UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

2 Ethical Data Science

Decision-Making, Values and Legitimacy

Jordi Vitrià


Does it
make smense
TO TO use
AI?

Legitimacy

When considering a possible use case of AI, we can ask this question:

Does it make sense to use AI?

The mother of all ethical AI questions

Legitimacy

According to institutional theory (politics), **legitimacy** refers to the **congruence between organizational activities and their cultural environment**.

The legitimacy of a political system depends on various factors: **how well** it achieves its goals, whether the subjects of the political system are **involved in developing the rules**, and whether the decision subject has the ability to **challenge decisions**.

It is the central problem of politics.

Legitimacy

Legitimacy represents an important form of **social evaluation**, as it is indispensable for the acceptance and diffusion of novel technologies.

The legitimacy question should precede other ethical questions such as discrimination or privacy.

Legitimacy

In these scenarios there are “legitimacy issues”:

- A student is proud of the creative essay she wrote for a standardized test. She receives a perfect score, but is disappointed to learn that the test had in fact been graded by a computer.

- A defendant finds that a criminal risk prediction system categorized him as high risk for failure to appear in court, based on the behavior of others like him, despite he had every intention of appearing in court on the scheduled date.

- An automated system locked out a social media user for violating the platform’s policy on acceptable behavior. The user insists that they did nothing wrong, but the platform won’t provide further details nor any appeal process.

Which are the issues?

Legitimacy

How do we evaluate the legitimacy of AI for taking **high-stake decisions**?

We need to understand which are the **critical issues** of a **high-stake decision** and how do they interact with a set of **values**.

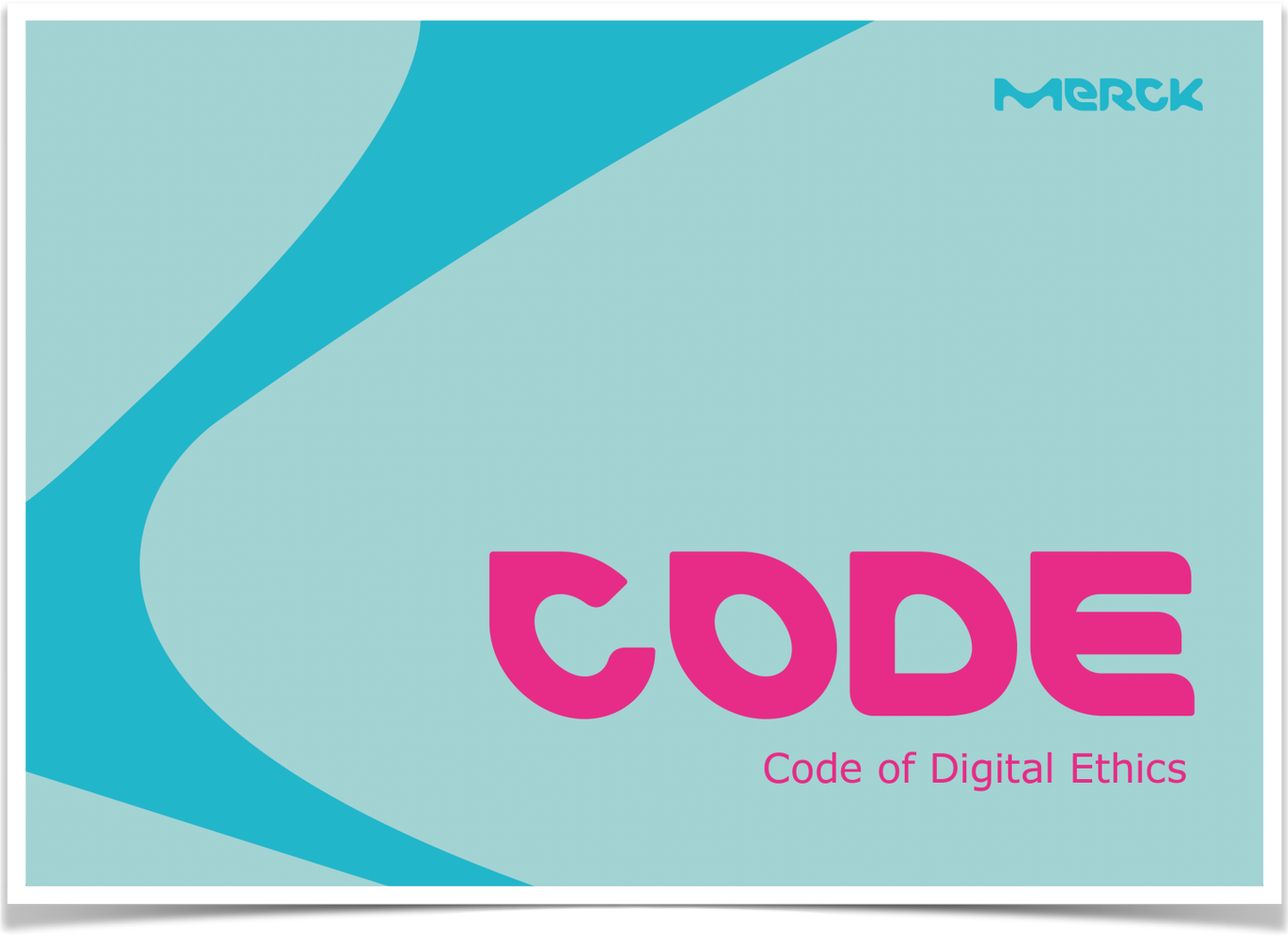
Values & Technology

Values/Principles and technology

Organizations define their values regarding AI through a multifaceted process. Here's a general outline of how this process might unfold:

- **Mission and Vision Alignment**: Companies begin by ensuring that their AI values align with their broader mission and vision. This means considering how AI can help achieve their goals while adhering to the ethical standards they've set for themselves.
- **Stakeholder Engagement**: They engage with various stakeholders, including employees, customers, partners, and potentially affected communities, to gather insights and perspectives on the ethical use of AI. This inclusive approach helps ensure that the company's AI values are considerate of diverse viewpoints and concerns.
- **Ethical Standards and Principles**: Many companies adopt ethical frameworks or principles specific to AI. These often include commitments to transparency, fairness, accountability, privacy, and ensuring that AI technologies do not cause harm. **These principles guide the development and deployment of AI systems**.
- **Regulatory and Industry Standards Compliance**: Companies also consider existing and anticipated regulations governing AI in their jurisdictions, as well as industry best practices and standards. This helps ensure that their AI values and practices are not only ethical but also legally compliant.

