"With artificial intelligence we're summoning *the demon*"

-Elon Musk

What can an artificial intelligence know?

Jörn Grahl

— How will AI change my life?

— Will it turn me into a paperclip?

— Will it take my job?

— How will AI change employment?

— What should my children learn?

— If we hire an AI, what can it do?

Fundamental questions

- How will AI change my life?
- Will it turn me into a paperclip?
- Will it take my job?
- How will AI change employment?
- What should my children learn?
- If we hire an AI, what can it do?
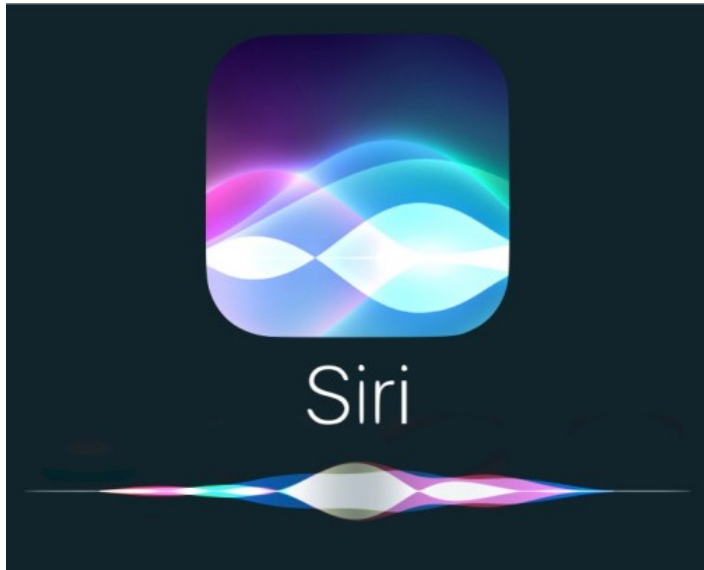
What can an artificial intelligence know?

Agenda

<u>Life</u>, consciousness, and intelligence

Is Siri alive?

„**Life** is a self-sustaining chemical system capable to undergoing Darwinian evolution "
(Joyce 1995)

Life, <u>consciousness</u>, and intelligence

Is Siri conscious (today)?

- = it feels a certain way to „be me"
- = you „know" what you are talking about
- = someone is „at home" inside a body

Thought experiment: fast forward 20 years.
Perfect language.
Is Siri conscious?
How can you know?

WHAT IS IT LIKE TO BE A BAT?

CONSCIOUSNESS is what makes the mind-body problem really intractable. Perhaps that is why current discussions of the problem give it little attention or get it obviously wrong. The recent wave of reductionist euphoria has produced several analyses of mental phenomena and mental concepts designed to explain the possibility of some variety of materialism, psychophysical identification, or reduction.[1] But the problems dealt with are those common to this type of reduction and other types, and what makes the mind-body problem unique, and unlike the water-H_2O problem or the Turing machine-IBM machine problem or the lightning-electrical discharge problem or the gene-DNA problem or the oak tree-hydrocarbon problem, is ignored.

Every reductionist has his favorite analogy from modern science. It is most unlikely that any of these unrelated examples of successful reduction will shed light on the relation of mind to brain. But philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and well understood, though entirely different. This has led to the acceptance of implausible accounts of the mental largely because they would permit familiar kinds of reduction. I shall try to explain why the usual examples do not

[1] Examples are J. J. C. Smart, *Philosophy and Scientific Realism* (London, 1963); David K. Lewis, "An Argument for the Identity Theory," *Journal of Philosophy*, LXIII (1966), reprinted with addenda in David M. Rosenthal, *Materialism & the Mind-Body Problem* (Englewood Cliffs, N. J., 1971); Hilary Putnam, "Psychological Predicates" in Capitan and Merrill, *Art, Mind, & Religion* (Pittsburgh, 1967), reprinted in Rosenthal, *op. cit.*, as "The Nature of Mental States"; D. M. Armstrong, *A Materialist Theory of the Mind* (London, 1968); D. C. Dennett, *Content and Consciousness* (London, 1969). I have expressed earlier doubts → "Armstrong on the Mind," *Philosophical Review*, LXXIX (1970), 394-403; "Brain Bisection and the Unity of Consciousness," *Synthèse*, 22 (1971); and a review of Dennett, *Journal of Philosophy*, LXIX (1972). See also Saul Kripke, "Naming and Necessity" in Davidson and Harman, *Semantics of Natural Language* (Dordrecht, 1972), esp. pp. 334-342; and M. T. Thornton, "Ostensive Terms and Materialism," *The Monist*, 56 (1972).

435

Life, consciousness, and <u>intelligence</u>

How humans use the term „intelligence"

