Learning Machines and measurement-related issues Workshop

Alessandro Giordani, Luca Mari, Mark Wilson

UC Berkeley, 4 August 2024

Abstract

When we say “Artificial Intelligence” today we mean Machine Learning (ML), particularly in its generative and conversational versions, and particularly under the spotlight as “chatbots.” Machines that learn, then: artificial agents whose behavior is the outcome of a combination of their programmed structure and their training process, in an unexpected reinterpretation of the “nature or nurture” adage. The very idea that a technological system can learn and thus deal with natural languages in a sophisticated and contextual way, i.e., is able to have conversations on practically any subject, is leading us toward a cognitive revolution, that raises many questions about us (the human beings), them (the machines), and our respective roles and relationships.

While questions like “do they think?”, “are they really intelligent?” may be left in the background, or simply dismissed as ill-posed, the evaluation of the quality of the behavior of chatbots is an increasingly important issue. Even though these are software entities with a formally specified structure, this behavior derives from the combination of such complex factors that it can be properly characterized as an empirical phenomenon. Hence, with the aim of acquiring sufficiently objective and intersubjective information on it, we are facing the challenge of defining measurable properties and developing measuring systems accordingly, as already acknowledged in particular by the EU Commission (e.g.: “... in cooperation with relevant stakeholders and organizations, such as metrology and benchmarking authorities, the Commission should encourage, as appropriate, the development of benchmarks and measurement methodologies for AI systems. In doing so, the Commission should take note and collaborate with international partners working on metrology and relevant measurement indicators relating to AI.” (AI Act, 2024, entry 74)).

In this workshop we introduce the technical concept of a “learning machine”, as grounded on an artificial neural network, and offer some preliminary hypotheses for a measurement-oriented conceptual framework about ML systems. This framework will be illustrated with some examples to make the presentation more concrete, from both non-generative and generative ML systems and applications. We plan for several discussions throughout the workshop, including topics such as (a) the basic formulation of the technical concepts, (b) our preliminary hypotheses, and (c) each of the examples. We hope and expect that participants will bring along their own hypotheses, and their own examples, thus ensuring a lively discussion. We will conclude by sketching some possible future research directions, both from our own work, and that of participants.

The context

The New York Times

A.I. Has a Measurement Problem

Which A.I. system writes the best computer code or generates the most realistic image? Right now, there's no easy way to answer those questions.

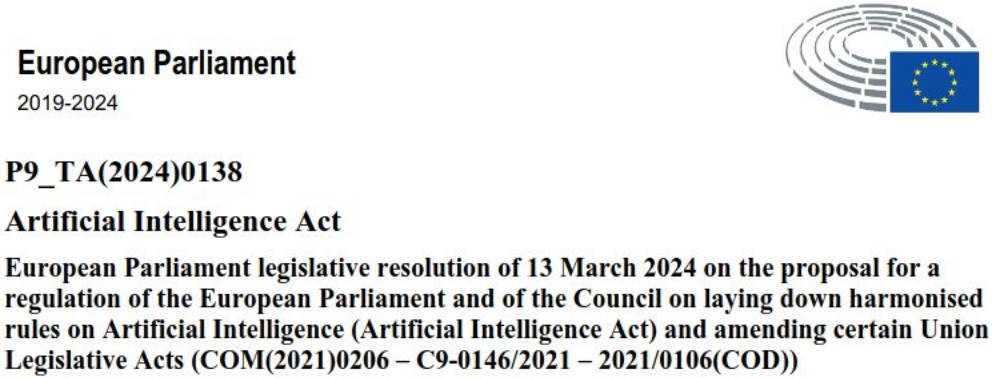
(15 April 2024, https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html)

The context

“Artificial Intelligence must be safe and secure. Meeting this goal requires robust, reliable, repeatable, and standardized evaluations of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use.”

(30 October 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence, Sec. 2a)

The context



European Parliament

2019-2024

P9_TA(2024)0138

Artificial Intelligence Act

European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))

“In cooperation with relevant stakeholders and organisation, such as metrology and benchmarking authorities, the Commission should encourage, as appropriate, the development of benchmarks and measurement methodologies for AI systems. In doing so, the Commission should take note and collaborate with international partners working on metrology and relevant measurement indicators relating to AI.”

(13 March 2024, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html , Sec. 74)

Scope and purpose

Let us better understand (AI and) this “measurement problem”
and explore together some ideas to operationalize some possible strategies
toward its solutions

Can we become active contributors to the solution of this “measurement problem”?

1. Background information on AI
2. AI for measurement science and measurement science for AI
3. Measurement science for AI: the received view
4. Toward an analytical framework
5. Sketches of an analytical framework
6. Open issues / main challenges

1. **Background information on AI**
2. AI for measurement science and measurement science for AI
3. Measurement science for AI: the received view
4. Toward an analytical framework
5. Sketches of an analytical framework
6. Open issues / main challenges

The basic position we propose

Will it be an **industrial revolution**?

Plausibly yes, but we cannot reliably predict its features yet

But it is, already today, a **cultural revolution**:

a paradigm shift that measurement science could help better understand

The example of a conversation with a chatbot

Chatting with an AI… (*not edited*)
A conversation simulating
a student-teacher relationship