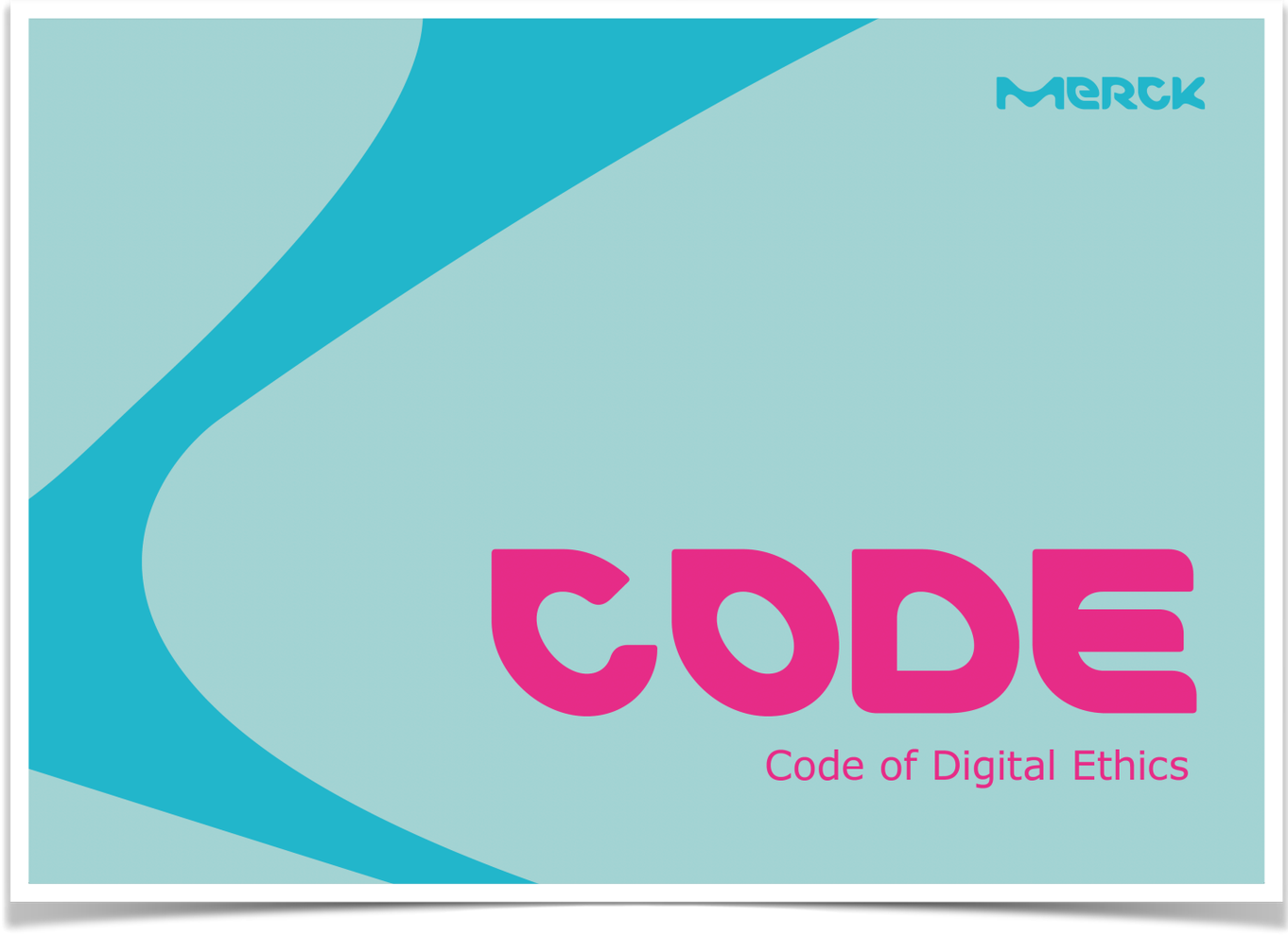
Example



It is based on five core principles: **justice**, **autonomy**, **beneficence**, **non-maleficence**, and **transparency**.

Example

- **Justice**:
  - Impartiality
  - Equality
  - Proportionality
- **Autonomy**:
  - Explainability
  - Privacy
  - Literacy
- **Non-Maleficence:**
  - Reliability
  - Controllability
  - Accountability
- **Transparency**:
  - Comprehensibility
  - Interactivity
  - Traceability
- **Beneficence**:
  - Security
  - Sustainability
  - Responsibility

