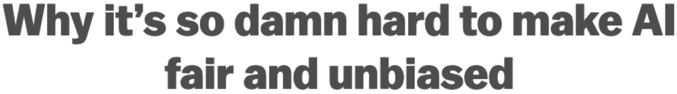
The political economy of AI:
Who controls the means of prediction?

Maximilian Kasy

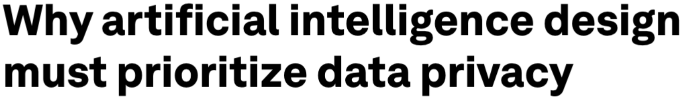
Department of Economics, University of Oxford

June 2024

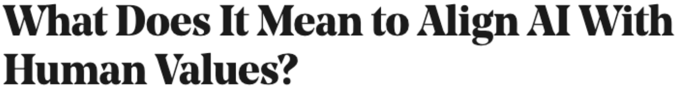
AI and its social impact in the news



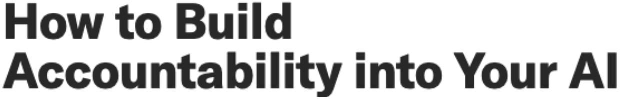
Why it’s so damn hard to make AI fair and unbiased



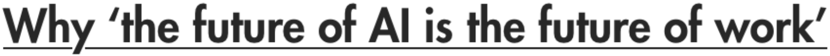
Why artificial intelligence design must prioritize data privacy



What Does It Mean to Align AI With Human Values?



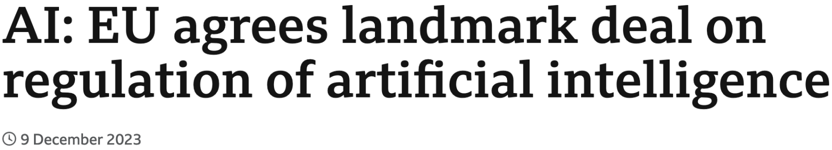
How to Build Accountability into Your AI



Why ‘the future of AI is the future of work’

Steps toward regulating AI

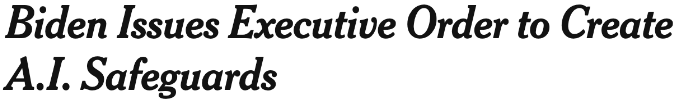
- European Union (BBC, Dec 2023):



AI: EU agrees landmark deal on regulation of artificial intelligence

🕓 9 December 2023

- United States (NYT, Oct 2023):



Biden Issues Executive Order to Create A.I. Safeguards

Introduction

- Concerns about the impact of AI:
    - Fairness, discrimination, and inequality.
    - Privacy, data property rights, and data governance.
    - Value alignment and the impending robot apocalypse.
    - Explainability and accountability.
    - Automation and wage inequality.
- Not just large language models and chatbots!
- Efforts to regulate AI.
- How can we think systematically about these questions?

Key takeaways of this talk

1. AI systems maximize a single, measurable objective.
2. In society, different individuals have different objectives.
   AI systems generate winners and losers.
3. Society-level assessments of AI
   require trading off individual gains and losses.
4. AI requires democratic control
   of algorithms, data, and computational infrastructure,
   to align algorithm objectives and social welfare.

Economics and machine learning

- Economics shares with AI and machine learning (ML) the languages of
  - optimization, and
  - probability.
- Economics, unlike AI and ML, considers
  - multiple agents
  - with unequal endowments,
  - conflicting interests, and
  - private information.
- Natural frameworks to think about the impact of AI:
  - Welfare economics,
  - social choice theory, and
  - causal inference.

Examples

- Algorithms for social networks / search engines select content to maximize user engagement, and ultimately ad revenue.
  - What about the impact on the public sphere and democracy?
  - What about (teenage) mental health?
- Algorithms for sales platforms set prices to maximize monopoly profits.
  - What about consumer welfare?
- Algorithms for hiring select job candidates who will contribute to profits; and who will not join a union.
  - What about equity, social mobility?
  - What about worker voice?

Roadmap

1. Background 1:
    - What is AI?
2. Background 2:
    - How do we measure social welfare?
    - Who could be agents of change?
3. The ethics, social impact, and regulation of AI:
    - Fairness, discrimination, and inequality.
    - Privacy, data property rights, and data governance.
    - Value alignment and the impending robot apocalypse.
    - Explainability and accountability.
    - Automation and wage inequality.

