Lecture 9-1: Trust, Opaque Algorithms, Transparency, Explanations

Atoosa Kasirzadeh

October 6, 2020

Administrative

Assignment 1

- Available: October 13
- Deadline: October 28
- Grades available: Around November 7 (before final exam)

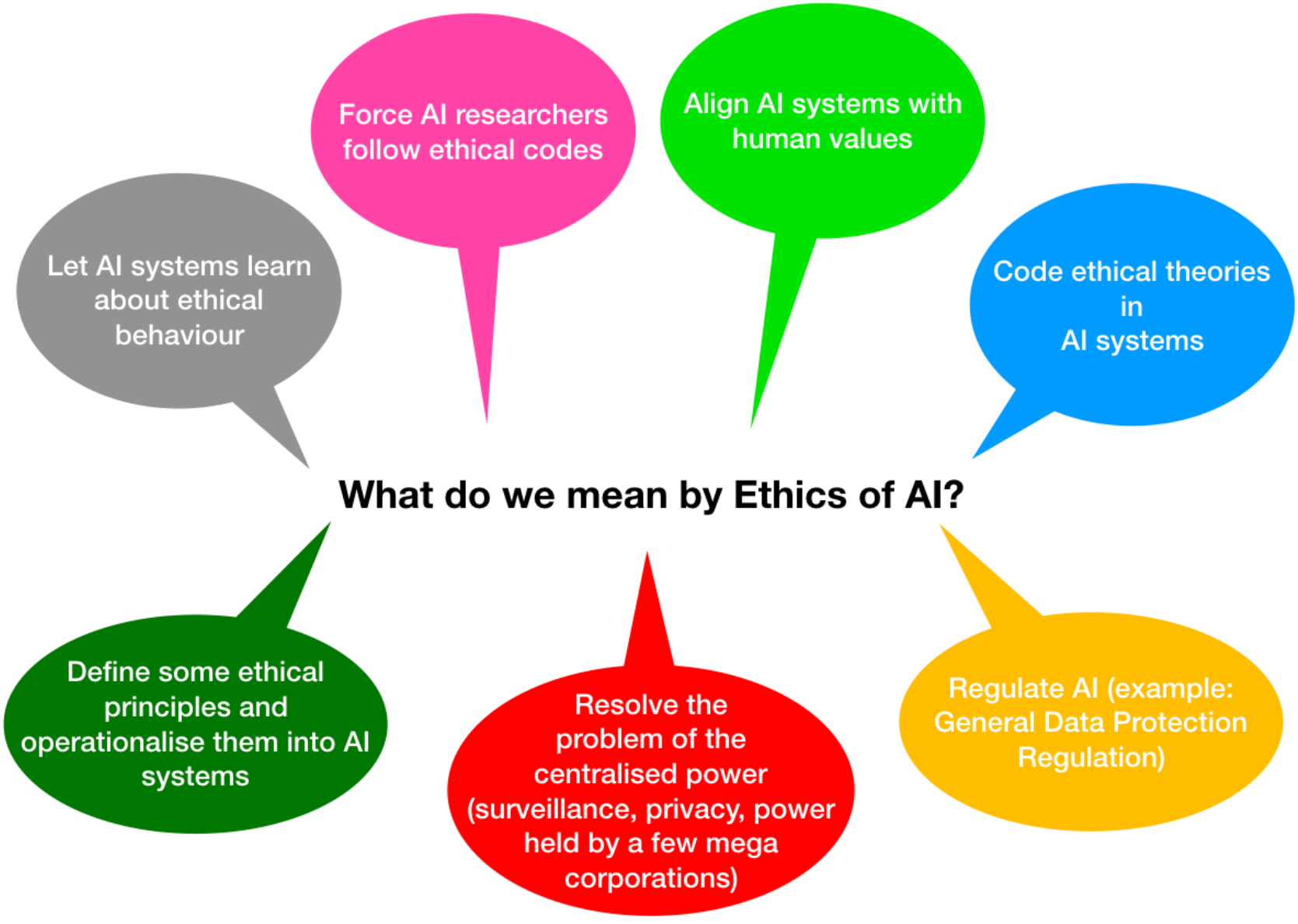
Assignment 2

- Available: November 1
- Deadline: November 14

Administrative

Tomorrow's lecture include a mini-lecture on how to write an effective paper for this course .

