The Social Impact of Artificial Intelligence - Lecture 18.2

Topic:

- Human-Centered Artificial Intelligence

Human-Centered Artificial Intelligence - 1

- Mary Wollstonecraft Shelley's *Frankenstein; or, The Modern Prometheus* is the first true science fiction novel.
- Prometheus stole fire from the gods and gave it to humanity.
- Zeus punished that theft, of technology and knowledge, by sentencing Prometheus to eternal torment.
- Dr. Frankenstein's creature attempted to assimilate into human society.
- His rejection by humankind caused him to exact revenge.
- Frankenstein's monster has now come to symbolize unbridled, uncontrolled technology, turning against humans.
- Concerns about the control of technology are now increasingly urgent as AI transforms our world.

Human-Centered Artificial Intelligence - 2

- Proposals for artificial general intelligence (AGI) envisage systems that outperform humans on a wide range of tasks, unlike so-called "narrow" AI that develops and trains systems for specific tasks.
- AGI might lead to a singularity when AGI bootstraps to a superintelligence, that could dominate humans, a possible existential risk [Good, 1965].
- Imagine an AGI system, given a goal to maximize the number of paperclips, could consume every resource available to it, including those required by humans [Bostrom, 2014].
- An AGI, without common sense, could pose an existential threat to humanity if its goals are misspecified or otherwise not aligned with the long-term survival of humans and the natural environment.
- This safety concern has come to be known as the alignment problem.

Human-Centered Artificial Intelligence - 3

- AI systems, such as self-driving cars and lethal autonomous weapons, may make life-or-death decisions without meaningful human oversight.
- AI systems may make harmful, even if not life-threatening, value-laden decisions impinging on human welfare, such as deciding who should get a mortgage, a job offer, or parole.
- A focus on autonomy and human control: How can designers create human-centered AI or human-compatible AI?
- Can human values be instilled in AI systems?
- Who's values? Who gets to decide what values?
- One technique for incorporating human values is Reinforcement Learning from Human Feedback (RLHF). Learns human values from samples.
- RLHF is the framework for a key module of ChatGPT.

Human-Centered Artificial Intelligence - 4

- Increasingly, especially in high-stakes applications, human decision-makers are assisted by semi-autonomous agents; this combination is known as human-in-the-loop.
- Real-time online intelligent systems are often structured as a hierarchy of controllers [AIFCA, Ch. 2].
- The lower levels operate very quickly, on short time horizons, while the higher levels have longer time horizons, and operate slowly on more symbolic data.
- Human interaction with hierarchically structured systems occurs at the higher levels.
- Human drivers can provide high-level navigation preferences to semi-autonomous vehicles.
- Humans can steer or brake to avoid accidents but only if they are paying attention.
- As vehicles become more automated distracted drivers may be unable to redirect their attention in time.

Human-Centered Artificial Intelligence - 5

- The concept of attention in neural networks is inspired by the concept of human attention.
- Key aspects of human attention include vigilance, the state of keeping careful watch for possible danger, and salience, the quality of being particularly noticeable or important.
- Designing AI systems so that humans can meaningfully interact with them requires designers who understand the critical roles of vigilance, salience, and attention.
- Designers of interactive AI systems must be versed in the practices of both human–computer interaction (HCI) and AI.
- Good interaction designs are needed for trustworthy AI systems. [Shneiderman, 2022]
- The "Guidelines for Human–AI Interaction" [Amershi, 2019] give strategies for doing less when the system is uncertain to reduce the costs and consequences of incorrect predictions.

Human-Centered Artificial Intelligence - 6

- Assistive technology for disabled and aging populations, as well as the general population, is an emerging area of beneficial AI applications.
- Assisted cognition, including memory prompts, is one set of applications. Assisted perception including visual and auditory aids and language translation is another.
- Assisted action, in the form of semi-autonomous smart wheelchairs, companions for older people, and nurses' assistants in hospitals and long-term care facilities, can be useful technologies.
- However, there are dangers in relying upon AI or robotic assistants as companions for the elderly and the very young.
- Researchers and developers of assistive technology, and other AI applications, should be aware of the dictum of the disability rights movement, "Nothing about us without us.".

Human-Centered Artificial Intelligence - 7

- A plethora of concepts are used to evaluate AI systems from a human perspective, including transparency, interpretability, explainability, fairness, safety, accountability, and trustworthiness.
- Transparency typically refers to the complete ecosystem surrounding an AI application, including the description of the training data, the testing and certification of the application, and user privacy concerns.
- Transparency is also used to describe an AI system whose outcomes can be interpreted or explained, where humans can understand the models used and the reasons behind a particular decision.
- Black-box AI systems, based, say, on deep learning, are not transparent in that sense. Systems that have some understanding of how the world works, using causal models, may be better able to provide explanations.

Human-Centered Artificial Intelligence - 8

- Enhancements in explainability may make an application more trustworthy.
- Enhanced transparency, interpretability, and fairness may also improve trustworthiness. Interpretability is useful for developers to evaluate, debug and mitigate issues. However, the evidence that it is always useful for end-users is less convincing.
- Understanding the reasons behind predictions and actions is the subject of explainable AI.
- It might seem obvious that it is better if a system can explain its conclusions.
- However, having a system that can explain an incorrect conclusion might do more harm than good. "Explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness." [Bansai, 2021]