

What is artificial intelligence?

Philosophical perspectives on AI

[Claus Beisbart](#)

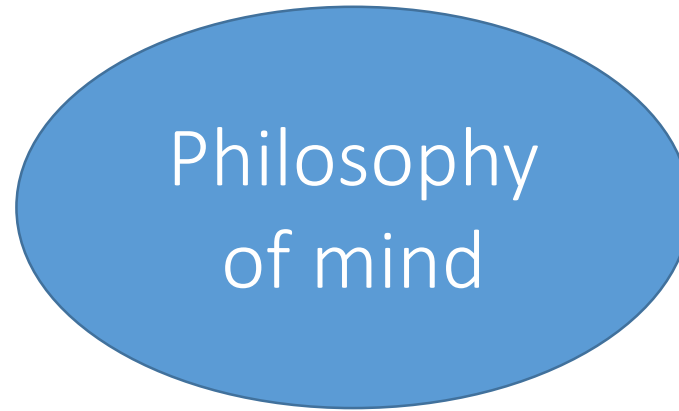
u^b

Institute of Philosophy
CAIM, MCID, OCCR

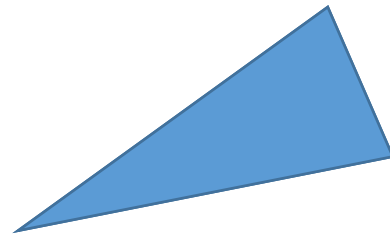
^b
**UNIVERSITÄT
BERN**

Münchenwiler Seminar 2024 „AI and Science“, 19.4.2024

Aim of this talk



Discuss basic questions about AI



Perspective: philosophy

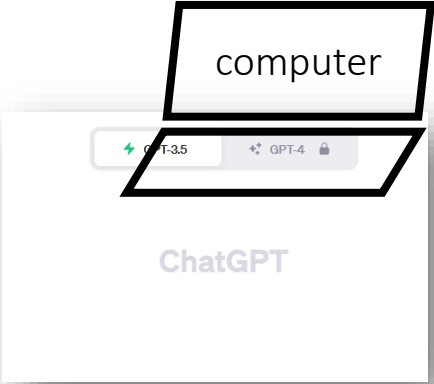


Structure of the talk

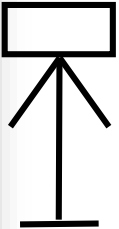

1. What is AI?
2. Does it really think?
3. Where will this get us?
4. What does AI mean for science?
5. What's the take-home message?

1. What is AI?

area of research that aims at building systems that ...



	like humans	ideally rational
think		
act		

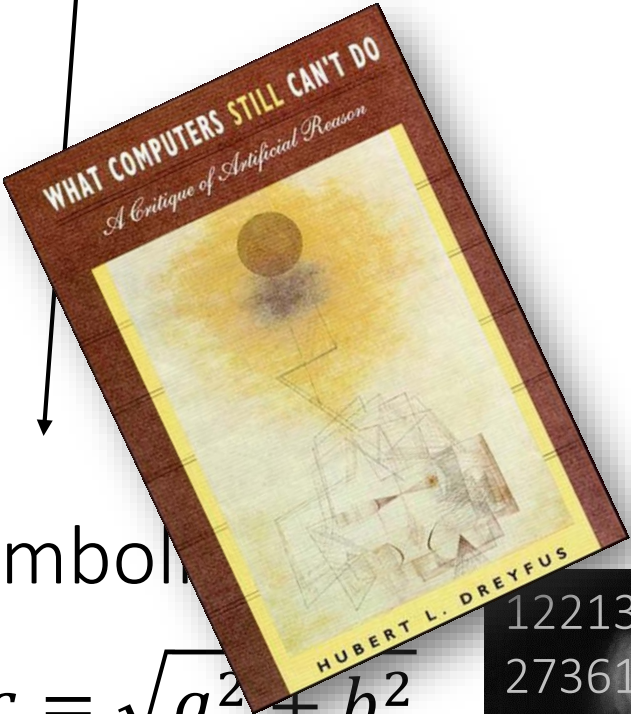


Bringsjord & Govindarajulu (2022) following Russel & Norvig (2009)

Remarks

1. This definition is at odds with current usage
“AI” often refers to a system built in this field, e.g. algorithm, robot etc.
2. The definition is on the cautious side

Paradigms



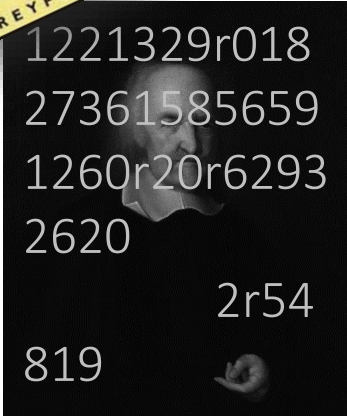
WHAT COMPUTERS STILL CAN'T DO
A Critique of Artificial Reason
HUBERT L. DREYFUS

GOFAI

symbol

$$c = \sqrt{a^2 + b^2}$$

Thomas Hobbes



1221329r018
27361585659
1260r20r6293
2620
2r54
819

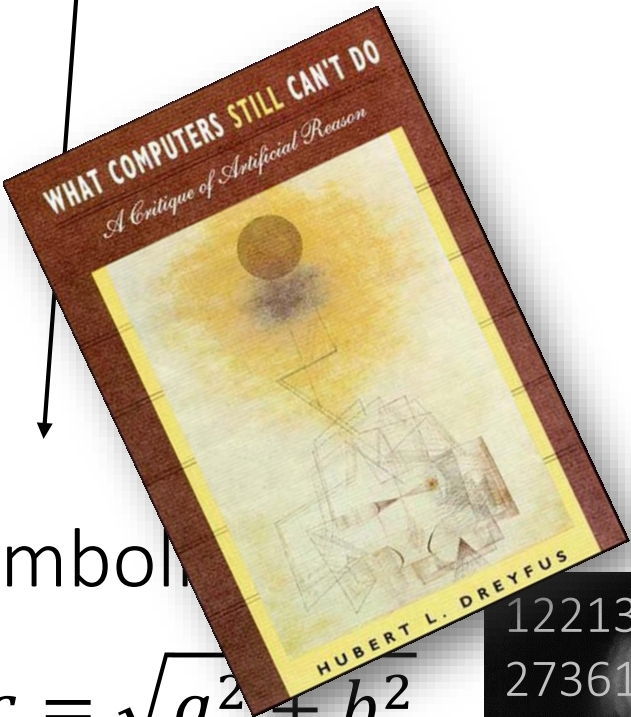
Hubert Dreyfus

“The psychological, epistemological, and ontological assumptions [behind GOFAI] have this in common: they assume that man must be a device which calculates according to rules on data which take the form of atomic facts. [...]

we shall explore three areas necessarily neglected in CS and AI but which seem to underlie all intelligent behavior: the role of the body in organizing and unifying our experience of objects, the role of the situation in providing a background against which behavior can be orderly without being rulelike, and finally the role of human purposes and needs in organizing the situation so that objects are recognized as relevant and accessible.”

Dreyfus 1992/94, 231, 234

Paradigms



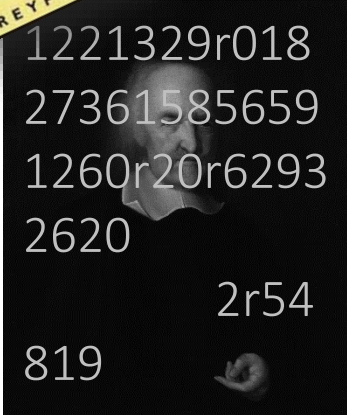
WHAT COMPUTERS STILL CAN'T DO
A Critique of Artificial Reason
HUBERT L. DREYFUS