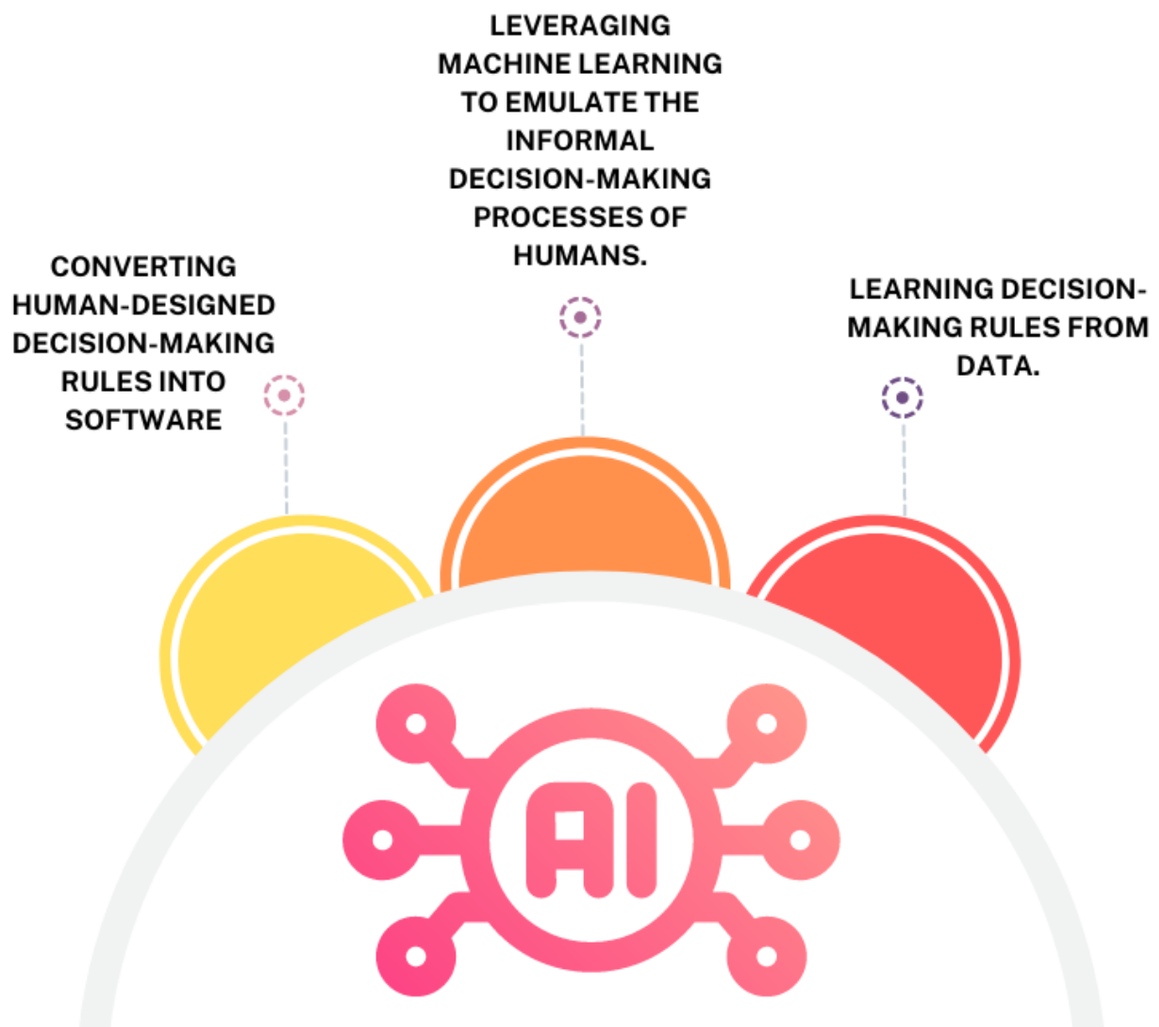


Decision-making automation

Decision-making automation refers to the use of technology, particularly software and algorithms, to automate the process of making decisions that were traditionally made by humans.

Kinds of automatic decision-making systems

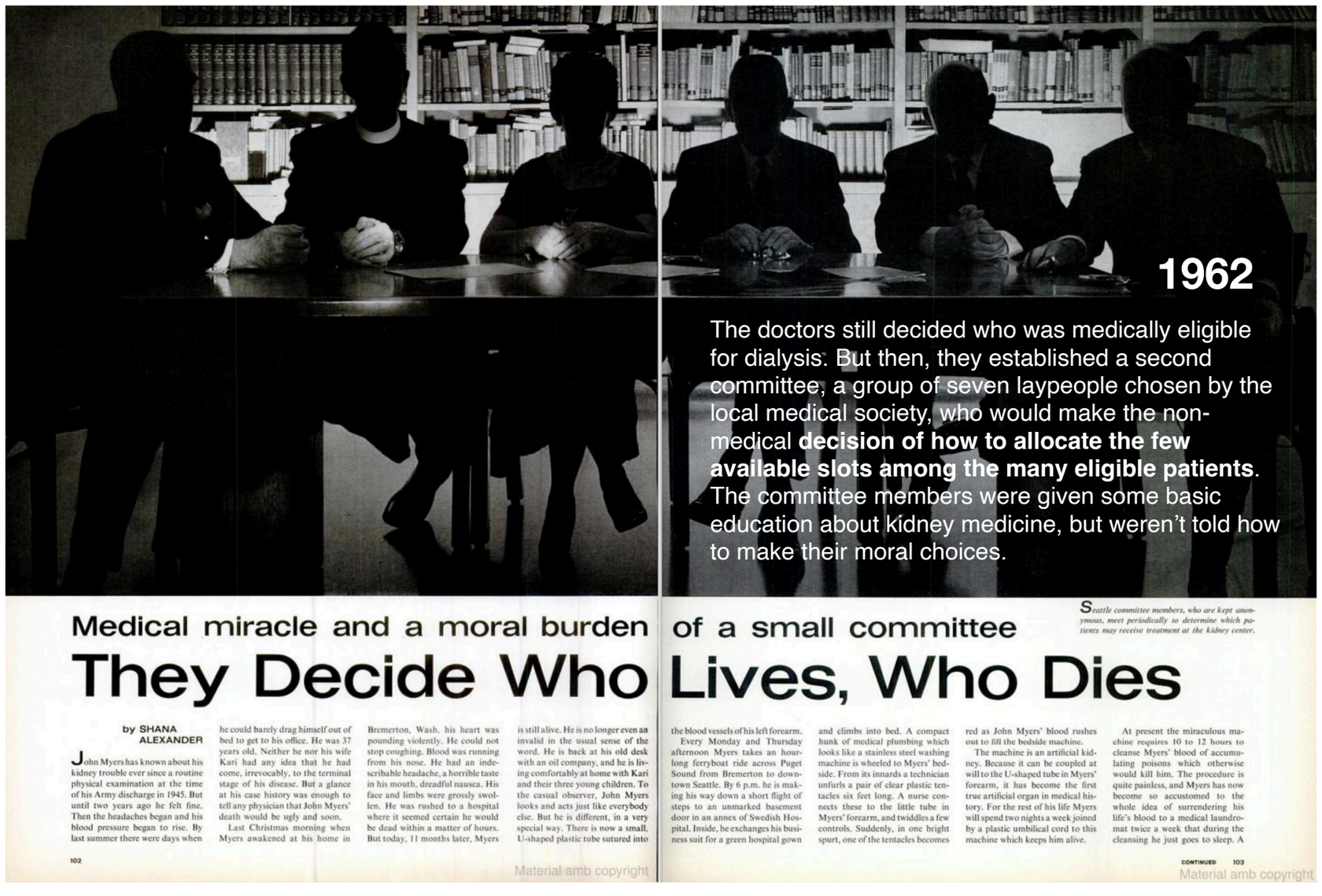
In the context of decision-making, **automation** can be categorized into three distinct types.



