Social foundations for statistics and machine learning

Maximilian Kasy

September 23, 2021

Today's argument

Single-agent **decision theory** provides the foundation for both statistics and machine learning (ML).

- **Examples** of decision problems:
  - Experimental design, estimation, inference, policy choice,
  - supervised learning, adaptive treatment choice.
- Components of decision problems:
  - Unknown **state** of the world,
  - which impacts the distribution of observed **data**,
  - as well as the **loss** function used to ultimately evaluate the decision.

The limits of decision theory

- This single-agent framework provides important insights.
- But it cannot address important scientific and societal challenges.

1. Replication crisis, publication bias, p-hacking, pre-registration, reforms of statistics teaching and the publication system.
2. The social impact of AI, algorithmic discrimination and inequality, value alignment of autonomous agents / robots.

Science and technology are social activities

- Scientific knowledge production, and the deployment of technology, are inherently social:
  1. Different agents have different objectives ("loss functions").
  2. The objectives of statistics and ML are socially determined – **who's objectives matter**?
  3. Replication crisis and reform proposals, and conflicts over the impact of AI can only be understood if we take this into account.
- This is well recognized by the **humanities**:
  - Philosophy, sociology, and history of science, science and technology studies.
  - But these fields do not provide **formal** and **prescriptive** recommendations for quantitative empirical researchers, or AI engineers.

Economics to the rescue

- Economics is well positioned to fill this gap:
  - We share the languages of **constrained optimization** and **probability theory** with statistics and ML.
  - But we are also used to considering **multiple agents** with unequal endowments, conflicting interests, private information.
- Today, I discuss two projects that are part of this general agenda:

Kasy, M. and Spiess, J. (2021). Rationalizing pre-analysis plans: Statistical decisions subject to implementability.
(work in progress)

Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision making.
(published, FAccT 2021)

Decision theory – a quick review

P-hacking and pre-analysis plans

Algorithmic fairness and economic inequality

Conclusion

AI as decision theory

The textbook “Artificial intelligence: a modern approach” (Russell and Norvig, 2016) defines the goal of AI as the derivation of

> *“ general principles of rational agents and on components for constructing them.”*

> *“An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.”*

> *“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.”*

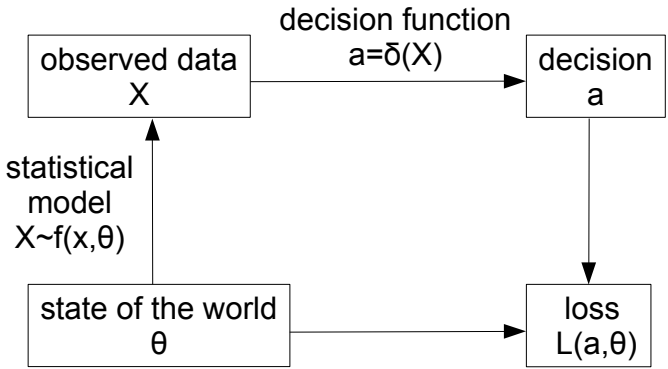
Statistics as decision theory

Similarly, the Bayesian statistics textbook by Robert (2007) states:

> *" Considering that the overall purpose of most inferential studies is to provide the statistician (or a client) with a decision, it seems reasonable to ask for an* ***evaluation*** *criterion* ***of decision procedures*** *that assesses the* ***consequences of each decision*** *and depends on the parameters of the model, i.e., the true* ***state of the world*** *(or of Nature)."*

> *" [...] implies a reinforced axiomatization of the statistical inferential framework, called Decision Theory. This augmented theoretical structure is necessary for Statistics to reach a coherence otherwise unattainable."*

Decision theory – General setup



Examples of decision problems

- **Estimation:**
  - Find an a which is close to some function μ of θ.
  - Typical loss function: $L(a, \theta) = (a - \mu(\theta))^2$.
- **Testing:**
  - Decide whether $H_0 : \theta \in \Theta_0$ is true.
  - Typical loss function: $L(a, \theta) = \mathbf{1}(a = 1,\ \theta \in \Theta_0) + c \cdot \mathbf{1}(a = 0,\ \theta \notin \Theta_0)$.
- **Targeted treatment assignment:**
  - Assign treatment W as a function of features X, $W = \delta(X)$.
  - Typical utility function: $E[\delta(X) \cdot (M - c)]$,
    for treatment effects M, treatment cost c.