GOFAI

symbolic

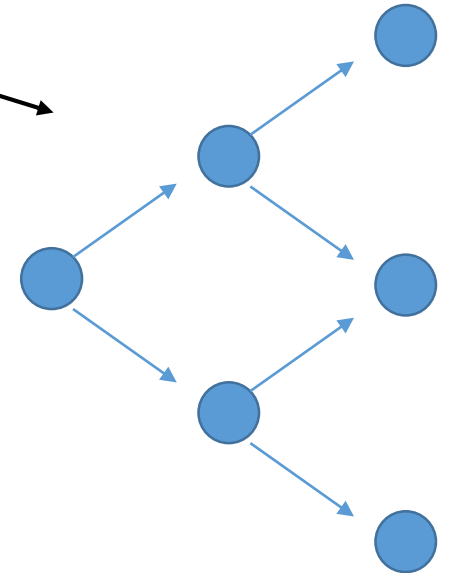
$$c = \sqrt{a^2 + b^2}$$

Thomas Hobbes

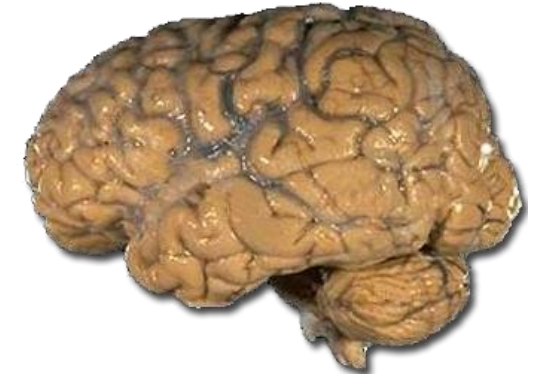


1221329r018
27361585659
1260r20r6293
2620
2r54
819

Subsymbolic/
connectionist

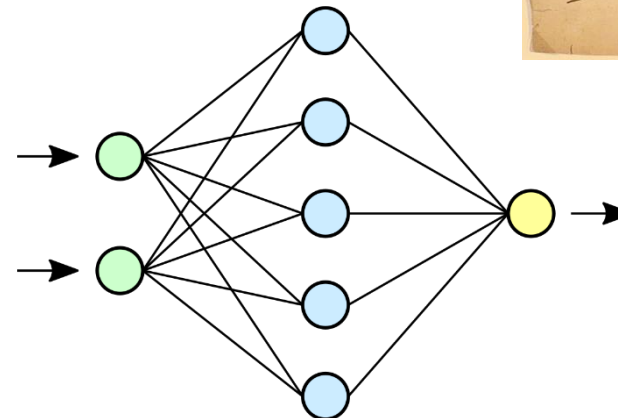
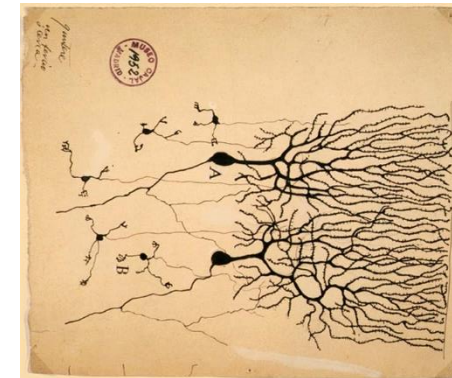


Underlying idea

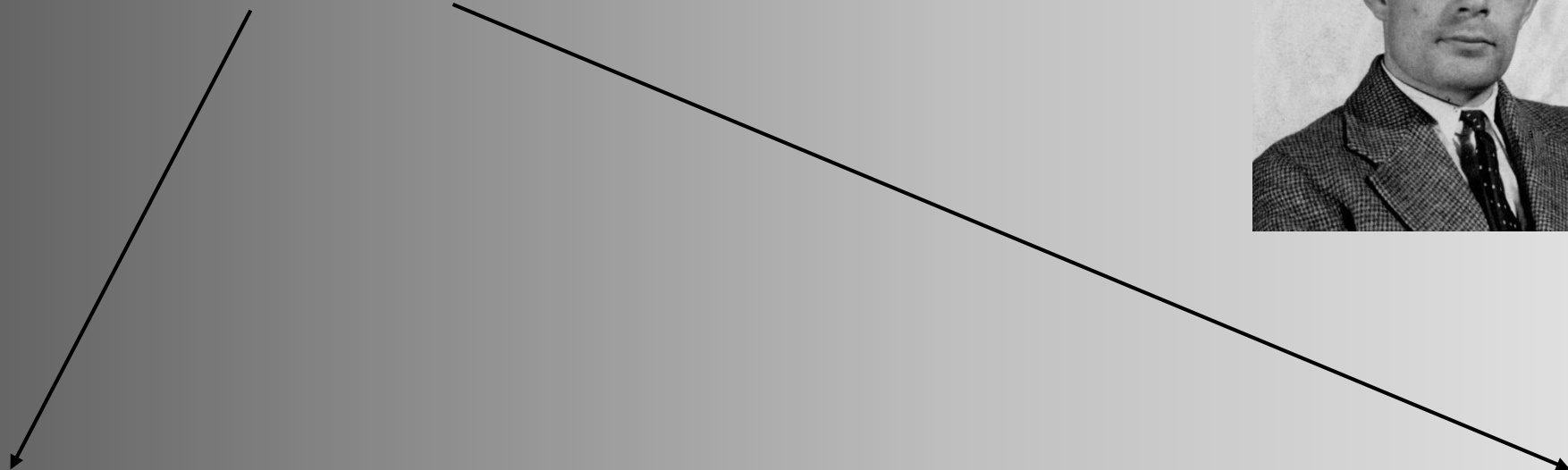
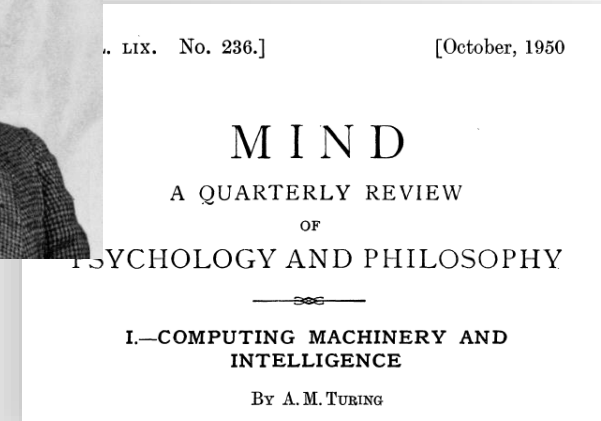


Connectionism

Thinking: processing of information in networks



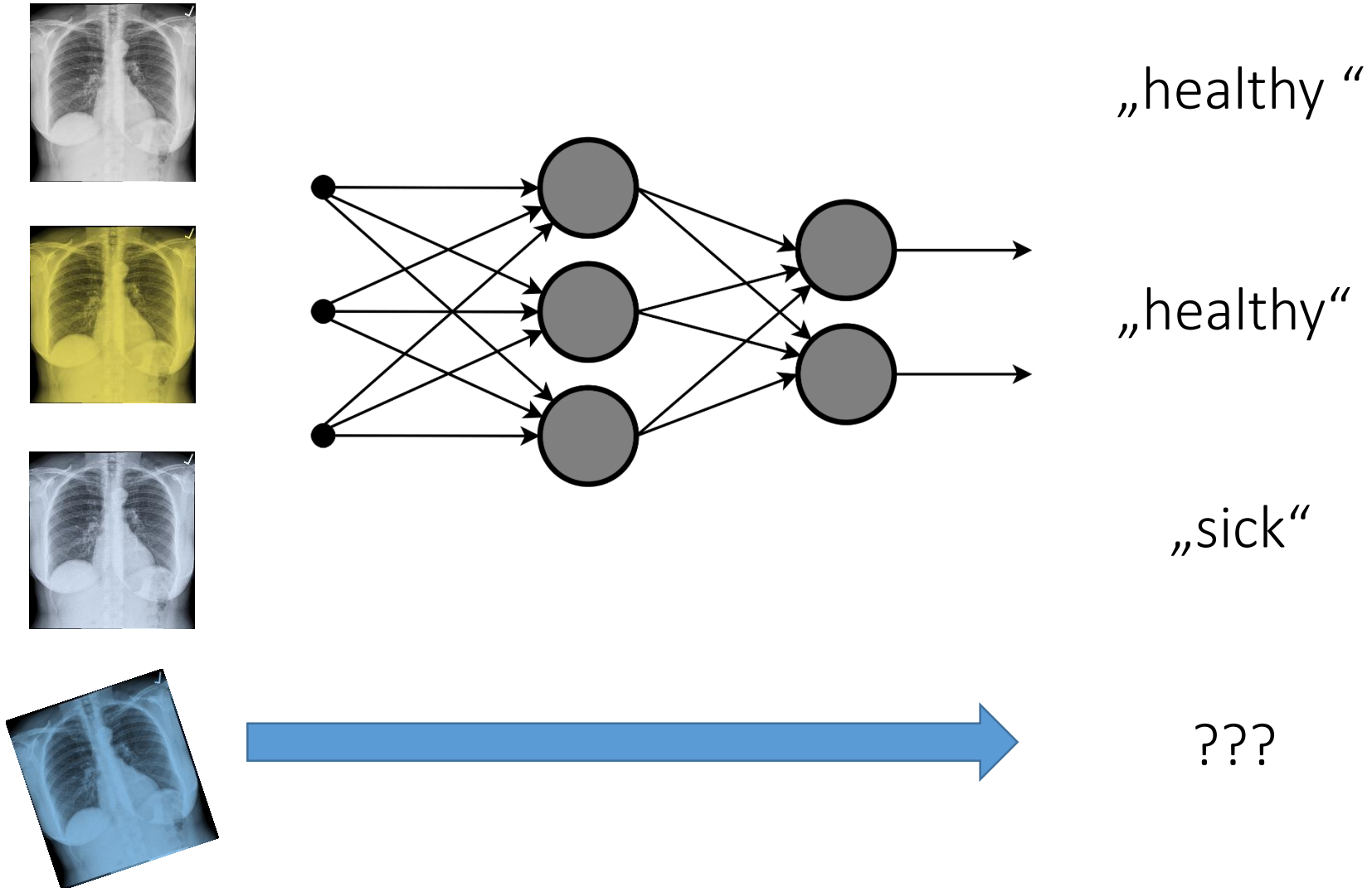
Algorithms



programmed

Inferred from data
„machine learning“

Supervised learning



AI status today

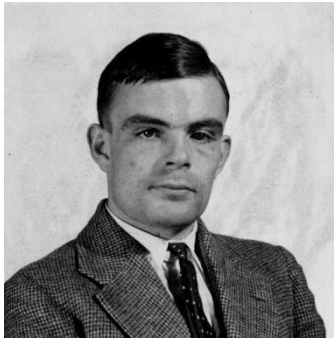
- Better than humans for many specific tasks
- No general intelligence yet
- Transfer learning might help

But is it really intelligent?
Does it really think?



