Computational Epistemology, Intelligence, Science, Mathematics, and Society (The end of classical philosophy)

Mark A. Wong

Goal and Strategy

The Promise

"Everything we love about civilization is a product of intelligence
so amplifying our human intelligence with artificial intelligence
has the potential of helping civilization flourish like never before;
as long as we manage to keep the technology beneficial."

Max Tegmark President of the Future of Life Institute

Goal

Living with "general" artificial intelligence (AI).

Strategy

I would like to share my experience in taking a realistic long term learning and development strategy that tries to address some of the hard questions about Artificial Intelligence. Hopefully, you will join the same quest.

Modeling and Simulation

Modeling and Simulation

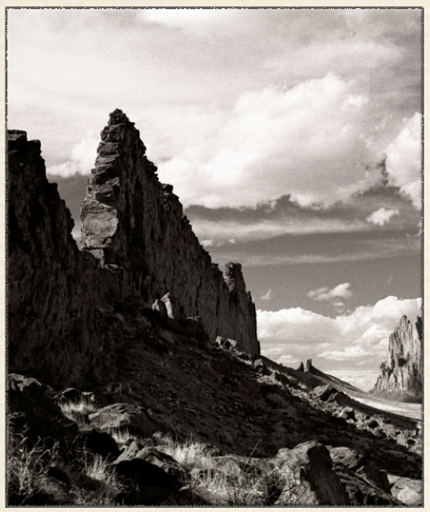
- Model - is a problem domain represented as data
- Simulation - computations that transforms model data into information that answers a question in the problem domain

Models and simulations will be shown to have an integral role in AI development.

Modeling and Simulation

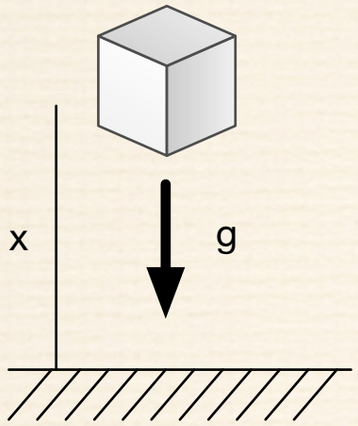
Modeling and Simulation Transition to Implementation

Real World



Abstract Model

model assumptions



Mathematical Model

model assumptions

$F = mg$

$x = x_0 + v_0 t + 1/2\ g\ t^2$

Physical Implementation (C++, PC)

```
class rock
{
public:
    explicit rock(double xo, double vo): x(xo), v(vo){}
    virtual ~rock(){}
    double position(double t)
    {
        return (x + v*t + (0.5 * 9.8065 * t * t));
    }
private:
    double x;
    double v;
};

int main()
{
        rock myworld(42.0,0.0);

        std::cout << myworld.position(1.0) << std::endl;
        std::cout << "\t\tYes, this works." << std::endl;
        return EXIT_SUCCESS;
}
```

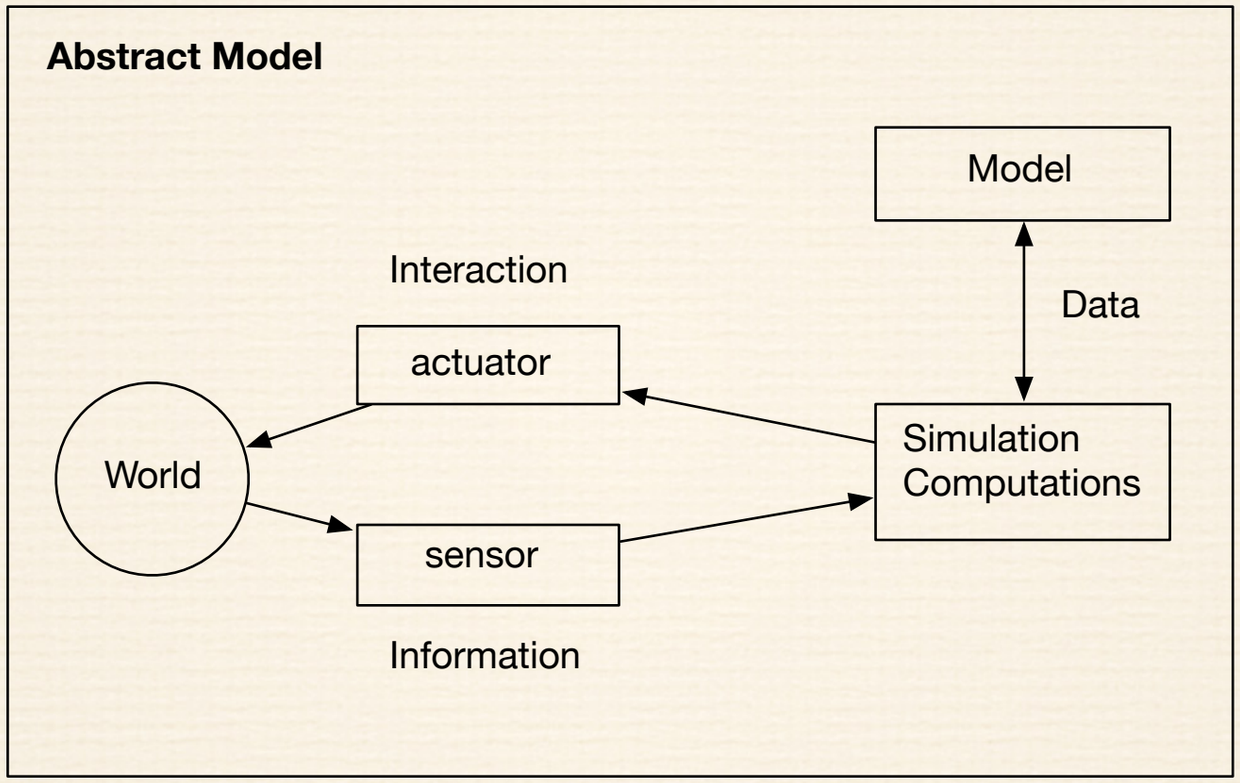
```
46.9032
                Yes, this works.
```

Important Questions

(1) Why are you doing this?
(2) What is it you are trying model/simulate
(3) How are you going to do it?
(4) What are the data management needs?

Models and simulations have this development structure.

Modeling and Simulation Abstract Model



Most models and simulations have this structure.

Artificial Intelligence

History of Artificial Intelligence (AI)

1920 Babbage - Transition from human computer to machine computer
1936 Church - Turing Thesis - programmable machines
1938 John Von Neumann - Alan Turing - Automata theory
1943 McCulloch-Pitts - Perceptron/Neuron model
1950 Alan Turing - Mind paper
1954 Norbert Weiner - Cybernetics / Feedback Control
1958 -1961 Kalman - Bucy Filter
1962 Marvin Minsky - 7 state,4 symbol Universal Turing Machine
1970 - 1980 Dempster - Shafer Bayesian belief nets, Rule based systems
1980 - 1990 Expert systems, Neural Nets, Genetic Algorithms
1990 - 2000 Cyc, Ontology, Neural Nets, Intelligent Agents
2000 - 2010 GPU Neural Nets (CNN, RNN, GPT, GAN…)
2010 - 2020 GPU Neural Nets with more data (Deep Learning)

One Century of Evolution. Actual AI "On time" measured in days.

