Advanced Machine Learning, Lecture 13 (week 14)

András Kornai

BME 2020 Dec 10

Ethics

- Semantics is the last missing piece for building true AI
- GPT4: perfect English syntax, very imperfect semantics
- What's the problem? Lack of internal model, GIGO
- It is being taught the wrong things 'toxic text'
- It is being taught with huge energy impact (versus 'green AI')

The person/algorithm boundary

- In creasingly important matters (financial, medical, legal) decisions are gradually taken over by algorithms
- The slavery model (Asimov's Laws of Robotics). Orthogonality (Bostrom 2012)
- Freedom, effectors
- Ethical rationalism (Gewirth 1972): Prospective purposive agent (PPA) **Definition:** can act with purpose and reason rationally. **Theorem:** PPA must obey the Principle of Generic Consistency (PGC): act in accord with the generic rights of your recipients [to freedom and well-being] as well as of yourself
- Contemporary algorithms don't quite meet PPA criteria (Kornai 2015)