2. Does it really think?

Alan Turing
1912-1954



Translates question:

Does it pass the Turing test?

VOL. LIX. No. 236.]

[October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND
INTELLIGENCE

By A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to

Turing, A. (1950), Computing Machinery and Intelligence, *Mind*, LIX: 433–460

Turing test/imitation game

Machine

(wants to deceive I that they
are human)

human

(wants to convince I
that they are human)

Who is machine/human?

X

Y

interrogator

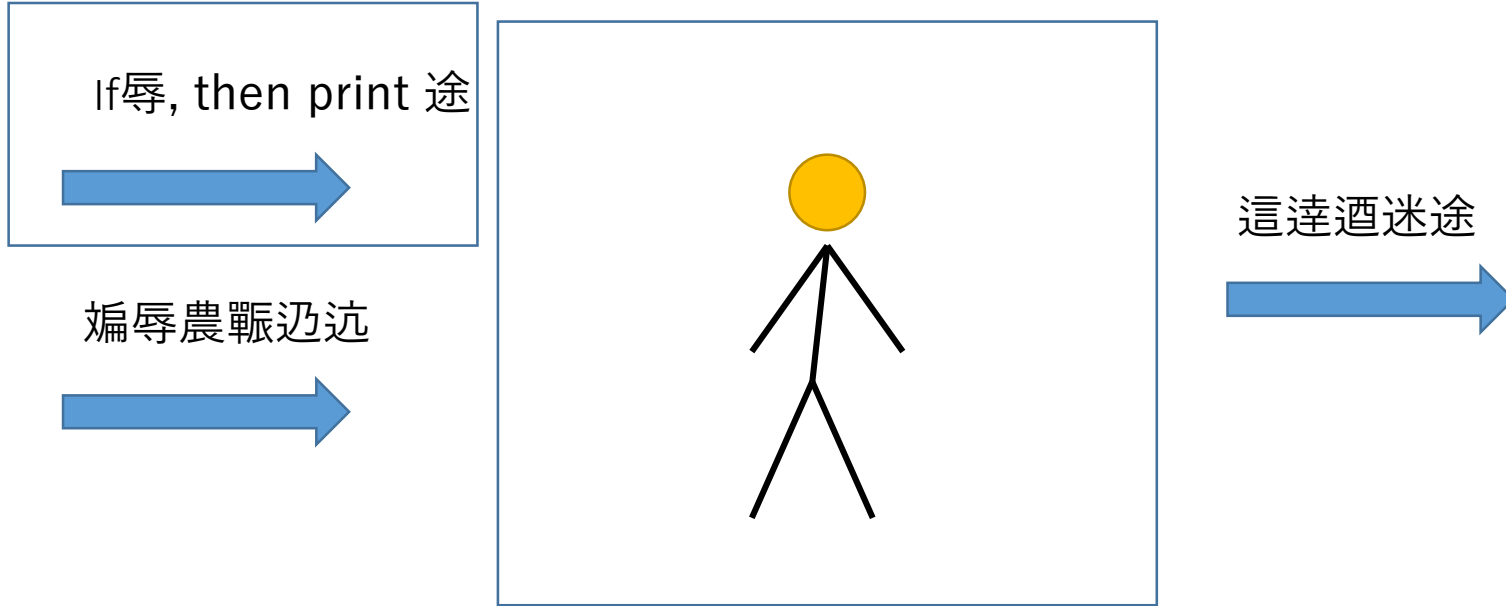
Objection

Weak AI

Maybe, the machine is simulating thinking, but
not really thinking!

Strong AI

Chinese room: step 1



Person does not understand Chinese

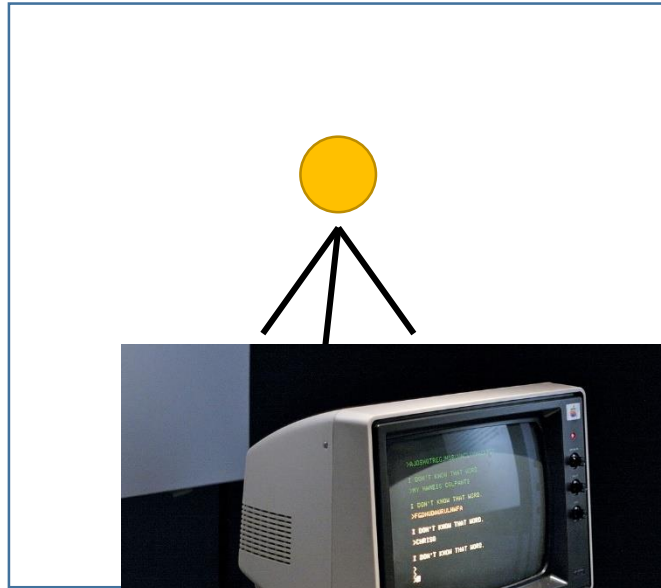
Searle (1980)

Chinese room: step 2

If辱, then print 途



爰辱農輶迈迄



這達迺迷途



Broader argument

1. Digital computers think only if they understand language.
2. Language understanding requires semantics, not just syntax.
3. Digital computers only manipulate symbols according to syntax.

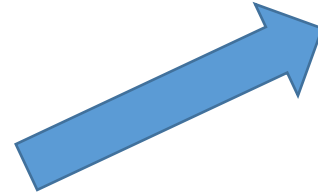
4. Thus, digital computers do not think.

Problems

- As stated, the argument only applies to GOFAI.
- System reply: In the scenario, the whole system may be said to understand Chinese.

Discussion today

local questions:



Global question:
Do digital
computers think?

Do they possess some
concepts?

Do they possess
explanations?

Can they discover laws?

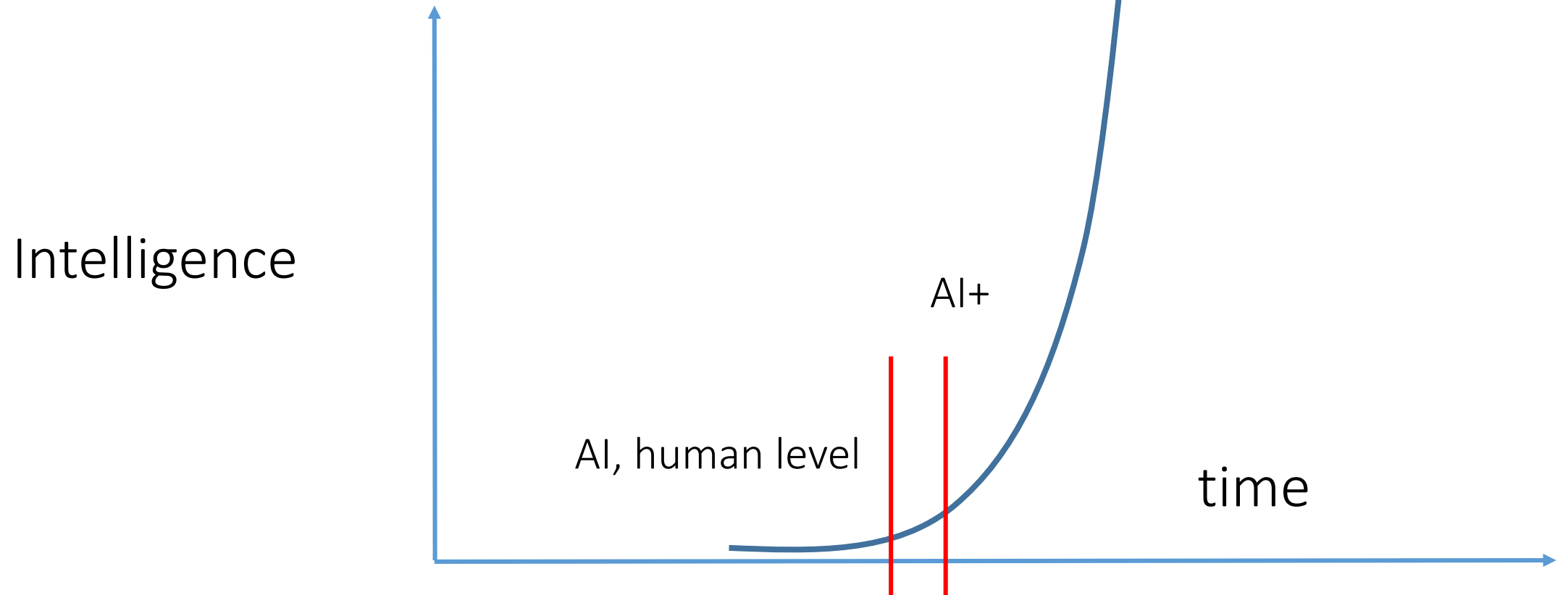
Do they perceive things?