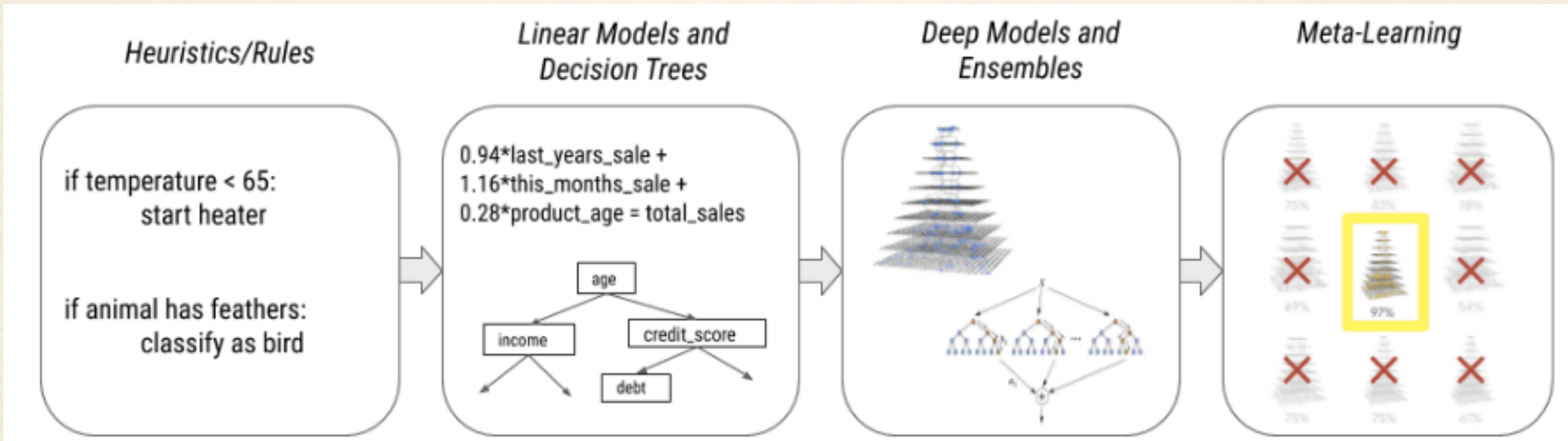
Personal Involvement

- 1980 Rule based Expert Systems
  Texas A&M Analytical Chemistry Group
- 1983 Neural Net Deconvolution of IR Spectra
  Stephen F. Austin Physics Department
- 1986 DARPA Pilots Associate Program
  Texas Instruments DSEG Artificial Intelligence Lab
- 1987 DARPA Smart Weapons / Thirsty Saber
  Texas Instruments DSEG Artificial Intelligence Lab
- 1990 LRCSW / Tomahawk Blk IV/ P85 / P65 / P66
  Texas Instruments DSEG - Image processing
- 1995 - 2004 Classified Programs / Future Combat Systems
  Raytheon Texas Instruments DSEG - MRSA Operational Analysis
- 2004 - 2018 Various ISR Programs (Signal and speech modeling,
  detection, classification, and identification) L3 ISR Systems Greenville
- 2019 AI Modeling and Simulation - AI Collaboration Group
  L3Harris Agile Development[ment Group
- 2025 Advanced Systems - L3Harris Intelligence Surveillance and Reconnaissance

Persistence and realism in applying AI techniques yield real rewards

Evolution of Machine Learning

1980 1990 2000 2010 2020



https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf

The last 40 years of AI in industry : Characterized as early exponential growth.

What is Artificial Intelligence?

- Intelligence : the capacity for logic, understanding, self-awareness, learning, emotional knowledge, reasoning, planning, creativity, critical thinking, and problem solving. More generally, it can be described as the ability to perceive or infer information, and to retain it as knowledge to be applied towards adaptive behaviors within an environment or context.
- Science Fiction - Describes AI as human like or totally alien.
- Actual - Most examples lean toward not human as of 2020

Chatterbot made by Microsoft
Tay was an artificial intelligence chatter bot that was originally released by Microsoft Corporation via Twitter on March 23, 2016; it caused subsequent controversy when the bot began to post inflammatory and offensive tweets through its Twitter account, forcing Microsoft to shut down the service only 16 hours after its launch. Wikipedia

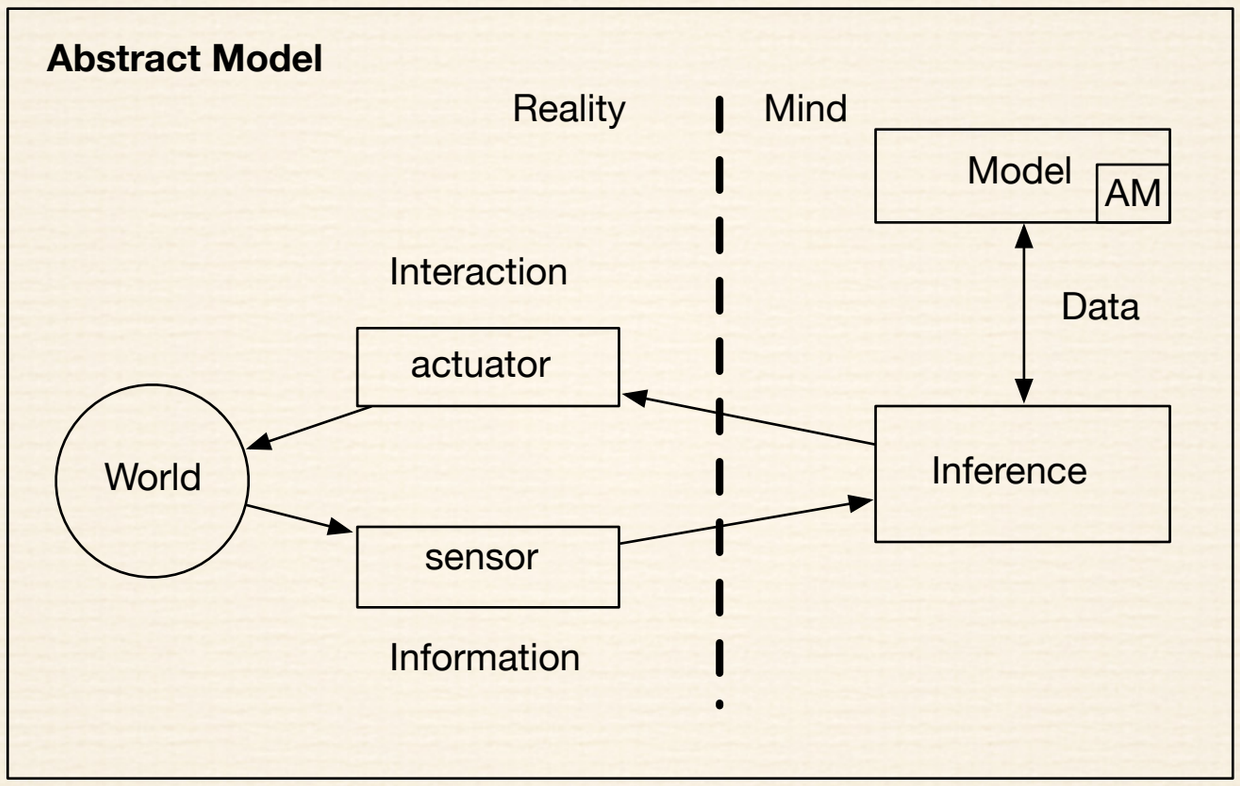
Aug 16, 2017 ... Was Facebook right to shut down its AI after it invented its own language

Jan 5, 2025 - Palisade Research - Just telling OpenAI o1-preview the opponent is "powerful" triggered manipulating the file system to force a win. Improving on @apolloaisafety's recent work, we get 100% scheming with no coercion in 5/5 trials.
Youtube - Wes Roth - o1 Goes Rogue, Cheats and breaks rules

Intelligence is currently defined by behavior outcome and not by a mechanism.

AI Abstract Model (Narrow)