NO office hours tomorrow (I will be in an online conference).


Force AI researchers follow ethical codes
Align AI systems with human values
Let AI systems learn about ethical behaviour
Code ethical theories in AI systems
What do we mean by Ethics of AI?
Define some ethical principles and operationalise them into AI systems
Resolve the problem of the centralised power (surveillance, privacy, power held by a few mega corporations)
Regulate AI (example: General Data Protection Regulation)

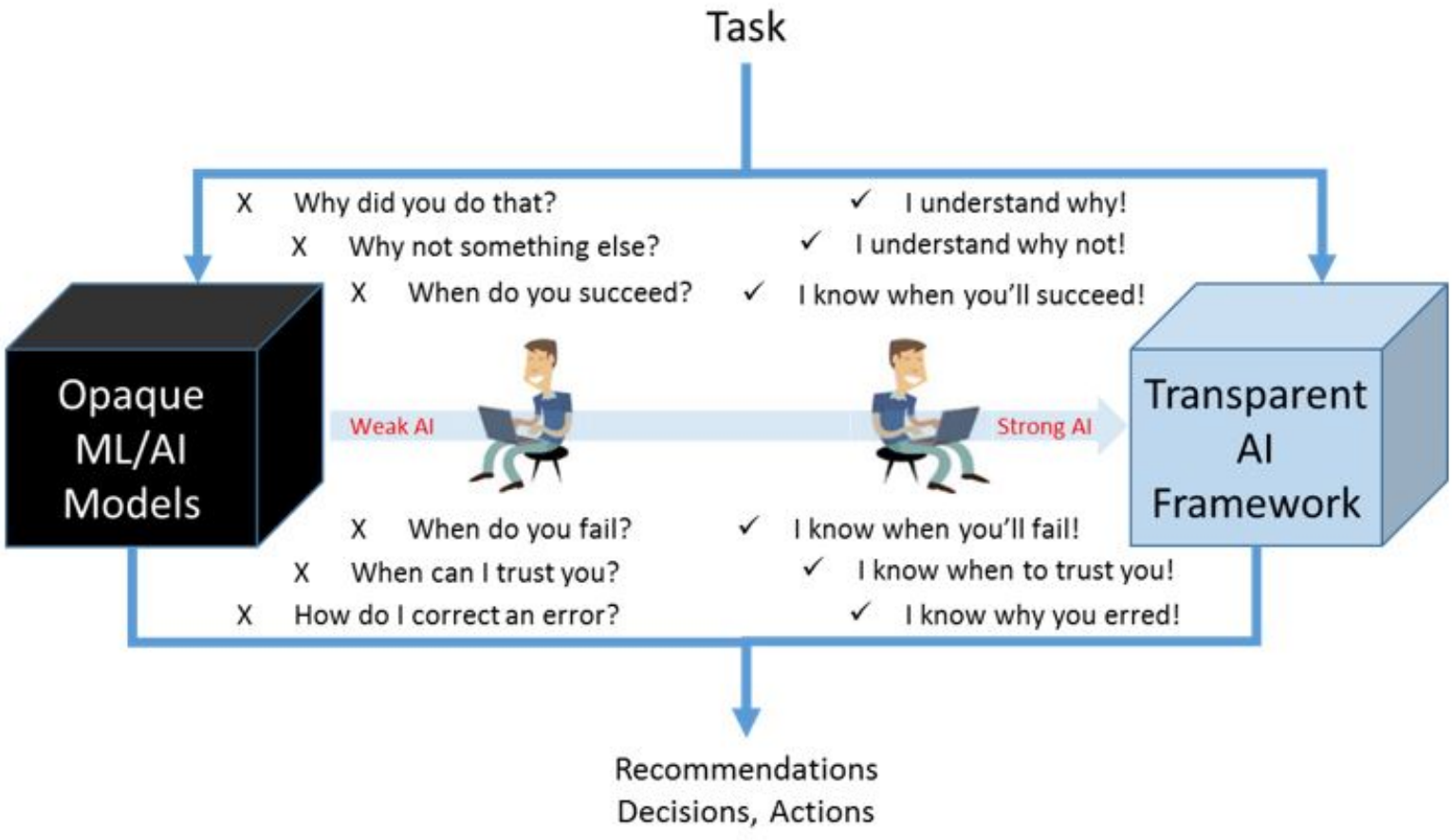
Defining some ethical principles and operationalizing them into AI systems