4. Where will this lead us?

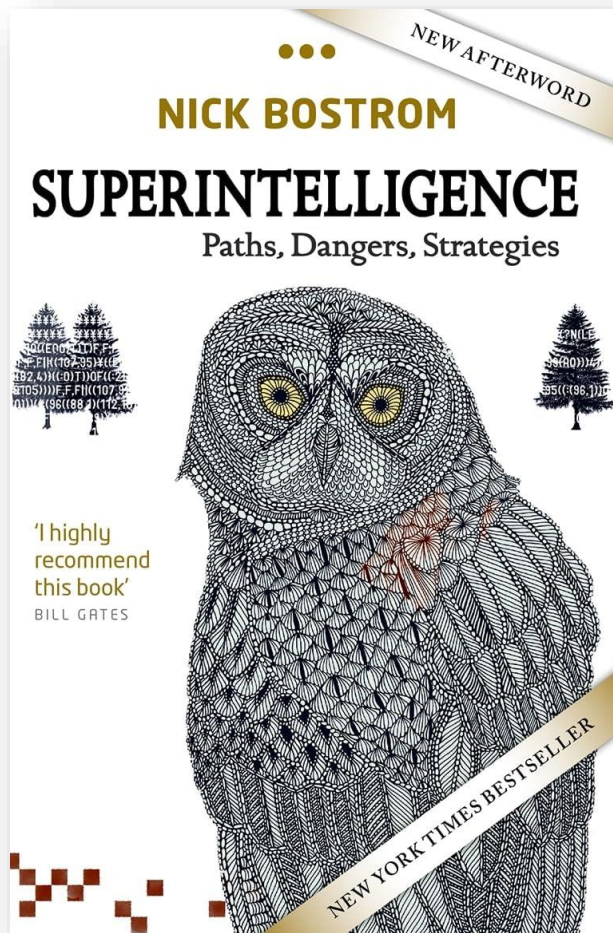
Questions:

Will there be human-level AI?
Will there be superintelligence?

Intelligence explosion



Important work



David Chalmers, *1966

David J. Chalmers

The Singularity *A Philosophical Analysis*

1. Introduction

What happens when machines become more intelligent than humans? One view is that this event will be followed by an explosion to ever-greater levels of intelligence, as each generation of machines creates more intelligent machines in turn. This intelligence explosion is now often known as the ‘singularity’.¹

Argument for human-level AI

- (i) Evolution produced human-level intelligence.
 - (ii) If evolution produced human-level intelligence, then we can produce AI (before long).
-
- (iii) Absent defeaters, there will be AI (before long).

Quoted from Chalmers 2010, p. 16/10

Argument for AI+

- (i) If there is AI, AI will be produced by an extendible method.
 - (ii) If AI is produced by an extendible method, we will have the capacity to extend the method (soon after).
 - (iii) Extending the method that produces an AI will yield an AI+.
-
- (iv) Absent defeaters, if there is AI, there will (soon after) be AI+.

Quoted from Chalmers 2010, p. 18/11

Iteration (Good's mechanism(

1. There will be AI (before long, absent defeaters).
 2. If there is AI, there will be AI+ (soon after, absent defeaters).
 3. If there is AI+, there will be AI++ (soon after, absent defeaters).
-
4. There will be AI++ (before too long, absent defeaters)

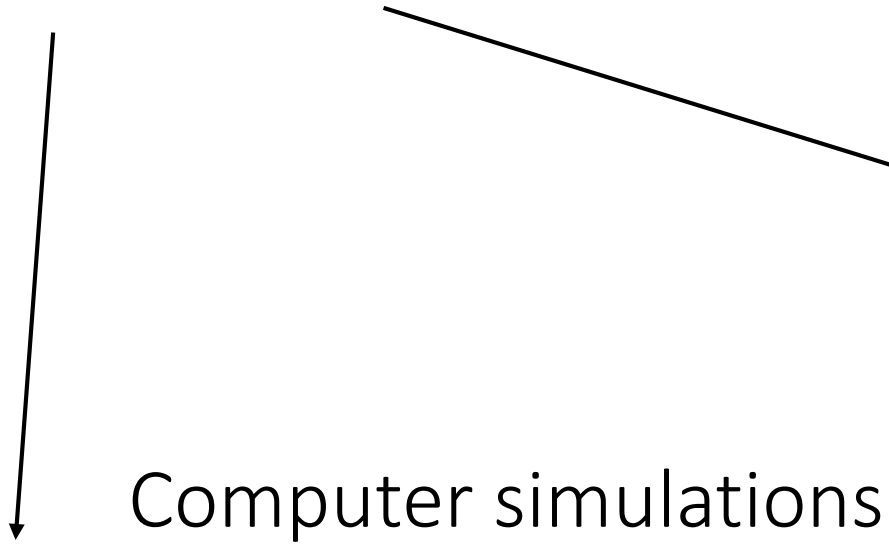
Quoted from Chalmers 2010, p. 12/6

Obstacles?

- Structural obstacles:
 - We are close to optimal intelligence.
 - We are not well positioned in intelligence space.
 - Diminishing returns.
- Correlation obstacles (see last argument)
- Manifestation obstacles
 - Motivational defeaters
 - Situational defeaters

Chalmers (2010)

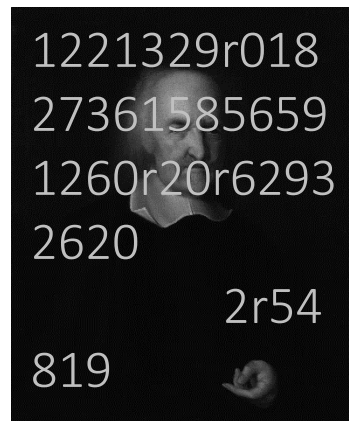
3. What are the implications for science?



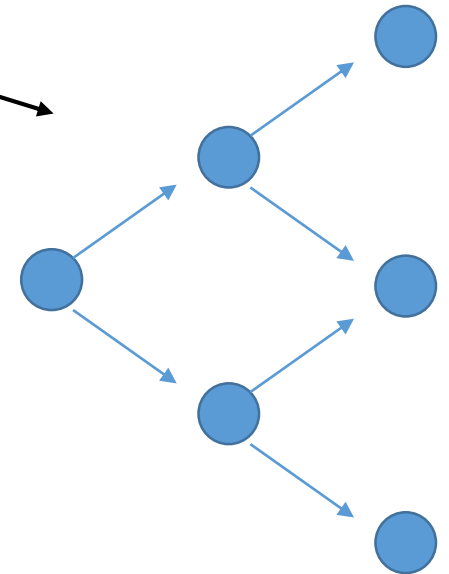
symbolic

$$c = \sqrt{a^2 + b^2}$$

Thomas Hobbes

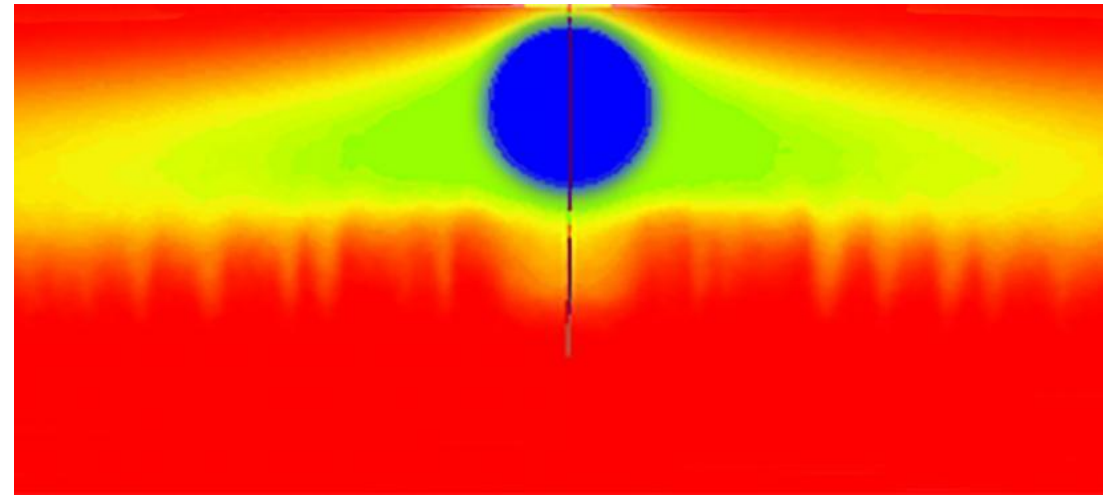


Subsymbolic/
connectionist



Neural networks

Computer simulations



Computer simulations:

- Obtain approximate and partial solutions to equations (fluid mechanics, Newtonian gravity, ...)
- Equations have physical meaning
- Equations trace time evolution of a system
- Computer obtains series of state descriptions of system