The novelty is not in **what** it knows,
but in **how** it (knows and) interacts

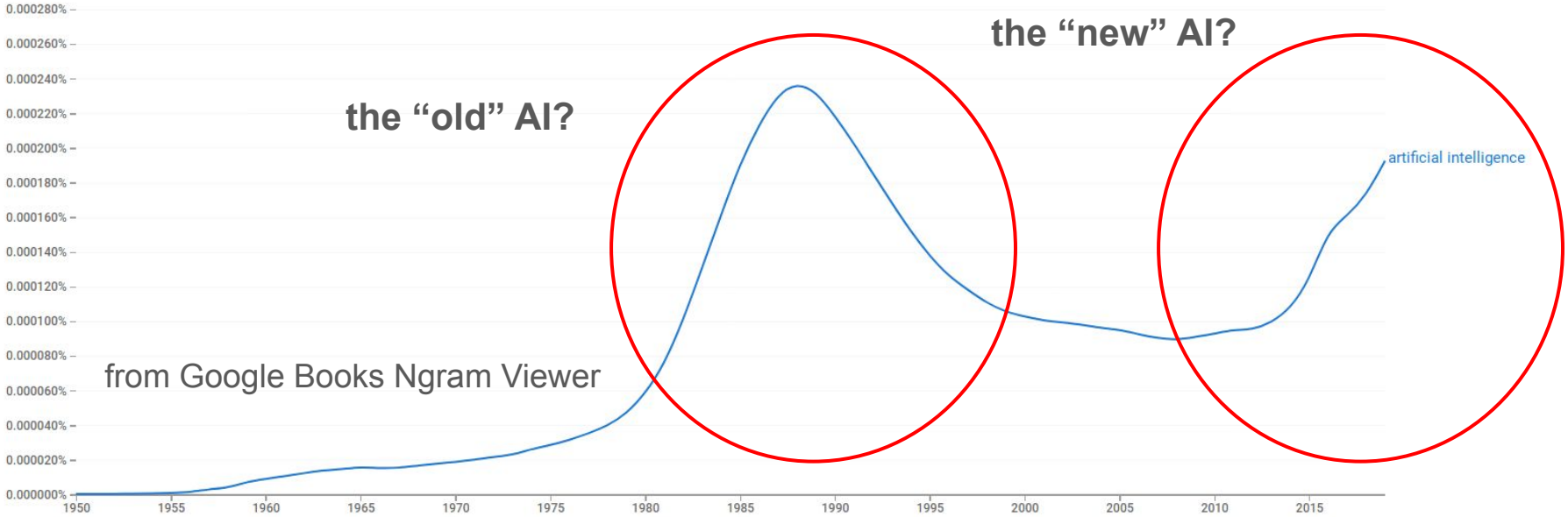
The entity with which we have had this conversation:

- writes a good English, and other languages
- produces original texts
- fulfills complex requests
- adapts its arguments to the context
- proposes creative contents
- analyzes and summarizes long texts
- shows sophisticated linguistic skills
- …

It is the first time that we may have such a kind of conversations with entities which are not human beings

How is it possible? How can chatbots exhibit such a behavior?

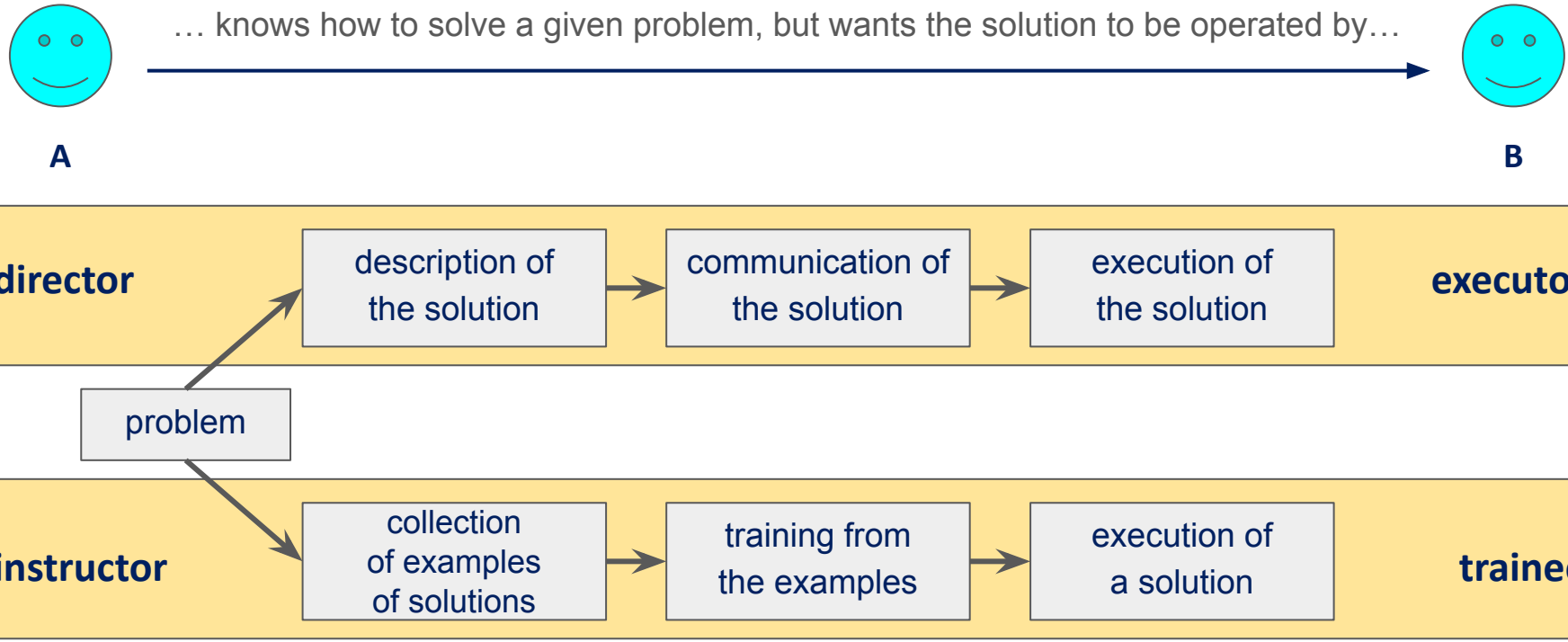
Artificial intelligence: a strange phenomenon



