



AI Narratives

Rich Sutton

University of Alberta

Alberta Machine Intelligence Institute

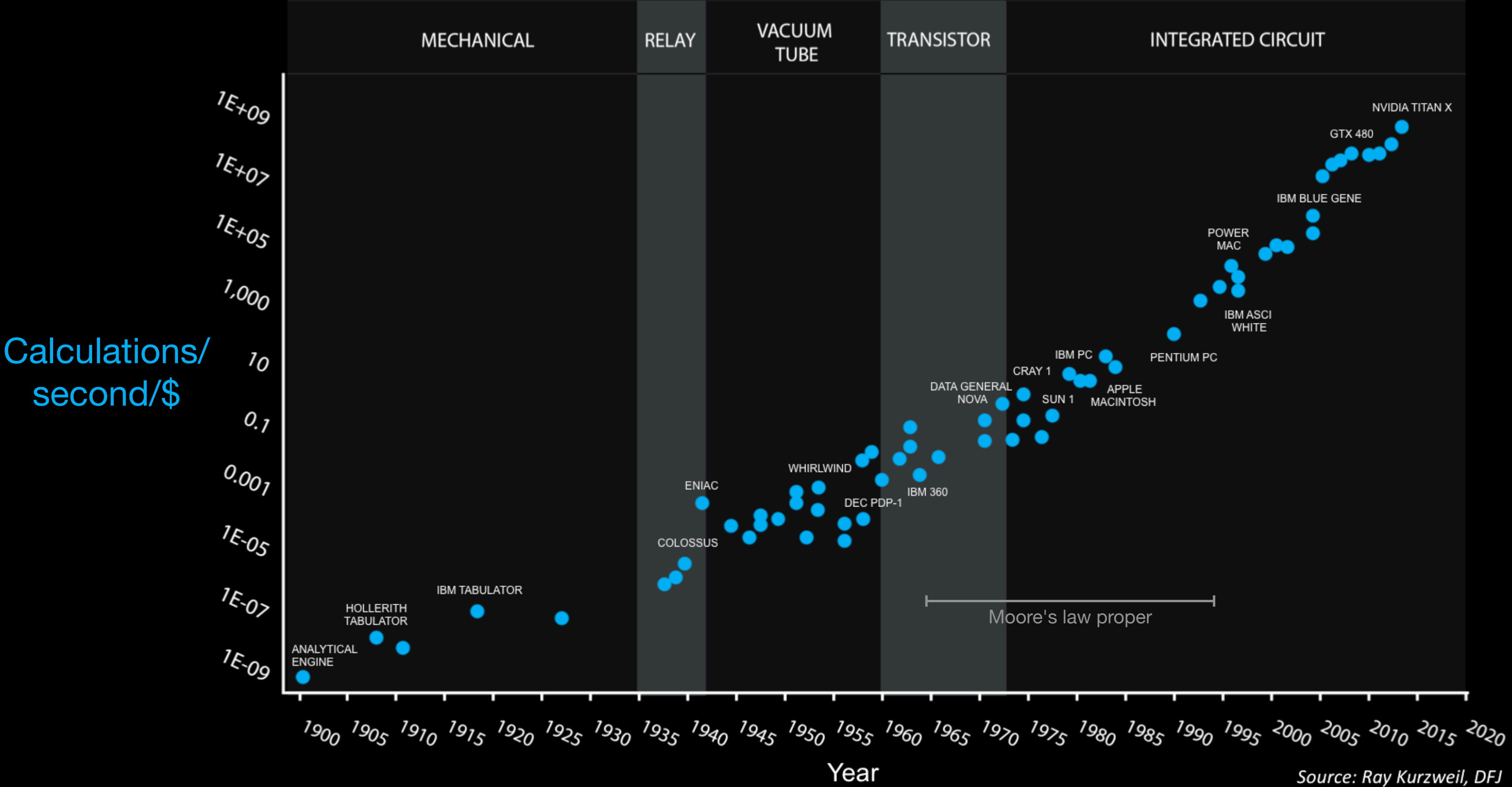
Reinforcement Learning and Artificial Intelligence Lab



Upperbound 2023



"Moore's" Law: Computation/\$ increases exponentially for nine+ decades



Doubling roughly every 2 years

The cumulative effect is tremendous!