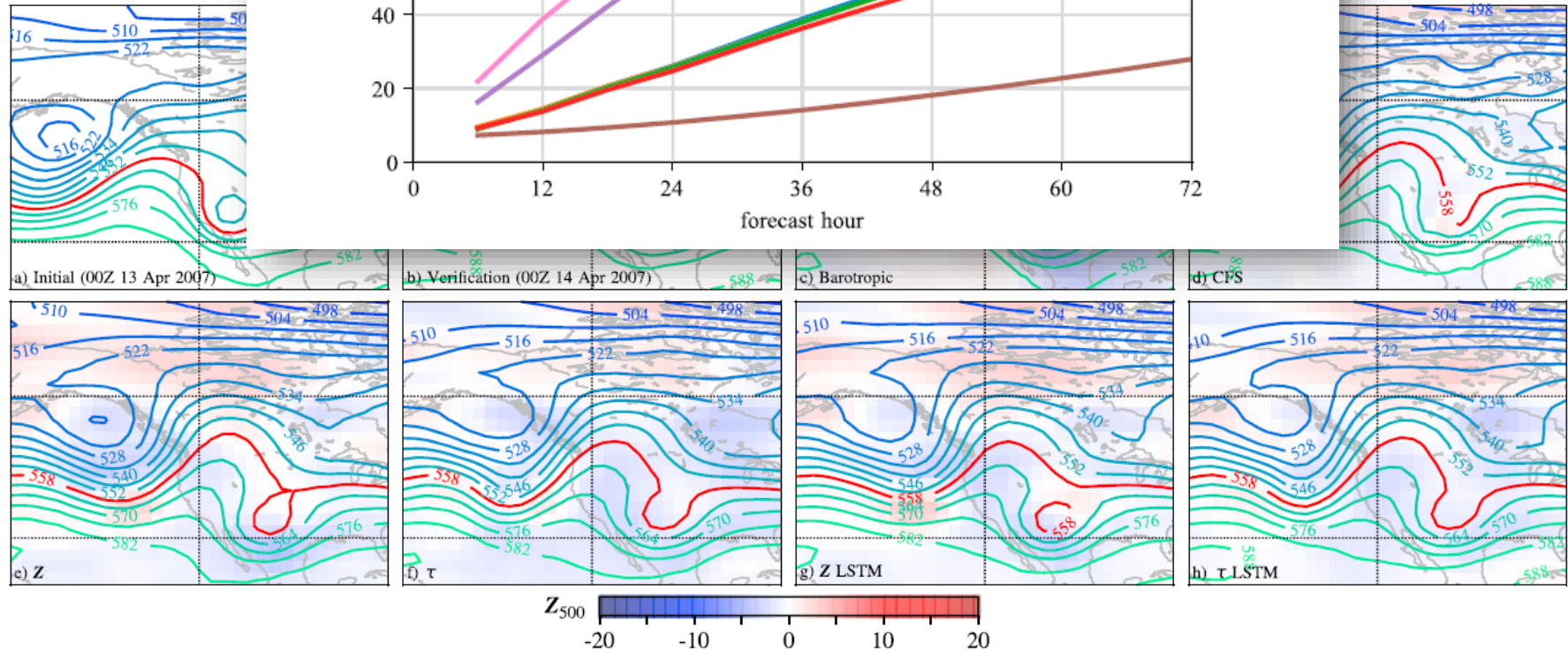
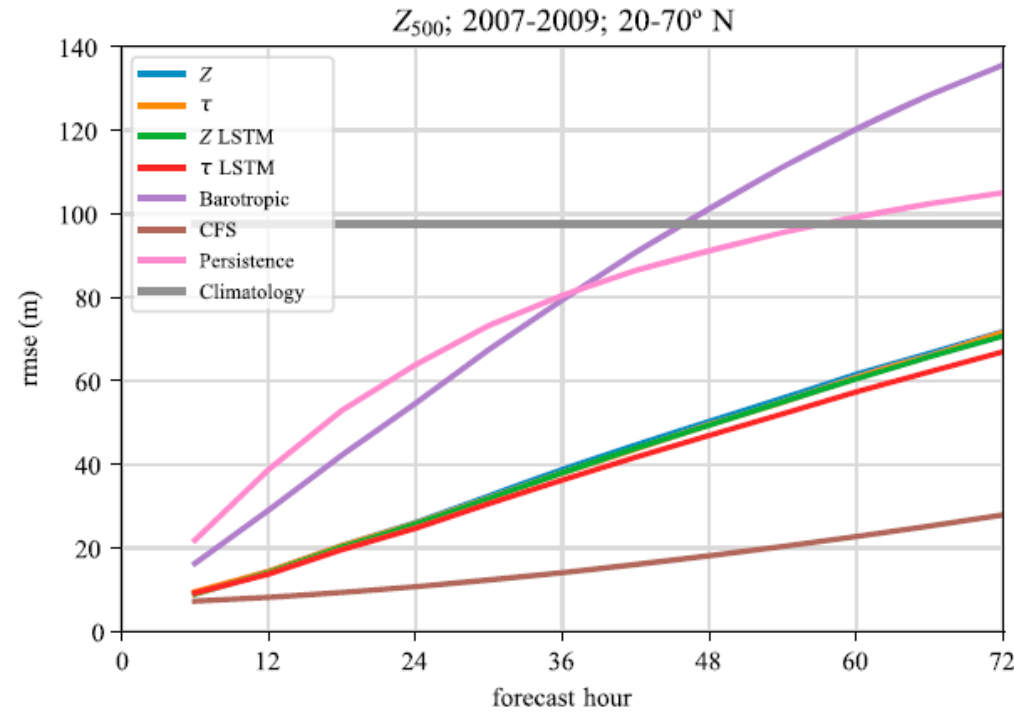
Networks

Can Machines Learn to Predict Gridded Height From

Jonathan A. Weyn¹

¹Department of Atmospheric Sciences, University of Washington, WA, USA

Abstract We develop deep convolutional neural networks (CNNs) trained on reanalysis data to predict the 500-hPa geopotential height from 2007 to 2009. The CNNs are trained on a Northern Hemisphere dataset up to 3 days, CNNs trained on climatological data, and the dynamical weather prediction model. The CNNs predict weather systems, which



Issues

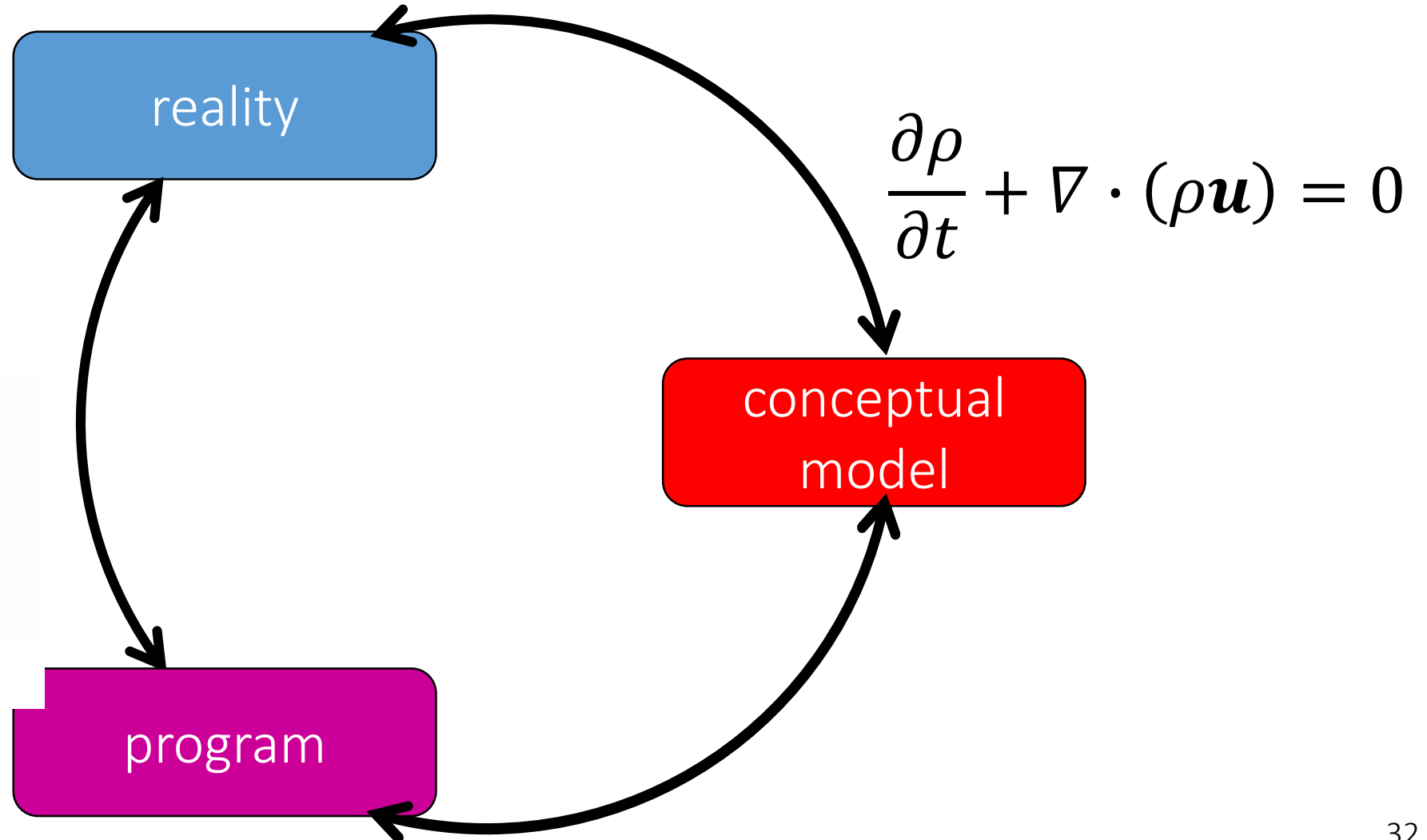
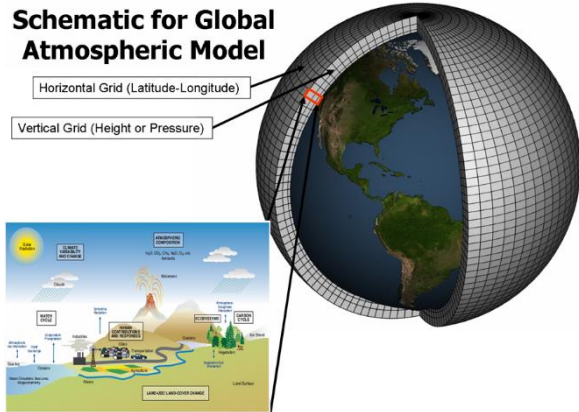
Can we trust the results?