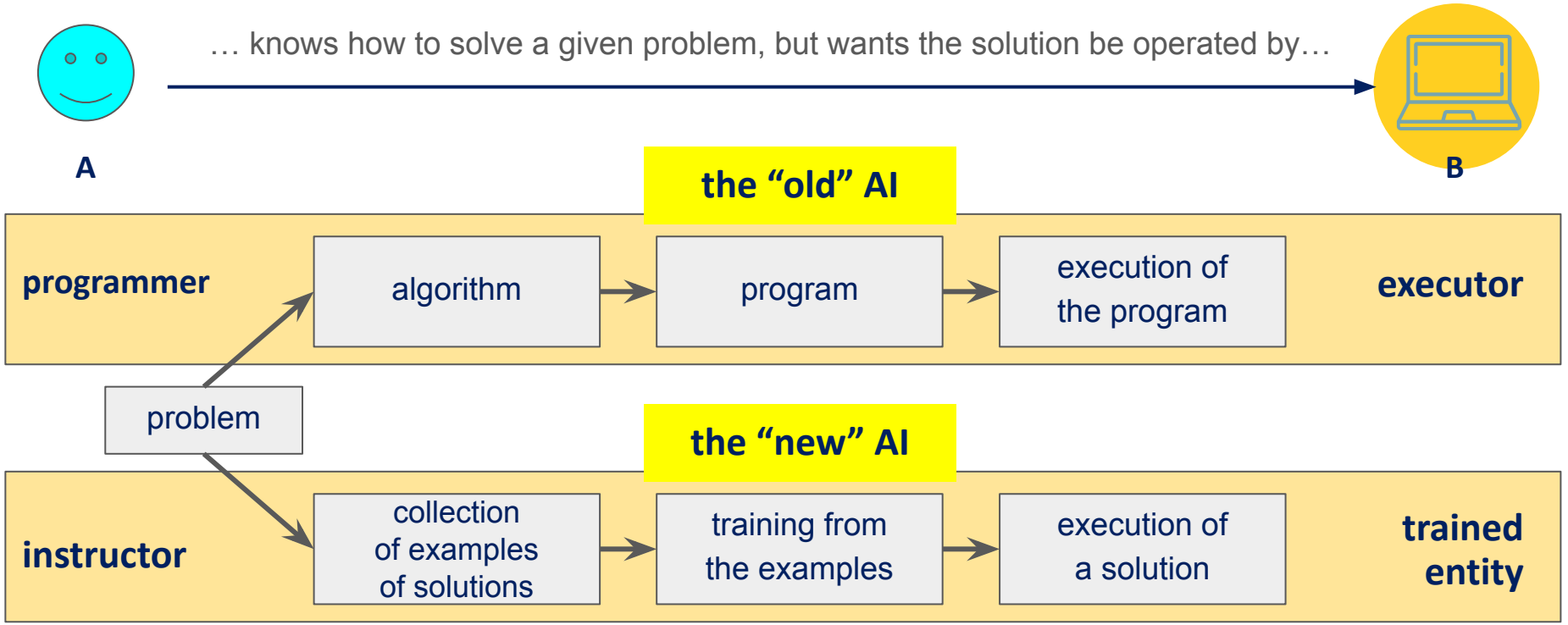
How can we explain it?

books.google.com/ngrams/graph?content=artificial+intelligence&year_start=1950&year_end=2019&corpus=26&smoothing=3#

Two strategies of problem solving



Two strategies of problem solving

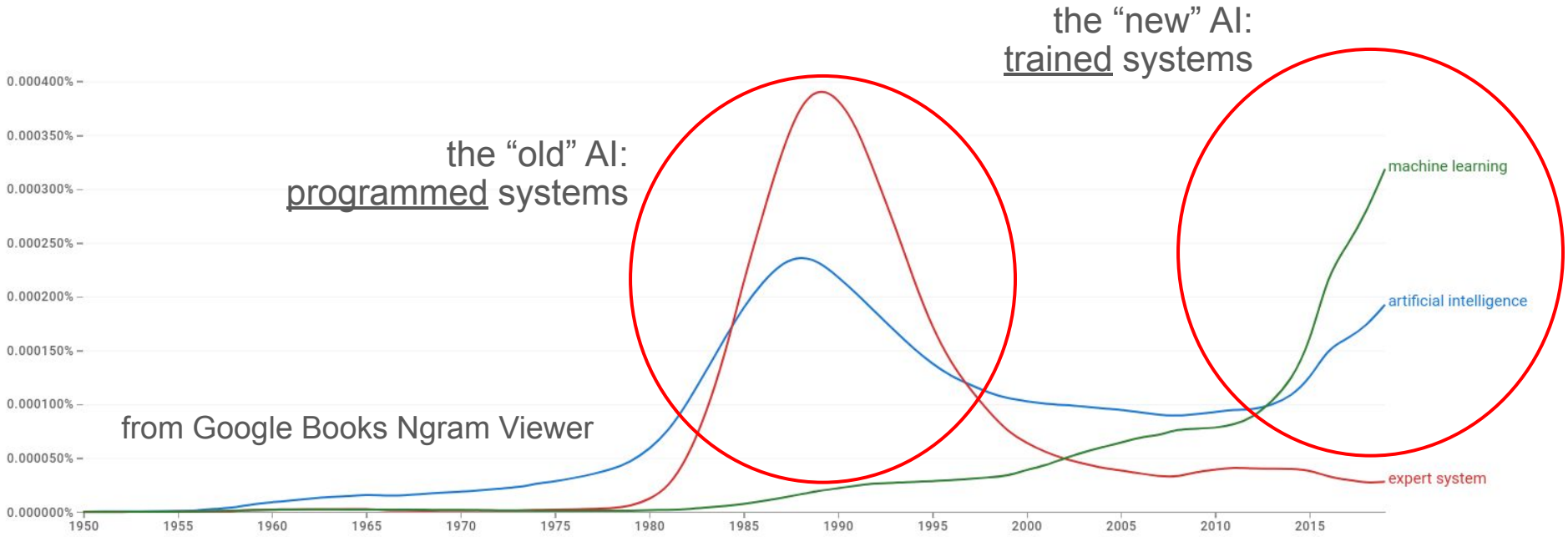


The “new” artificial intelligence: machine learning

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed”

