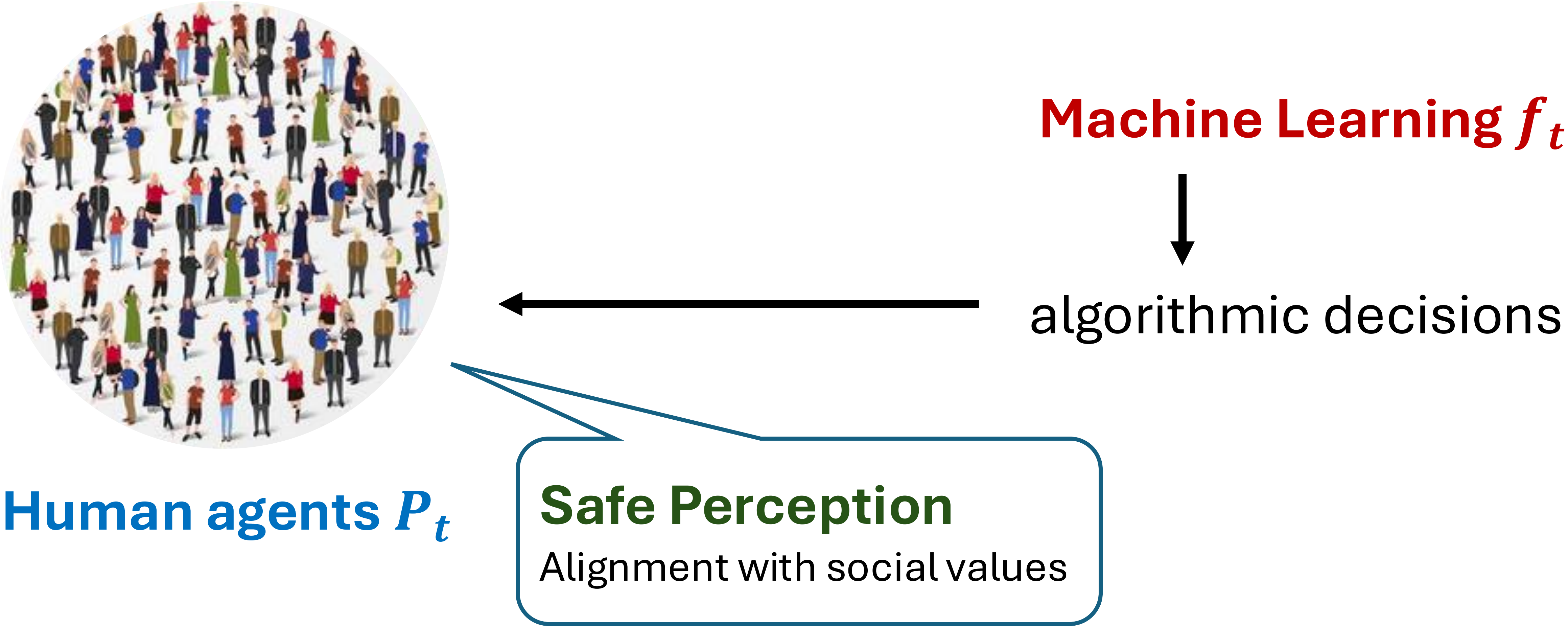
algorithmic decisions

Human agents P_t

Safe Deployment

Robust to distribution shifts and changing environments

ML interacting with humans



Machine Learning f_t

algorithmic decisions

Human agents P_t

Safe Perception

Alignment with social values

ML interacting with humans

Machine Learning f_t

algorithmic decisions

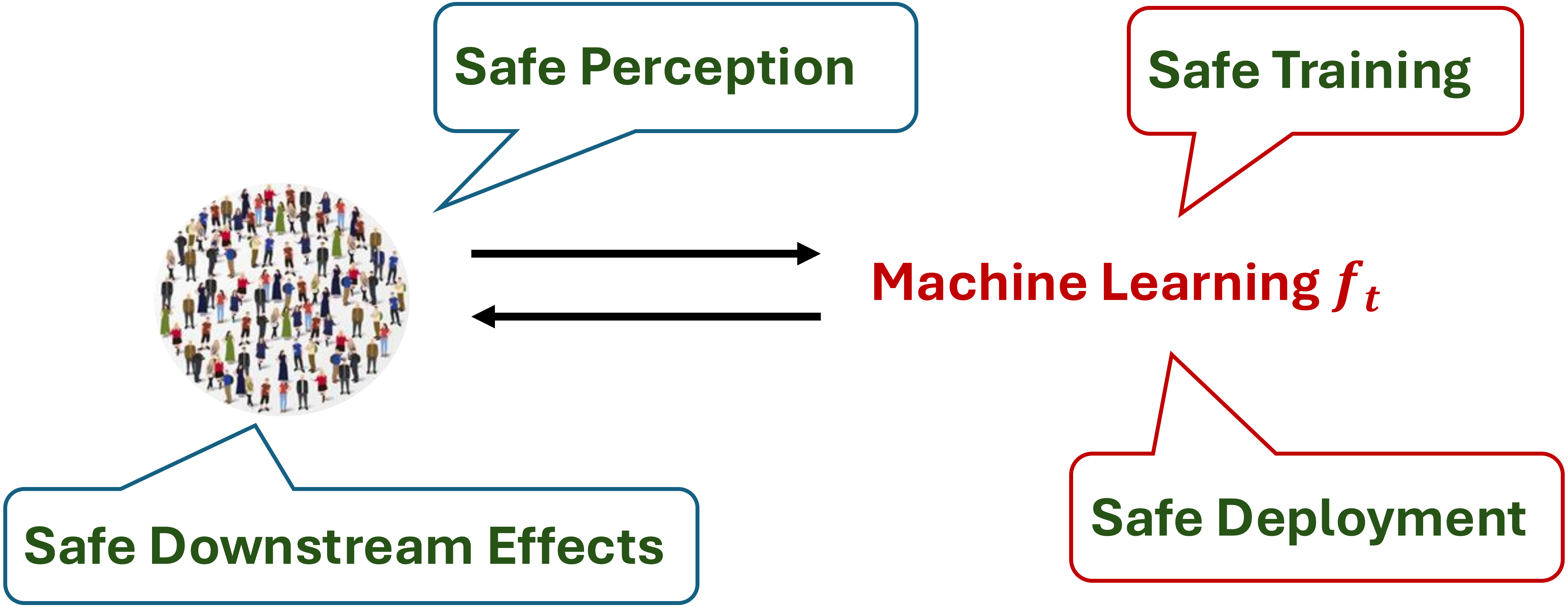
Human agents

$$\boldsymbol{P_{t+1} \sim \mathbb{T}_t(P_t, f_t)}$$

Safe Downstream Effects

Induce desirable and safe human behavior

Long-term safety for human-ML ecosystem



Safe Perception

Safe Training

Machine Learning f_t

Safe Downstream Effects

Safe Deployment

Long-term safety for human-ML ecosystem

- Modeling and incorporating feedback effects between humans and learning system
- Examining long-term impact of short-term safety assurance on human-ML system
- Developing learning methods and intervention mechanisms to enhance long-term safety
- Developing mechanisms to better understand agent dynamics

…

Happy to chat more!