Decision theory – a quick review

P-hacking and pre-analysis plans

Algorithmic fairness and economic inequality

Conclusion

P-hacking and pre-analysis plans

- Trial registration and pre-analysis plans (PAPs):
  A standard requirement for experimental research.
  - Clinical studies / medicine: Starting in the 1990s.
  - (Field) experiments in economics: More recently.
- Standard justification: Guarantee validity of inference.
  - P-hacking, specification searching, and selective publication distort inference.
  - Tying researchers' hands prevents selective reporting.
  - "PAPs are to frequentist inference what RCTs are to causality."
- Counter-arguments:
  - Pre-specification is costly.
  - Interesting findings are unexpected and flexibility is necessary.

No commitment (pre-registration) in decision theory

- Two alternatives, in a generic decision problem:
  1. We can commit to (**pre-register**) a rule $\delta(\cdot)$ before observing X.
  2. We can pick $\delta(X)$ after observing X.
- By the **law of iterated expectations**

$$\begin{aligned} E[L(\delta(X), \theta)] &= E[E[L(\delta(X), \theta)|X]] \\ &= \sum_x E[L(\delta(x), \theta)|X = x] \cdot P(X = x). \end{aligned}$$

- Therefore:
  - Picking the optimal $\delta(\cdot)$ (to minimize the sum) is the same
  - as picking the optimal $\delta(x)$ for every value of x (each term of the sum).
  - The decision-problem is **dynamically consistent**.

A mechanism design perspective

Claim: Concerns about p-hacking, publication bias, pre-registration are at their core about **divergent interests** between **multiple actors**.

Q: How to incorporate this social dimension into prescriptive methodology?

A: Model statistical inference as a **mechanism design** problem!

- Take the perspective of a reader of empirical research who wants to implement a statistical decision rule.
- Not all rules are implementable when researchers have divergent interests and private information about the data, and they can selectively report to readers.
- Agenda: Characterize optimal decision rules subject to implementability.

Setup

- Two agents: Decision-maker and analyst.
- The analyst observes a vector

$$X = (X_1, \ldots, X_{\bar{n}}),$$

where

$$X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta).$$

- Analyst: Reports a subvector X_I to the decision-maker, where

$$I \subset \{1, \ldots, \bar{n}\}.$$

- Decision-maker: Makes a decision

$$a \in \{0, 1\},$$

based on this report.

Prior and objectives

- Common prior:
$$\theta \sim \text{Beta}(\alpha, \beta).$$
- Analyst's objective:
$$u^{\text{an}} = a - c \cdot |I|.$$
$|I|$ is the size of the reported set,
c is the cost of communicating an additional component.
- Decision-maker's objective:
$$u^{\text{d-m}} = a \cdot (\theta - \underline{\theta}).$$
$\underline{\theta}$ is a commonly known parameter.
Minimum value of θ beyond which the decision-maker would like to choose $a = 1$.

Timeline

1. The decision-maker commits to a decision rule
$$a = a(J, I, X_I).$$
2. The analyst reports a PAP
$$J \subseteq \{1, \ldots, \bar{n}\}.$$
3. The analyst next observes X, chooses $I \subseteq \{1, \ldots, \bar{n}\}$, and reports
$$(I, X_I).$$
4. The decision rule is applied and utilities are realized.

Implementability

- Let x denote values that the random vector X may take.
- Reduced form mapping (statistical decision rule)

$$x \mapsto \bar{a}(x).$$

- $\bar{a}(x)$ is implementable
  if there exist mappings $I(x)$ and $a(I, x_I)$
  such that for all x

$$\bar{a}(x) = a(I(x), x_{I(x)}),$$

  and

$$I(x) \in \operatorname*{argmax}_I \; a(I, x_I) - c \cdot |I|.$$

Notation

- Successes among all components: $s(X) = \sum_{i=1}^{\bar{n}} X_i$.
  Successes among the subset I: $s(X_I) = \sum_{i \in I} X_i$.
  Failures among the subset I: $t(X_I) = |I| - s(X_I)$.