A. Samuel, 1959

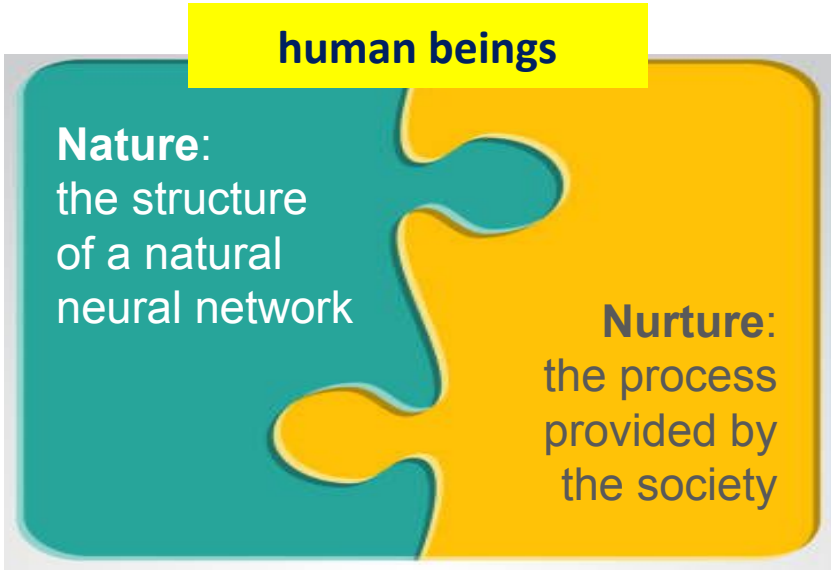
Two kinds of artificial intelligence, then…

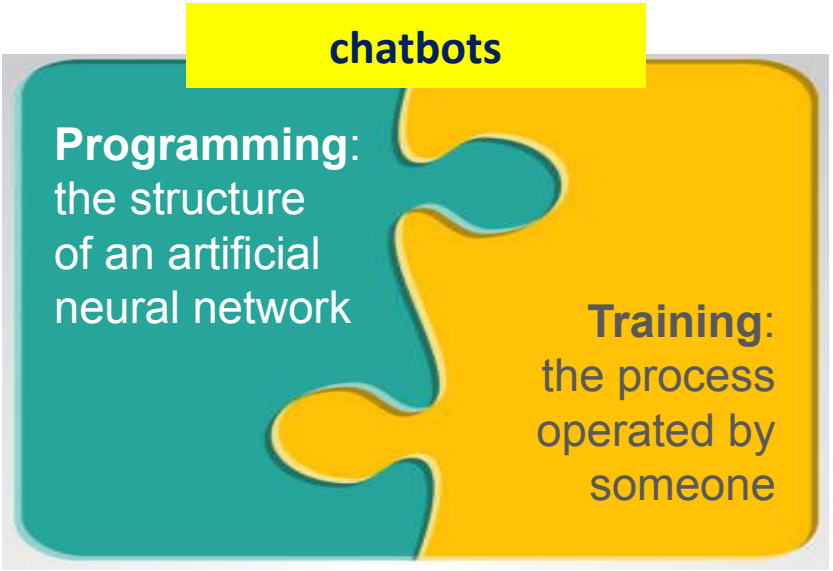


https://books.google.com/ngrams/graph?content=artificial+intelligence%2Cexpert+system%2Cmachine+learning&year_start=1950&year_end=2019&corpus=en-2019&smoothing=3

Learning machines: a tentative interpretation

"**Nature versus nurture** is a long-standing debate in biology and society about the relative influence on human beings of their genetic inheritance (nature) and the environmental conditions of their development (nurture)."
https://en.wikipedia.org/wiki/Nature_versus_nurture





Learning machines: something on which we still have a lot to learn!

But perhaps is it only hype, or worse?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 

March 2021, Proc. ACM Conf. on Fairness, Accountability, and Transparency, https://dl.acm.org/doi/10.1145/3442188.3445922

Noam Chomsky: The False Promise of ChatGPT

8 March 2023, New York Times, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models

10 March 2023, Philosophy & Technology, https://link.springer.com/article/10.1007/s13347-023-00621-y

An interpretation…

… to avoid what could be a pseudo-problem:

The Fathers of the field had been pretty confusing: John von Neumann speculated about computers and the human brain in analogies sufficiently wild to be worthy of a medieval thinker and Allan M. Turing thought about criteria to settle the question of whether Machines Can Think, a question of which we now know that it is about as relevant as the question of whether Submarines Can Swim.

E. Dijkstra, 1984 http://www.cs.utexas.edu/users/EWD/ewd08xx/EWD898.PDF

1. Background information on AI
2. **AI for measurement science and measurement science for AI**
3. Measurement science for AI: the received view
4. Toward an analytical framework
5. Sketches of an analytical framework
6. Open issues / main challenges

Framing our exploration

The relation between AI and measurement science (MS) is twofold:

- **AI for MS**: how can AI help improving measurement?
  (“smart” meters, automated test generation and evaluation, …)
- **MS for AI**: how can MS help improving learning machines (LMs)?

We focus here on the latter, and specifically about evaluation of LM behavior:

- can we measure the quality of the behavior of a LM? how?
- for a given kind of task, what are the best LMs?

1. Background information on AI
2. AI for measurement science and measurement science for AI
3. **Measurement science for AI: the received view**
4. Toward an analytical framework
5. Sketches of an analytical framework
6. Open issues / main challenges

The received view

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?'

https://academic.oup.com/mind/article/LIX/236/433/986238

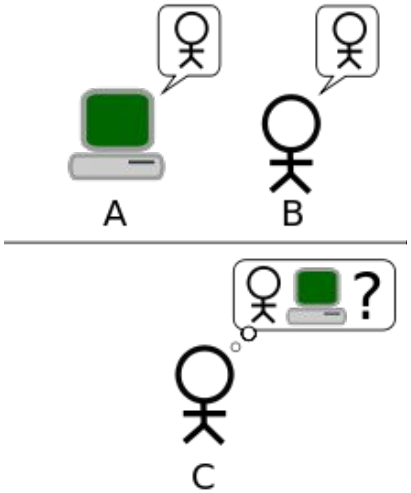
Being intelligent, behaving intelligently

"The **Turing test**, originally called the imitation game by Alan Turing in 1950, is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human."
https://en.wikipedia.org/wiki/Turing_test

C asks the same questions to A and B

and receives the answers from both,
with no information of who / which answered what

If C cannot identify the computer by its answers, then the behavior of A is not distinguishable from the one of (the supposedly intelligent) B



https://plato.stanford.edu/entries/turing-test

Beyond the Turing test?