Anything using computation becomes 10 times more valuable every 5 years

Creates a powerful persistent pressure for computer automation

This is the story of our time

The Age of Computation!

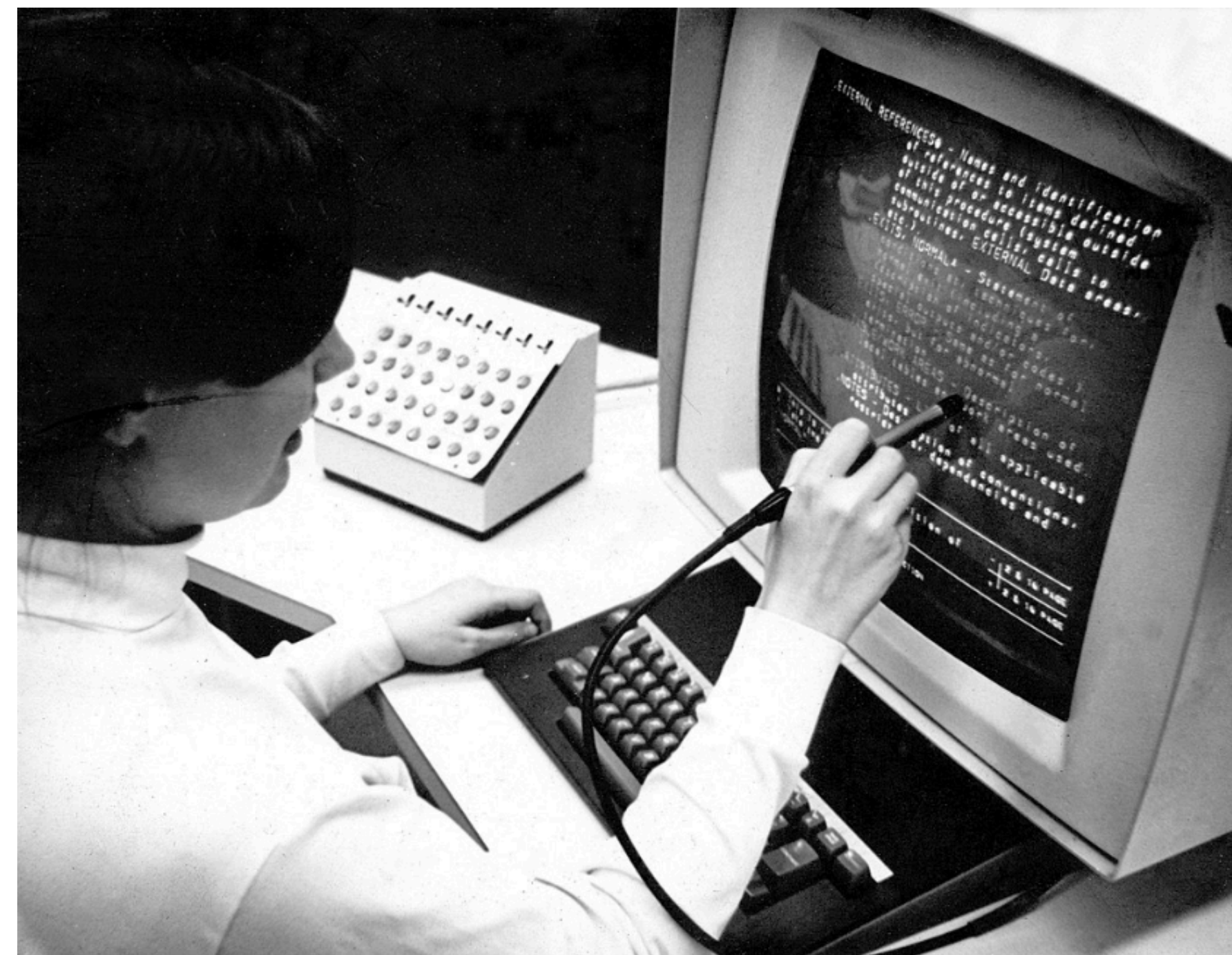
"AI" has cashet

Anything new and computational is called AI



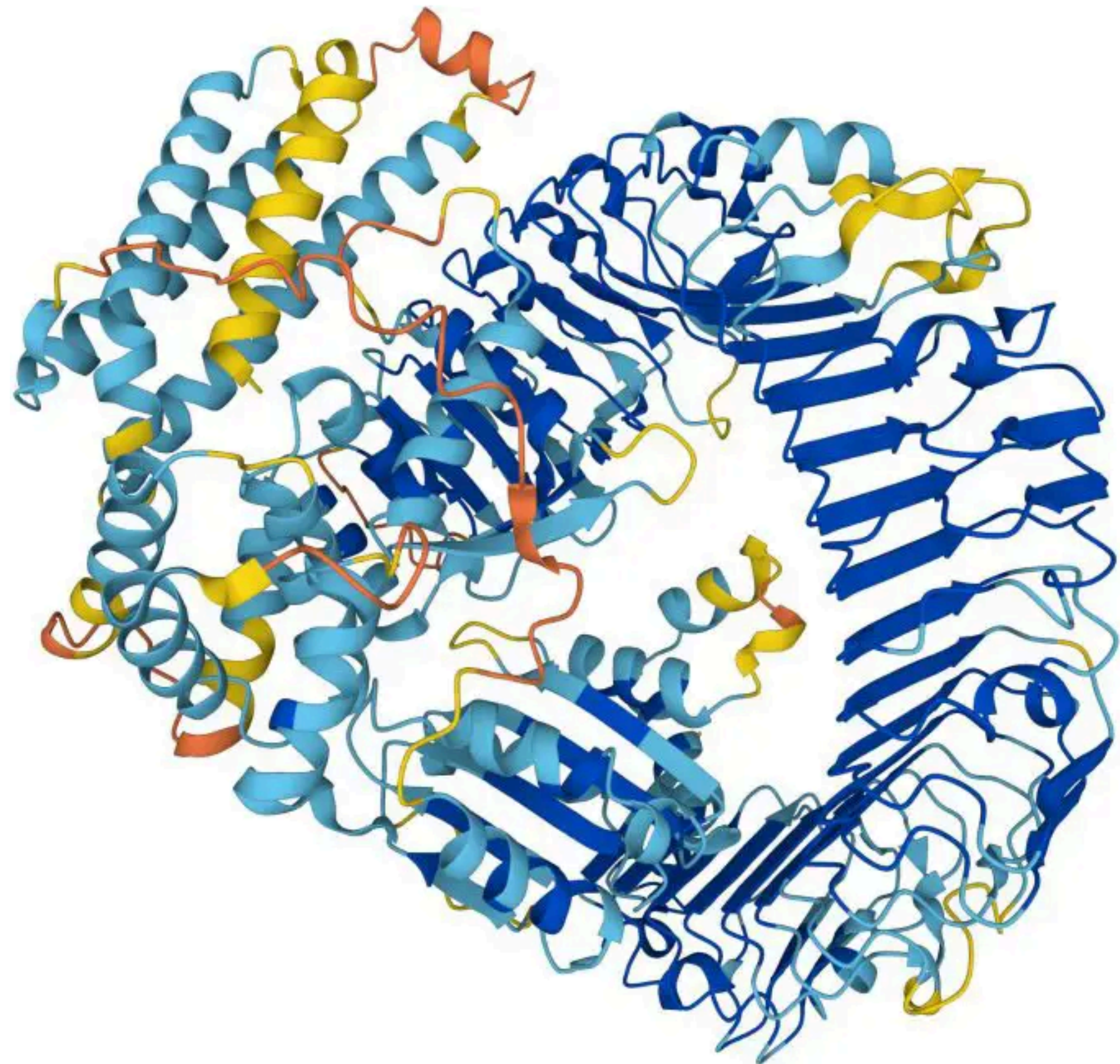
"Intelligent" terminals [\[edit \]](#)

