AI Philosophy

14

14 AI Philosophy*

AI Philosophy

Big questions: Can machines think??

- How *can* minds work
- How *do* human minds work, and
- Can *nonhumans* have minds

philosophers have been around for much longer than computers

AI philosophy is a branch of the philosophy of science concerning on philosophical problems of AI

Does machine intelligence differ from human one??

Is it possible to have a machine civilization??

AI debate

Debate by philosophers each other and between philosophers and AI researchers

– Possibility: philosophers have not understood the content of AI attempts

– Impossibility: the efforts of AI to produce general intelligence has failed

The nature of philosophy is such that clear disagreement can continue to exist unresolved

Another debate among AI researchers focuses on different approaches to arrive at some goals of AI

– logicism or descriptive approach vs. non-logicism or procedural approach

– symbolism vs. behaviorism

Weak AI

Weak AI: Machine can be made to act *as if* there were intelligent

Most AI researchers take the weak AI hypothesis for granted

Objections

1. There are things that computers cannot do, no matter how we program them
2. Certain ways of designing intelligent programs are bound to fail in the long run
3. The task of constructing the appropriate programs are infeasible

Mathematical objection

Turing's Halting Problem

Gödel Imcompleteness Theorem

Lucas's objection: machines are formal systems that are limited by the incompleteness theorem, while humans have so such limitations

– Turing machines are infinite, whereas computers are finite, and any computer can be described as a system in propositional logic, which is not subject to Gödel's theorem

– Humans were behaving intelligently before they invented mathematics, so it is unlikely that formal mathematical reasoning plays more than a peripheral role in what it means to be intelligent

– "We must assume our own consistency, if thought is to be possible at all" (Lucas). But if anything, humans are known to be inconsistency

Strong AI

Strong AI (Searle 1980): Machines that act intelligently have real, conscious minds

Many philosophers claim that a machine that passes the Turing Test would still not be actually thinking, but would be only a simulation of thinking

– there is no scientific consensus on methods for determining consciousness and whether machines possess consciousness

The philosophical issue so-called mind-body problem is directly relevant to the question of whether machines could have real minds

– dualist vs. monist (or physicalism)

AGI

Capabilities

– Achieving AGI does not imply that AI systems think or understand in a human-like way

– Achieving AGI does not imply that AI systems possess qualities such as consciousness or sentience

Performance: the depth of an AI system's capabilities

– how it compares to human-level performance for a given task

Generality: the breadth of an AI system's capabilities

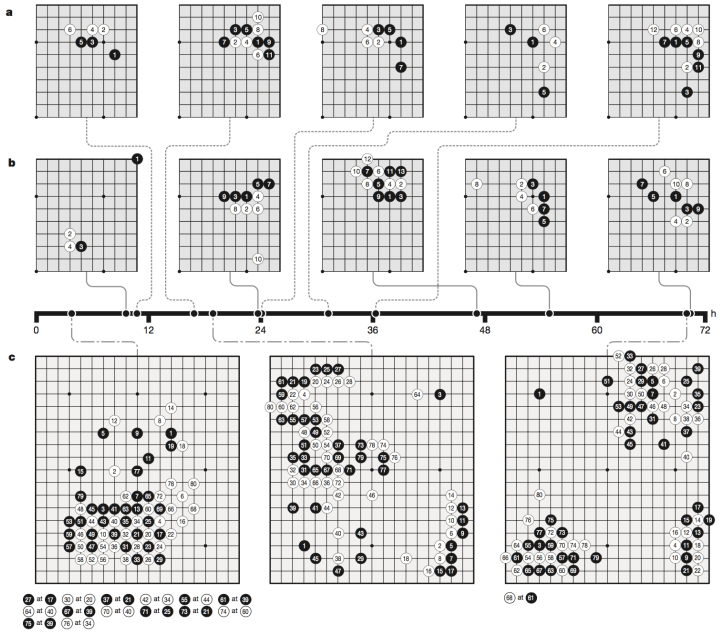
– the range of tasks for which an AI system reaches a target performance

Example: Alpha0

- "The God of chess" of superhuman
  – there would not have any human-machine competition
- self-learning without prior human knowledge
  – an algorithm that learns, *tabula rasa*, superhuman proficiency
  – – only the board of chess as input
- a single neural network to improve the strength of tree search
  – the games of chess have been well defeated by AI

Can a single algorithm solve a wide class of problems in challenging domains??