Evaluating the quality of behavior through a black box strategy, then

It is simple to put in operation but

- it is anthropocentric
  (there can be useful non-human-like forms of intelligence)
- the measurand is only implicitly defined
  (‘intelligence’ defined as what is assessed by Turing test?)
- its outcomes are strongly contextual
  (see Eugene Goostman’s affair, https://en.wikipedia.org/wiki/Eugene_Goostman)
- only relates to some components of quality of behavior
  (which is not only about “intelligence”, but also responsibility, etc.)

1. Background information on AI
2. AI for measurement science and measurement science for AI
3. Measurement science for AI: the received view
4. **Toward an analytical framework**
5. Sketches of an analytical framework
6. Open issues / main challenges

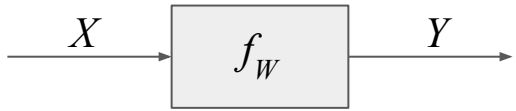
Looking inside the box of a Learning Machine

LMs are software systems, but their behavior is **not programmed**

They are neither search engines nor databases: they neither search nor store data

In their current “reference implementations” (artificial neural networks),
they are parametric functions trained by adapting parameter values to fit the provided examples

$$Y = f_W(X)$$



1. **Training**: adapt the weights W so that

$$known\ expected\ output = f_W(known\ given\ input)$$

(typically by means of gradient descent of a loss function, as in this tiny example)

2. **Inference**: predict an output by computing f_W, with the fitted weights W, on the given input

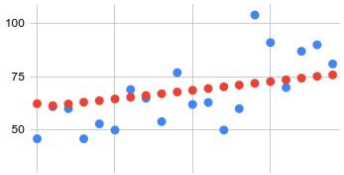
$$predicted\ output = f_W(known\ given\ input)$$

Looking inside the box of a Learning Machine

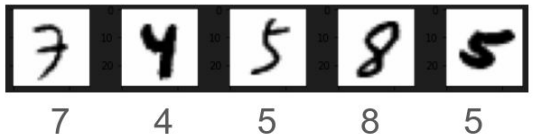
An example: let’s teach a neural network how to read digits!
https://lmari.github.io/chatting/activities/mnist_en.html

Orders of magnitude…

Linear regression: 10^0 params



Reading handwritten digits: 10^4-10^5 params



SOTA Transformers: 10^9-10^{12} params

Human brain: 10^{15} params

LM behavior becomes more complex when the number of its parameters increases, and this makes its evaluation more complex in turn

Learning Machines and Measuring Systems

Some interesting structural analogies:

- LMs perform inference (forward mode) only after have been trained (backward mode)
- MSs perform measurement (forward mode) only after have been calibrated (backward mode)

Training	**Calibration**
training set	calibrated measurement standards
labels in training set	reference values of quantities of measurement standards
Inference	**Measurement**
input data	measurand
prediction	measurement result

Evaluating the quality of behavior of Learning Machines

The analogy with measuring systems is again suggestive

The quality of the behavior of

a LM depends on
- **its designed structure**
- **the quality of its training**

  which is about
  - the training set (correctly sampled and unbiased)
  - the training process

a MS depends on
- **its designed structure**
- **the quality of its calibration**

  which is about
  - measurement standards (correctly sampled and unbiased)
  - the calibration process