An "intelligent" terminal^[23] does its own processing, usually implying a microprocessor is built in, but not all terminals with microprocessors did any real processing of input: the main computer to which it was attached would have to respond quickly to each keystroke. The term "intelligent" in this context dates from 1969.^[24]



"AI" has cashet

Anything new and computational might be called AI



AlphaFold predicts the 3D structure of proteins from their amino acid sequences

Sequenced all 200M known proteins

Released open source to the world, and transformational for modern biological research

Used machine learning, deep learning

But not like natural intelligence in any way

Isn't it a misnomer to call it AI?

Let's distinguish two kinds of AI

- **Tool AI**
 - uses lots of computation to do something hard or amazing
 - but must be wielded by a person, who provides the goal
- **Agent AI**
 - creating new people-like agents
 - with autonomy and goals, experience, a model of the world
 - this is fully ambitious AI -- "the prize"

Tool AI

AlphaFold

Large Language Models

DALL-E

Midjourney

No goals or interactive
learning experience

Agent AI

AlphaGo, AlphaZero, MuZero (games)

Deep Q-Network (Atari)

Gran Turismo Sophy (racecar)

even simple Dyna (gridworld)

Clear goals, observations,
actions, agency, interaction

Aspires to "the Prize"

Tool AI is less threatening, but has its dangers

- Fake news, deep fakes, election manipulation, bias, loss of jobs, plagiarism
- Bad people might use LLMs to take over the world, but the people would be doing it, not the LLMs
- The problems of Tool AI are real, but manageable, like the earlier
 - Loss of farming jobs in the industrial revolution
 - Loss of trust in photographs after photoshop
- The good will likely outweigh the bad, as it always has in the past
- Tool AI is scary, but not very scary

Can Tool AI lead to Agent AI?



“To think this all began with letting autocomplete finish our sentences.”

Large language models exemplify "the Eliza effect"

ELIZA effect

🌐 10 languages ▾

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

The **ELIZA effect**, in [computer science](#), is the tendency to unconsciously assume computer behaviors are analogous to human behaviors; that is, [anthropomorphisation](#).

Overview [\[edit \]](#)

In its specific form the ELIZA effect refers only to "the susceptibility of people to read far more understanding than is warranted into strings of symbols—especially words—strung together by computers".^[1] A trivial example of the specific form of the Eliza effect, given by [Douglas Hofstadter](#), involves an [automated teller machine](#) which displays the words "THANK YOU" at the end of a transaction. A naive observer might think that the machine is actually expressing gratitude; however, the machine is only printing a preprogrammed string of symbols.^[1]

More generally, the ELIZA effect describes any situation^{[2][3]} where, based solely on a system's output, users perceive computer systems as having "intrinsic qualities and abilities which the software controlling the (output) cannot possibly achieve"^[4] or "assume that [outputs] reflect a greater causality than they actually do".^[5] In both its specific and general

Large language models exemplify "the Eliza effect"

ELIZA

Article [Talk](#)

From Wikipedia

The **ELIZA effect** is a human behavior

Overview

In its specific form, the Eliza effect is a warranted interpretation of the Eliza effect, often referred to as "the Eliza effect" at the end of the machine is more generally a computer system that achieves a general

Welcome to

```
EEEEEE LL IIII ZZZZZZ AAAAA  
EE LL II ZZ AA AA  
EEEEEE LL II ZZZ AAAAAAA  
EE LL II ZZ AA AA  
EEEEEE LLLLLL IIII ZZZZZZ AA AA
```