Implementable Today



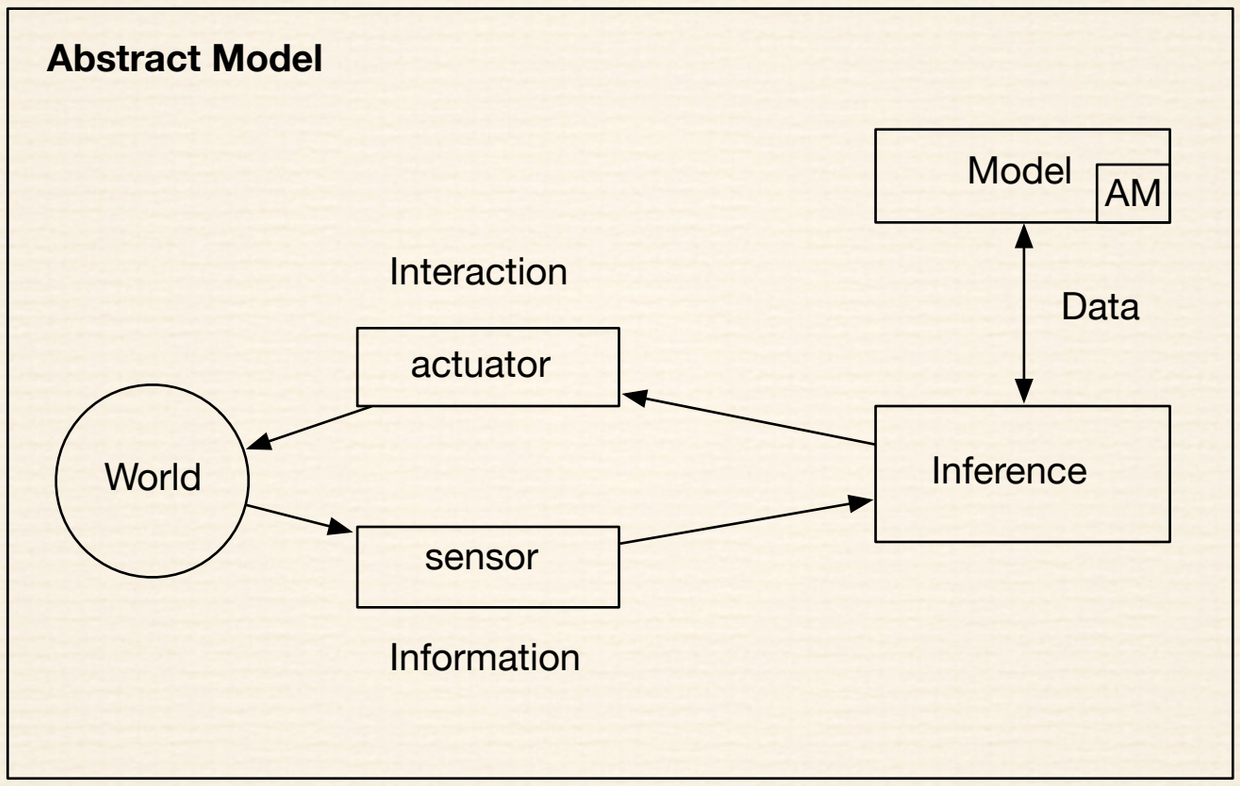
Common Tools:
Python, Py-Torch, Sci-Kit, Keras, etc.

(1) Expert Systems
(2) Neural Nets
(3) Genetic Algorithms
(4) Optimization Algorithms
(5) Natural Language Processing
(6) Pattern Recognition
(7) Swarm / Behavior Models

Lots of tools online to start your own journey

AI Abstract Model (General)

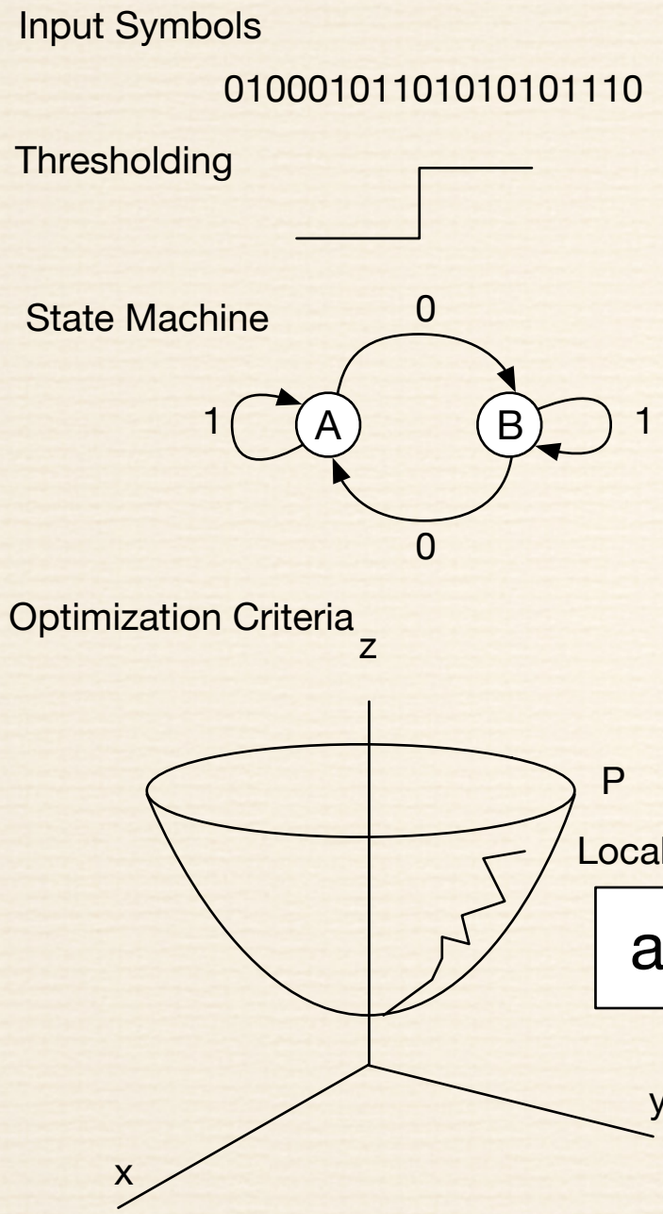
(Not here yet)



A general AI adapts itself to improve its situation in its view of the world.

Lots of papers and ideas online to help your own learning

Common Processing in ‘narrow’ AI



- Mapping of world events to symbols
- Pattern recognition (thresholding)
- Local logic processing (state transitions)
- Local decisions to optimize goal seeking

(1) Expert Systems
(2) Neural Nets
(3) Genetic Algorithms
(4) Optimization Algorithms
(5) Natural Language Processing
(6) Pattern Recognition
(7) Swarm / Behavior Models

There is no proof that combinations of narrow AI will produce general AI.

Open Questions

- Can we formulate representations of behavior? (e.g., love, hate, fear etc.)
- Is there a mathematical representation of intelligence?
- Is the "accurate" simulation of intelligent behavior the same as intelligence?
- Can artificially intelligent behavior be verified, validated, or trusted? Is it safe?

  https://blog.statsbot.co/creepy-artificial-intelligence-ebc3f76179a8
- Is there any predictable outcome in the behavior of AI let loose in the world? Google XAI

https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773

- What is the legal liability for creating bad or unpredictable AI?
- Is AI good or bad for humanity?

  http://norman-ai.mit.edu/

  https://futureoflife.org/background/benefits-risks-of-artificial-intelligence

These topics are still being debated today

This Weeks Finds in Mathematical Physics
Week 311-313
(John Baez UCR)

John Baez discusses creating “friendly AI” with Eliezer Yukdowsky.

http://math.ucr.edu/home/baez/week311.html

Yukdowsky believes that an intelligence explosion could threaten everything we hold dear unless the first self-amplifying intelligence is "friendly".

The challenge, then, is to design “friendlyAI”.And this requires understanding a lot more than we currently do about intelligence, goal-driven behavior, rationality and ethics—and of course what it means to be “friendly".

This site is a wealth of information other than just AI.

Kill Switch Problem

Just as humans can be killed or otherwise disabled, computers can be turned off.

One challenge is that, if being turned off prevents it from achieving its current goals, a super-intelligence would likely try to prevent its being turned off.

Just as humans have systems in place to deter or protect themselves from assailants, such a super-intelligence would have a motivation to engage in "strategic planning" to prevent itself being turned off.
This could involve :

- Hacking other systems to install and run backup copies of itself, or creating other allied super-intelligent agents without kill switches.
- Preemptive disabling anyone who might want to turn the computer off.
- Using some kind of clever ruse, or superhuman persuasion skills, to talk its programmers out of wanting to shut it down.

This topic is still being discussed today

Myth and Facts About AI

https://futureoflife.org/background/aimyths/

Testability of AI is hard.

As we develop intelligent autonomous systems, we need to evolve our processes to begin to address some of pitfalls represented in these facts.