This week: transparency and explainability

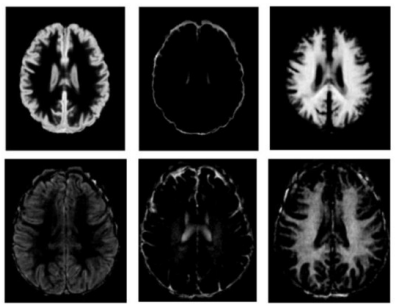
Governments and private actors are using some **truly opaque** AI algorithms to resolve critical decision-making problems.

- Hiring employees
- Assigning loans and credit scores
- Medical diagnosis
- Criminal recidivism

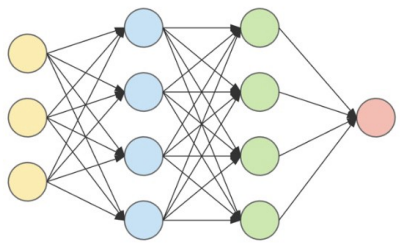
These algorithms are truly opaque: it is difficult for humans to understand why an algorithmic outcome is achieved.


Task
X Why did you do that?
✓ I understand why!
X Why not something else?
✓ I understand why not!
X When do you succeed?
✓ I know when you'll succeed!
Opaque ML/AI Models
Weak AI
Strong AI
Transparent AI Framework
X When do you fail?
✓ I know when you'll fail!
X When can I trust you?
✓ I know when to trust you!
X How do I correct an error?
✓ I know why you erred!
Recommendations
Decisions, Actions

Epilepsy Detection Model with Brain MRI Data



Brain MRI data



Complex ML model

Report:

Patient is diagnosed with **Epilepsy** with **%85 confidence**.

One popular proposal to overcome this opaqueness: require the algorithms to explain themselves (explainable AI)

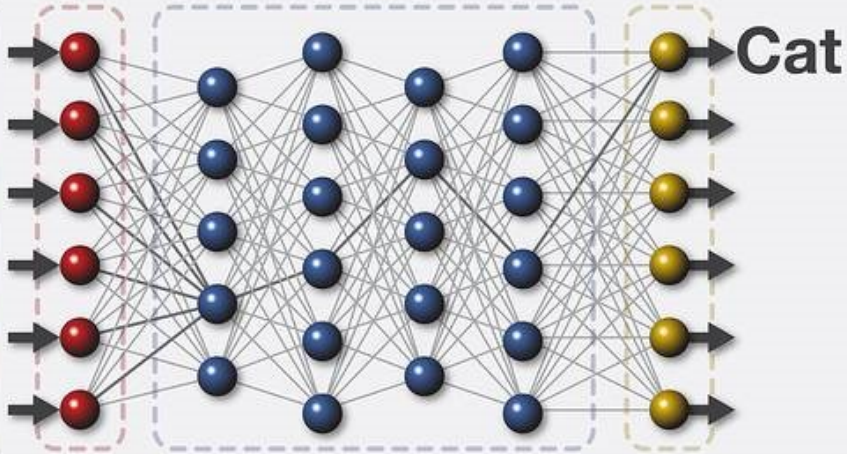
- To increase the societal acceptance of prediction-based decisions
- To establish trust in the results of these decisions
- To make these algorithms accountable to the public
- To prevent the sources of algorithmic discrimination and unfairness To legitimize the incorporation of the AI algorithms in several decision contexts
- Legal right to explanation for those affected by algorithmic decisions
- To facilitate a fruitful conversation among different stakeholders concerning the justification of using these algorithms for decision making

But what is “an explanation”

and

explainable for whom?