Computer simulations: Sargent cycle

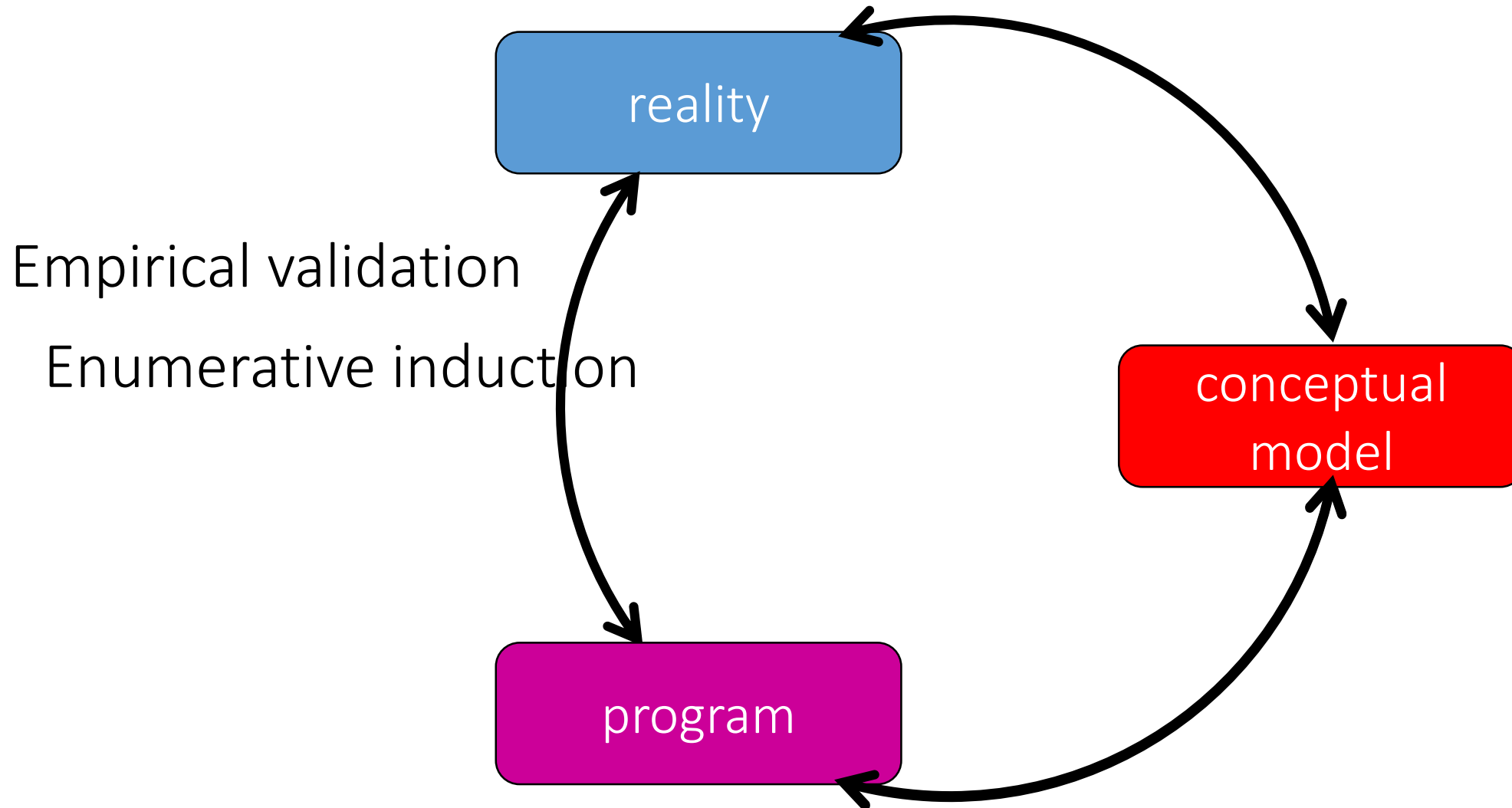


Schematic for Global Atmospheric Model

Horizontal Grid (Latitude-Longitude)
Vertical Grid (Height or Pressure)



