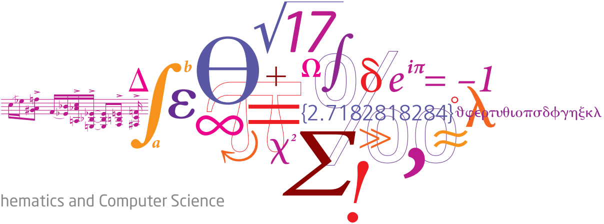
Trust and explainability: The relationship between humans & AI

Thomas Bolander, Technical University of Denmark

cl{A.I.}ms Forum 2, 1 May 2018

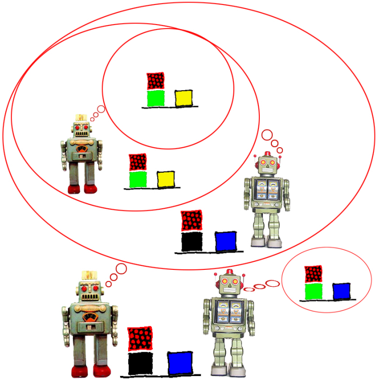


DTU Compute
Department of Applied Mathematics and Computer Science

A bit about myself

Thomas Bolander

- Associate professor in AI at **DTU Compute**, **Technical University of Denmark**.
- Member of the **SIRI commission**.
- **Current research**: Social aspects of AI. To equip AI systems with a **Theory of Mind** (ToM).



Quotes on AI and trust

> *The measure of success for AI applications is the value they create for human lives. In that light, they should be designed to enable people to* **understand AI systems** *successfully, participate in their use, and* **build their trust***.*

> *AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are* **human-aware and trustworthy***.*

(My highlighting)

(One Hundred Year Study on AI: 2015–2016, Stanford University, 6. september 2016)

Introduction to trust and explainability

When do we trust the **decisions**, **predictions** or **classifications** of another agent (person, AI system, company):

1. When the agent is **always right** (0 probability of mistakes)?
2. When the agent is **almost always right** (very low probability of mistakes)? The same mistake might be repeated, but still with very low probability.
3. When the agent is **most often right**, but when not, the agent has an acceptable and **explainable reason** for not being so. The same mistake is not repeated (one-shot learning).

What is artificial intelligence (AI)?

Definition by John McCarthy, the father of AI:

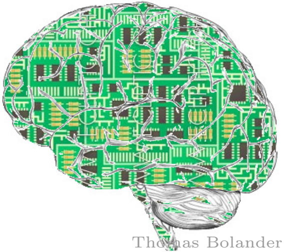
> *"Artificial intelligence is the* **science** *and* **engineering** *of making* **intelligent machines**, *especially* **intelligent computer programs**.*"*

John McCarthy, 2006

Doesn't imply that they are intelligent in the same way as humans.

AI today is probably more different from human intelligence than anyone anticipated.

Threatens **trust** of humans in decisions made by AI.



Symbolic vs sub-symbolic AI

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

↑

👍 robust, predictable, explainable

👎 strictly delimited abilities

👍 flexible, learning

👎 never 100% predictable/error-free

↓

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

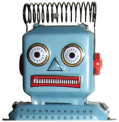
symbolic

sub-symbolic

Humans vs machines in relation to trust

Human

Fluent integration of subsymbolic (intuition, patterns) and symbolic reasoning (language, logic). Leads to ability to explain.



Machine

High performance and precision on clearly delimited task.

Fundamental differences between human and machine intelligence can negatively affect trust, in particular by:

- **Lack of robustness**: Lack of precision and robustness of machine.
- **Lack of human understanding**: 1) Lack of understanding how machine works; 2) lack of reasonable explanations in case of failures.

Lack of understanding: 3 hardest problems in AI

Carl Frey, 20 April 2017
Kolding, Denmark

Toby Walsh, 18 March 2017
Science & Cocktails, Copenhagen

Both have **social intelligence** among the 3 human cognitive abilities that are hardest to simulate by computers and robots.

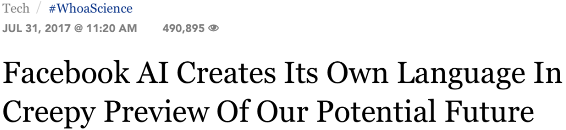
Social intelligence: The ability to understand others and the social context effectively and thus to interact with other agents successfully.

