Explaining Your Failures—A View From AI

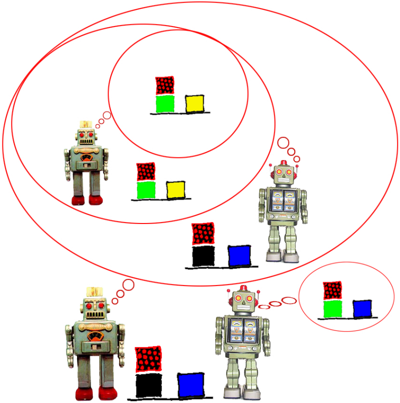
Thomas Bolander, Associate Professor, DTU Compute

Copenhagen University, 24 January 2018

A bit about myself

Thomas Bolander

- Associate professor in **logic** and **artificial intelligence** (AI) at **The Technical University of Denmark**.
- Member of the **SIRI commission**.
- **Current research**: Social aspects of AI. How to equip AI systems with a **Theory of Mind** (ToM)?

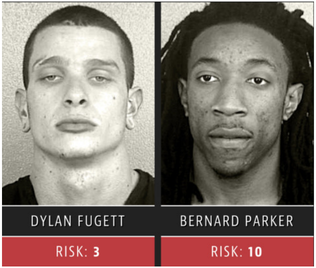


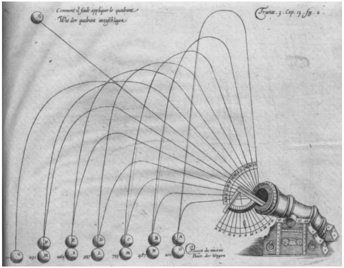
Mathematical models

Good **mathematical models** together with **powerful computers** can be used to do classification and prediction.

- **Classification examples**: cat/dog images; good/bad customers.
- **Prediction examples**: the weather; whether an inmate will commit crime during parole.

Clearly the **precision/quality** of a prediction will be limited by the precision of the mathematical model, e.g. in a model of ballistic trajectories (how well does the model approximate the real physical phenomenon).





Explicit vs implicit mathematical models

But maybe even more crucial than **precision** of a mathematical model is its **type**: explicit or implicit.

- **Explicit model example**: using the laws of physics to predict ballistic trajectories.
- **Implicit model example**: training an artificial neural network to distinguish between pictures of cats and dogs (or predict the horisontal range of a ballistic trajectory).

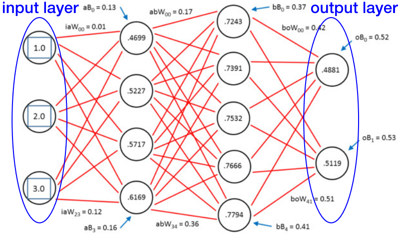
The **current trend** in AI and big data moves towards implicit models.

Challenge: When they fail, we often can't find the source of failure, and can't fix it.

$$a_x = \frac{-kv_x}{m} = \frac{dv_x}{dt} \text{ (1)},$$

and

$$a_y = \frac{1}{m}(-kv_y - mg) = \frac{-kv_y}{m} - g = \frac{dv_y}{dt} \text{ (2)}$$



Symbolic vs sub-symbolic AI

The symbolic paradigm (1950–): Simulates human symbolic, conscious reasoning. Search, planning, logical reasoning. **Ex**: chess computer.

$\uparrow$

- 👍 robust, predictable, explainable
- 👎 strictly delimited abilities

- 👍 flexible, learning
- 👎 never 100% predictable/error-free

$\downarrow$

The sub-symbolic paradigm (1980–): Simulates the fundamental physical (neural) processes in the brain. Artificial neural networks. **Ex**: image recognition.

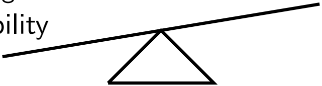
symbolic

$\downarrow$

$\uparrow$

sub-symbolic

Challenges in sub-symbolic AI



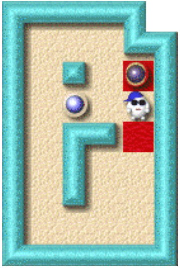
If a model can't be 100% precise, we should at least be able to **explain why/where it fails** when it fails, and find out how to improve. Ideally the model itself should be able to **explain** this: "I believed the trailer was a road sign because it was a big white rectangle with text."

When can we expect explanations of failures?

Is it realistic to expect a system to be able explain failures in classification/prediction?

Why did you believe this was red?

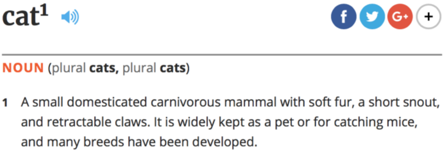
Why did you believe this was a horse?



Why did you move the marble into the red square?

Modelling: input vs output

Mathematical modelling: To produce a model (output) from some input.


cat[1]

NOUN (plural **cats**, plural **cats**)

1 A small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.

Symbolic input

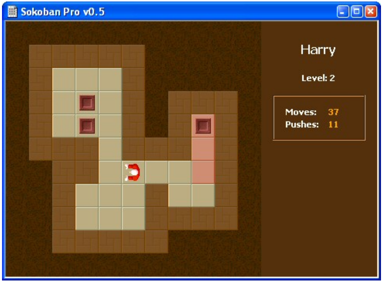
Subsymbolic input

- **Symbolic AI**: Input is symbolic, output is symbolic (explicit model).
- **Subsymbolic AI**: input is raw data (subsymbolic), output is subsymbolic (implicit model).

What we really need for **explainability**: input is raw data, output is explicit model (symbolic). Requires combining symb. and subsymb. AI.

My work

- Learning to create **symbolic plans from raw data** (in Sokoban and similar environments). With Andrea Dittadi (DTU).
- Learning **symbolic representations of actions** (not yet from raw data, though). With Nina Gierasimczuk and Andrés Libermann (DTU).
- Abductive reasoning to produce **explanations of failed plan execution**. With Sonja Smets (ILLC, Amsterdam).
- Explaining the **failures of other agents**: goal recognition, theory of mind, multi-agent planning.



Human child, 18 months old

http://www2.compute.dtu.dk/~tobo/children_cabinet.mpg

The child is *not* given any instructions beforehand.

(Warneken & Tomasello, Science, vol. 311, 2006)