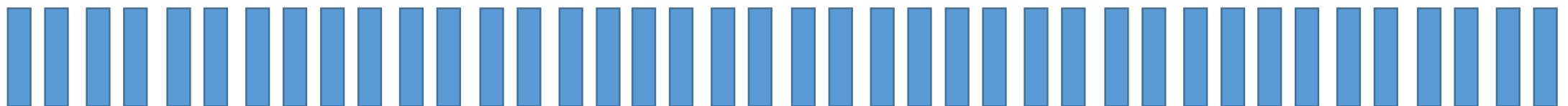
Computer simulations: Sargent cycle



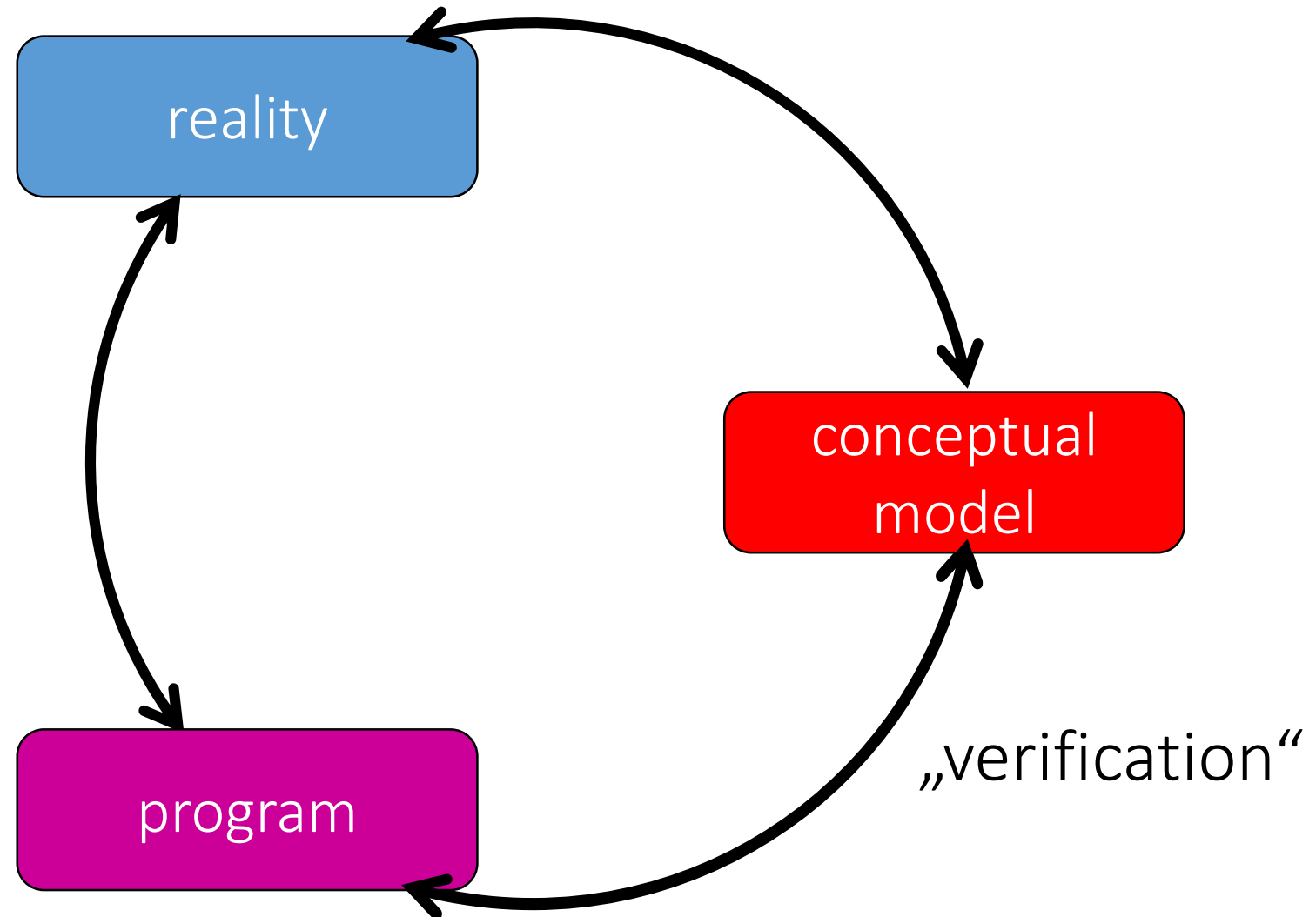
A challenge: opacity

Humphreys (2009, p. 618):

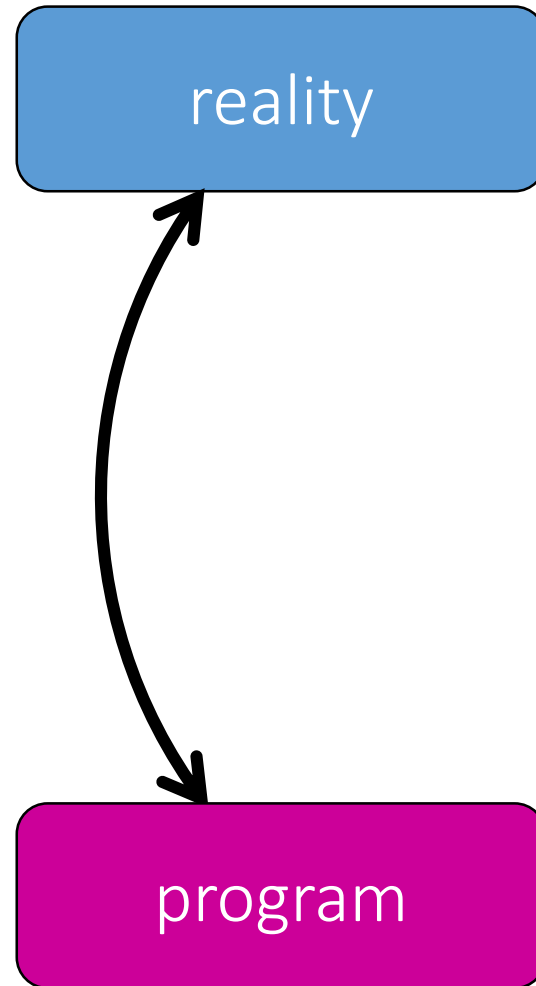
„Here a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process“



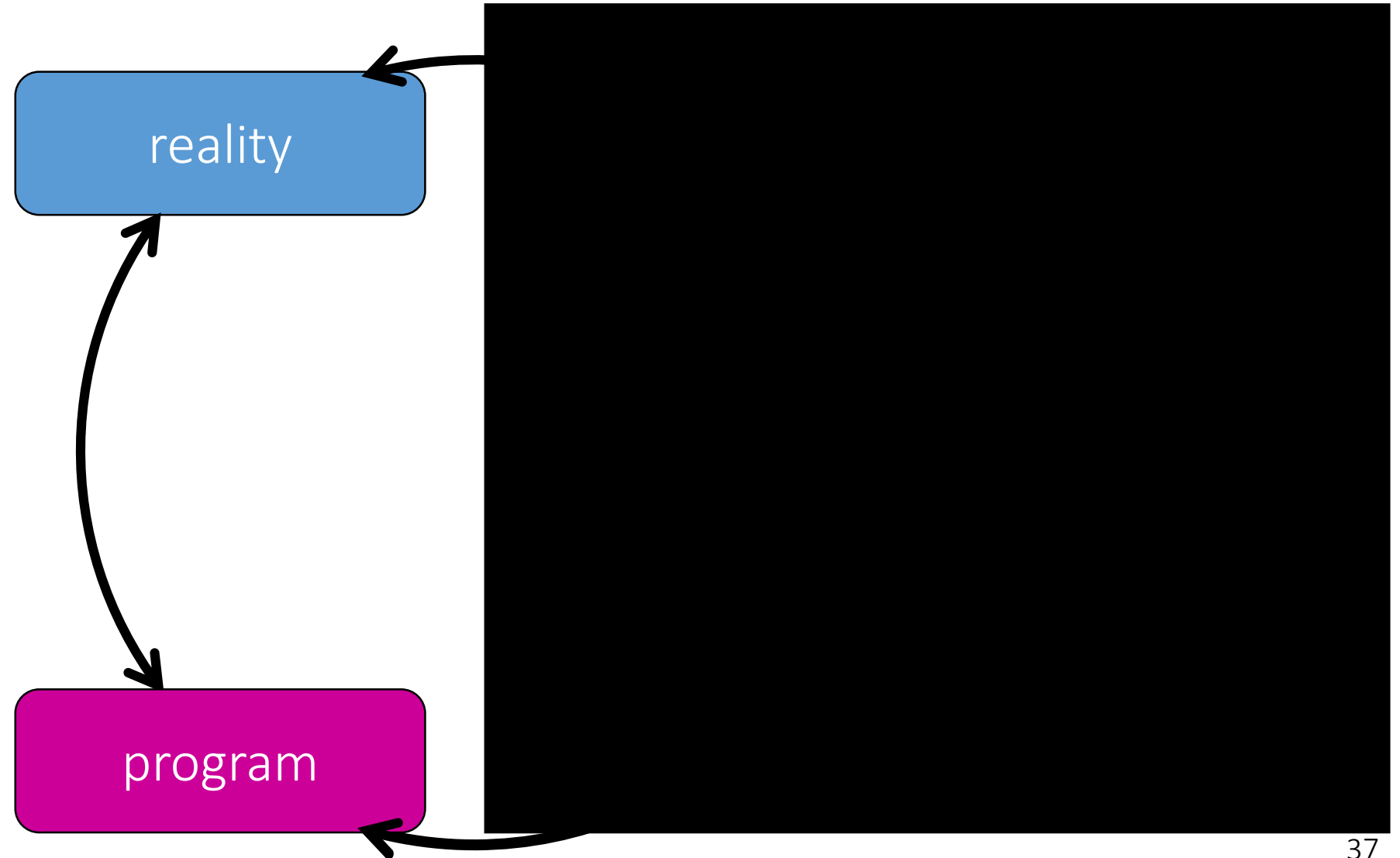
Computer simulations: Sargent cycle



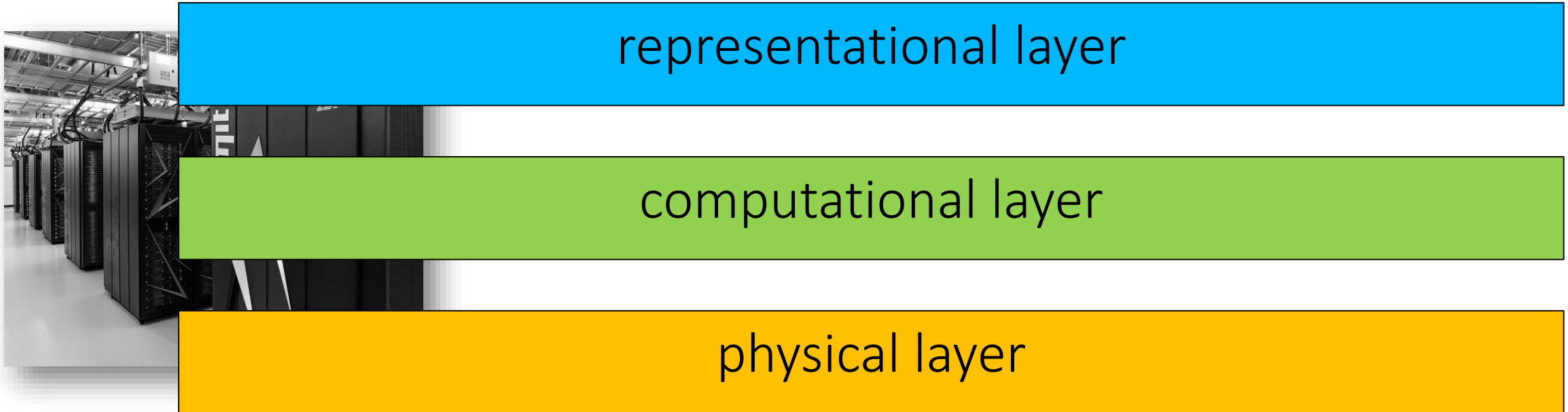
Neural networks



Maybe, there is more???