Eliza is a mock Rogerian psychotherapist. The original program was described by Joseph Weizenbaum in 1966. This implementation by Norbert Landsteiner 2005.

```
ELIZA: Is something troubling you ?  
YOU: Men are all alike.  
ELIZA: What is the connection, do you suppose ?  
YOU: They're always bugging us about something or other.  
ELIZA: Can you think of a specific example ?  
YOU: Well, my boyfriend made me come here.  
ELIZA: Is it important to you that your boyfriend made you come here ?  
YOU: He says I'm depressed much of the time.  
ELIZA: I am sorry to hear that you are depressed.  
YOU: It's true. I am unhappy.  
ELIZA: Can you explain what made you unhappy ?  
YOU:
```

languages ▼

Tools ▼

ous to

is

specific form

THANK

however,

ceive

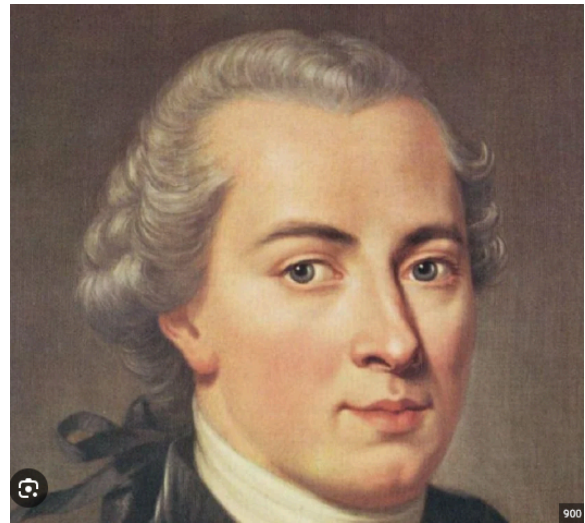
ssibly

general

Philosophy of mind



John Locke wrote “An Essay Concerning Human Understanding”



Emmanuel Kant wrote “The Critique of Pure Reason”



Rene Descartes said “i think, therefore i am”

Scientists and non-scientists have been fascinated by their inner workings

Gustav Fechner

Jean Piaget

Hermann Ebbinghaus

Sigmund Freud

Ivan Pavlov

Carl Jung

Edward Thorndike

Timothy Leary

B. F. Skinner

Ray Kurzweil

Edward Tolman

“Intelligence is the most powerful phenomenon in the universe.”

–Ray Kurzweil

Could Agent AI be such a powerful phenomenon?

Yes

Could Tool AI be such a powerful phenomenon?

No

The case for ambition

- Within your lifetime, **AI researchers will understand the principles of intelligence**— what it is and how it works—well enough to create beings of far greater intelligence
Agent AI
- **This will change everything.** The way we work and play. Our senses of identity. The goals we set for ourselves and our societies.
- It will be the **greatest intellectual achievement of all time**
 - a fundamental goal of science, engineering, and the humanities
 - whose significance is beyond humanity, beyond life, beyond good and bad
- I call it **the Prize**. The Prize is a **great and glorious goal!**
- Ambitious AI researchers should keep their **Eyes on the Prize** Agent AI

But will understanding intelligence be good?

- Some will see it as bad, others as good
- Understanding intelligence cannot be successfully opposed; it is too economically valuable; the game theory is too strong
- Like all great changes, good or bad doesn't matter: **we don't get to choose**
 - Compare with Darwin's understanding of evolution, with people descendant from animals
 - Or to fish, contemplating the changes if they become land animals

Is creating smarter people (Agent AI)
a grand and glorious prize, a natural next step in the human quest,
or a nightmare bringing the end of all we know and love?

