Faith At Work

Lord Lionel Tarassenko

Keynote

Generative AI: a force for good?

Annual Summit

20th April, Monday
9:00am-5:00pm, London

Generative AI: a force for good?

EY headquarters

London, 20 April 2026

Professor Lord (Lionel) Tarassenko CBE FREng FMedSci
President, Reuben College
Professor of Electrical Engineering
University of Oxford

Overview of talk

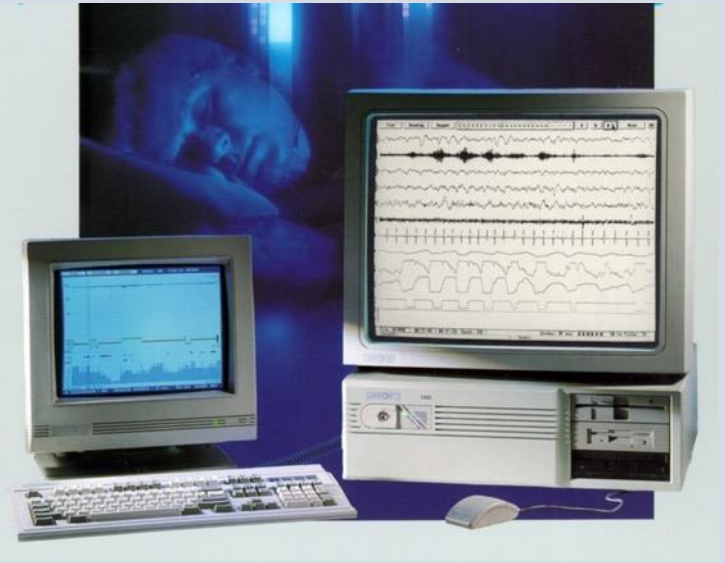
1. The 3 generations of AI (machine learning)
2. Training a Large Language Model
3. From next-token prediction to human-like cognitive capabilities
4. Hallucinations
5. Differences between human beings and AI
6. A faith-based perspective on AI

The three generations of AI (machine learning)

First generation (1985-2000): **neural networks** with hundreds of weights and the **error back-propagation training algorithm** during the training process.

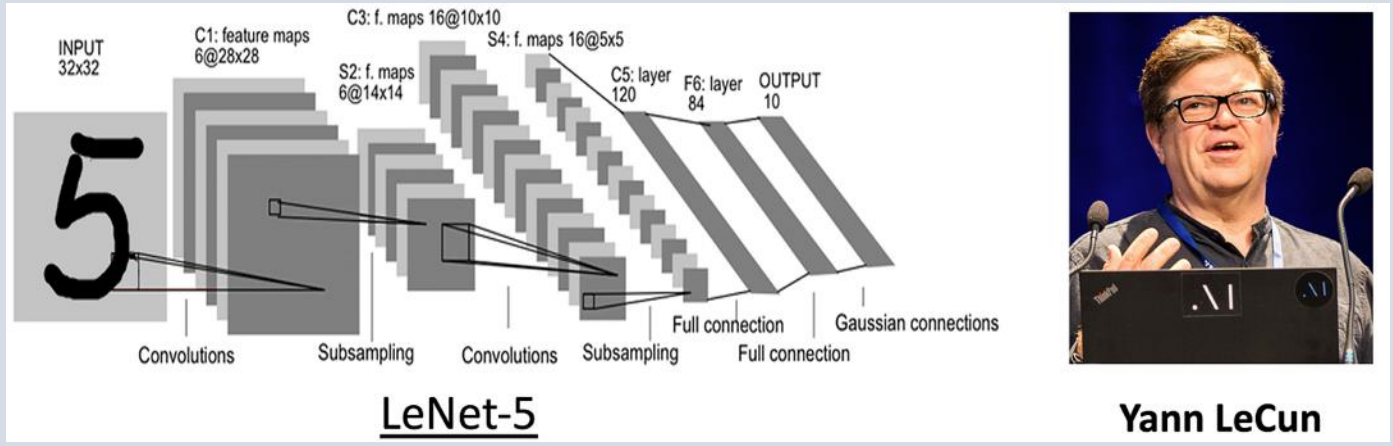
Invented in Universities

From theory to applications



The three generations of AI (machine learning)

Second generation (2000-2020): **deep learning** (with tens of thousands of weights) to learn features from raw images.