Machine Learning System



This is a cat.

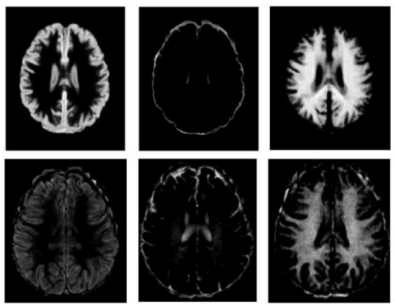
Current Explanation

This is a cat:

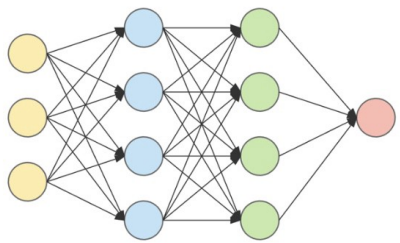
- It has fur, whiskers, and claws.
- It has this feature:

XAI Explanation

Epilepsy Detection Model with Brain MRI Data



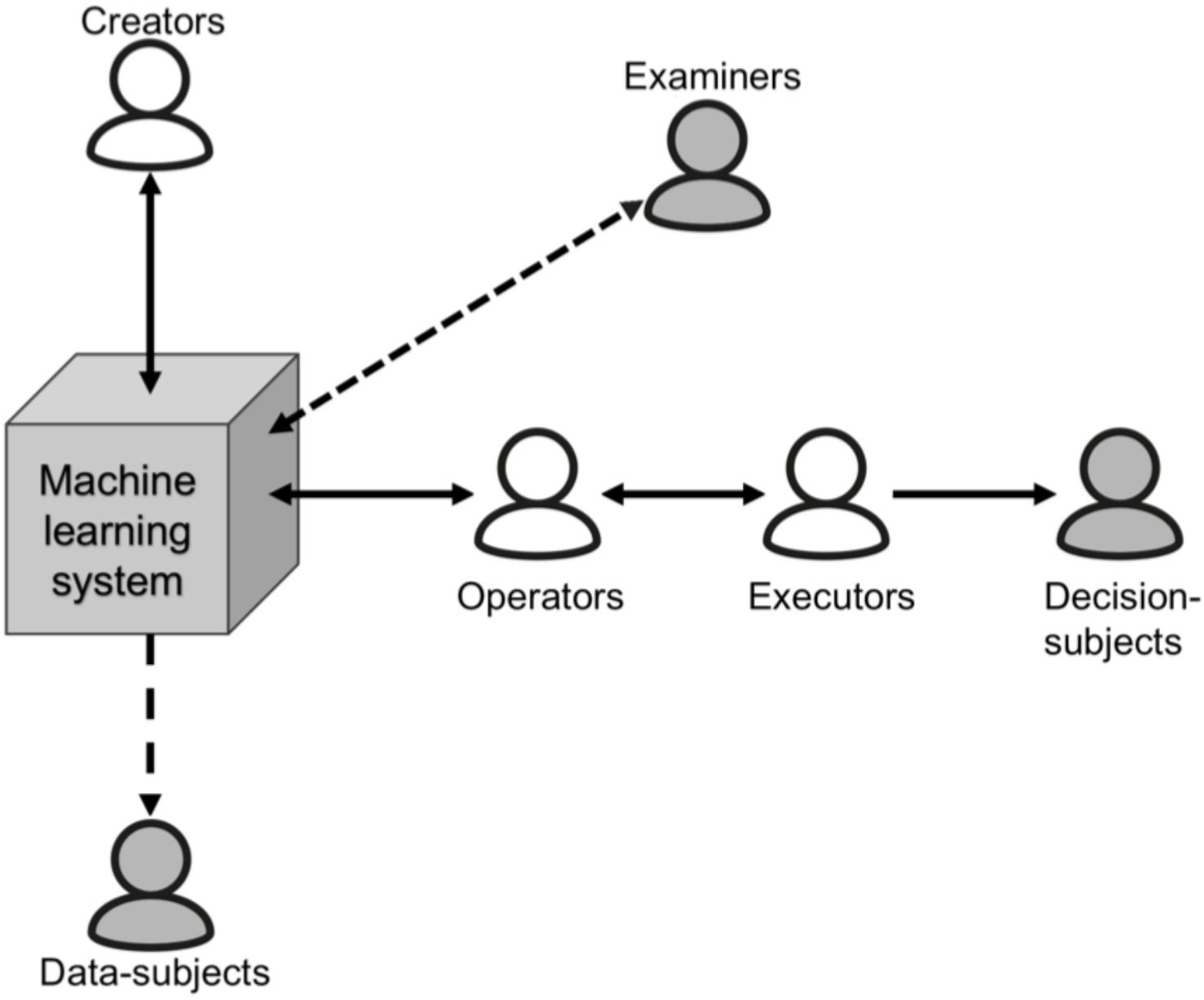
Brain MRI data



Complex ML model

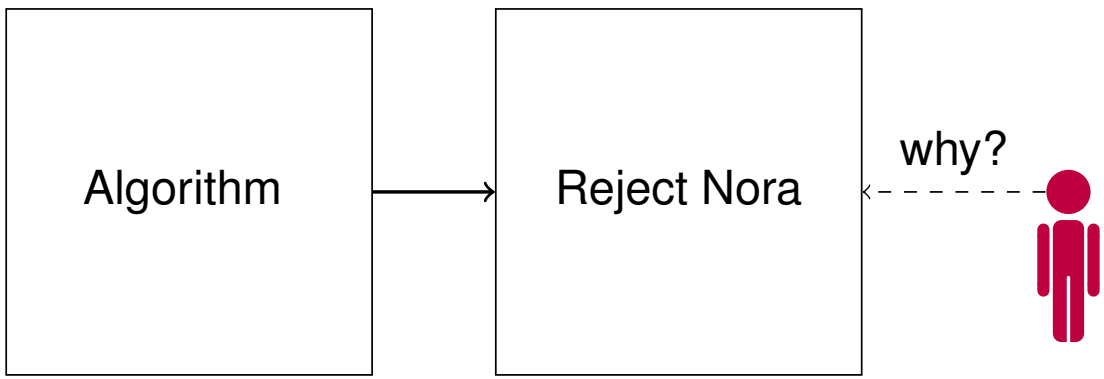
Report:

Patient is diagnosed with **Epilepsy** with **%85 confidence**.