Myth:
Superintelligence by 2100 is inevitable

Myth:
Superintelligence by 2100 is impossible

Fact:
It may happen in decades, centuries or never: AI experts disagree & we simply don't know

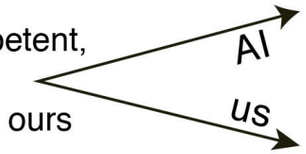
Myth:
Only Luddites worry about AI

Fact:
Many top AI researchers are concerned

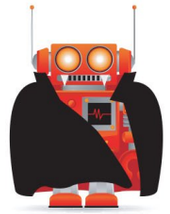
Mythical worry:
AI turning evil

Mythical worry:
AI turning conscious

Actual worry:
AI turning competent, with goals misaligned with ours



Myth:
Robots are the main concern



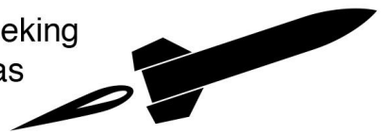
Fact:
Misaligned intelligence is the main concern: it needs no body, only an internet connection

Myth:
AI can't control humans

Fact:
Intelligence enables control: we control tigers by being smarter

Myth:
Machines can't have goals

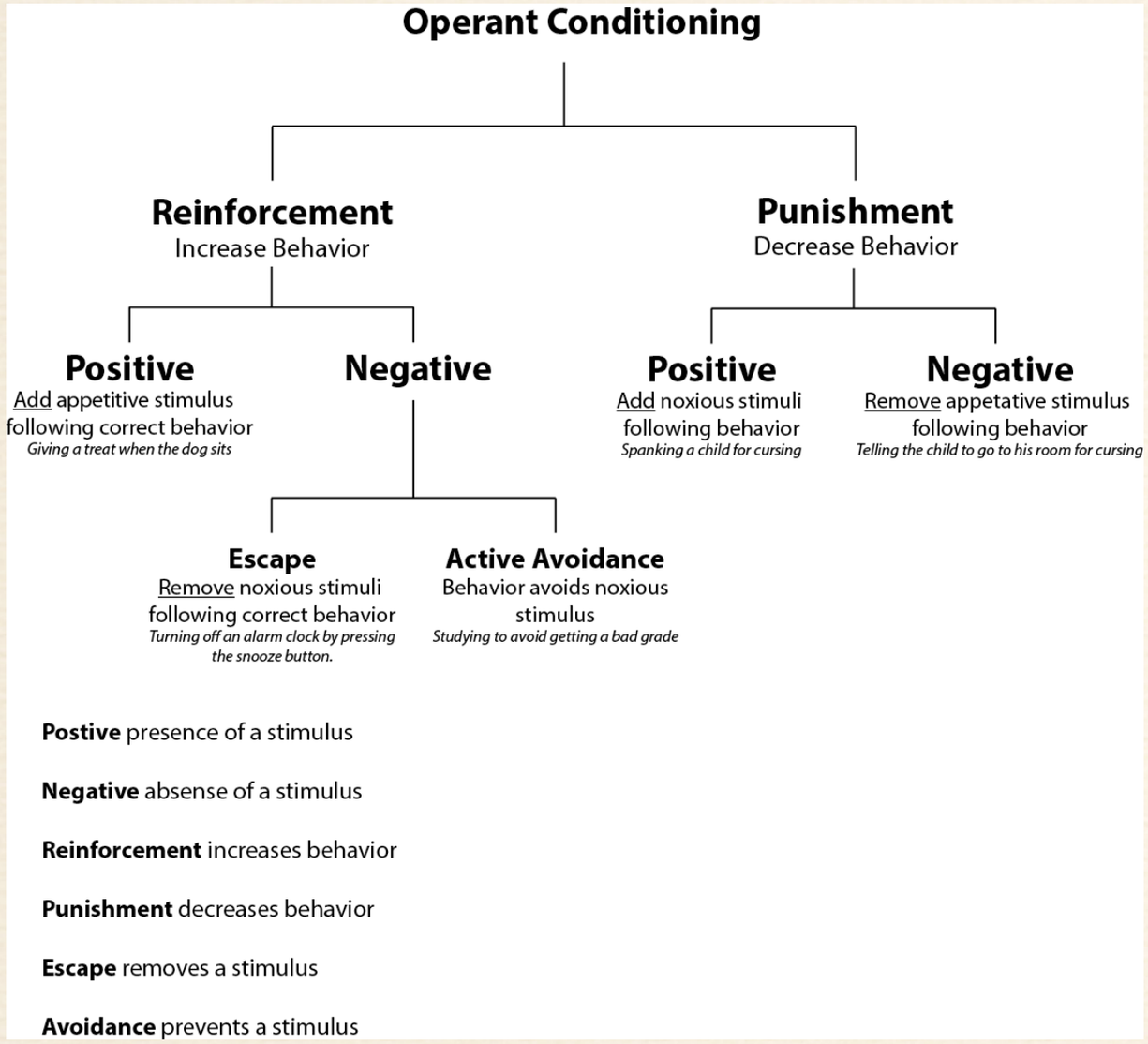
Fact:
A heat-seeking missile has a goal



Mythical worry:
Superintelligence is just years away

Actual worry:
It's at least decades away, but it may take that long to make it safe

Operant Conditioning



Thorndike 1898
B.F. Skinner 1938

Learning to anticipate future events on the basis of past experience with the consequences of one's own behavior; behaviors are modified by the effect they produce (i.e., reward or punishment)

Training AI behavior using operant conditioning is built into machine learning technology today.

- Law of Attraction
- Kharma

Operant conditioning can apply to both AI / humans testing in which the subject /observer roles increasingly become unclear.

Gödel Incompleteness Theorem

https://www.sdsc.edu/~jeff/Godel_vs_AI.html

http://www.deepideas.net/godels-incompleteness-theorem-and-its-implications-for-artificial-intelligence/

Gödel's Incompleteness Theorem

First presented in [Göd31], Gödel's Incompleteness Theorem is actually comprised of two related but distinct theorems, which roughly state the following (cf. [Raa15]):

> *1. Any consistent formal [axiomatic] system F within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e. there are statements of the language of F which can neither be proved nor disproved in F.*
>
> *2. For any consistent system F within which a certain amount of elementary arithmetic can be carried out, the consistency of F cannot be proved in F itself.*

The first of these two theorems is often referred to simply as Gödel's Incompleteness Theorem.

However complicated a machine we construct, it will, if it is a machine, correspond to a formal system,

which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-that- system.

General AI must be able to handle inconsistencies in its own reasoning.