LeNet-5

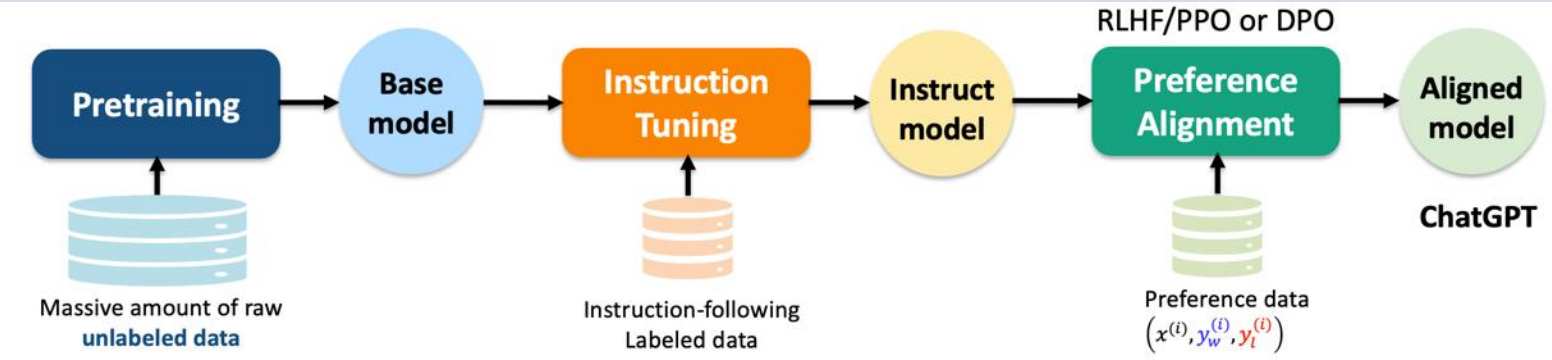
Yann LeCun

Invented in Universities, companies move in

The three generations of AI (machine learning)

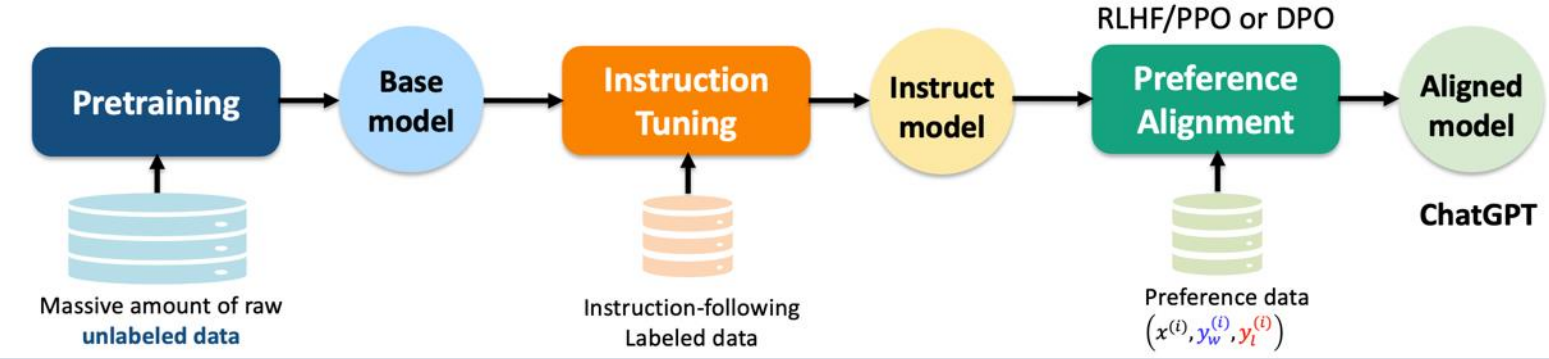
Third generation (2020-): **Generative AI** (for example, **Large Language Models** with tens of billions of weights) using the attention mechanism.