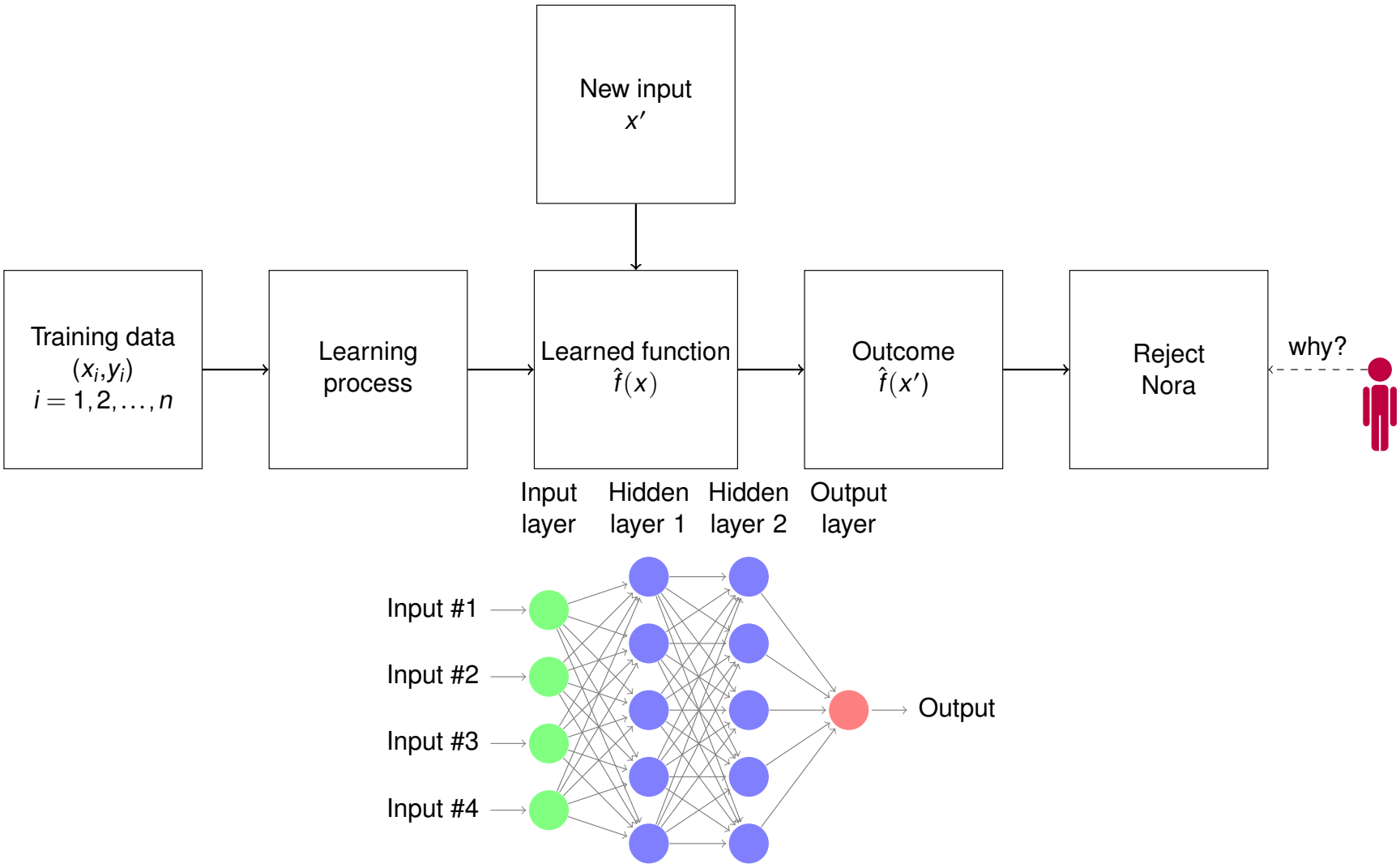

Creators
Examiners
Machine learning system
Operators
Executors
Decision-subjects
Data-subjects