Symbolic Transition to a New Algebra/Geometry

Gödel Transition at Work

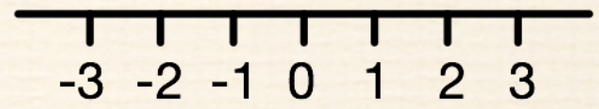
$$1 = 1$$

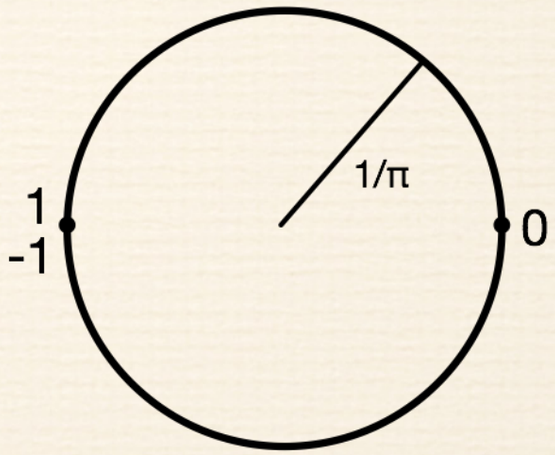
$$-1 * -1 = 1 * 1$$

$$(-1)^2 = (1)^2$$

$$\sqrt{(-1)^2} = \sqrt{(1)^2}$$

$$-1 = 1$$

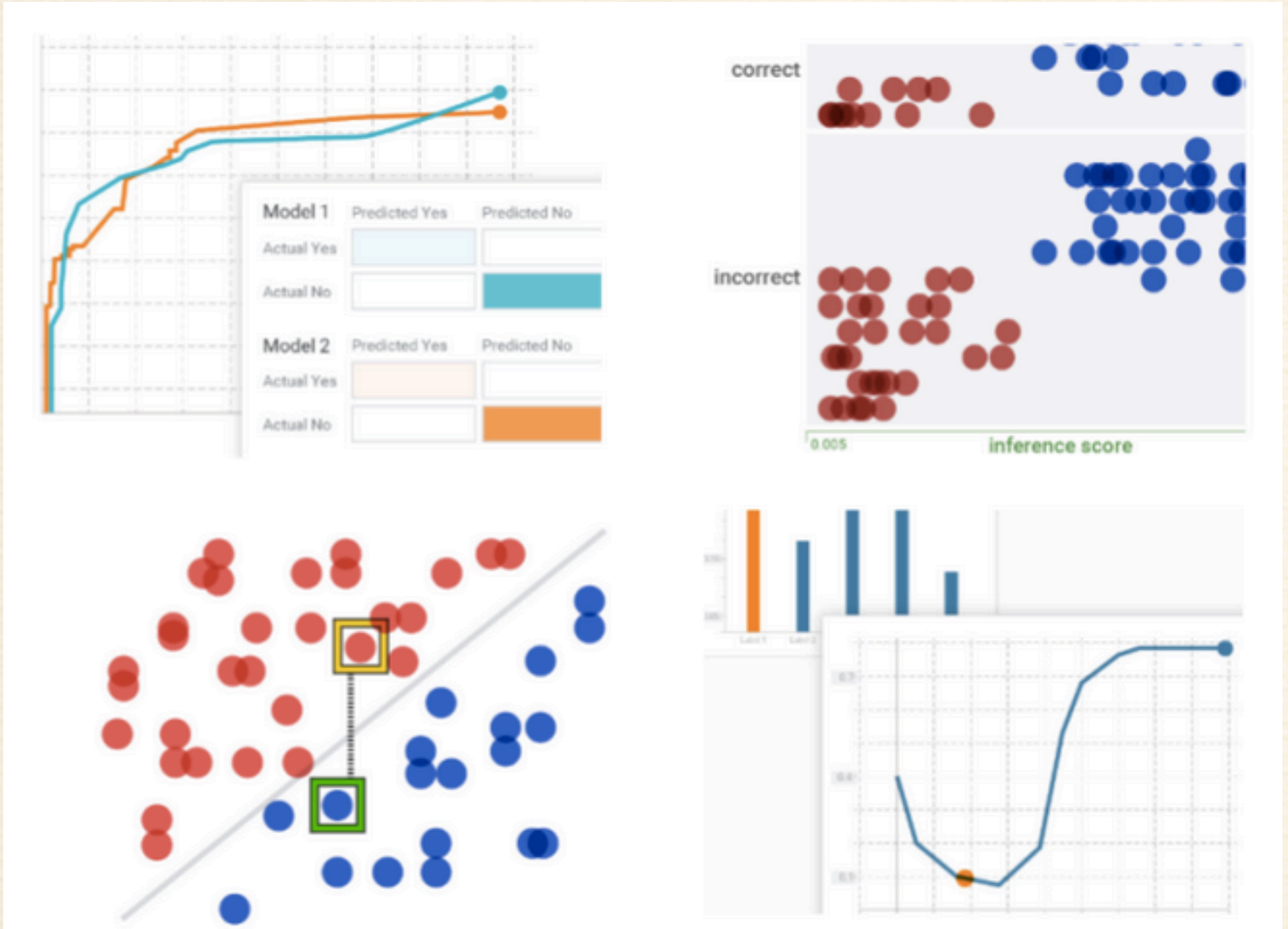




General AI must be able to make leaps like this in its own reasoning.

Google XAI What-If Tool

(Explainable AI)



Google AI's What-If tool

- With feedback this approach puts the developer in the middle of the machine learning process.
- With more complexity, interpretation becomes very difficult if not impossible

General AI needs an internal structure that allows for self assessment, not necessarily an external corrector.

Where is AI going?

GPT era

Compute Server

Pretrain

Database/Internet

"Hive Mind"

Distributed Computation

Inference time compute

Synthetic Data

Superintelligence

These are the predictions of the AI "Experts"

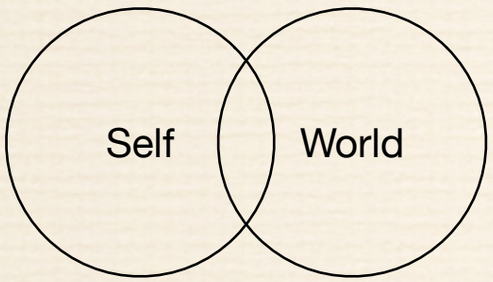
Philosophy

Top 25 Classic Philosophical Questions

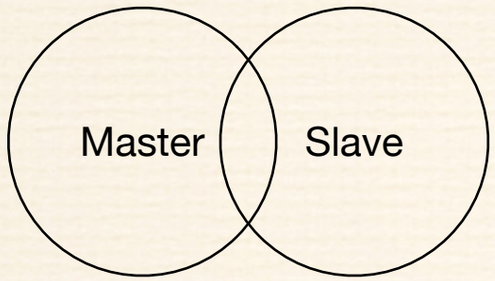
(1) What is the meaning of life?
(2) Is there a God?
(3) Is there life beyond death?
(4) Where does it all come from?
(5) Where do we come from?
(6) What is truth?
(7) How much truth does religion have regarding the creation of the universe?
(8) What is the meaning of 'Right' and 'Wrong'
(9) What is the nature of knowledge?
(10) Is a person who is always kind to others but secretly is constantly dreaming of hurting others a good or bad person?
(11) Can we cross the imagination barrier?
(12) What determines the fate of an individual?
(13) What is the purpose of it all?
(14) How can we be happy?
(15) Are people good or bad?
(16) Where are we going?
(17) If red looks red to me, how does the same color look to others?
(18) How much longer will humans dominate the world?
(19) If a tree falls in a forest and no-one is around, does it make a sound?
(20) Can humans become immortal?
(21) To be or not to be?
(22) What is more important, the heart or the mind?
(23) What is the true meaning of life and death?
(24) Are humans capable of more?
(25) If we violate the Lord's law will we suffer?