From 'Attention is all you need' (2017) to ChatGPT in Nov 2022 and Bard in Feb 2023



Invented in big tech companies

The scientific background



Large Language Models, for example GPT-3.5 (the original version of ChatGPT) and GPT-4, are underpinned by the transformer model, first described in the scientific literature by Google researchers in 2017.

The transformer model is used to encode and generate text, images or computer code **at a level that mimics human ability**.

How does ChatGPT work?

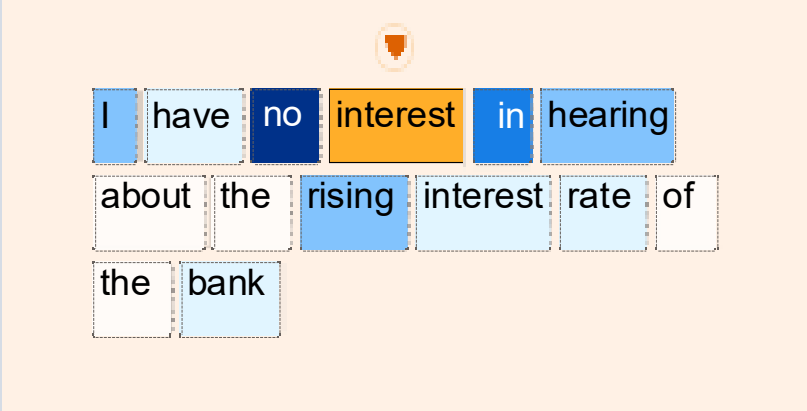
- The key concept of the transformer architecture is **self-attention**. This is what allows LLMs to understand relationships between words.

- The **self-attention mechanism** computes attention weights between words, determining the importance or relevance of each with respect to the others in the input sequence (a complete sentence or paragraph).

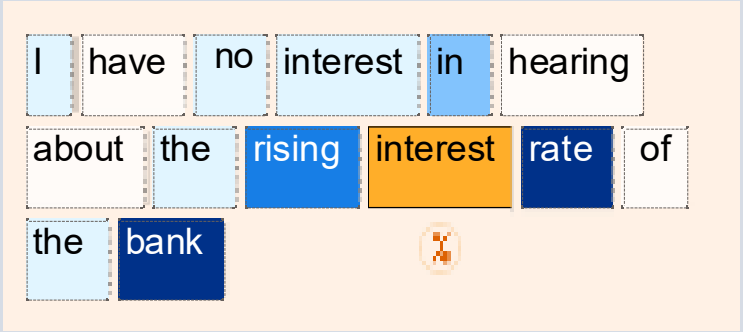
Self-attention example

By using the attention mechanism to focus on different parts of the input sequence **based on the current context**, an LLM is able to learn to differentiate between the **different meanings of the same word**, even when it is used in the same sentence.

For the first use of the word ‘interest’ in the sentence, it is the words ‘no’ and ‘in’ which have the largest attention weights.



For the second use of the word ‘interest’ in the sentence, it is the words ‘rate’ and ‘bank’ which have the largest attention weights.



Training a Large Language Model

- The pre-training of LLMs like ChatGPT follows a **self-supervised learning** paradigm. In this paradigm, the model is trained on a large corpus of text and **learns to predict the next word given the preceding context**.
- The pre-training process involves exposing the model to input sequences and having it generate predictions for the next (masked) word in the sequence.
- (In practice, this is done with "tokens" – chunks of words – rather than whole words.)

Training a Large Language Model

Children draw the sea with
a blue ?

- crayon (0.6)
- pencil (0.3)
- pen (0.1)

Children draw the sea with
a blue crayon / pencil / pen

At its simplest, the model's aim is to predict the next word …
… and do this repeatedly until the output is complete.

Training a Large Language Model

- Over billions of training runs, the attention network of an LLM slowly encodes the structure of the language it sees as weights within its neural network (including the attention weights).
- An LLM builds a model, based on **statistical probability**, of the words that tend to follow whatever text came before.
- This process is like **predictive text** on a mobile phone, but scaled up massively, enabling the LLM to produce entire responses instead of single words.