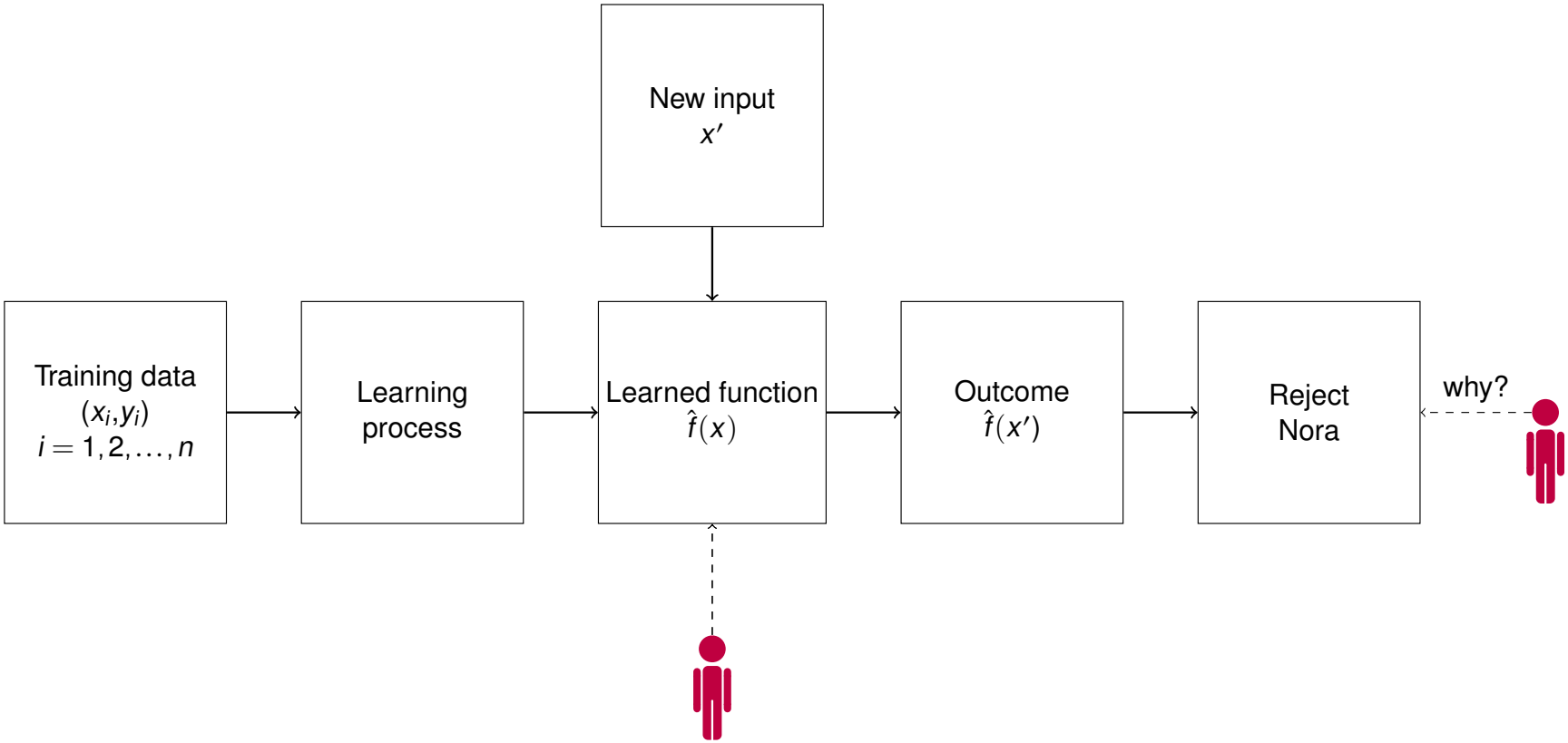
A machine-learning algorithm sifts through several job applications to recommend a hire for company X.

Nora, a competent candidate, applies for the job. Her application gets rejected by the algorithmic decision. Nora wants to know why she is rejected: she searches for an explanation.