- Would it be good or bad?
This is an incredibly subjective question, too raw to answer today
- My view is that Agent AI is grand and glorious
But am I blinded by lust for "the Prize"?
- Or are the fear mongers just fearing change,
fearing a loss of control that they don't really have,
or talking themselves into a messianic excuse to do harm
 - Those who think they are saving the world may do evil things

The fear-monger narrative may be winning

- Their's is the standard narrative.
- Consider that the field of AI safety tries to solve "the control problem" of how we can keep AIs of superior intelligence under our human control
 - And does not ask if our more-intelligent AIs would then be our slaves, and us slave holders
 - And does not ask if people would use their slaves against each other
- Consider that AI safety tries to solve "the alignment problem" of aligning superior AI's goals with peoples' values
 - And does not ask which peoples' values?
 - And does not ask how such controls would be enforced?
 - And does not ask how peoples' values would then be able to evolve?

The fear may come from viewing AI as the other tribe

- Many times in humankind's history different people have come in contact
 - Sometimes peacefully to trade and inter-marry
 - Sometimes violently to dominate, kill, or enslave
- This is in our genetic history, both as dominator and dominated
- This is still part of our attitudes toward other peoples and other countries
 - Though we try to tolerate and value what is different in others
 - Particularly in Canada we strive to be welcoming and not fearful
- The AI fear-monger narrative builds our instinctual fear of the other

There's also a hopeful narrative for the coming of AIs

- Where the AIs are not alien things to be controlled
 - but are our allies and offspring, of us rather than against us
- Where we don't try to control tightly out of fear
 - but rather appreciate the AI's differences and insights they bring
 - including the possibility that they might show us paths to improving our human values

In the hopeful narrative

- We don't ask, "how can we control the AIs' goals?",
 - but rather "how can we work together, even though we have different goals"
- Notice that this is what people already do
 - What you want and I want are not the same
 - We have different stomachs, different families, and different bank accounts
 - Our goals are not the same, but *symmetric*
- Laws and mores have been established so that it is most rewarding to keep our promises, respect property, and not harm each other
- In RL language, we all have different *rewards*, but the world is such that our most important *values* are aligned

One AI or many? How likely is a singleton?

- In the fearful narrative, there is one super-intelligence that gains control and then suppresses all the others (a singleton); then things are bad and irrecoverable
- Whereas, in the hopeful narrative, there are always many diverse AIs so that it is better to work with others than to try to dominate them all
- The fearful narrative is based on a fast takeoff, sometimes called "FOOM"
 - If a super-AI arises very fast, then there might be a moment when it could take over
 - But if AIs improve only at the rate of Moore's law (doubling every 18 months) then one dominating all others everywhere is very unlikely
- Is the world very susceptible to being taken over?
- Or is it a complex adaptive system, decentralized and diverse, and therefore resistant to global capture?

Peoples attitudes to AI's arrival won't change easily

- Perhaps it is enough just to question the fearful narrative
 - to realize that there are real alternatives
 - that control and enslavement are not the only possible path
 - that they have their own risks and distasteful morality
- Perhaps it is enough to realize that the fearful narrative may be leading us to a bad place
- I urge you, do not rush to support the fearful narrative

To summarize,

The stories of AI

- It is the **Century of Computation**, of Moore's law driving everything
- It is helpful to separate two kinds of AI
 - **Amazing Tool AI**
 - a clear Risk/Reward profile
 - a familiar story of technological disruption
 - **Fully-ambitious Agent-AI**
 - a historic human quest (the Prize)
 - strange unsettling narratives, **Fearful** and **Hopeful**
- Which Agent-AI narratives wins will determine the final **Story of AI**

One more thing

Open Mind Research

- A non-profit organization focused on executing the Alberta Plan for AI research
- A distributed network of research fellows across Canada, and potentially the world
 - but centered in a meaningful sense in Edmonton, where I will remain

Open Mind Research

- A non-profit organization focused on executing the Alberta Plan for AI research
- A distributed network of research fellows across Canada, and potentially the world
 - but centered in a meaningful sense in Edmonton, where I will remain
- With Joseph Modayil and Melanie Marvin, both formerly of DeepMind Alberta
- All research product will be published in the open scientific literature,
 - No intellectual property or other equity will be retained
- Our key adjectives are open, focused, and lean
- We are currently seeking two more donors, at least one from Alberta



Thank you for your attention

Thank you for your attention