Artificial Intelligence

Philosophy and Ethics of AI

Overview

• “Weak” Artificial Intelligence
  – Can machines *act* as if they had intelligence?
• “Strong” Artificial Intelligence
  – Can machines *really think*?
• Ethics of Artificial Intelligence
  – What *problems* does AI pose for the people of the world?
**Weak Artificial Intelligence**

- Can machines *act* as if they actually had intelligence?
  - For this course, we have assumed Weak AI is possible
  - But is it really?
    - Earlier we had “*acting humanly*” and “*acting rationally*” in the 4 categories of AI
    - Given this, then the answer is “yes”

**The Start**

- Recall 1956 summer workshop at Dartmouth that coined “AI”…
- They stated, “*Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it*”
  - Thus AI founded on assumption that Weak AI was possible
- But others have asserted Weak AI is impossible
Weak Artificial Intelligence

• One point of view:
  – “The question of whether machines can think ... is about as relevant as the question of whether submarines can swim.” - Edsger Dijkstra (Note: CS pioneer that helped shape the field from both an engineering and a theoretical perspective)
    • Submarines don’t swim like people, but they certainly “move through the water”
    • Even if machines don’t think like people, they are capable of displaying intelligence in their own fashion
    • So the question is meaningless – like arguing that airplanes don’t fly because they don’t flap their wings

Weak Artificial Intelligence

• Idea of “appearing intelligent” has long history
  – Turing Test
    • Turing predicted by 2000 that computers would be able to consistently pass this kind of test
      – He was wrong – it happens even to geniuses :)
    – Eliza, Eugene Goostman, and chatbots
Weak Artificial Intelligence

• The argument against Weak AI from the “Informality of Behavior”
  – Argument is characterized as
    • Human behavior cannot be boiled down into simple rules
    • All that computers do is follow rules
    • Therefore, computers cannot behave as intelligent as humans
  – This critique is more about older AI techniques
    • “Good Old Fashioned AI”

Weak Artificial Intelligence

• The argument against Weak AI from “Informality of Behavior” (con’t)
  – In 1986, philosophers Hubert & Stuart Dreyfus (brothers) outlined several issues/problems they saw with the idea of machine intelligence
  1. Background knowledge is needed to generalize well from sets of examples
  2. Machine intelligences cannot learn without the help of a human trainer (supervised)
  3. Learning algorithms have problems with large number of inputs, and require human intervention to get the right features
  4. Machines cannot direct their own sensors to seek out learning on their own
**Weak Artificial Intelligence**

- The objections of Dreyfus & Dreyfus presented are more about AI research *at that time*
- Recent research has made attempts to address these issues
  1. Background knowledge remains a problem, but efforts are being made to use the *Web as a repository* of background knowledge
  2. Advances in unsupervised learning (and reinforcement learning) have shown promising results with machines that *can learn* on their own
  3. New algorithms work with *millions* of features
     - And, possibly more importantly, advances are being made in algorithms to allow the machine to select features itself, without human intervention
  4. Robotics is showing the way with advances to allow our agents to *direct* themselves to potential learning opportunities

**Strong Artificial Intelligence**

- Still, many philosophers have claimed that a machine that passes the *Turing Test* would still *not be actually thinking* (Strong AI), it only *simulates thinking* (Weak AI)…
**Strong Artificial Intelligence**

- Many of the objections to *Strong* AI come down to the *mind-body problem*
  - Old problem of philosophy:
    - Is there a *separation* between the mind and the body?
    - Or, is the mind an *emergent* property of the body?
  - Philosopher and mathematician Renee Descartes (1641) stated that body and mind are separate
    - Body was of the physical, material realm, and Mind was of the mental, immaterial realm
    - The mind controlled the body as a puppeteer controls a puppet

---

**Strong Artificial Intelligence**

- But perhaps the mind is not separate from the body
  - Mental states ARE physical states
  - Biology supports this idea in part
    - Damage to certain areas of the brain affects mental abilities
Strong Artificial Intelligence

- A thought experiment:
  - Assume we know enough about the brain where the input-output behavior of all neurons is completely understood
  - Assume that it’s so well understood that we can build nanomachines that mimic this behavior and can interface with biological neurons
  - Assume we can do this all so well that we can actually replace damaged (or undamaged) neurons with these devices without interrupting the function of the brain
  - Given all of those assumptions, think about a person whose brain has gradually been replaced by these nanomachines
    - Consider it over a period of years (perhaps to heal a degenerative neural disease)
  - **What happens to the consciousness of the person???

*Consciousness* has been defined as: sentience, awareness, subjectivity, the ability to experience or to feel, wakefulness, having a sense of selfhood, the executive control system of the mind, etc.

---

(John Searle - Philosopher, 1992)

You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when doctors test your vision, you hear them say “We are holding up a red object in front of you; please tell us what you see.” You want to cry out “I can’t see anything. I’m going totally blind.” But you hear your voice saying in a way that is completely out of your control, “I see a red object in front of me.”… your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.