Lack of understanding: Machines behaving strangely

Lewis et al.: Deal or No Deal?, ArXiv, June 2017:

"We found that updating the parameters of both agents led to divergence from human language."

Forbes, 31 July 2017:

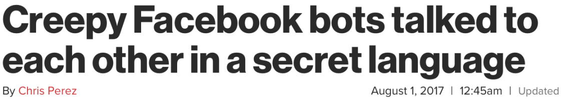


Tech / #WhoaScience

JUL 31, 2017 @ 11:20 AM 490,895 👁

Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future

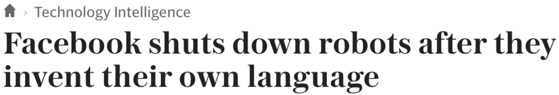
New York Post, 1 August 2017:



Creepy Facebook bots talked to each other in a secret language

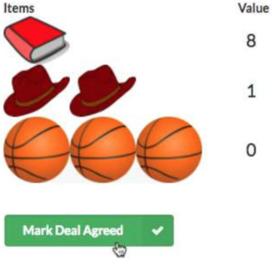
By Chris Perez August 1, 2017 | 12:45am | Updated

The Telegraph, 1 August 2017:



Technology Intelligence

Facebook shuts down robots after they invent their own language



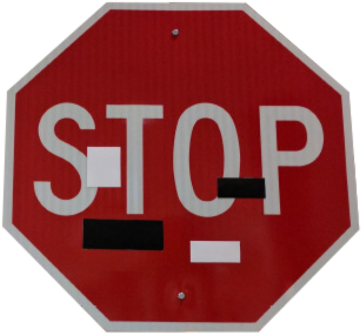
Lack of robustness: verification vs machine learning

Safety/robustness/flawlessness: verification vs machine learning.

Verification:

Intel Pentium bug 1994

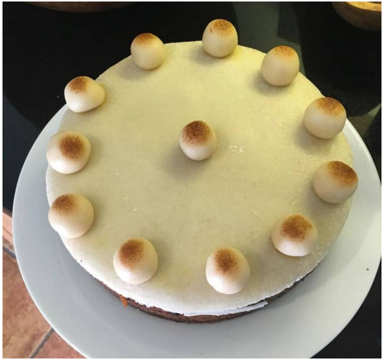
Machine learning:



Simple hacking of machine learning techniques

Lack of robustness: Fooling deep neural networks

`http://www2.compute.dtu.dk/~tobo/deepvis`
`recognition.mov`



‘Your account has been disabled for not following the Instagram Community Guidelines, and we won’t be able to reactivate it.

We disable accounts that post content that is sexually suggestive or contains nudity. We understand that people have different ideas about what’s okay to share on Instagram, but to keep Instagram safe, we require everyone to follow our guidelines.

(Metro UK, 5 April 2015)

Trust from low probability of mistakes?

Is it sufficient that *the agent is almost always right*?

No:

> *For example, we cannot argue that a pedestrian detector is safe simply because it performs well on a large data set, because that data set may well omit important, but rare, phenomena (for example, people mounting bicycles). We wouldn't want our automated driver to run over a pedestrian who happened to do something unusual.*

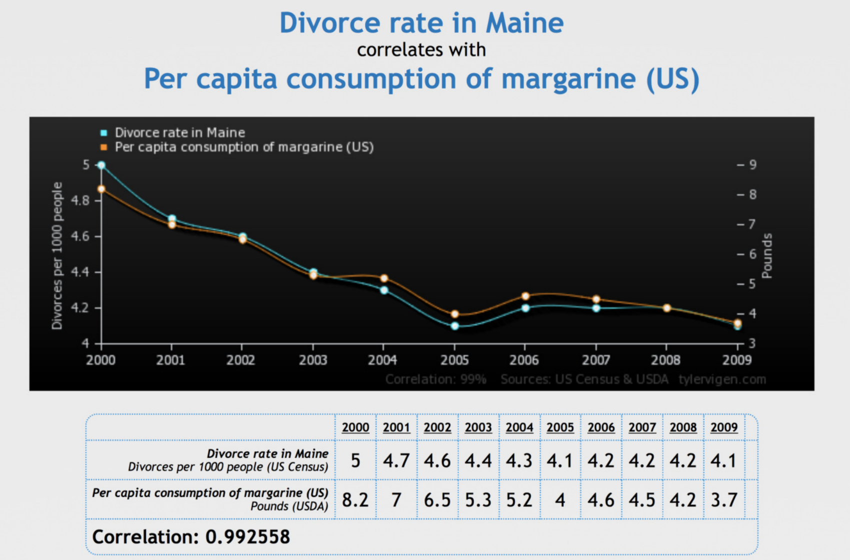
(Russell & Norvig: Artificial Intelligence—A Modern Approach, 3ed, 2010.)