Kinds of automatic decision-making systems

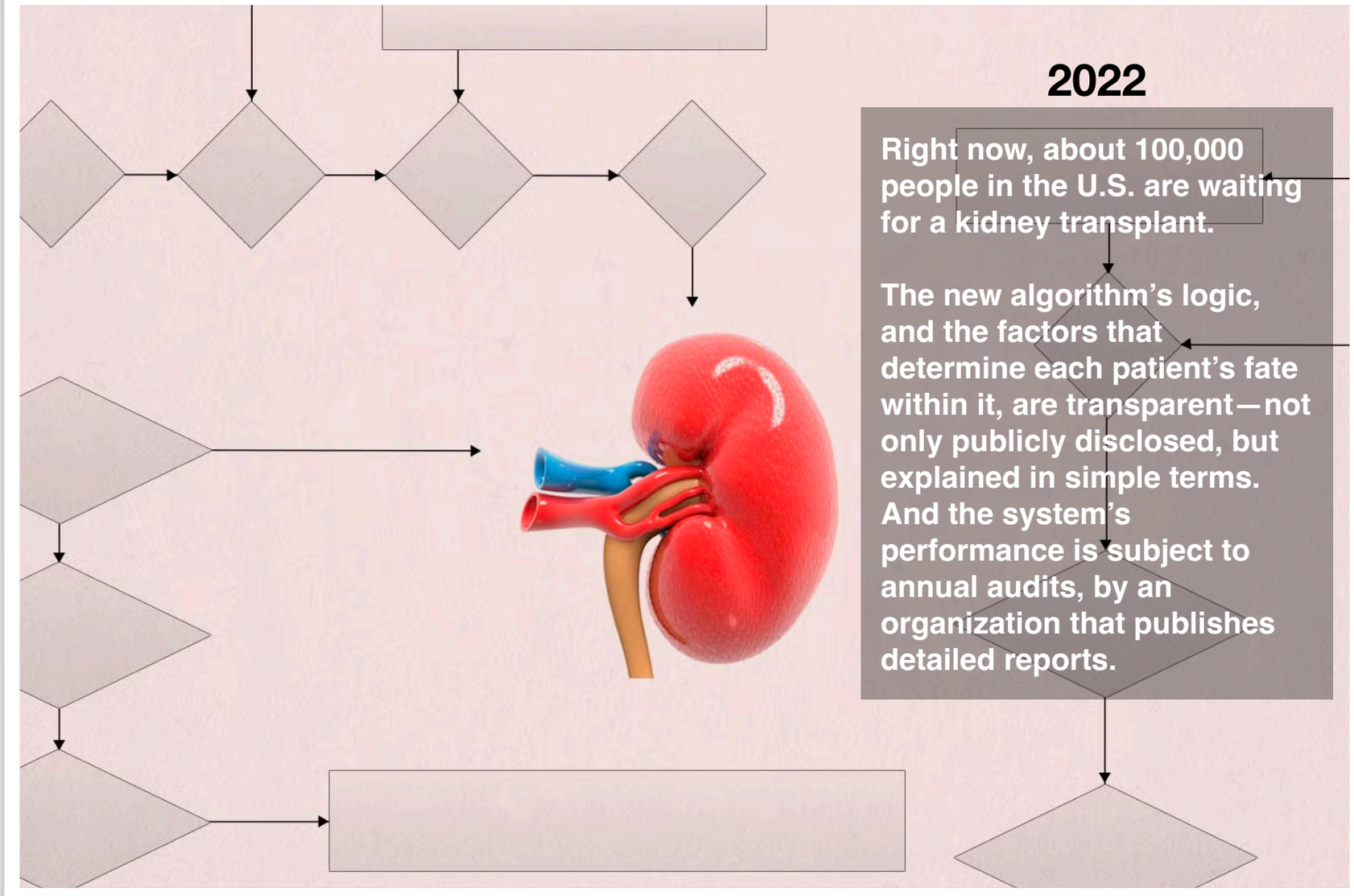
Converting human-designed decision-making rules into software.

Rules that have been set down by hand.



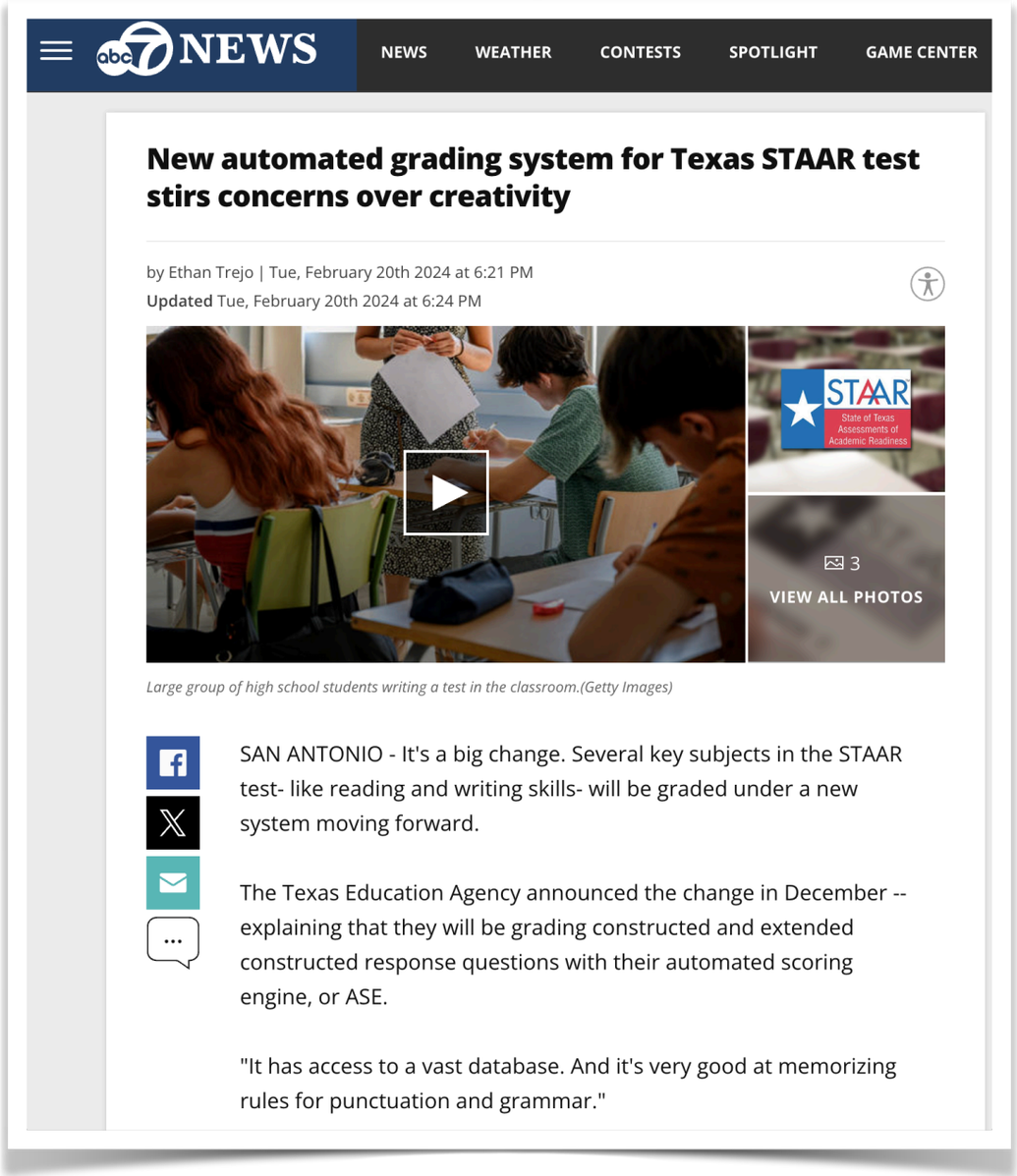
Kinds of automatic decision-making systems

Converting human-designed decision-making rules into software.