Computer simulation



Cf. Barberousse et al. (2009)

Artificial neural networks



computational layer

physical layer

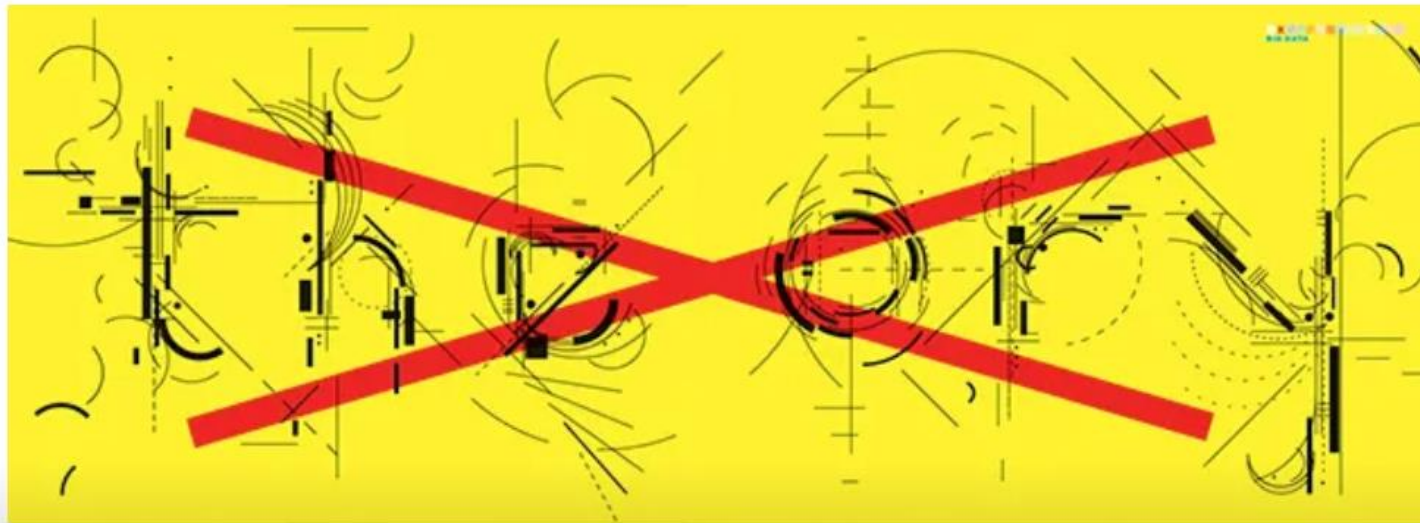
Consequences

- Trust in computer simulations rests on trust in conceptual model and empirical validation
- Trust in neural networks rests only on empirical validation (enumerative induction)
- Schubbach (2019): justification and explanation part company
- Networks are even more opaque than computer simulations
- Networks similar to expert judgment/tacit knowledge (Schubbach 2019)

Where does this lead us?

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

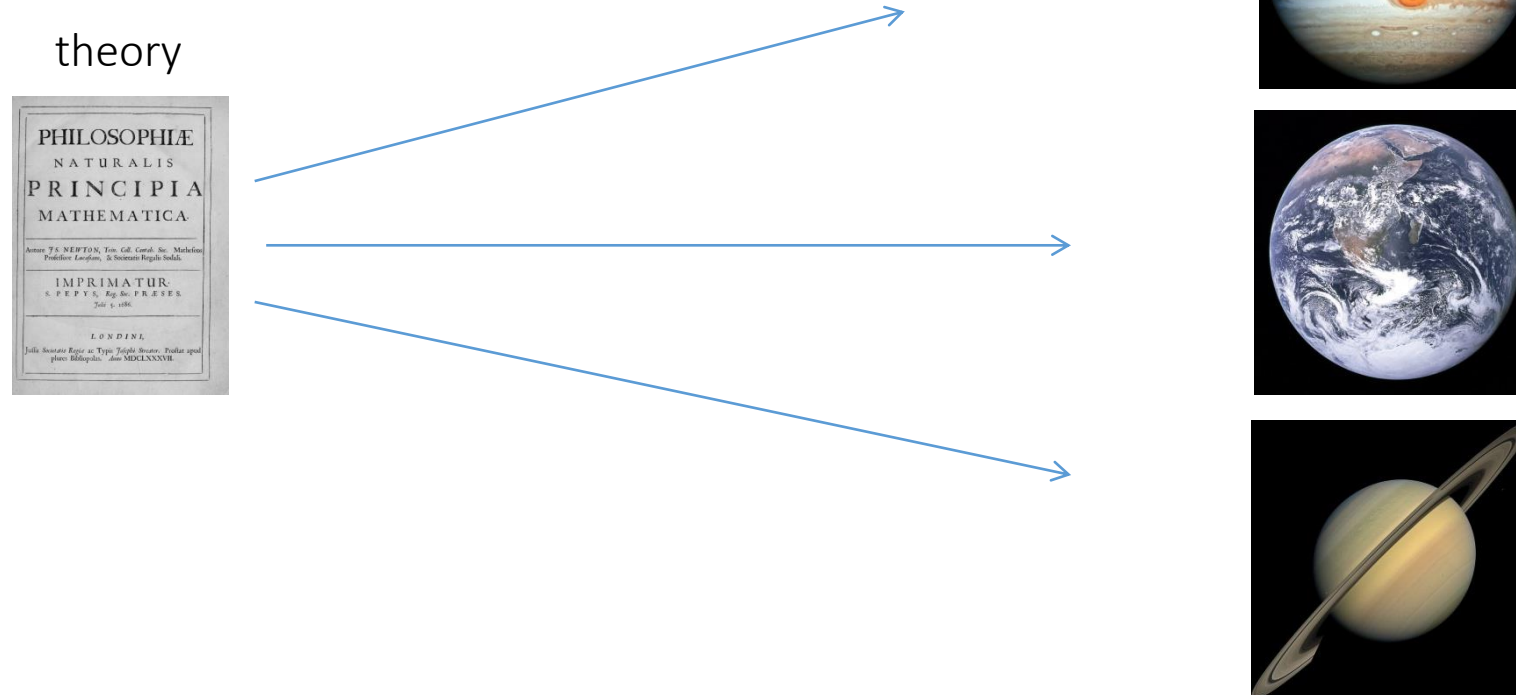
Illustration: Marian Bantjes “All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]



“The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.”

Theories

Real world systems



A theory contains lots of information in a very condensed manner.

Reactions

- Embrace Anderson's argument
- Argue that network-based science is poor
- Argue that good networks etc. contain theories

5. What's the take home message?

- AI delivers the outputs of many cognitive tasks.
- It outperforms humans in some tasks.
- It's arguable that AI doesn't really think, but this matters only for ethics.
- An intelligence explosion is possible, if not plausible.
- AI need not lead to a "flat science".

Merci viumau!

Literature

Bringsjord, S. & Govindarajulu, N. S. (2022), Artificial Intelligence, *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), E. N. Zalta & U. Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/>

Barberousse, A., Franceschelli, S., & Imbert, C., Computer Simulations as Experiments, *Synthese* 169 (2009), 557–574.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press.

Chalmers, David J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9-10):9 - 10.

Dreyfus, Hubert L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press (here imprint 1994).

Humphreys, P. 2009, The philosophical novelty of computer simulation methods, *Synthese* 169, 615–626.

Müller, V. C. (2021), [Ethics of Artificial Intelligence and Robotics](#), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.).

Räz, T.; Beisbart, C. (2022), The Importance of Understanding Deep Learning, *Erkenntnis*, <https://doi.org/10.1007/s10670-022-00605-y>

Literature

Russell, S. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach 3rd edition*, Saddle River, NJ: Prentice Hall

Sargent, R., Validation of simulation models. In Proceedings of the 1979 Winter Simulation Conference, ed. H. J. Highland, M. F. Spiegel, and R. E. Shannon, 497-503. Piscataway, New Jersey: IEEE.

Schlesinger, S., Crosbie, R. E., Gagné, R. E., Inis, G. S., Lalwani, C. S., Loch, J., Sylvester, R. J., Wright, R. D., Kheir, N., & Bartos, D., Terminology for Model Credibility, *Simulation* 32 (1979), 103–104.

Schubbach, A. Judging machines: philosophical aspects of deep learning. *Synthese* 198, 1807–1827 (2021).
<https://doi.org/10.1007/s11229-019-02167-z>.

Searle, John (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3):417-57.

Sullivan, E. (2022), Understanding from machine learning models, *British Journal for the Philosophy of Science* 73(1), 109–133, <https://doi.org/10.1093/bjps/axz035>.

Turing, A. (1950), Computing Machinery and Intelligence, *Mind*, LIX: 433–460

Weyn, J. A.; Durran, D. R.; Caruana, R. (2019), Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data, *J. Adv. Model. Earth Syst.* 11, 2680–2693, <https://doi.org/10.1029/2019MS001705>.