What is AI?

Social welfare and agents of change

The ethics and social impact of AI

AI is automated decisionmaking

- AI systems maximize measurable objectives:

  Russell and Norvig (2016), chapter 2:

  > *For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.*

- Leading approach: Machine learning (ML).
  Based on statistical inference.

- Other paradigms exist:
  Expert systems, automated reasoning.

Supervised learning

- Predicting outcomes Y given features X.
- Prediction $g(X)$, prediction loss $l(g(X), Y)$.
- Key ideas:
  Variance / bias tradeoff.
  Tuning using cross-validation.

Examples:

- Image recognition, voice recognition, automatic translation.
- Evaluation of job candidates / university applicants, bail setting in courts, credit scoring.
- Predicting ad clicks, user engagement.

Objective:

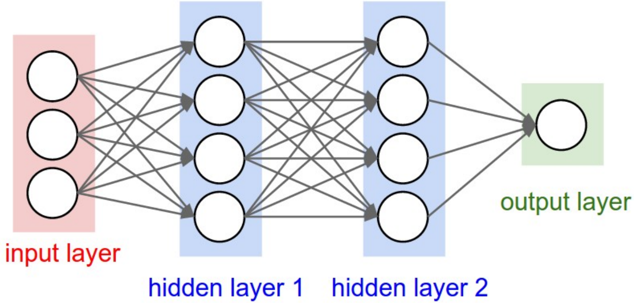
$$E[l(g(X), Y)]$$

Chihuahua or Muffin?

Deep learning

- One approach to supervised learning.
- Building prediction functions $g(\cdot)$ from many simpler functions ("neurons").
- Successful for large, rich data sets.

A neural net



Sloth or chocolate croissant?

Targeted treatment assignment

- Typically, prediction is only the first step.
- Often used to assign a treatment $W = h(X)$ based on features X.
- Maximize average outcomes Y among the treated.
  $\Rightarrow$ Treat if $g(X) > 0$.

Examples:

- Hiring job candidates.
- Giving credit.
- Admitting students.
- Choosing medical treatments.

Objective:

$$E[h(X) \cdot Y]$$

Multi-armed bandits

- Often we need to learn while taking actions.
- Maximize average outcomes over time.

⇒ Tradeoff between

1. *exploration*
   (experimenting to figure out what works),
2. and *exploitation*
   (using what we have learned).

Examples:

- Use a new medical treatment?
- Show a particular ad?
- Provide a training to an unemployed worker?

Objective:

$$\frac{1}{T}\sum_{i=1}^{T} Y_i$$

Key takeaways

- AI constructs systems
  which maximize a measurable objective (reward).
- Such systems take data as an input,
  and produce chosen actions as an output.

What is AI?

Social welfare and agents of change

The ethics and social impact of AI

Social welfare

Common presumption for many theories of justice:

- Normative statements about society
  are based on statements about individual welfare
- Formally:
  - Individuals $i = 1, \ldots, n$.
  - Individual i's welfare v_i.
  - Social welfare is a function of individuals' welfare

$$F(v_1, \ldots, v_n).$$

Many questions

- **Who is to be included** among $i = 1, \dots, n$?
  - All citizens? All residents? All humans on earth?
  - Future generations? Animals?
- **How to measure individual welfare** v_i?
  - Opportunities or outcomes?
  - Utility? Resources? Capabilities?
- **How to aggregate** to social welfare? How much do we care about
  - Millionaires vs. homeless people?
  - Sick vs. healthy people?
  - Groups that were victims of historic injustice?

How to measure individual welfare

Utilitarian approach:

- Dominant in economics
- Formally:
  - Choice set C_i.
  - Utility function $u_i(x)$, for $x \in C_i$.
  - Realized welfare
$$v_i = \max_{x \in C_i} u_i(x).$$
- Double role of utility
  - Positive: Individuals choose utility-maximizing x.
  - Normative: Welfare is realized utility.

Aggregating to social welfare

Welfare weights:

- Social welfare $F(v_1, \ldots, v_n)$.
- Define:
$$\omega_i := \frac{\partial}{\partial v_i} F(v_1, \ldots, v_n).$$
- Welfare weight ω_i measures how much we care about increasing welfare of i.
- There is no “objective” way to pick welfare weights.