Kinds of automatic decision-making systems

Leveraging machine learning to emulate the informal decision-making processes of humans.



abc7 NEWS | NEWS | WEATHER | CONTESTS | SPOTLIGHT | GAME CENTER

New automated grading system for Texas STAAR test stirs concerns over creativity

by Ethan Trejo | Tue, February 20th 2024 at 6:21 PM
Updated Tue, February 20th 2024 at 6:24 PM

Large group of high school students writing a test in the classroom.(Getty Images)

SAN ANTONIO - It's a big change. Several key subjects in the STAAR test- like reading and writing skills- will be graded under a new system moving forward.

The Texas Education Agency announced the change in December -- explaining that they will be grading constructed and extended constructed response questions with their automated scoring engine, or ASE.

"It has access to a vast database. And it's very good at memorizing rules for punctuation and grammar."

Decision makers have primarily relied on informal judgment rather than formally specified rules.

Kinds of automatic decision-making systems

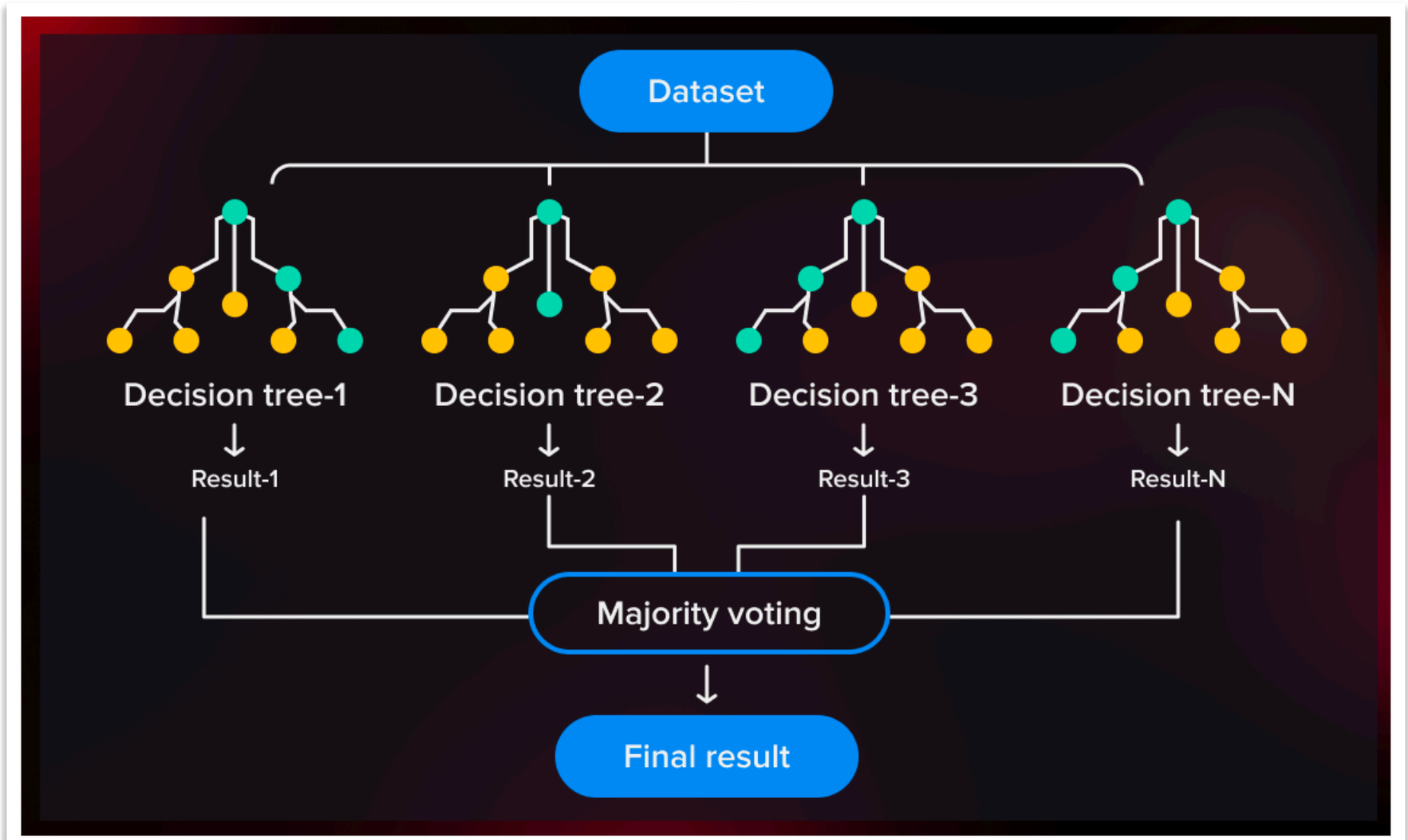
Learning decision-making rules from labeled data (based on a loss function that is a proxy of the policy).

Predictor variables	Men (n = 3989)	Women (n = 4953)
Age (years), mean ± SD	82.7 ± 2.7	82.6 ± 2.7
Height (cm), mean ± SD	172.1 ± 6.7	156.9 ± 6.1
Weight (kg), mean ± SD	79.4 ± 12.7	65.4 ± 12.7
Current weight as % age 25 weight, mean ± SD	111 ± 17	117 ± 21
Number of frailty components (0 to 5), mean ± SD	1.8 ± 1.3	2.5 ± 1.3
Number of medications (0 to 20, truncated), mean ± SD	7.8 ± 4.5	6.6 ± 3.8
Grip strength (kg), mean ± SD	35.3 ± 9.7	18.4 ± 5.8
Gait speed (m/s), mean ± SD	1.07 ± 0.24	0.85 ± 0.26
Chair stand speed (#/s), mean ± SD	0.37 ± 0.18	0.32 ± 0.19
Total hip BMD (g/cm^2), mean ± SD	0.929 ± 0.148	0.727 ± 0.141
Femoral neck BMD (g/cm^2), mean ± SD	0.756 ± 0.134	0.630 ± 0.125
Femoral neck BMD T-score, mean ± SD	−0.8 ± 1.1	−1.9 ± 1.0
Race/ethnicity, n (%)		
White	3478 (87.2)	4402 (88.9)
Black/Other	511 (12.8)	551 (11.1)
Number of chronic conditions, n (%)		
None	868 (21.8)	1268 (25.6)
1	1266 (31.7)	1651 (33.3)
2+	1855 (46.5)	2034 (41.1)
Mobility score (0 to 6), n (%)		
0	2986 (75.5)	3104 (63.8)
1	440 (11.1)	612 (12.6)
2+	528 (13.4)	1153 (23.7)
Fall history (number of falls in 12 months), n (%)		
None	2632 (66.0)	3272 (66.3)
1	703 (17.6)	968 (19.6)
2+	654 (16.4)	697 (14.1)
Fracture history, n (%)		
None after age 50 years	2854 (71.5)	2423 (50.4)

Uncovering patterns in a dataset that predict an **outcome** or **property of policy interest** (such as risks of cardiovascular disease, life expectancy, etc.)— and then bases decisions (such as transplant priority) on those predictions.