- Maximal number of components the analyst is willing to submit:

$$\bar{n}^{PC} = \max\{n : \ 1 - cn \geq 0\} = \lfloor 1/c \rfloor .$$

- First-best cutoff for the decision-maker:

$$\underline{s}^*(n) = \min\left\{\underline{s} : \ E[\theta | s(X_{1,\ldots,n}) = \underline{s}] \geq \underline{\theta}\right\} .$$

- Minimal cutoff for the decision-maker:

$$\underline{s}^{min}(n) = \min\left\{\underline{s} : \ E[\theta | s(X_{1,\ldots,n}) \geq \underline{s}] \geq \underline{\theta}\right\} .$$

Symmetric decision rules

- Consider symmetric rules of the form

$$a(s(X_I), t(X_I)),$$

Proposition (Optimal symmetric decision rule)

The optimal reduced-form decision rule that is symmetrically implementable takes the form

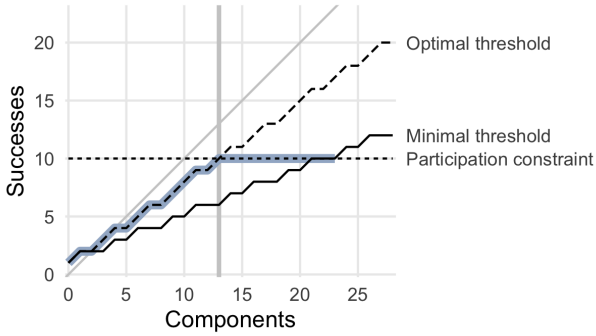
$$\bar{a} = \mathbf{1}(s(X) \geq \min(\underline{s}^*, \bar{n}^{PC})),$$

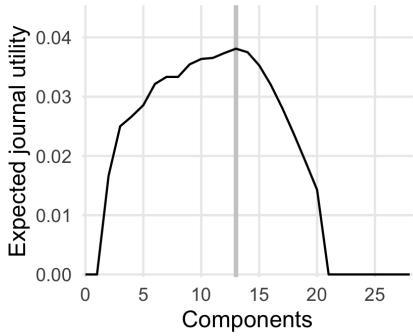
if $\bar{n}^{PC} \geq \underline{s}^{min}$*, and can be implemented by*

$$a = \mathbf{1}(s(X_I) \geq \min(\underline{s}^*, \bar{n}^{PC})).$$

Otherwise the optimal decision rule is given by $a \equiv 0$.

Symmetric cutoff without PAP, uniform prior





If the number of components $\bar{n}$ is to the right of the maximum $\bar{n}^*$:

- PAPs increase decision-maker welfare
- by forcing the analyst to ignore all components $i > \bar{n}^*$.

Decision theory – a quick review

P-hacking and pre-analysis plans

Algorithmic fairness and economic inequality

Conclusion

Algorithmic fairness and economic inequality

1. Standard definitions of algorithmic **fairness**:
   - Absence of "bias."
   - Similar to "taste based discrimination" in economics, which is defined as a deviation from profit maximization.
   - Fairness as a **decision problem**, *aligning treatment and latent merit*.
2. Social choice theory (in economics), theory of justice (in political philosophy):
   - **Social welfare** is typically defined based on individuals' welfare.
   - Points of contention: How to measure individual welfare, how to **aggregate** / trade off welfare across individuals.
   - Policies (and algorithms!) are evaluated based on their **consequences** for social welfare.

"Bias" versus "social welfare" have very different implications!

Fairness in algorithmic decision making – Setup

- Binary treatment W, treatment return M (heterogeneous), treatment cost c.
  Decision maker's objective

$$\mu = E[W \cdot (M - c)].$$

- All expectations denote averages across individuals (not uncertainty).