This analogy suggests a blueprint for our analysis

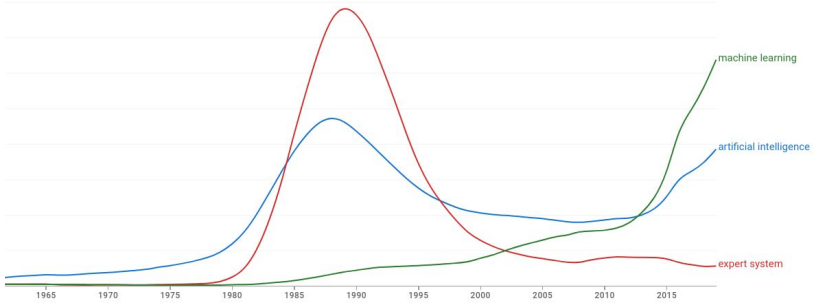
Our context: refined version

artificial intelligence

programmed vs learned behavior

knowledge/rule-based, expert systems

data-driven, machine learning systems



(relatively) simple vs complex validation

classifiers / regressors

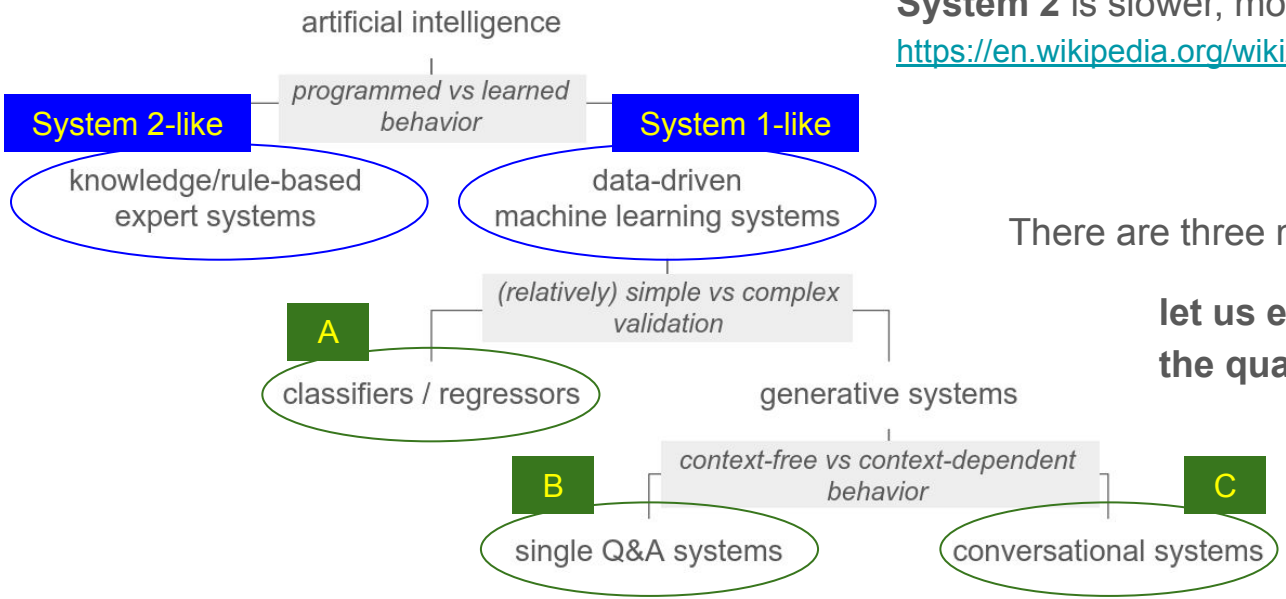
generative systems

context-free vs context-dependent behavior

single Q&A systems

conversational systems

Our context: preliminary hypotheses



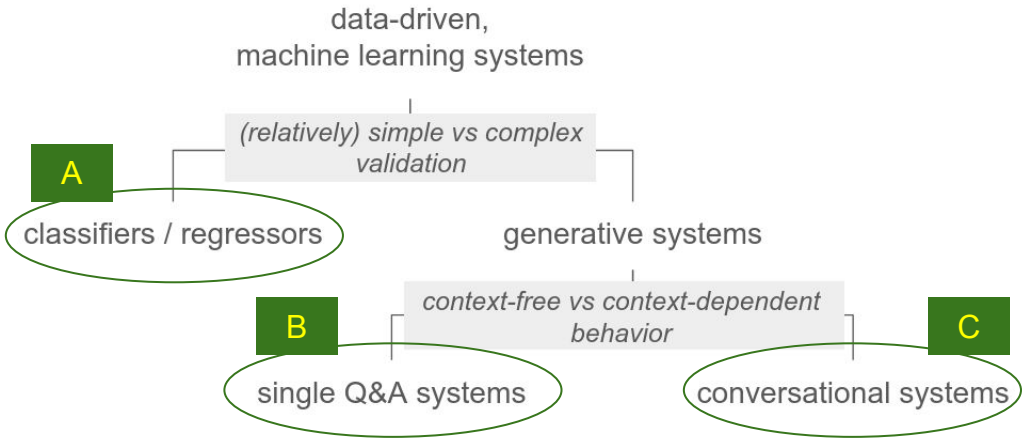
"Thinking, Fast and Slow [by] Daniel Kahneman [distinguishes] between two **modes of thought**:
System 1 is fast, instinctive and emotional;
System 2 is slower, more deliberative, and more logical."
https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow

There are three main types of System 1-like LMs:

let us explore how to evaluate the quality of their behavior

1. Background information on AI
2. AI for measurement science and measurement science for AI
3. Measurement science for AI: the received view
4. Toward an analytical framework
5. **Sketches of an analytical framework**
6. Open issues / main challenges

An evaluation-oriented framework (draft!)



Once a LM has been trained, it can be put in operation for inference

For evaluating the quality of its behavior, two main issues need to be considered:

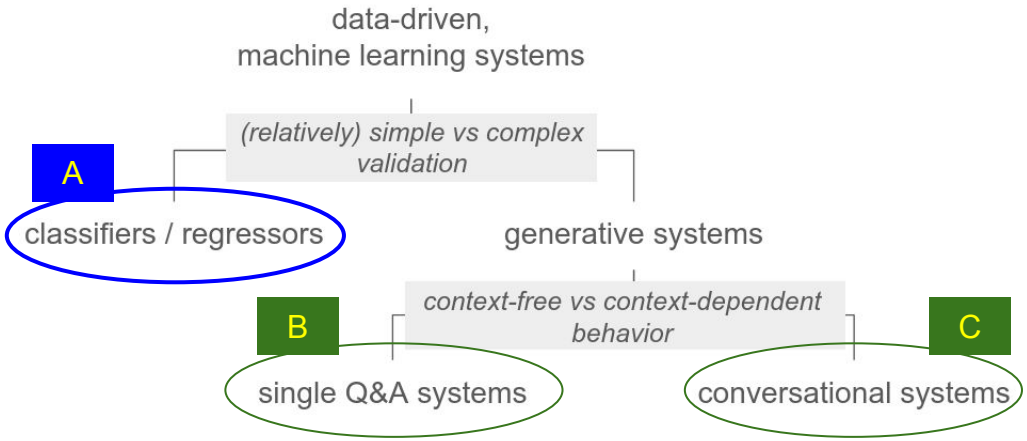
– do we know **what** we want to measure?

→ is the measurand (“quality of behavior”) well defined?

– do we know **how** we want to measure?

→ is the measuring system well designed?

An evaluation-oriented framework: Type A



Traditional ML systems,
for classifications or regressions

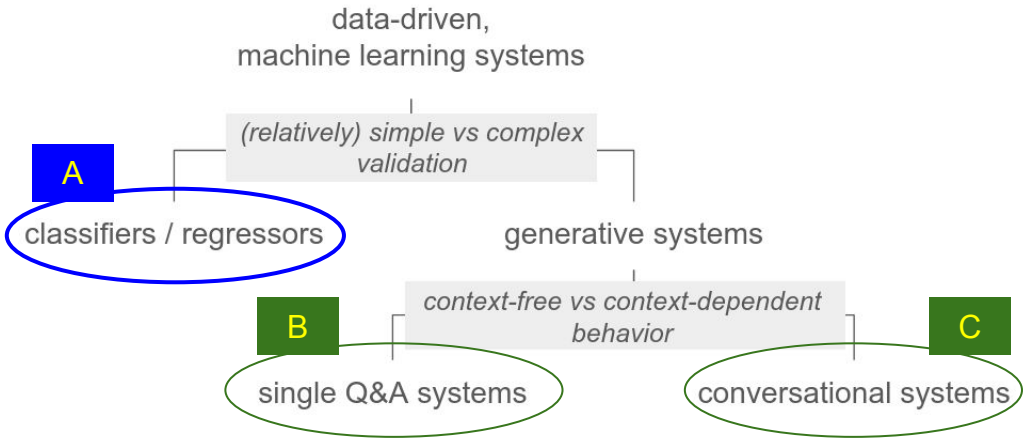
- the measurand is well defined
- labels / true values for training / calibration are available