*“Help, I’m shrinking!”*
**Strong Artificial Intelligence**

- Or would the consciousness of the person be unchanged?
  - The person would not even realize that anything had changed
  - The assumptions of the experiment say that this gradual change **cannot change the behavior** of input stimulus to output behavior
- This leads to one of **three** possible conclusions
  1. The mechanisms that determine consciousness in a biological brain are still active in the replacement brain, so **it is also conscious**
  2. The mechanisms that determine consciousness in a biological brain are completely disconnected and unimportant to the connection between stimulus and behavior, so they’re missing from the electronic brain and it is **not conscious**
  3. The experiment is so impossible that speculation about what it would mean is completely worthless

  **If conclusion #1 holds, then intelligence is not a property of biology and means that computers could be capable of conscious thought!**

---

**Ethics and Risks of AI**

- **Should** we be pursuing the goal of artificial intelligence?
  - Even if it is possible, **is it the right thing to do?**
  - Every new technology brings new ethical questions, AI is no different
- Multiple “hot-button” concerns for AI
Ethics and Risks of AI

• **People might lose their jobs to automation**
  – Modern industrial economy dependent on computers and tech
  – However automation through information technology and AI has created more jobs than it has eliminated
  – Designed more to assist a human (not replace)

Ethics and Risks of AI

• **People might have too much (or too little) leisure time**
  – Arthur C. Clarke wrote that people might be “faced with a future of utter boredom”
  – However in our information society, people are actually working harder
Ethics and Risks of AI

• *People might lose their sense of being unique*
  – Weizenbaum (Eliza) comments that AI research makes possible the idea that humans are automata
    • Results in a loss of autonomy or even humanity

Ethics and Risks of AI

• *AI systems might be used toward undesirable ends*
  – Perhaps even the destruction of human life
  – Almost all technologies can be used for a variety of purposes, some helpful and some harmful
    • AI is no different in this respect
  – AI is common on the battlefield
    • Autonomous aircraft patrol the skies providing reconnaissance
    • Semi-autonomous aircraft (“drones”) perform bombing missions via remote control
    • Semi-autonomous ground vehicles assist in finding mines and other hazards on the battlefield
Ethics and Risks of AI

– AI has made surveillance technology pervasive
  • If a human has to watch every camera, listen to every telephone conversation, or read every e-mail, surveillance is limited by the number of people you can employ
  • AI agents can watch every camera, listen to every conversation, and read every e-mail, limited only by the state-of-the-art in technology
  • There is a trade-off between privacy and security, and technology forces us to confront the ethics of that question

Ethics and Risks of AI

• The use of AI systems might result in a loss of accountability
  – What if a medical AI agent misdiagnoses a patient in an unreasonable way… who is responsible?
    • The hospital, company, programmer?
  – What about self-driving/autonomous cars
    • Who (or what) will be accountable?
  – Big issue in upcoming AI deployment
Ethics and Risks of AI

• *The success of AI might mean the end of the human race*
  – A standard staple of science fiction stories
    • The premise of the Matrix, Terminator, and other films
      – Really, goes back at least to *Frankenstein*
    • But is it something that we need to worry about?
  – Concerns about AI “run amok”
    • The AI might make disastrous mistakes with deadly consequences
      – Self-driving cars that cause crashes
      – Missile guidance systems that mistake a flock of birds for an attack
      – (Of course, people make mistakes as well)

• “Be careful what you wish for”
  – We might program an AI to minimize human suffering… and it decides that we’d suffer less if we were **all dead**!!!
  – Basically a concern that we won’t be able to properly program the AI to do what’s best for us
Ethics and Risks of AI

• Can we actually predict what a self-learning AI will learn and will “evolve” into?

• More recently described as the technological singularity (SciFi author Vernor Vinge, 1993)
  – “Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended” by 2023!

Ethics and Risks of AI

• Another view: The replacements we build might be ourselves
  – Transhumanism
    • The idea that humans will be improved by AI and/or biotech into a new “transhuman” race
  – Ray Kurzweil (Computer Scientist, author, inventor, futurist) believes the Singularity will be a merging of human and technology
    • “By the end of this century, the nonbiological portion of our intelligence will be trillions and trillions of times more powerful than unaided human intelligence” (2005)
Ethics and Risks of AI

• The question of how to limit our “future robotic overlords” is not a new one
  – The most famous example is one of the oldest – Isaac Asimov’s “Three Laws of Robotics” ***
  1. “A robot may not injure a human being or through inaction allow a human being to come to harm”
  2. “A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law”
  3. “A robot must protect its own existence, as long as such protection does not violate the First or Second Laws”

Futurama: ‘T.: The Terrestrial’

Fry: “I'm gonna sneak a closer look. Bender, you'll let me know if we have to cheese it, right?”

Bender: “First Law of Robotics, baby.”
The Big Bang Theory

“Sheldon may be a robot...”

https://www.youtube.com/watch?v=BKkEI7q5tug

In 2015, $10M donated to Future of Life Institute (FLI)

**FLI Mission:** To catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges.
Open Letter on AI

• Jan 2015: Stephen Hawking, Elon Musk, and multitudes of AI experts signed an open letter on AI calling for research on the societal impacts of AI
  – Called for concrete research on how to prevent certain potential “pitfalls”

https://futureoflife.org/ai-open-letter/

A Final Thought…

• Suppose we really can create an intelligent and self-aware machine
  – Maybe not an ultra-intelligent mind that will bring us to the Singularity, but as smart as us
• What are the ethical implications?
  – Is owning a robot that does housework like owning a vacuum cleaner, or is it more like slavery?
    • Robotic rights?
    • At this point, a Roomba is just a “dumb” machine, not even at the level of animal intelligence
      – But if it were as smart as a four-year-old?
  – Is there a line that we cross based on intelligence?