The Right to be an Exception to a Data-Driven Rule

Sarah H. Cen & Manish Raghavan

AI Ethics Reading Group

November 17, 2022

We make sense of our world through **rules**.

But, to every rule, there are **exceptions**.

What happens to individuals on which the rule fails?

Sentencing decisions

Mandatory minimum sentences (1970s)

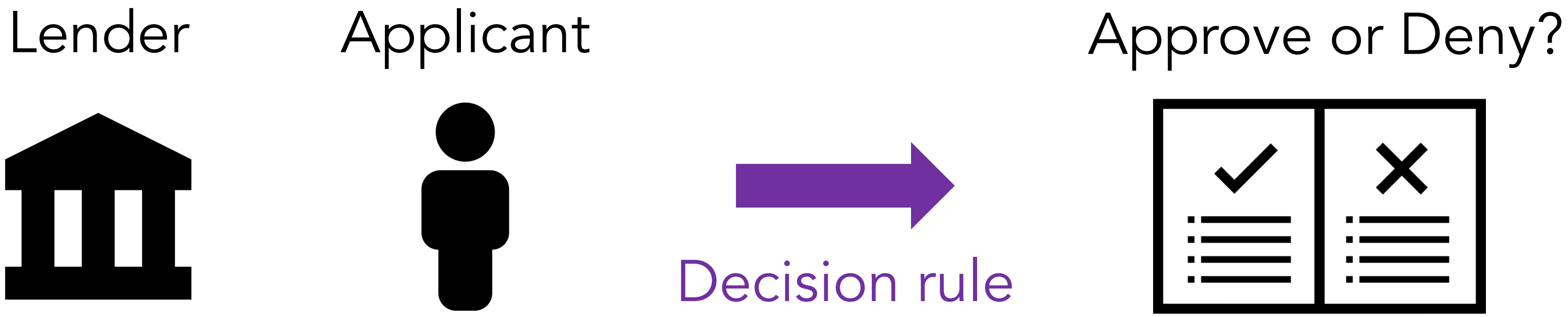
- Standardized set of rules
- Intended to improve fairness, predictability, & consistency

Lockett v. Ohio (1978)

- No mandatory minimum sentences for capital cases
- Requires consideration of a case's particular circumstances
- Due to "seriousness and irrevocability of the death penalty"

Data-driven exceptions

Data-driven rule: decision rule behind data-driven decision aid.



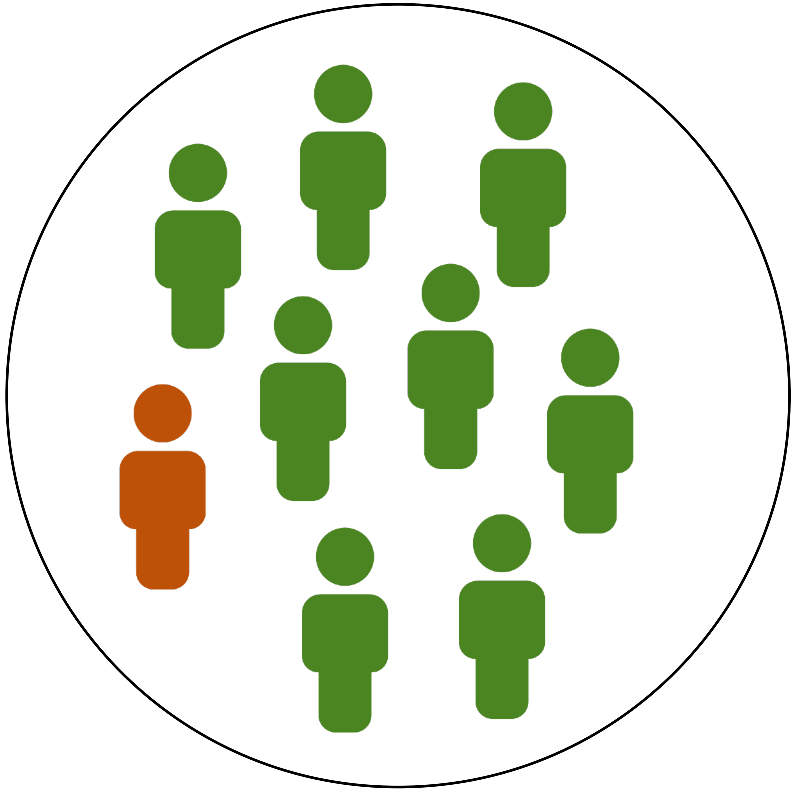
An applicant may be approved under some rules but not others.