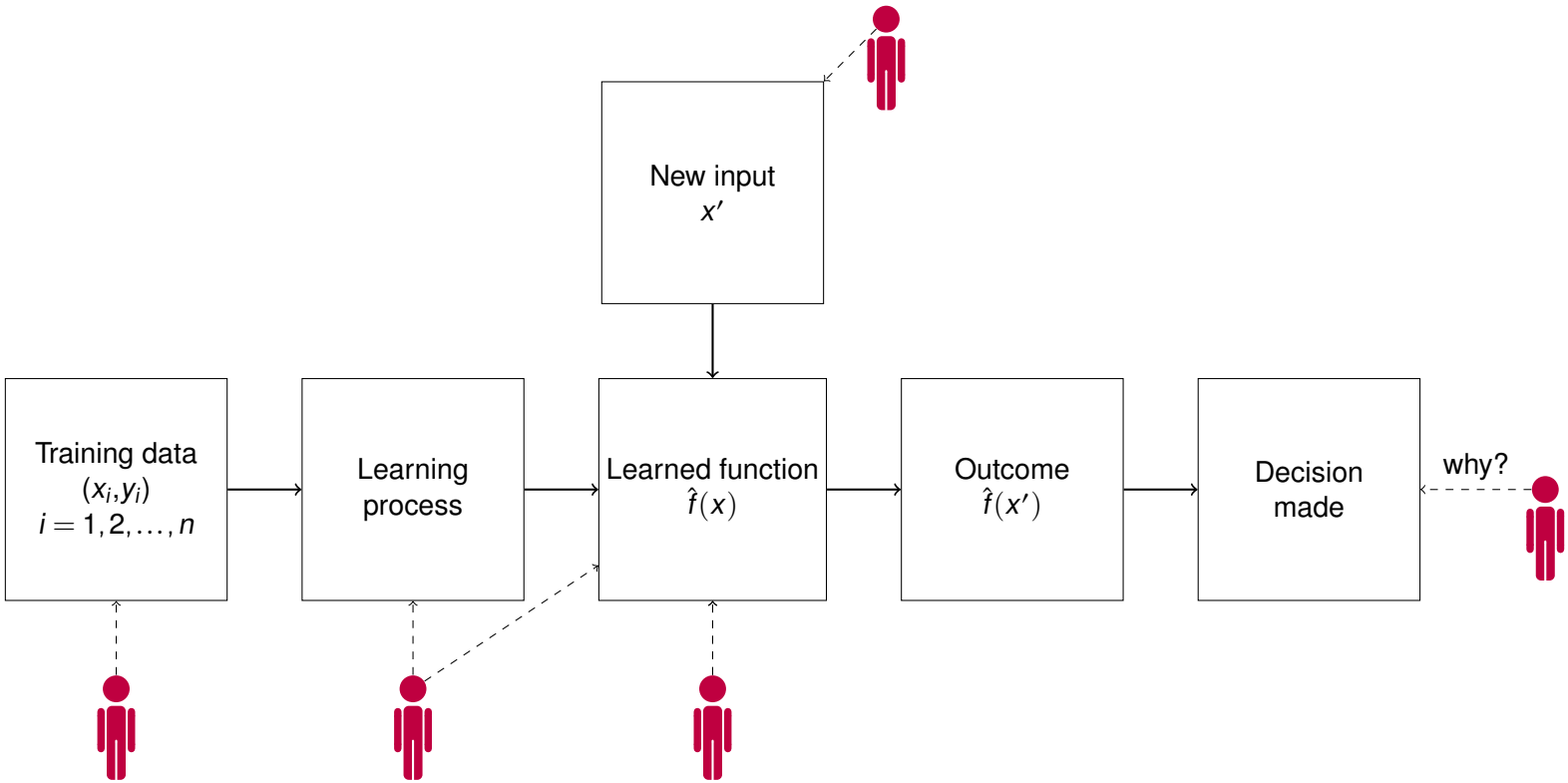


How nice would it be if the algorithm were able to explain its decision?


New input
x'
Training data
(x_i, y_i)
$i = 1, 2, \ldots, n$
Learning process
Learned function
$\hat{f}(x)$
Outcome
$\hat{f}(x')$
Reject Nora
why?
Input layer
Hidden layer 1
Hidden layer 2
Output layer
Input #1
Input #2
Input #3
Input #4
Output



What kinds of explanations do you seek?


New input
x'
Training data
(x_i, y_i)
$i = 1, 2, \ldots, n$
Learning
process
Learned function
$\hat{f}(x)$
Outcome
$\hat{f}(x')$
Decision
made
why?