Tools: k-nearest neighbors, logistic regression, decision trees, neural networks, …
Examples: handwritten character recognition, antispam filtering, recommendation systems, sentiment analysis, time series forecast, …

Well-known statistical / data mining techniques: distinction between features and targets; training vs test set split; bias vs variance (underfitting vs overfitting); …

Well-known statistical / data mining quality parameters: precision, recall, accuracy, …

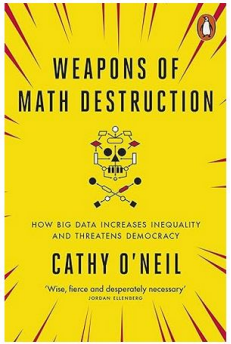
An evaluation-oriented framework: Type A



Traditional ML systems,
for classifications or regressions

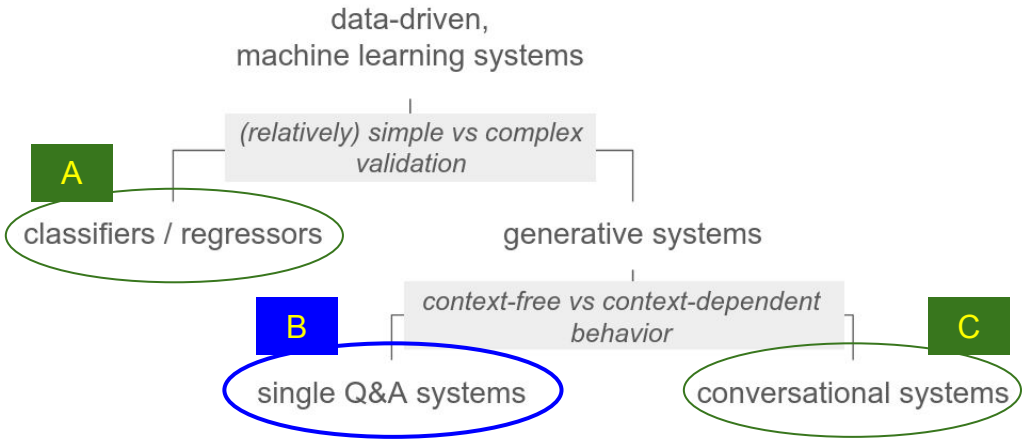
Two main measurement-related problems:

- **instability**, if the relation between the features and the targets changes in time
  ("independent and identically distributed variables" (IID) condition not fulfilled)
- **bias**, if the training set is not sufficiently representative of the population
  (incorrect choice of measurement standards used for calibration)



In education this is analogous to evaluation by means of multiple choice tests:
if known problems are solved (bias ("studying for the exam"), undersampling, …),
assessing students' skills using such tests is unproblematic

An evaluation-oriented framework: Type B



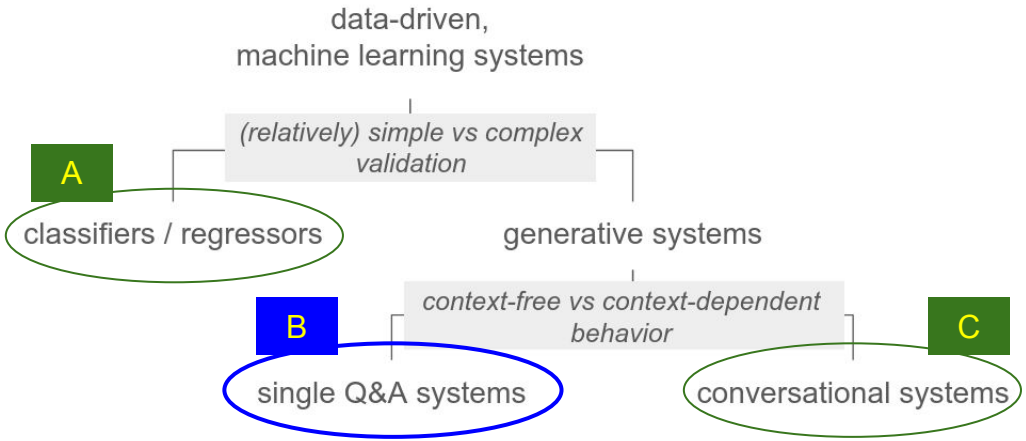
Single Q&A GenAI systems,
for context-free tasks

- the measurand is not so well defined
- labels / true values used in training / calibration may be controversial in inference

Tools: RNNs, Transformers
Examples: translation, summarization, …

All common benchmarks for language models assume single Q&A
(see, e.g., https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard)
– MMLU (Massive Multitask Language Understanding): performance on a wide range of tasks (“the SAT for chatbots”)
– HellaSwag: commonsense reasoning
– PIQA (Physical Interaction Question Answering): comprehension of physical interactions
– WinoGrande: common sense reasoning, complex pronoun disambiguation

An evaluation-oriented framework: Type B



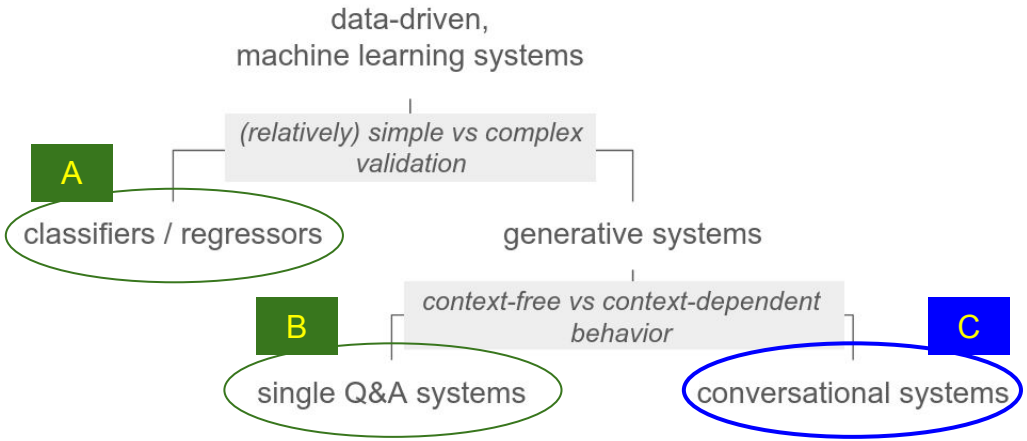
Single Q&A GenAI systems,
for context-free tasks