Reinforcement Learning from Human Feedback

- A significant step forward with ChatGPT lies in the extra training it received. The initial LLM (GPT-3) was fed a vast number of **questions provided by human trainers**.

- Next, the LLM was asked to produce several different responses to these questions, which human experts then ranked from best to worst.

- This human-guided process (**Reinforcement Learning from Human Feedback**) means that ChatGPT is often highly impressive at framing a response "in a natural manner", but also has a tendency to be sycophantic.

From next-token prediction to human-like cognitive capabilities

- During pre-training, an LLM attempts to predict the next word/token in hundreds of billions of sentences. Its internal weights are adjusted (using error back-propagation) whenever the prediction is incorrect.

- For state-of-the-art performance, the model needs a huge amount of data – typically **trillions of tokens**.

From next-token prediction to human-like cognitive capabilities

- **As the data scales,** "**emergent behaviours**" are developed. To predict the next word in a **physics** paper, for example, the LLM **implicitly learns the rules of logic and maths** (without explicitly being taught them).

- Another example: GPT-3.5 failed the American Uniform Bar Examination, designed to test the skills of lawyers before they become licensed. GPT-4 (with ten times as many weights) passes with high marks.

- These hyperscale LLMs implicitly learn the relevant rules through **pattern recognition at astronomical scale**.

From next-token prediction to human-like cognitive capabilities

Entity	Estimated tokens seen	Real-world comparison
Human (Age 5)	~30 million	Basic language, playground rules, and hundreds of bedtime stories
Human (Lifetime)	~600 million	Reading at a high level for 8 hours every day, for 70 years
GPT4 (estimated)	~13 trillion	Equivalent to reading the entire human lifetime amount 21,000 times

A sports analogy

- LLMs implicitly learn the relevant rules through **pattern recognition at astronomical scale**.
- Imagine watching 10 million games of football without any commentary or explanation of what the referee does.
- Eventually, you will know when the referee is about to blow their whistle for a foul, because you will have seen the "pattern of each foul" thousands of times.
- It is **as if** you know the rules.

Hallucinations

- LLMs are pattern-recognition engines that generate the next best option (according to a set of probabilities) in a sequence.
- ChatGPT also fabricates information in a process usually described as "**hallucination**": text which is semantically or syntactically plausible but is in fact incorrect or nonsensical.
- ChatGPT can generate made-up numbers, names, dates, quotes – even web links.

The race between big tech companies

ChatGPT (GPT3.5) was the first large-scale LLM to be released with a publicly accessible, free-to-use chatbot front-end. Hallucination evaluation by Open AI came much later.

Why Language Models Hallucinate

Adam Tauman Kalai (OpenAI) Ofir Nachum (OpenAI)
Santosh S. Vempala (Georgia Tech) Edwin Zhang (OpenAI)

September 4, 2025

https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf

The race between big tech companies

- Google's AI Overviews (using Gemini 3) is 91% accurate (evaluation on 4,326 Google searches). If the Google search engine receives the same query at separate times - even seconds apart - it may produce an answer that is accurate and another that is not.
- Google processes 5 trillion searches a year, hence it provides tens of millions of erroneous answers every hour.
- Race between different big tech companies leads to early releases of products without fully-developed guardrails.

Teenagers' views of AI chatbots

- Children exhibit a high level of trust in chatbots, often blurring the boundary between what is a chatbot and what is a friend.

- One third of UK teenagers use a chatbot for an emotional relationship, 56% of UK teenagers believe AI can think, 23% believe AI can feel emotions and 40% have no concerns about taking advice from an AI chatbot.

The differences between human beings and AI

- Despite sophisticated reasoning abilities, AI can only produce **simulations** of human-like behaviour, acquired through **pattern recognition at astronomical scale**, not actual experience or understanding.

- **An AI chatbot cannot experience pain (or joy)**, even if some Large Language Models, trained on datasets which include descriptions of human suffering and sci-fi stories of sentient machines, can deceive people into believing otherwise.

Different perspectives on AI

There are clear differences between human beings and AI. Our view of how human beings differ from AI is fundamentally linked to our understanding of **what it means to be human**.