Kinds of automatic decision-making systems

Learning decision-making rules from labeled data (and using a proxy loss function).



Kinds of automatic decision-making systems

Learning decision-making rules from labeled data (and using a proxy loss function).

Example: To apply a policy for selecting “the student who will benefit the most from studying at your university” you can employ several proxy concepts. Some key proxy concepts include:

- Academic Performance (best predicted scores).
- Engagement and Participation (best predicted engagement).
- Socio-Economic Impact: (best predicted socio-economic impact, either personally (e.g., first-generation college students) or on their communities.

Legitimacy & Decision Making

About the process

Lesson of History

Up to now, in our society, **critical decisions** were often made by **bureaucratic systems**.

Bureaucracies arose, in part, to counteract the **subjectivity**, **randomness**, and **inconsistency** inherent in human decision-making, that can result in **arbitrary** decisions.

Its established rules and procedures are designed to reduce the impact of **weaknesses** found in individual decision makers.

About the process

Arbitrariness has two sides.

The first view of arbitrariness is primarily concerned with **procedural regularity**: whether a decision-making scheme is executed **consistently**.

When decision-making is arbitrary in this sense of the term, individuals may find that they are subject to different decision-making schemes and receive different decisions simply because they go through the decision-making process at different times.

This **principle** is based on the **belief** that people are entitled to similar decisions unless there are reasons to treat them differently.

About the process

The second view of arbitrariness refers to the basis for making decisions without **reasoning**, even if decisions are consistently made on that basis.

This **principle** is based two beliefs:

- the **belief** that random decision-making (in general) shows a lack of respect for people &
- the **belief** that subjective decision-making can lead to unfairness, errors, and lack of quality.

About the decisions

The results of the decisions must be:

- **Accurate** (the system must provide results that are "**correct**" in most cases or very close to the ideal result). **Correctness** of the results must be defined in a way that is **compatible with the values** of those affected by the decisions.
- **Reliable** (the system offers stable and consistent results in different scenarios, is invariant to some kinds of changes in the environment).
- **Effective** (the system delivers results that affect/ impact the real world in the expected way).

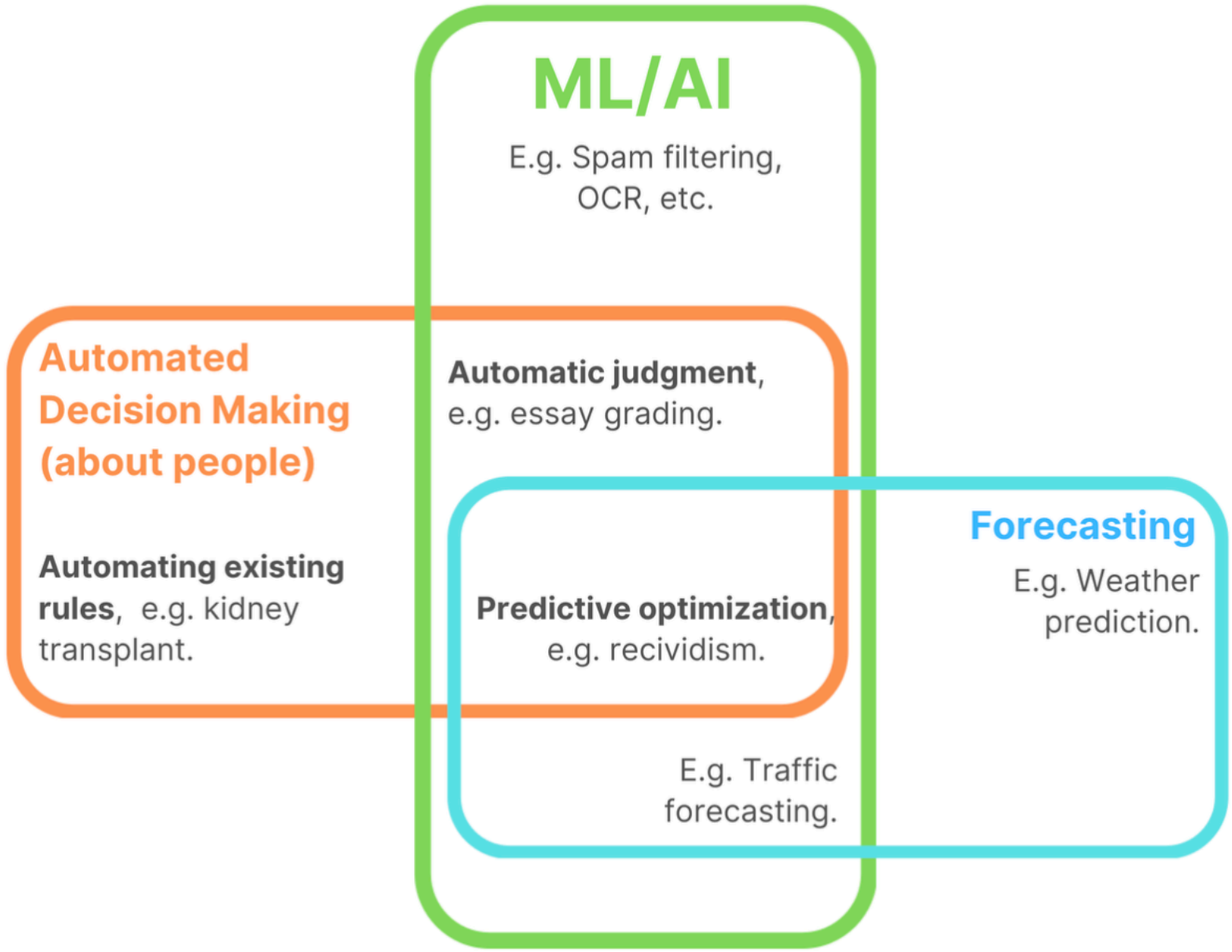
Legitimacy & Decision Making

We can consider the application of automated decision-making processes in a specific scenario to be legitimate if it meets several criteria at two different levels:

1. The results of the decisions are:
    - **Accurate & Aligned**
    - **Reliable**
    - **Effective**

2. The decision-making process is:
    - **Well executed**
    - **Well justified**

Legitimacy & Decision Making

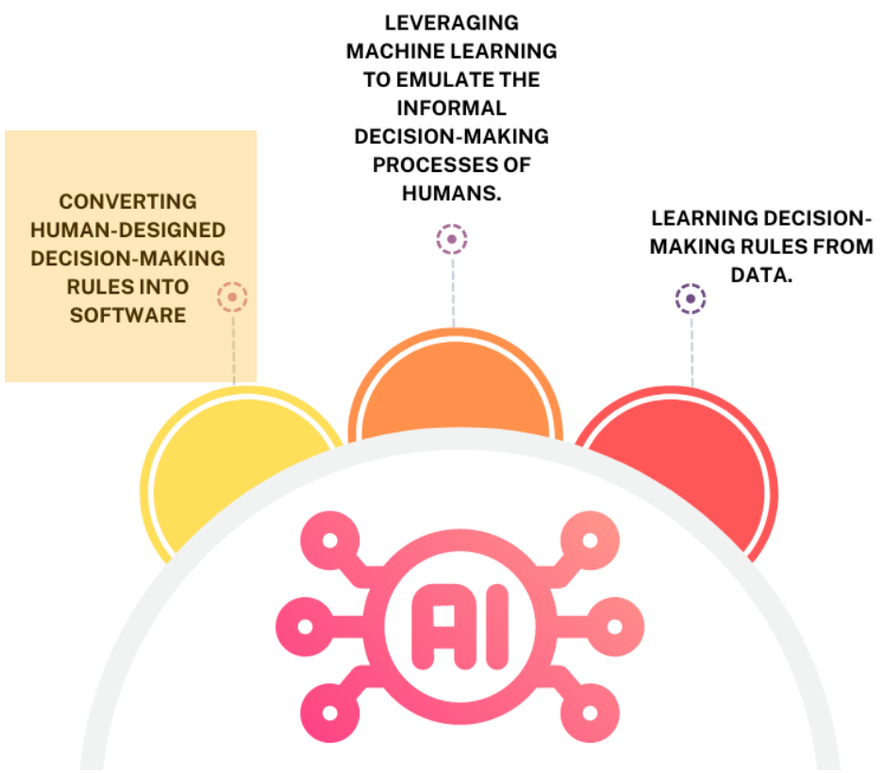


Credit: https://predictive-optimization.cs.princeton.edu/

Legitimacy & Decision Making

The first form of automation is a direct response to **arbitrariness as inconsistency**. It **allows procedural regularity**.

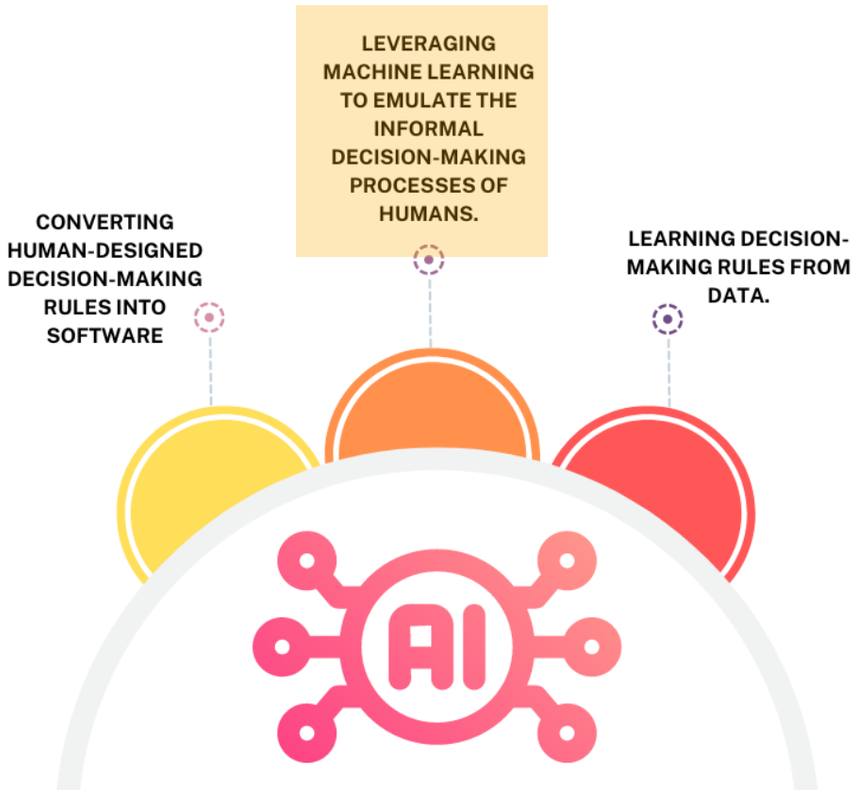
However, several problems can arise: policies intended to be automated may lack clarity or specificity, leading programmers to make subjective decisions and thus overstep their bounds in policy definition. Also, the software may be prone to errors. Automation also requires an institution to pre-determine all decision-making criteria, leaving no flexibility for unforeseen or unforeseen details. In addition, automation poses a significant risk as it potentially reduces accountability and intensifies the impersonal nature of bureaucratic interactions.



Legitimacy & Decision Making

In the second case, this form of automation could help solve problems of **arbitrariness** in human decision-making by formalizing and fixing a decision-making scheme similar to what humans might have used in the past .

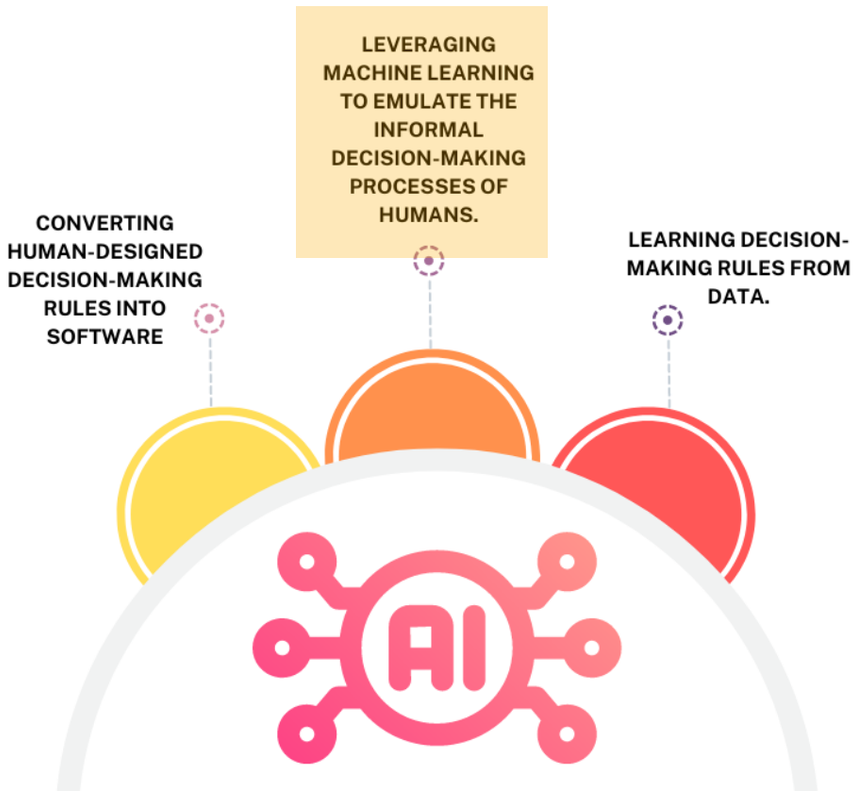
In this sense, machine learning can be **desirable** because it can help smooth out any **inconsistencies** **or** **subjectivities** in human decisions.