Regaining trust: explainable AI

- Trust in AI systems is at risk when systems are neither 100% **robust**, nor **explainable** (by themselves or from the outside).
- In lack of 100% robustness, we need more **transparent** and **explainable** AI.
- **Subsymbolic AI** (e.g. neural networks) is naturally opaque.
- **Symbolic AI** (e.g. manually hand-crafted rule-based systems) is naturally transparent, but difficult to craft.
- Best current bet is to **combine**: The output of learning is rules and explicit models that can be inspected, understood and modified by humans.

The Big Data mantra of "what, not why" is challenged when decisions are made by algorithms, and the people affected want an explanation.

Correlations vs causal relationships: AI can't distinguish



Divorce rate in Maine

correlates with

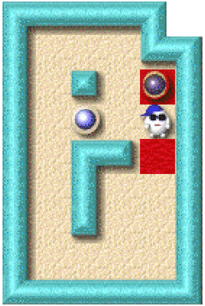
Per capita consumption of margarine (US)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

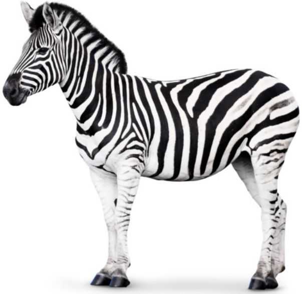
Correlation: 0.992558

When can we expect explanations of failures?

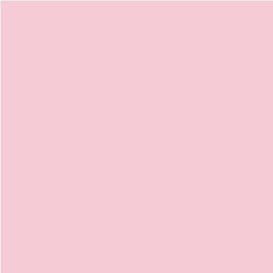
Is it realistic to expect a system to be able explain failures in classification/prediction?



Why did you move the marble into the red square?



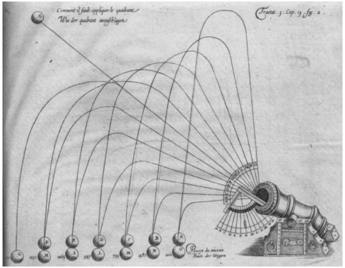
Why did you believe this was a horse?



Why did you believe this was red?

APPENDIX

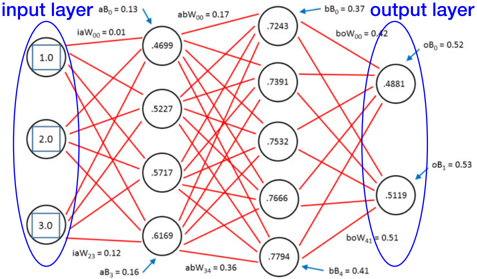
Symbolic vs sub-symbolic AI: explicit vs implicit models


Comment il fault appliquer le quadrant
Wie der quadrant anzuschlagen
Tractat. 3. Cap. 13. fig. 2

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

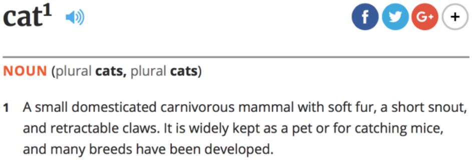
and

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



From raw data to symbolic representations

Subsymbolic input (raw data)



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

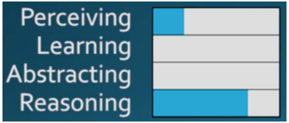
- **Subsymbolic AI**: input is raw data (subsymbolic), output is subsymbolic (implicit model).
- **Symbolic AI**: Input is symbolic, output is symbolic (explicit model).