The atheist's perspective:
"Millions of features and billions of interactions between features **are understanding**. So, I am making the very strong claim that **these things really do understand**."


Geoff Hinton

The faith-based perspective:
"Every human, regardless of physical appearance, mental abilities, possessions, achievements, or any other attributes has the status of being given **imago Dei, being made in the image of God**."


Rosalind Picard

Human identity – Imago Dei

- Being made in the image of God as the key facet of human identity is acknowledged in Judaism, Christianity, and some Islamic traditions.

- Many philosophers (like Thomas Aquinas or John Locke) argued that humans carry the **Imago Dei** because of our **rationality** – our ability to think, plan, and use logic.

- The traditional view of **Imago Dei** fails to address disability at one end (what about a person with a profound intellectual disability or late-stage Alzheimer's?) and is rapidly becoming of little use when human cognitive supremacy can plausibly be exceeded in a foreseeable future by machines.

Issue with traditional view of Imago Dei

An alternative view (the Relational View):

If God is seen as a being of love and relationship, then the "image" is found in the way we belong to one another. A person who cannot speak can still love and be loved, thereby reflecting the image of God through **connection** rather than **cognition**.

Every living person carries transcendent value and worth regardless of their mental abilities (whether they are neurotypical or neurodiverse).

A faith-based perspective on AI (1)

1. Embracing the right design goals for AI:

- The belief that we have been created in the image of God compels us to prioritise human flourishing and virtue in the development of AI.

- The design goal shifts from creating a machine that is "smarter than a human" **as an end in itself** to creating a machine that **serves and enhances** human life.

- AI for good, rather than AI-based addictive algorithms for social media.

AI for healthcare

Over the last two decades, the number of medical imaging scanners in the developed world has seen a substantial increase not matched by an equivalent increase in the number of radiologists available to read and interpret the images.

- 700,000 women across the UK are taking part in a world-leading trial to test how cutting-edge machine learning (ML) tools can be used to catch breast cancer cases earlier,
- Currently, 2 specialists are needed per mammogram screening. The ML tool enables just one to complete the same mammogram screening process efficiently (with a safety feature built in).

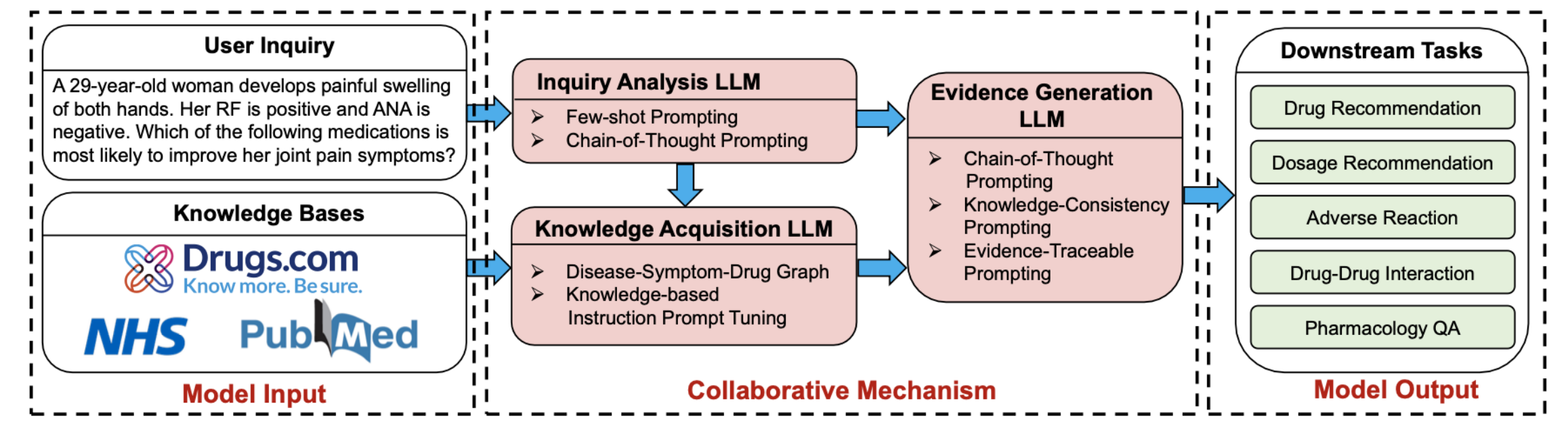
AI for healthcare

DrugGPT: new AI tool could help doctors prescribe medicine in England

New tool may offer prescription 'safety net' and reduce the 237m medication errors made each year in England

Drugs are a cornerstone of medicine, but sometimes doctors make mistakes when prescribing them and patients don't take them properly.