Exceptions are natural.

Data-driven exceptions matter because:

1. ML ↔ statistical **averages**
2. ML can be applied **rapidly** and **repeatedly**
3. Data-driven rules are **non-intuitive**

Example: Exceptions in healthcare



Common cold

Fatal disease

Treated as average of statistically similar individuals?

vs.

Rule out exceptional (high-risk) cases?

Right to be an Exception

To a Data-Driven Rule

Individual rights

Rights in the age of AI

- Right to be forgotten (EU, 2014)
- Right to reasonable inferences (Wachter, 2019)
- Right to rectification (GDPR, 2016)
- Right to access (GDPR, 2016)
- ...

Goal: redistribute power back to decision subjects.

Right to be considered an exception to a data-driven rule

When the risk of harm is high, a data-driven decision-maker must adopt the presumption that the subject may be an exception to the data-driven rule.

They must inflict harm only if they have applied the appropriate care and diligence in ruling out the possibility that the decision-subject is a data-driven exception.

Right to be considered an exception

1. **Harm**
2. **Individualization**
3. **Uncertainty**

Foundations

Data-driven decision-making

Data-driven rule: decision rule behind data-driven decision aid.

Decision subject: individual directly impacted by decision.

Why exceptions arise:

1. **Sampling bias**: small number of observations.
2. **Model (in)capacity**: can only do well on some individuals.
3. **Distribution shift**: learns model on different population.
4. **Partial observability**: minorities look like majority to model.
5. **Initialization**: sensitivity to random weights initially assigned.
6. ...

Moving away from averages

Are there protections for individuals who fall through the cracks?

Surprisingly few.

Most still rely on average-based notions.

Ex: Some believe improving accuracy justifies a method.

But accuracy is an average-based notion!

Loomis v. Wisconsin (2017)

Loomis v. Wisconsin

From the ruling:

1. Although algorithm is secret, no relevant information is hidden from Loomis because he knows inputs and outputs.
2. Use of gender by algorithm was not discriminatory and promoted accuracy to the benefit of defendants.

Loomis: algorithm is secret → violates right to due process

If argue on basis of accuracy (*average* notion), an *individual* will always lose.

Need new language: **harm, individualization, uncertainty**!

The three ingredients

Harm, individualization, and uncertainty

Element #1: Harm

Measurement stick: What level of care, skill & diligence required?

Weighs right against other stakeholder interests.

Ex: Individualized sentencing vs. judicial economy.

How to measure harm?

"Significant effects" (Kaminski & Urban, 2021)

"High-risk inferences" (Wachter & Middelstadt, 2019)

"Risk methodology" (EU AI Act, 2021)

Element #2: Individualization

Individualization: tailoring a rule to specific circumstances.

Shifts from aggregate to individual.

An information concept → considering totality of circumstances.

Limitations to individualization in data-driven rules.

Even if a data-driven rule were fully individualized (incorporated all relevant features), would this be enough?

Element #3: Uncertainty (Part I)

Exceptions defy general rules.

So, is can we just improve individualization? No.

(This is where we differ from existing proposals.)

Why? Always sources of uncertainty.

Ex: Suppose individualized by incorporating more info.

The more tailored, the less data (i.e., less evidence).

Even if sufficient data, unremovable sources of doubt.

Element #3: Uncertainty (Part II)

Two types of uncertainty:

1. **Epistemic**: reducible uncertainty from lack of knowledge.
2. **Aleatoric**: irreducible uncertainty from "unknowability"

    e.g., randomness or too many factors

Individualization reduces epistemic, but not aleatoric, uncertainty.

Ex: College student's performance is not predetermined.

Computational irreducibility (Wolfram, 2002)

Example: Parole decisions (Part I)

Problem: Average outcomes over those who look statistically similar.

- Washes out details that make defendant unique
- Defendant judged based on actions of others, not their own

(This uncertainty matters when risk of harm is high!)