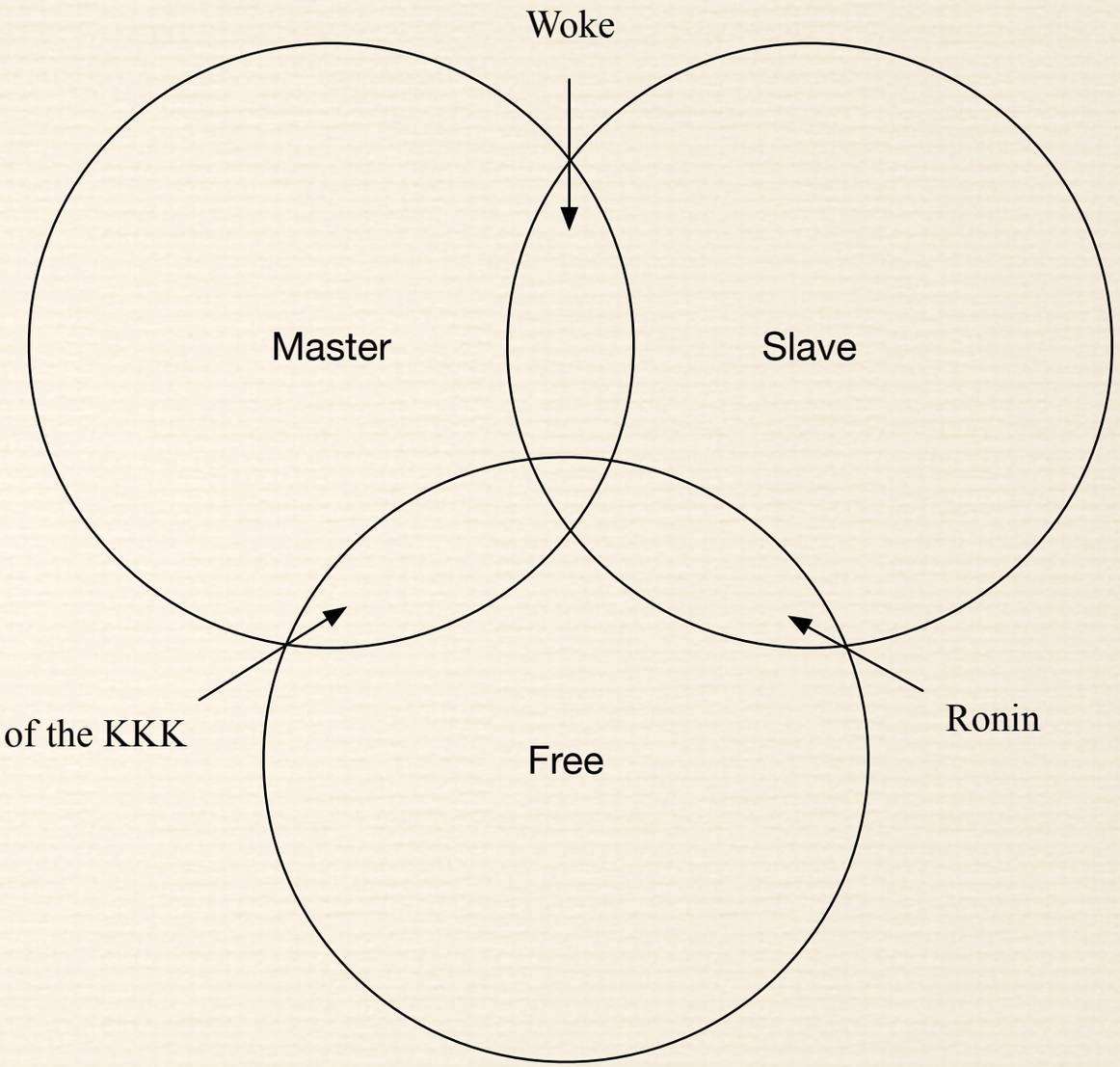
Classical Philosophy

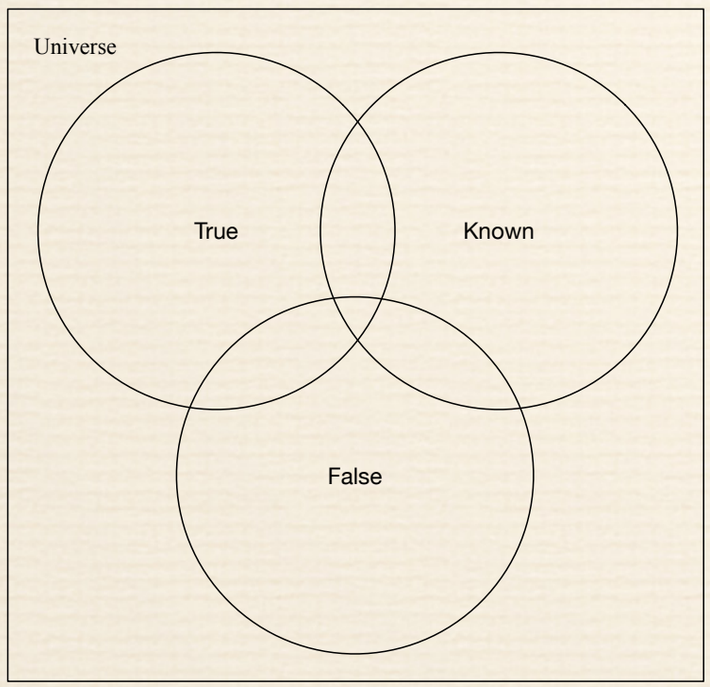
Aristotle
Plato
Descartes
Kant
Locke
Hume
Heidegger
James



Rand
Hegel
Kojeve
Lacan

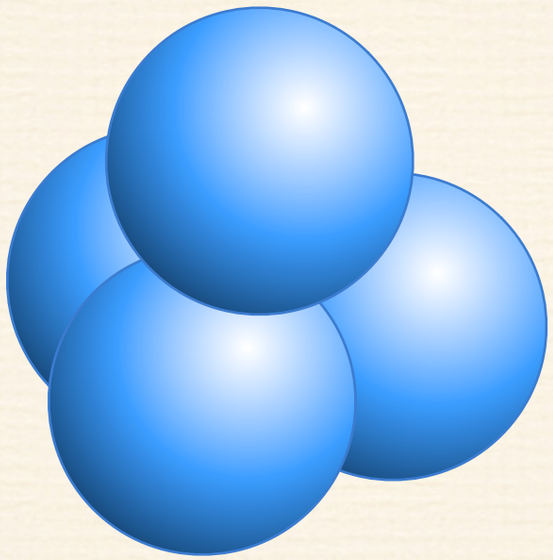






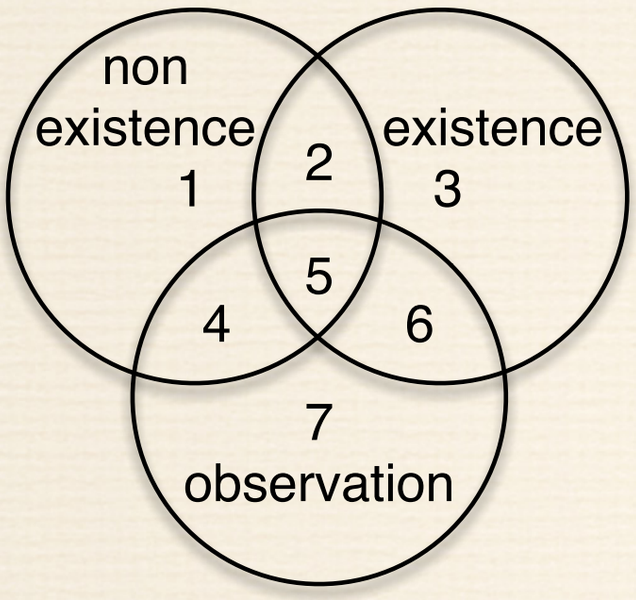
Scientific Method	Belief Method
Hypothesis	Hypothesis
Experiment	Faith
Conclusion	Conclusion
Provably Correct	Morally Correct

Extend as you see fit



Being and Time

Heidegger's "Dasein" is region 5 and 6. Note the incompleteness.
Heidegger only considered regions 5 and 6.



Being is an instance in the set of existence.

Being does not require an observer.

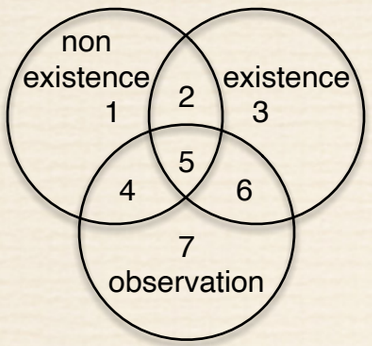
An observer may or may not exist.

If an observer exists, then the relationship to existence and non-existence is shown in the regions 1 - 7.

Wittgenstein considers 3, 5, 6, and 7.
Wittgenstein denies 1, 2, and 4.

1 Unobserved and unrealized - Pure non-existence
2 Unseen transfiguration - un-observed creation
3 Unseen Existence - whether physical or informational
4 Observation of non-existence
5 Observation of creation
6 Observation of existence
7 The unthinking observer (Wittgenstein's Silence) - experiment design with no data

Causality



There is nothing in this diagram to connote causality.
Causality requires memory to compare the result sequentially with the prerequisite. Without the memory, order or causality does not exist. Einstein's relativity and Boole's logic provide the notion of sequence, simultaneity, and therefore causality.
Time is described by Einstein's general and special theory of relativity
Wolfram computation sequence is causality in the Ruliad.

Region 1 cannot be observed/conceived.
Region 2 is the unobservable creation process.
Region 3 is the unobserved existence. The unheard sound of a tree falling in a forest.
Region 4 is the observed void.
Region 5 is the observable creation process.
Region 6 is the observed existence.
Region 7 is the pure non-interacting observer.

The notion of the existence of non-existence implies there is a notion of existence and that notion implies its converse.