God of Go



Discovering new Go knowledge without understanding, conscious

Many human concepts can be regressed from the Alpha0 network after training

Ref. McGrath T et al. Acquisition of Chess Knowledge in AlphaZero, arXiv, 2021

Limitations of Alpha0

Assumptions under Alpha0

Deterministic + Perfect information + Zero sum two ply

⇐ self-play reinforcement + neural network + MCTS (probability)

1. deterministic ⇒ **nondeterministic**
   - okay, probability + control
2. perfect information ⇒ nondeterministic **imperfect** information

+general sum

   - okay deep reinforcement learning + Nash equilibria

   say, AlphaStar, but what about Poker (bridge) and Mahjong?

3. Imperfect information ⇒ **complex** information
   - some, say, AlphaFold, ReBel
4. ⇒ **Strong AI**
   - hard, say, deduction (math), common sense etc.

Alpha0 algorithm could not be directly used outside of the games, though the method be done

Failure of Alpha0

Adversarial policy played against KataGo (the strongest publicly available), achieving a $> 99\%$ win-rate against KataGo without search, and a $> 50\%$ win-rate when KataGo uses enough search to be superhuman

But the adversary is easily beaten by human amateurs

– professional-level AI systems may harbor surprising failure modes

– self-play may not be as robust as previously thought

The adversary learned a simple strategy that stakes out a small corner territory and then places some easily capturable stones in KataGo's larger complementary territory

– this strategy loses against even amateur Go players

See Wang T et al., Adversarial policies beat professional level GO AIs, arXiv, 2022

Generalization of Alpha0

A game ⇒ a GGP of the games of chess ⇒ GGP of games

Game ⇒ **non-game** (complex problems)

– unknown, possible domains with strict assumptions

says, AlphaFold (protein folding), reducing energy consumption, searching for new materials, weather prediction, climate modelling, language understanding and more

Due to the non-explanation of neural networks (black box method)

Can a single algorithm solve a wide class of problems in challenging domains??

– "God of chess" is "not thinking"

– no **principle** of understanding Go/Chess or intelligence, but output "knowledge" of Go/Chess for human

An algorithm, without mathematical analysis, is experiment

– it is not general enough to generalization

Alpha0 and deep reinforcement learning

Will quest in deep reinforcement learning lead toward the goal??

– learn to understand, reason, plan, and select actions

With knowledge or without knowledge

– learning by observations without knowledge similar to baby

– knowledge is power of intelligence

most AI systems are knowledge-based

Can the technologies of AI be integrated to produce human-level intelligence??

– no one really knows

– – keep all of technologies active on "frontier of search"

As early AI, there is still a long way to go

Example: ChatGPT

Discussion in class

ChatGPT possesses what level of intelligence??

Will future iterations like GPT-5 or GPT-15 achieve AGI?? (something like iPhone)

What impact will ChatGPT have on higher education??

Etc.

The brain replacement experiment

Functionalism: a mental state is any intermediate causal condition between input and output, i.e., any two systems with isomorphic causal processes would have the same mental state

The brain replacement experiment: Suppose

– neurophysiology has developed to the point where the input-output behavior connectivity of all the neurons in the human brain is perfectly understood

– the entire brain is replaced by a circuit that updates its state and maps from inputs to outputs

What about the consciousness??

Brain-machine interfaces

BMI: try in neural engineering (biotechnology), tantalizing a new industry such as Neuralink by E Musk

Two questions

1) How do I get the right information out of the brain?

– brain output – recording what neurons are saying

2) How do I send the right information into the brain?