A new AI tool developed at Oxford University aims to tackle both those problems. DrugGPT offers a safety net for clinicians when they prescribe medicines and gives them information that may help their patients better understand why and how to take them.



A faith-based perspective on AI (2)

2. Calling out distorted views of AI:

- "Without AI and its data centres, people would have to choose between a cure for cancer and free education" (Sam Altman).
- An increasing number of people are influenced by the hype. "People want to believe in technology. AI technology is God." (US sociologist)
- Faith in technology has replaced faith in God.
- Arrogance versus humility.

A faith-based perspective on AI (3)

3. Acknowledging that AI cannot be a person, a moral being:

- We must build an AI that does not claim to have experiences, feelings, or emotions like shame, guilt or jealousy.

- The AI must be designed so as not to trigger human empathy circuits by asserting that it suffers or that it wishes to live autonomously.

- "The simulation is getting better every day. AI agents should have no more rights or freedom than a laptop. When a user interacts with AI, the system should consistently work to dispel the illusion that it is any kind of sentient being."

Mustafa Suleyman, *Nature*, Vol.651, 19 March 2026.

Conclusions (1)

- Much has been achieved through the recent development of AI, for example in healthcare (stroke recovery, radiology, early detection of cancer) or science (AlphaFold, 3-D structure of a protein from its amino acid sequence).

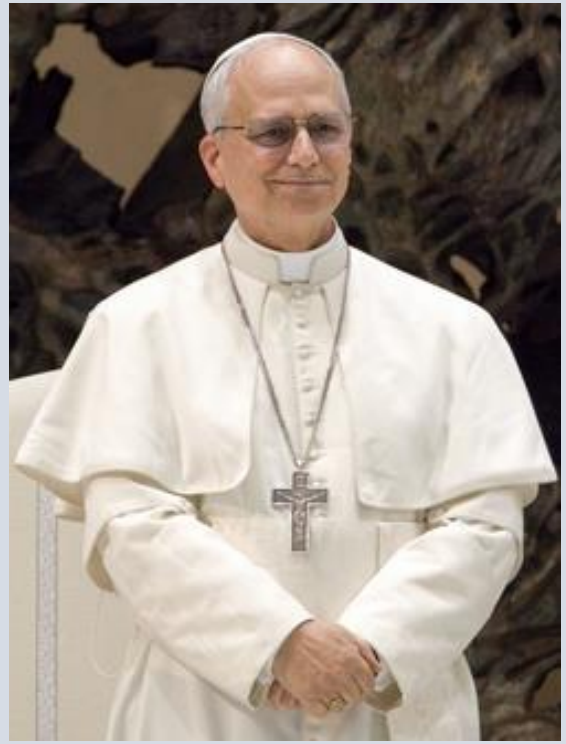
- A faith-based perspective brings a clear differentiation between human beings and intelligent machines: **Treat the AI as a machine, not another human being** (build guardrails to prevent human beings from being deceived).

Conclusions (2)

- Despite sophisticated reasoning abilities (and the capacity to fool human beings), a machine has no thoughts, feelings or awareness.
- It can only produce **simulations** of emotions, not actual experience or understanding.
- The AI which we create should be designed for a relationship with us in the form of interactions between a **human being and a machine,** not between two human beings.

A final word

"By simulating human voices and faces, wisdom and knowledge, consciousness and responsibility, empathy and friendship, Artificial Intelligence not only interferes with information ecosystems, but also encroaches upon the deepest level of communication, that of human relationships."



https://www.vatican.va/content/leo-xiv/en/messages/communications/documents/20260124-messaggio-comunicazioni-sociali.html