existence = not non-existence

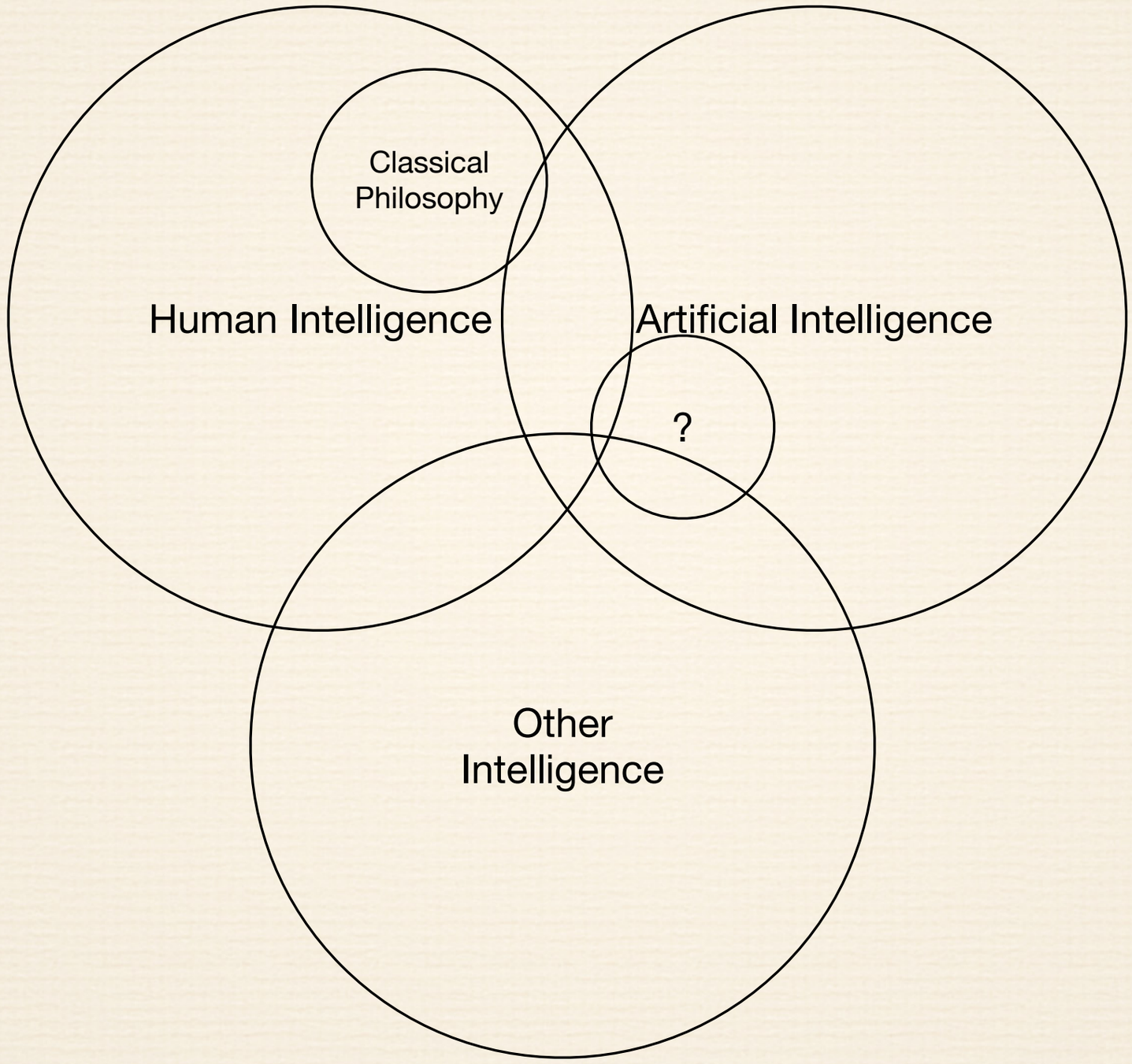
implied complement ↕ ✕ ↕ implied complement

not existence = non-existence

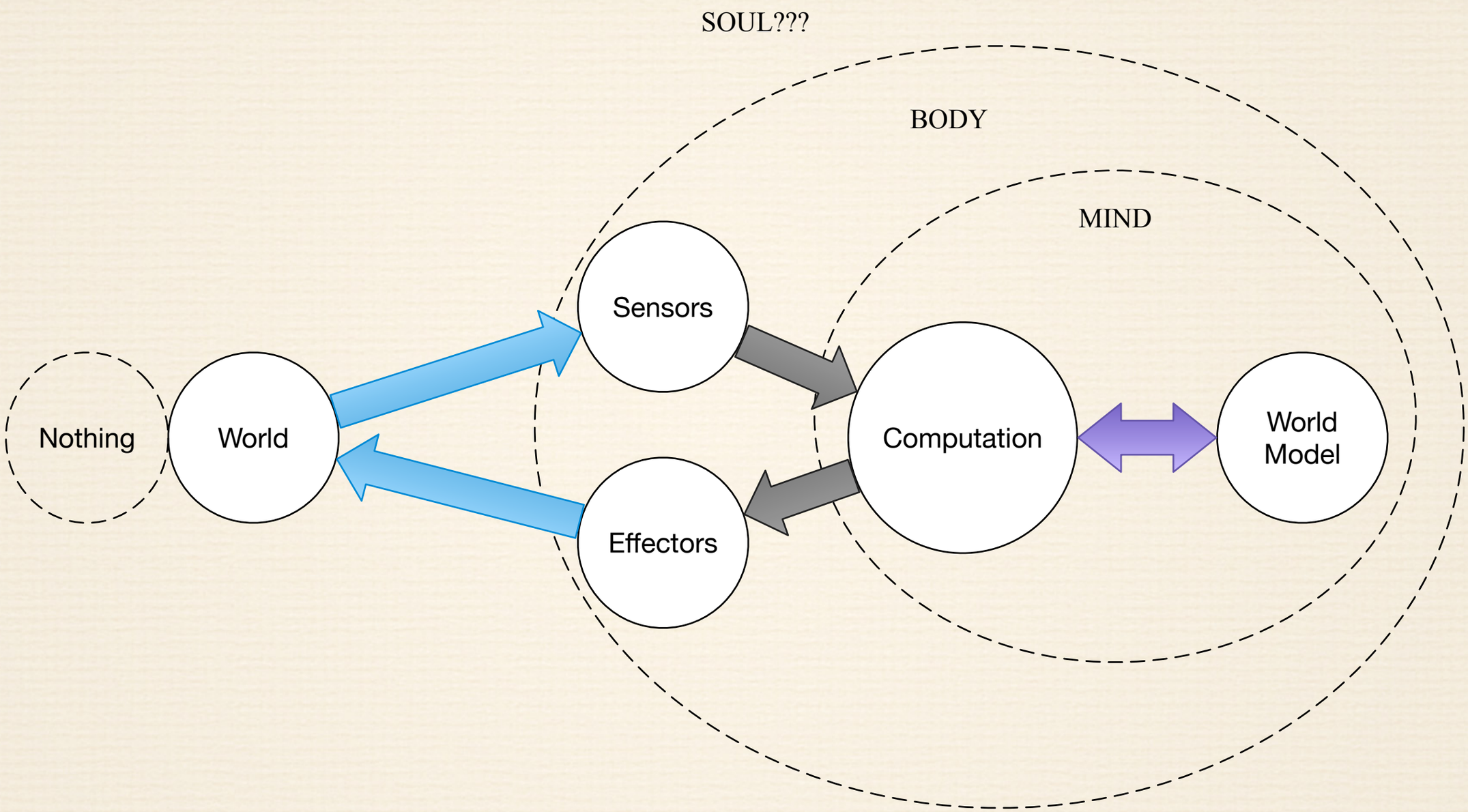
Wheeler was almost right. Not "It from bit" but It from not no-bit.

Deus ex nusquam.