Together with what was mentioned for Type A systems, the key measurement-related problem: there could be no intersubjective criteria to assess the quality of inference results

(see the case of BLEU (bilingual evaluation understudy): “the closer a machine translation is to a professional human translation, the better it is”, https://en.wikipedia.org/wiki/BLEU)

In education this is analogous to evaluating the quality of essays, summaries, translations, ..., a process for which establishing sufficiently objective and intersubjective criteria can be hard, but which has already been studied and for which psychometrics has already developed tools, such as construct maps

An evaluation-oriented framework: Type C



Sequential Q&A GenAI systems,
for context-sensitive tasks

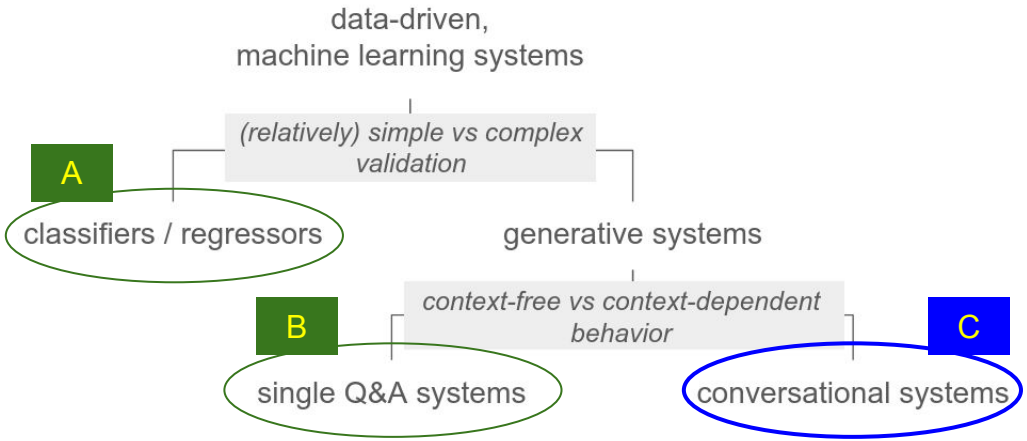
- the measurand is not so well defined
- labels / true values used in training / calibration may be controversial in inference

Tools: Transformers
Examples: like for Type B, plus conversations

We are not aware of any benchmark / metric specifically devoted to context-sensitive tasks

A widely assessment tool is LMSYS Chatbot Arena Leaderboard (https://chat.lmsys.org/?leaderboard), based on direct comparison and using the Elo rating system

An evaluation-oriented framework: Type C

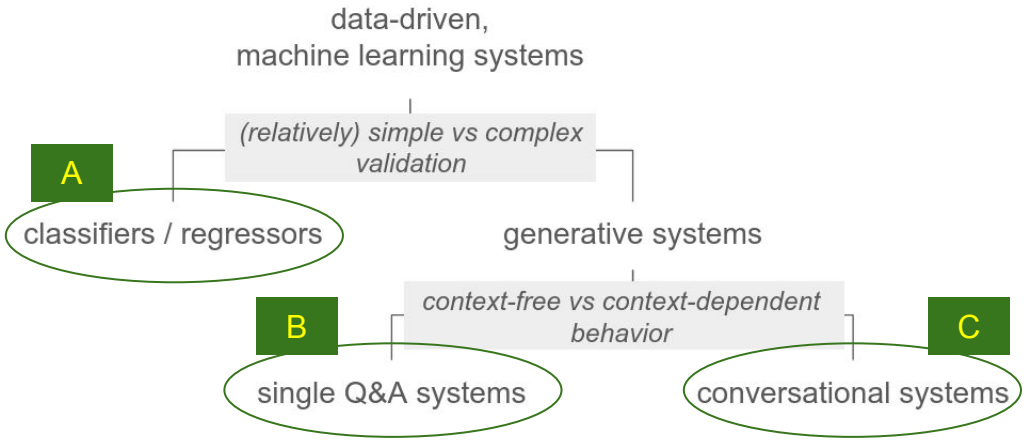


Sequential Q&A GenAI systems,
for context-sensitive tasks

The key measurement-related problem is the same as the one for Type A systems,
and even much harder to solve:
there could be no intersubjective criteria to assess the quality of inference results

In education this is analogous to evaluating the quality of an (interactive) oral examination,
a process for which establishing sufficiently objective and intersubjective criteria is very hard

In summary



Types of systems

A, traditional ML systems

B, single Q&A GenAI systems

C, sequential Q&A GenAI systems

Measurement-related problems

solved or well-known

partially solved, hard

unsolved, very hard

1. Background information on AI
2. AI for measurement science and measurement science for AI
3. Measurement science for AI: the received view
4. Toward an analytical framework
5. Sketches of an analytical framework
6. **Open issues / main challenges**

Open issues / main challenges

The evaluation of the quality of behavior of Type C systems (chatbots…),
as a context-sensitive task, is still an open issue

Moreover:

- chatbots can be trained to operate with functions / programmed tools,
  and therefore hybrid System 1 - System 2 entities
- chatbots can be enabled to interact with each other in agent-based architectures

How to evaluate the quality of behavior of these systems is still an open issue

Finally, chatbots are inevitably ideological in their interaction:
how to decide whether a certain ideology is appropriate is a extra-metrological question

Thanks for your participation!

Alessandro Giordani, Luca Mari, Mark Wilson