Agents of change

- How do we ensure that the objectives maximized by AI align with maximizing social welfare $F(v_1, \ldots, v_n)$?
- Which agents have the interests, the values, and the capacity, to move technology and policy?
- Voluntary ethical behavior by corporate managers and engineers?
- Economics: Corporations are primarily profit maximizing.
  Profit maximization might not be aligned with social welfare maximization.
- Democratic control is necessary.
  Those affected by AI decisions need to have effective control over the objectives that are maximized.

Key takeaways

- Different individuals have different objectives.
  In terms of these objectives, AI systems generate winners and losers.

- Going from individual gains and losses to society-level assessments of AI requires aggregation, trading off individual gains and losses.

What is AI?

Social welfare and agents of change

The ethics and social impact of AI

Fairness, discrimination, and inequality

Standard view:
(Pessach and Shmueli, 2020)

- Fairness ≈ treating people of the same “merit” independently of their group membership.
- If an algorithm is maximizing firm profits then its decisions are fair by assumption.
- No matter how unequal the resulting outcomes within and across groups.
- Only deviations from profit-maximization are “unfair.”

Alternate view:
(Kasy and Abebe, 2021)

- Welfare / equality ≈ (counterfactual / causal) consequences of an algorithm for the distribution of welfare of different people.
- Fairness vs. equality:
  1. Improved prediction ⇒ Treatments more aligned with “merit.”
     Good for fairness, bad for equality.
  2. Affirmative action / redistribution:
     Bad for fairness, good for equality.

Privacy, data property rights, and data governance

Standard view:
(Dwork and Roth, 2014)

- Differential privacy.
  - It should make (almost) no observable difference whether your data are in a dataset.
  - No matter what other information is available to a decisionmaker.
- Machine learning performance is unaffected by differential privacy.
- Related:
  Individual property rights over data.

Alternate view:
(Viljoen, 2021)

- Primary use of data in ML is to learn *relationships*, not individual data.
  ⇒ Informational externalities.
  (Acemoglu et al., 2022)
- Privacy / property rights cannot prevent harms from AI.

⇒ Only democratic governance can address harms, not individual property rights.

Value alignment and conflicts of interest

Standard view: (Russell, 2019):

- Value alignment is a gap between human and machine objectives.
- Possible solutions:
  1. More careful engineering of objective functions.
  2. Infer objectives from observed human behavior (“inverse reinforcement learning”).

Alternate view:

- Value alignment is a gap between the objectives of those controlling the algorithm and the rest of society.
- Additionally:
  Not everything is observable, imposing fundamental limits on optimization.
- Possible solutions:
  1. Democratic control to align algorithm objectives with society.
  2. Refrain from deploying AI in some consequential settings.

Explainability and accountability

Standard view:

- Which algorithmic decisions can be "explained?" (Vredenburgh, 2022)
    - "Simple" mapping from data to decisions.
    - "Simple" is a moving target.
- Related: Who is responsible for algorithmic decisions?

Alternate view:

- We need transparency on objectives and constraints, not on algorithms.
- Complicated algorithms can have simple objectives.

⇒ Possibility of public debate on legitimate objectives.

⇒ Democratic control, rather than plutocracy, in the choice of objectives.

Automation and wage inequality

Standard view:
(Acemoglu and Autor, 2011)

- Production function framework :
  - Total output is a function of inputs: Workers, capital, technology.
  - Wage = marginal productivity.
- Technical progress without shared prosperity:
  - Change in technology such that
  - output increases, but
  - marginal productivity decreases.

Alternate view:

- AI is more than just another shifter of the production function.
  - Optimization of rewards,
  - by choosing actions
  - based on available data.
- Political economy:
  1. Who chooses the objective (reward function)?
  2. Who controls the data?
  3. Who controls the hardware and software to do the optimization?

Key takeaways

- Issues raised by AI:
  Fairness, privacy, value alignment, accountability, and automation.

- Resolving them requires democratic control of
  - algorithm objectives,
  - and of the means to obtain them:
    Data and computational infrastructure.

- Democratic control requires
  - public debate and
  - binding collective decision-making,
  - at many different levels of society.

Further reading

- **How AI works (and for whom)**
  Book forthcoming with University of Chicago Press.

- **The political economy of AI:**
  Towards democratic control of the means of prediction.
  Workingpaper available on my website.

Thank you!