- Bias: we call systems intelligent that behave like humans
- Machines that can speak, see, listen (perception)
- Self-learning machines (AI for GO, Chess… -- cognition)
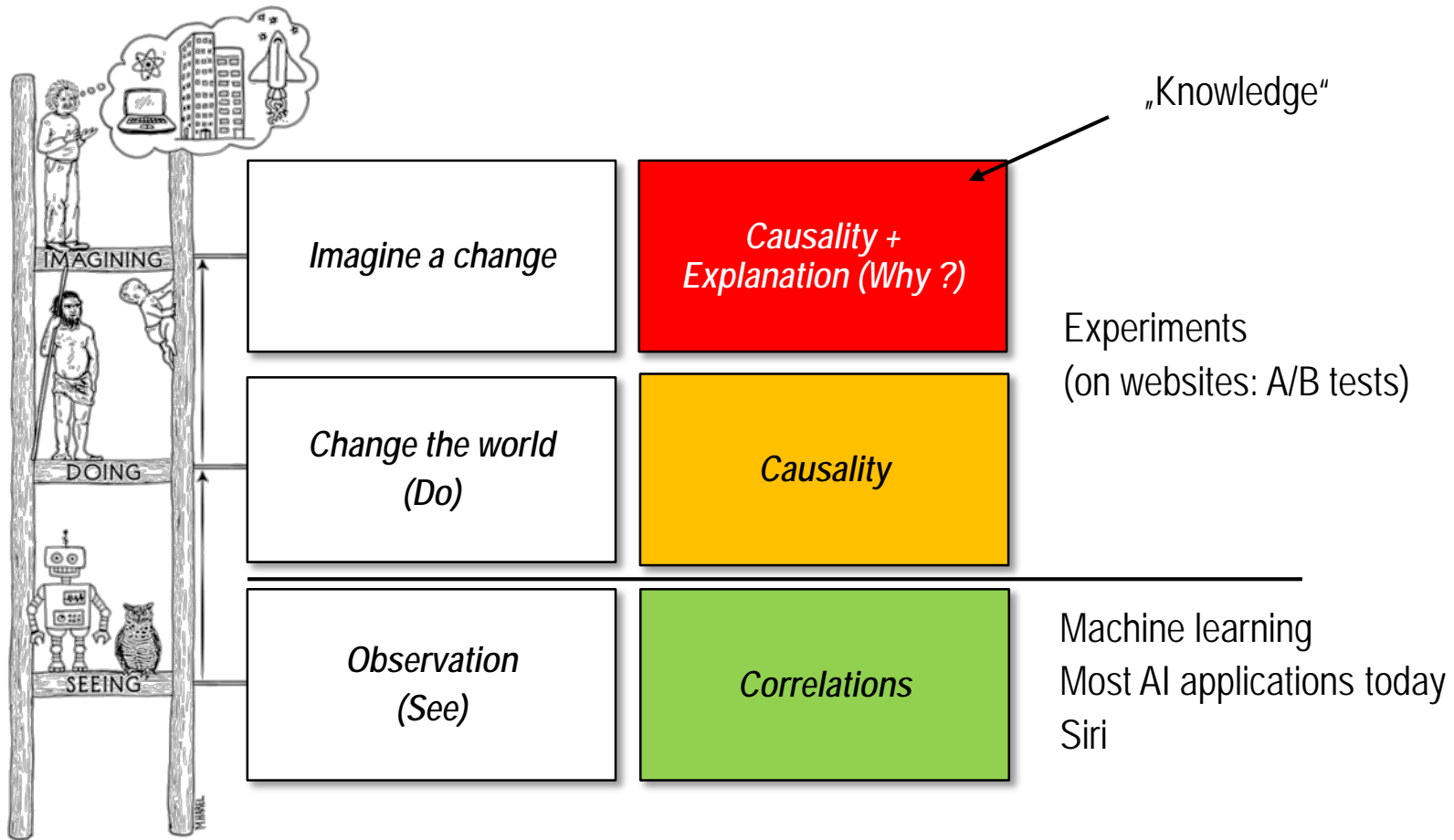- Autonomous behavior (vehicles, robots,…)

Agenda

Ladder of causality

		"Knowledge"
Imagine a change	**Causality + Explanation (Why ?)**	
Change the world (Do)	**Causality**	Experiments (on websites: A/B tests)
Observation (See)	**Correlations**	Machine learning Most AI applications today Siri

Aus: Pearl (2018) The book of why

Ladder of causality
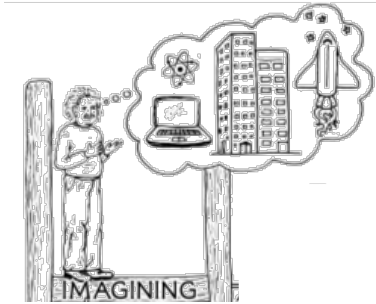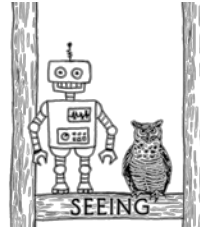
	Bottom	Middle	Top
Activity	Seeing	Doing	Imagining
Association	Correlation	Causality	Causality
Methods	Machine Learning	Experiments	Causal mediation
Motivation	*You want to predict something*	*You want to evaluate, compare, optimize*	*You want (middle) + explain the effect. Why did it happen?*
What are you modeling?	Probabilities	Effects (changes)	Effects and imagined universes
How?	Learn model from many examples	Randomize, intervene, measure	Middle + isolate direct and indirect effects
Direction	Undirected	Directed	Directed
What is moving?	Everything as it whishes	Intervention only	Intervention (through action), mediator (through intervention)

What is Machine Learning (ML)?

Simplest explanation

ML is just a <u>thing labeler</u>
(it matches items from two groups)

… based on many, many examples.

Siri does not know what she is talking about.
Siri matches audio files to your spoken words.
Siri need access to the internet, a library of examples.

Agenda

Back to fundamental questions

Application/ limitations of the ladder

Questions that can partly be answered using the ladder:
- If we „hire" an AI, what can it do?
  - With success: label stuff. Later: more.
- Will it take my job?
  - Yes, if you label things all day long
- How will AI change employment?
  - Subtitution / automation of tasks, not jobs, starting at bottom of ladder, then move up over time
- What should my children learn?
  - Hard to tell: STEM, but: Jack Ma: Learn an instrument. Dance. Sing!

Question that can not be answered using the ladder:
- How will AI change my life?
- Will it turn me into a paperclip?
  - Possible, but not likely. Engineering has a strong safety tradition.

Take home message

In Judea Pearl's words:

– Current AI is *not* very powerful (often bottom of ladder)
– But many applications are *easier* than expected (also bottom of ladder)
– To build truly intelligent machines, teach them *cause and effect* (middle and top of ladder)