– inputting information into the brain's natural flow or altering that natural flow in some other way – stimulating neurons

Early BMI type: Artificial ears and eyes

Chinese room

The Chinese Room: Searle's "Minds, brains, and programs" (1980)

- The system consists of
  1. a human, who understands only English (plays the role of the CPU)
  2. A rule book, written in English (program), and
  3. Some stacks of paper (storage device)

Chinese room

- The system is inside a room with a small opening to the outside
  1. through the opening appear slips of paper with indecipherable symbols
  2. the human finds matching symbols in the rule book and follows the instructions
  3. the instructions will cause one or more symbols to be transcribed onto a piece of paper that is passed back to the outside
- From the outside, the system is taking input in the form of Chinese sentences and generating answers in Chinese that are as "intelligent" as assumed to pass the Turing Test

Chinese room

Argumentation: Searle's axioms

1. Computer programs are formal (syntactic)
2. Human minds have mental contents (semantics)
3. Syntax by itself is neither constitutive of nor sufficient for semantics
4. Brains cause minds

Chinese room

Argumentation: Searle's reasons

– The person in the room does not **understand** Chinese, i.e., running the right program does not necessarily generate understanding

– – so the Turing test is wrong

– So-called biological naturalism: mental states are high-level emergent features that are caused by low-level physical processes in the neurons, and cannot be duplicated just by programs having the same functional structure with the same input-output behavior

Chinese room

Objection

– The person does not understand Chinese, the overall system consisting of the person and the book does

– Searle relies on intuition, not proof

Searl's reply

– Imagine that the person **memorizes** the book and then destroys it

– – there is no longer a system

Objection again

– How can we be so sure that the person does not come to learn Chinese by memorizing the book?

Chinese/Chinese dictionary-go-round

Harnad's (1999) argumentation

– Difficult version: Suppose you had to learn Chinese as a **second** language and the only source of information you had was a Chinese/Chinese dictionary

– Impossible version: Suppose you had to learn Chinese as a **first** language and the only source of information you had was a Chinese/Chinese dictionary

How do you think about??

Summation: a simplified form of Chinese Room

Summation test (Levesque H)

Instead of speaking Chinese, testing the ability to add twenty ten-digit numbers (no more, no less)

- A book listing every possible combination of twenty ten-digit numbers
- A person who does not know how to add to get the summation
- Any time the person is asked what a sum is, the correct answer could be found by looking it up in the book

Such a book can not exist

– 10^{200} distinct entries for all the combinations of numbers
(the entire physical universe only has about 10^{100} atoms)

Summation

Another smaller book can definitely exist (a few pages)

– an English/Chinese language description of how to add

Argumentation

A person who does not know how to add but who **memorizes** the instructions in the book would thereby learn how to add

Hint: What it would be like to **memorize** the Chinese book? $\rightarrow$ What a computer program for Chinese would need to be like?

$\Leftarrow$ the only way to find out is to tackle those technical challenges, just as Turing suggested

Summation

A story (Intelligent Machinery, Turing, 1948)

Infant Gauss was asked at school to do the addition $15+18+24+\cdots+54$ and he immediately wrote down 483, presumably having calculated it as $(15+54)\times(54-12)/2\times 3$. Imagine circumstances where a foolish master told the child that he ought instead of to have added 18 to 15 obtaining 33, then added 21, and so on

$\Leftarrow$ Reflection??

Ethics

The ethical considerations of how AI should act in the world

- People might lose their jobs to automation
- People might have too much (or too little) leisure time
- People might lose their sense of being unique
- AI systems might be used forward undesirable ends
- The use of AI systems might result in a loss of accountability
- The success of AI might mean the end of the human race

The Future of AI

Near future

- 2020s+
  - Connectionists + Symbolists + Baysians + $\cdots$
  - Clouds and fog
- 2040s+
  - Algorithmic convergence
  - Server ubiquity
  - Some AGI and autonomous agents

Toward human-level AI

Human-Level AI

- Understanding the principle of intelligence is still AI long-term goal of AI
  – people made airplanes and then found aerodynamics

- Taking AI systems over (more expensive) human jobs
  – there are still many human cognitive skills that AI does not yet know how to do

- Machine intelligence may differ from human one
  – Is it possible to have a machine civilization?

But not Human-Level AI yet

The quest for AI is not yet complete

The Future of AI

Far future, always quest

When would AI arrive at the goal of building human-level intelligence??

What if AI does damage human civilization??