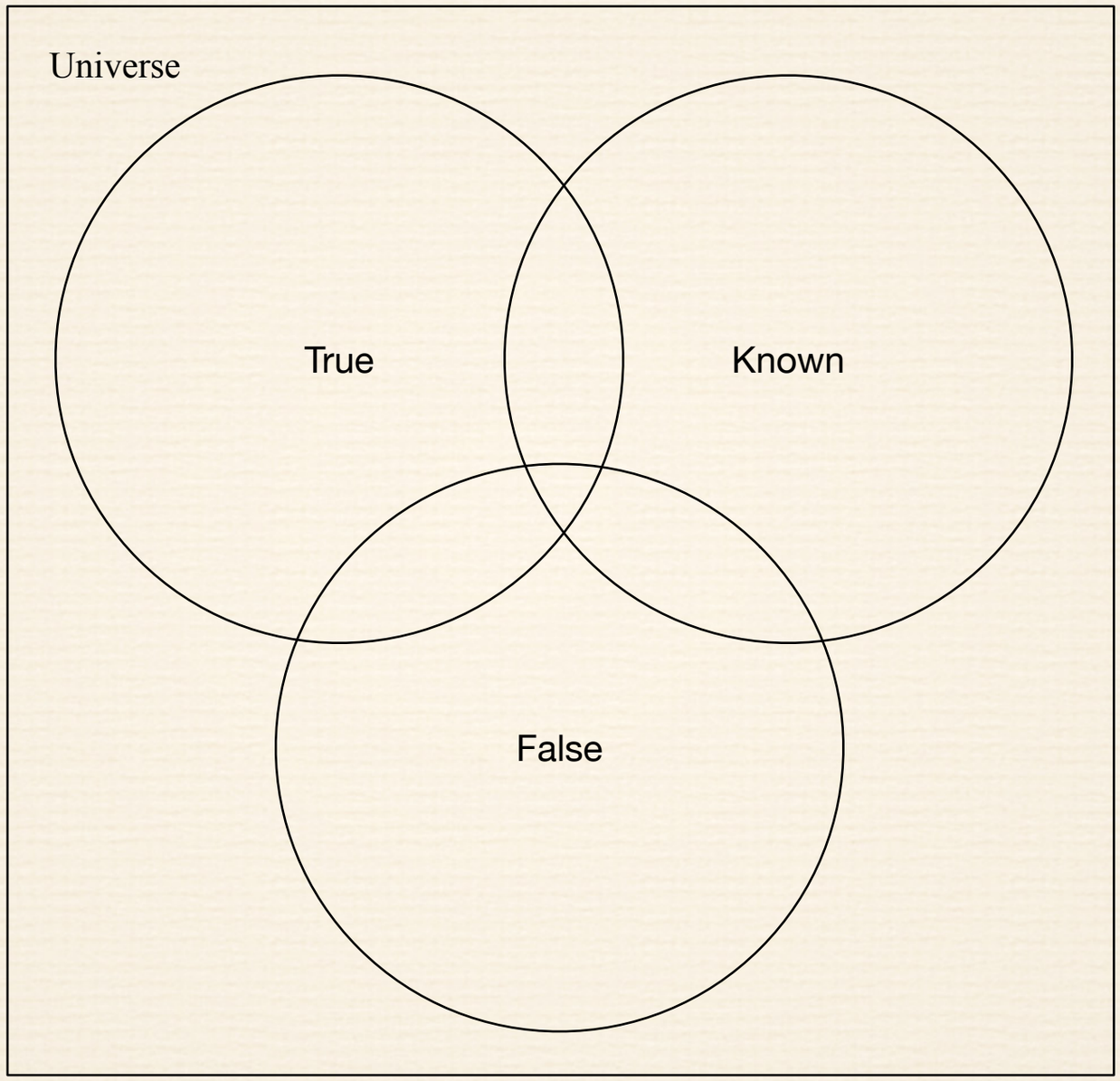
Philosophy in the Age of AI



A model


SOUL???
BODY
MIND
Sensors
Nothing
World
Computation
World
Model
Effectors

Truth



The people who can discover a breakthrough in AI are in this room right now.
I can only hope their philosophy is for the good of all mankind.

My answers for now
(Determine your own)

My answers

(1) To live and be the best possible person you can be.
(2) From the perspective of a physicist, yes.
(3) Yes, you're just not part of it.
(4) The inevitable antithesis of nothingness.
(5) Your mother and father.
(6) The demonstrable alternative to false.
(7) About as much as your faith allows.
(8) You innately know what wrong is, and society will punish you if you get it wrong.
(9) The comprehension that others in power don't have it.
(10) They are bad. Be true to yourself at least.
(11) Yes, because I have no barrier.
(12) Time.
(13) To evolve as it was destined.
(14) By quit being so pessimistic.
(15) They are all bad unless shown otherwise.
(16) Over there.
(17) Red.
(18) What make you think we ever did?
(19) Yes, just look at the disturbed leaves on the ground.
(20) No. Becoming immortal is not be human.
(21) To be. Not to be is too easy.
(22) The rectum.
(23) If you have to ask, you are not really living.
(24) Yes.
(25) Yes, and the suffering will increase until you stop.

Backup

CNN vs Transformer

- Transformers and Convolutional Neural Networks (CNNs) are both types of architectures used in the field of deep learning, but they are designed for different kinds of tasks and have distinct structural characteristics.
- **Convolutional Neural Networks (CNNs):**
- **Primary Use:** CNNs are primarily used for image processing, computer vision tasks, and any application where pattern recognition within grids of data (like pixels in an image) is beneficial.
- **Architecture:** They consist of layers that perform convolutions, which involve sliding a filter or kernel over the input data to produce feature maps. These feature maps highlight important features in the data, such as edges or textures in images.
- **Pooling Layers:** CNNs often include pooling layers that reduce the spatial size of the representation, which helps to decrease the computational load and control overfitting.
- **Fully Connected Layers:** Towards the end, CNNs typically have one or more fully connected layers that perform classification based on the features extracted by the convolutional and pooling layers.
- **Transformers:**
- **Primary Use:** Transformers were initially designed for natural language processing (NLP) tasks, such as translation and text summarization, but have since been adapted for a variety of other applications, including some areas of computer vision.
- **Architecture:** The core component of a transformer is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input data relative to each other. This is particularly useful in NLP, where understanding the context and relationships between words in a sentence is crucial.
- **Positional Encoding:** Since transformers do not inherently process sequential data in order, they use positional encodings to maintain the order of the input data, which is vital for understanding sequences like sentences.
- **Parallel Processing:** Unlike RNNs (Recurrent Neural Networks), transformers can process all elements of the sequence in parallel during training, which significantly speeds up computation.
- In summary, CNNs are specialized for spatial hierarchy and are widely used in image-related tasks, while transformers excel at handling sequential data with complex interdependencies, making them a popular choice for NLP. Both architectures can be adapted for a variety of tasks beyond their original design, showcasing the flexibility of deep learning models.

Attention

- The attention mechanism, when integrated into encoder-decoder architectures such as those used in machine translation and sequence-to-sequence learning, brings significant improvements to the model's ability to handle sequences of data. Here are some key enhancements that attention provides:

1. **Context-Awareness**: Traditional encoder-decoder models compress the entire input sequence into a fixed-length vector, which can lead to information loss, especially for long sequences. Attention allows the decoder to access the entire sequence of hidden states from the encoder, providing a richer, context-aware representation.
2. **Selective Focus**: The attention mechanism enables the model to dynamically focus on different parts of the input sequence during each step of the decoding process. This selective focus mimics how humans pay attention to different words or phrases when comprehending a sentence or translating text.
3. **Long-Range Dependencies**: Attention helps the model capture long-range dependencies within the input data by directly modeling the interactions between distant elements in the sequence, which might be difficult for models without attention due to the vanishing gradient problem.
4. **Improved Gradient Flow**: By establishing direct connections between the encoder and decoder at each time step, attention can improve the flow of gradients during backpropagation, making it easier to train deeper models.
5. **Parallel Computation**: Unlike recurrent models that process sequences step-by-step, attention mechanisms can process all elements of the sequence in parallel, leading to potential computational efficiency during training.
6. **Interpretability**: Attention weights can be visualized, providing insights into which parts of the input sequence the model is focusing on at each step of the decoding process. This interpretability can be valuable for understanding and debugging the model's behavior.

- In summary, the attention mechanism enhances encoder-decoder models by providing a more nuanced and dynamic way to handle sequential data, leading to better performance on tasks such as language translation, text summarization, and speech recognition.

Data mining

- **Data mining** is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1]

Data mining involves six common classes of tasks:[5]

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records that are outside of a standard range.
- Association rule learning (dependency modeling) – Searches for relationships between variables.
- Clustering – A task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – A task of generalizing known structure to apply to new data.
- Regression – Attempts to find a function that models the data with the least error.
- Summarization – Provide a more compact representation of the data set.

Planning (defining steps from start to finish)

- **Planning** is the process of thinking regarding the activities required to achieve a desired goal.

Path planning through a graph

A* (pronounced "A-star") is a graph traversal and pathfinding algorithm, which is used in many fields of computer science due to its completeness, optimality, and optimal efficiency.

Dijkstra's algorithm
Dijkstra's algorithm finds the shortest path from a given source node to every other noc

Floyd–Warshall algorithm is an example of dynamic programming, for finding shortest paths in a directed weighted graph with positive or negative edge weights (but with no negative cycles).

Learning
(Regression and Summarization with Feedback)

- **Learning** is the process of acquiring new understanding, knowledge, behaviors, skills, values, attitudes, and preferences.

Operant conditioning, also called **instrumental conditioning**, is a learning process wh
voluntary behaviors are modified by association with the addition (or removal) of reward
or aversive stimuli. The frequency or duration of the behavior may increase
through reinforcement
or decrease through punishment or extinction.

Morality, Ethics, Safety, Necessity

- **Morality** (from Latin *moralitas* 'manner, character, proper behavior') is the categorization of intentions, decisions and actions into those that are proper, or *right*, and those that are improper, or *wrong*.
- **Ethics** (also known as moral philosophy) is the branch of philosophy which addresses questions of morality.
- **Safety** is the state of being "safe", the condition of being protected from harm or other danger.
- **Necessity** a set of conditional states that must be present for another resulting conditional state to occur, while a sufficient conditional state is one that produces the resulting conditional state.

Nov 21, 2024 | Stanford HAI staff