"**Treat[s] the wrongdoing by some as justification for imposing extra costs on others**"

(Jorgensen, 2021)

Individualization only ensures that instead of paying for wrongs of everyone, a defendant pays for wrongs of people increasingly similar to them. Uncertainty & harm matter!

Example: Parole decisions (Part II)

When the risk of **harm** is high, level of **individualization** & **uncertainty** matter.

“It is morally negligent or reckless to intentionally harm someone unless we have not only reasonably high credence [...] Very roughly: our present evidence must be such that little if any new information [...] would cause our credence to drop.”

(Jorgensen, 2021)

Rejecting the presumption that the defendant is law-abiding should follow only if the judge’s belief is so strong that very little if any new information would sway it.

This is inherently a balance of harm, individualization & uncertainty.

Tying them together

Three elements: (1) Harm, (2) Individualization, & (3) Uncertainty

Harm: Determines the level of consideration.

Individualization: Shifts attention away from aggregate.

Uncertainty: Emphasizes limits of data.

When a decision may inflict harm, should only inflict harm when certainty is high enough. More risk → more certainty.

So, what's the point?

Harm, individualization, and uncertainty **map between legal and machine concepts**.

We depart from previous discussions in two ways:

1. Going beyond individualization
2. Accuracy not the right notion

Operationalizing the right

Legal measures

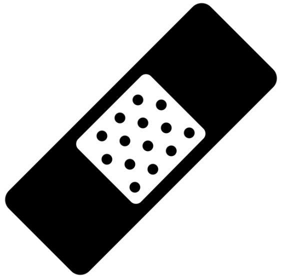
<u>Ex ante</u> measures

Responsibility of decision makers *before* deploying an algorithm

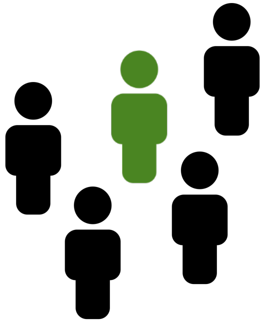
<u>Ex post</u> measures

Post-deployment rights of individuals affected by the algorithm

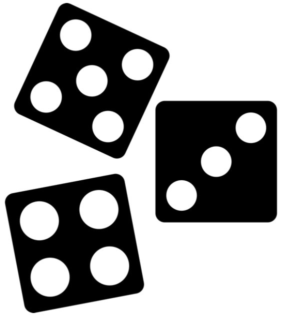
Ex ante legal measures



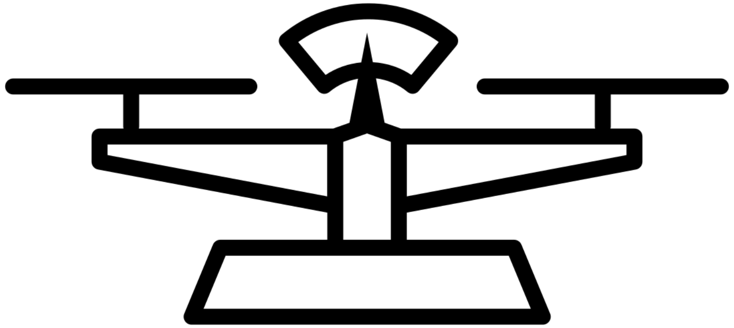
Harm



Individualization



Uncertainty



Ex post legal measures

Accountability through contestation.

cf. Kaminski & Urban (2018)

Ex: Title VII of US Civil Rights Act ("disparate impact" clause)

1. Disparate impact

2. Business necessity

3. Alternative rule

Technical concepts

1. **Causal inference**: Shifts way from frequency analyses

   Instead determines what factors led to outcome.

2. **Robust optimization**: Accounts for unlikely outcomes.

   Emphasizes uncertainty.

3. **Algorithmic fairness**: Aligns algorithmic values w/ ours.

   Recent shift toward individual fairness.

Takeaways

Exceptions are natural in data-driven decision-making.

Averages + systemic + non-intuitive → need protections

Does not imply every individual *is* an exception. But when decision inflicts harm, consider the possibility the subject may be an exception.

Three elements:

1. **Harm** provides measure of risk.
2. **Individualization** ensures fine-grained consideration.
3. **Uncertainty** capture inherent limits.

Thank you!

Questions?