- M is unobserved, but predictable based on features X.
  For $m(x) = E[M|X = x]$, the optimal policy is

$$w^*(x) = \mathbf{1}(m(X) > c).$$

Examples

- Bail setting for defendants based on predicted recidivism.
- Screening of job candidates based on predicted performance.
- Consumer credit based on predicted repayment.
- Screening of tenants for housing based on predicted payment risk.
- Admission to schools based on standardized tests.

Definitions of fairness

- Most definitions depend on **three ingredients**.
  1. Treatment W (job, credit, incarceration, school admission).
  2. A notion of merit M (marginal product, credit default, recidivism, test performance).
  3. Protected categories A (ethnicity, gender).
- I will focus on the following **definition of fairness**:

$$\pi = E[M|W = 1, A = 1] - E[M|W = 1, A = 0] = 0$$

"Average merit, among the treated, does not vary across the groups $\boldsymbol{a}$*."*

This is called "predictive parity" in ML,
the "hit rate test" for "taste based discrimination" in economics.

Fairness and $\mathscr{D}$'s objective

Observation

Suppose that W, M *are binary ("classification"), and that*

1. $m(X) = M$ *(perfect predictability), and*
2. $w^*(x) = \mathbf{1}(m(X) > c)$ *(unconstrained maximization of* $\mathscr{D}$*'s objective* μ*).*

Then $w^*(x)$ *satisfies predictive parity, i.e.,* $\pi = 0$.

In words:

- If $\mathscr{D}$ is a firm that is maximizing profits and observes everything then their decisions are fair by assumption.
  - No matter how unequal the resulting outcomes within and across groups.
- Only deviations from profit-maximization are "unfair."

Three normative limitations of “fairness” as predictive parity

Notions of fairness of this form have several key limitations:

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

3. Fairness-based perspectives **focus on categories** (protected groups)
   and ignore within-group inequality.

Social welfare as an alternative framework

- The framework of fairness / bias / discrimination contrasts with perspectives focused on *consequences for social welfare*.
- Common presumption for most theories of justice:

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

$$SWF = F(Y_1, \ldots, Y_n).$$

- Key points of contention:
  1. Who is included among the individuals i? Who's lives matter?
  2. How to measure individual welfare Y_i?
  3. How to trade off welfare across individuals i?

The impact on inequality or welfare as an alternative to fairness

- Outcomes are determined by the **potential outcome equation**

$$Y = W \cdot Y^1 + (1 - W) \cdot Y^0.$$

- The **realized outcome** distribution is given by

$$p_{Y,X}(y,x) = \left[p_{Y^0|X}(y,x) + w(x) \cdot \left(p_{Y^1|X}(y,x) - p_{Y^0|X}(y,x) \right) \right] \cdot p_X(x).$$

- What is the impact of $w(\cdot)$ on a **statistic** ν?

$$\nu = \nu(p_{Y,X}).$$

  Examples: Variance, quantiles, between group inequality, **social welfare**.

- Cf. Distributional decompositions in labor economics!

When fairness and equality are in conflict

- Fairness is about **treating** people of the same "**merit**" independently of their **group** membership.
- Equality is about the (counterfactual / causal) **consequences** of an algorithm for the distribution of **welfare** of different **people**.

Examples when they are in conflict:

1. Increased surveillance / **better prediction** algorithms:
   Lead to treatments more aligned with "merit"
   Good for fairness, bad for equality.
2. Affirmative action / **compensatory interventions** for pre-existing inequalities:
   Bad for fairness, good for equality.

Conclusion

- These two projects discuss settings
  showing the limitations of decision theory
  for understanding important current debates.

- Other authors (in econometrics, economic theory, computer science, and elsewhere) are exploring related ideas!

The road ahead:

1. Reconceptualize statistics as knowledge production and communication in a social context, involving diverse actors with divergent interests.

   Provide formal, prescriptive guidance for researchers based on this perspective.

2. Develop an understanding of algorithmic decision making / ML / AI
   based on the awareness that different people have different objectives,
   and control rights over data and algorithms make a difference.

Thank you!