Legitimacy & Decision Making

These decision-making schemes can be considered equivalent to those employed by humans, and are therefore likely to perform similarly, even though the model may make its decisions differently and produce quite different error patterns.

Worse, the models could also learn to base themselves on criteria in ways that humans would find troubling or objectionable, even if doing so still produces a set of decisions similar to what humans would make.

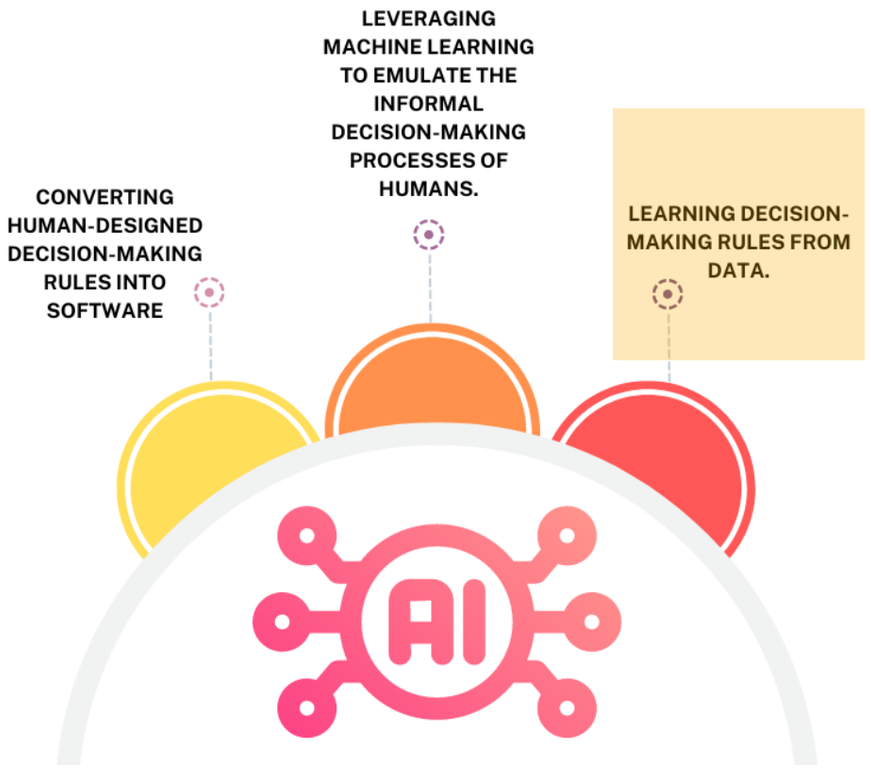


Legitimacy & Decision Making

The third form of automation, which we'll call **predictive optimization**, speaks directly to concerns with **reasoned decision making**.

Predictive optimization attempts to provide a more rigorous basis for **decision-making based only on criteria to the extent that they demonstrably predict the outcome or quality of interest.**

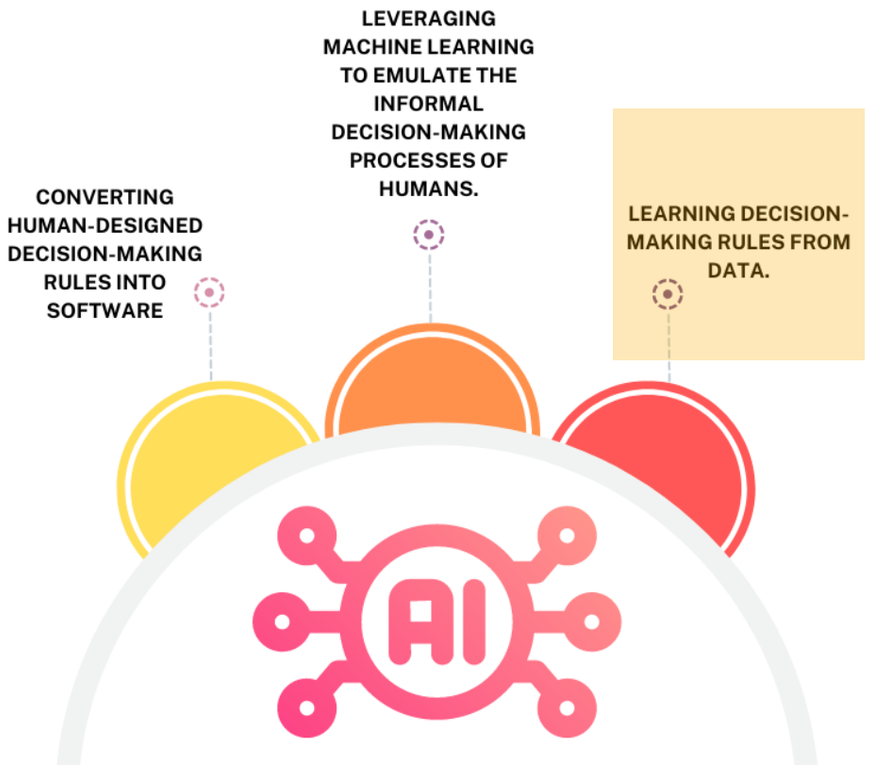
(Consistently execute a pre-existing policy through automation does not ensure that the policy itself is reasoned. Nor does relying on past human decisions to induce a decision-making rule guarantee that the basis of automation decision-making will reflect reasoned judgments).



Legitimacy & Decision Making

But it has flaws:

- Good predictions may not lead to good decisions (causality),
- It's hard to measure what we really care about (proxy loss functions),
- Training data rarely matches the deployment configuration (drift of distribution),
- Social outcomes are not predictable with precision (predictability), with or without machine learning.



Conclusions

In consequential **applications** of AI, to establish **legitimacy**, the decision-makers must be able to affirmatively justify their scheme according to the dimensions we have set out: the level of accuracy, reliability and effectiveness of their predictions, of their potential ethical problems, and that is also well executed and well justified.

To these properties we could add a condition of prudence, **irreducibility**: that there is no comparable solution based on human-designed algorithms and that therefore the decision system cannot be based on converting human-designed decision-making rules into software.