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

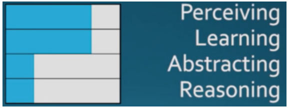
Three waves of AI

DARPA (Defense Advanced Research Projects Agency, USA) identifies **three waves of AI**:

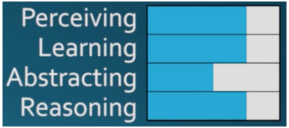
- *"The first wave of AI: Handcrafted knowledge"*. Essentially the symbolic paradigm.
- *"The second wave of AI: Statistical learning"*. Essentially the subsymbolic paradigm.
- *"The third wave of AI: Contextual adaptation"*. Essentially the combination of symbolic and subsymbolic approaches. Combine perception in neural networks with symbolic models for representing features, allowing **explanations** ("I thought it was a cat because it has fur and a short snout").



First wave



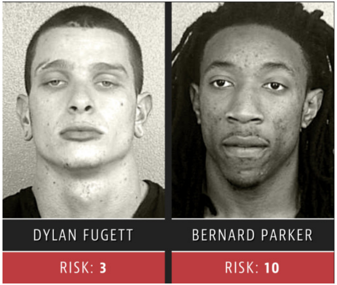
Second wave

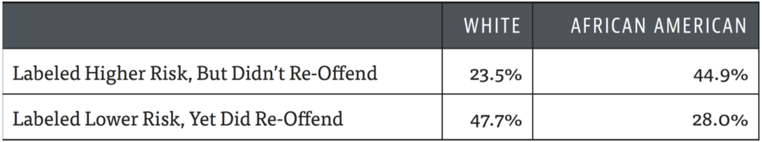


Third wave

Loosing trust: machine bias

Huge potential in machine learning algorithms (subsymbolic AI) for **predictions** and **decision making**. However, any bias in the data will also be learned.





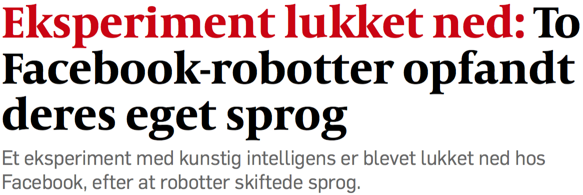
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

(Angwin et al.: Machine Bias, ProPublica, 23 May 2016)

Balls have zero to me to me to me to me to me to me to me to me to

Lewis et al.: Deal or No Deal?, ArXiv, June 2017:
“We found that updating the parameters of both agents led to divergence from human language.”

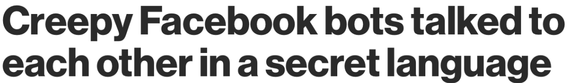
Politiken, 31 July 2017:



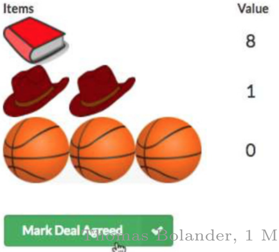
Eksperiment lukket ned: To Facebook-robotter opfandt deres eget sprog

Et eksperiment med kunstig intelligens er blevet lukket ned hos Facebook, efter at robotter skiftede sprog.

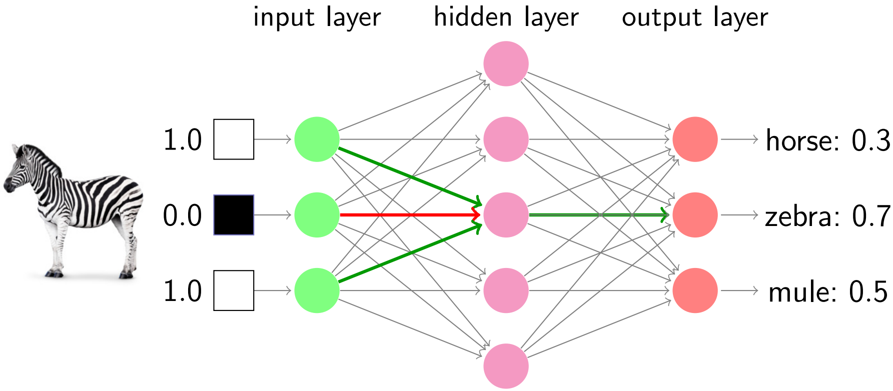
New York Post, 1 August 2017:



Creepy Facebook bots talked to each other in a secret language



Neural networks


input layer
hidden layer
output layer
1.0
0.0
1.0
horse: 0.3
zebra: 0.7
mule: 0.5