Runaround

Isaac Asimov

28 February 2023

Conduct

- **From Ethos of MLC…**
  - https://mlcollective.org/wiki/code-of-conduct/
- **Highlights**
  - Learn Humbly. Ask dumb Questions!
  - Be Curious.
  - Expectation of Confidentiality
  - Reporting -> send me a direct message

Communication Question

Fundamental Rules of Robotics

1. A robot may not injure a human being, or allow a human being to come to harm (no harm)

2. A robot must obey the orders given to it by human beings. This must not conflict with *Rule 1 (obedience)*

3. A robot must protect its own existence so long as such protection does not conflict with *Rule 1 or 2 (avoid danger)*

Popular Areas of Debate

1. Privacy and Surveillance
2. Manipulation of Behavior
3. Opacity of AI systems
4. Bias in AI
5. Human-AI Interaction
6. Autonomous Systems
7. Machine Ethics
8. Moral Agents
9. singularity

Source: https://plato.stanford.edu/entries/ethics-ai

Sides of AI Ethics

- *They were in the radio room now-with its already subtly antiquated equipment, untouched for the ten years previous to their arrival. Even ten years, technologically speaking, meant so much. Compare Speedy with the type of robot they must have had back in 2005. But then, advances in robotics these days were tremendous.*

Current AI doesn’t reflect our values in society (e.g., government, research)

Asimov's POV and US Governance

- Informed Consent (e.g., HIPAA)
- Right to Privacy (4th Amendment)
- One Party Consent States

1. Privacy and Surveillance
2. Manipulation of Behavior
3. Opacity of AI systems
4. Bias in AI
5. Human-AI Interaction
6. Autonomous Systems
7. Machine Ethics
8. Moral Agents
9. Singularity

Asimov's POV and Previous Discussions

- **Paper:** *Differentially Private Fair Learning*
  - **Idea:** It's ideal to have systems that combine different elements that build trust. Having a System that is both Private and Fair has pros/cons

- **Paper:** *Causal Parrots: LLMs may talk causality but are not causal*
  - **Idea:** LLMs appear to be more causal than than we can actually prove them to be. LLMs may be able to reason

1. Privacy and Surveillance
2. Manipulation of Behavior
3. Opacity of AI systems
4. Bias in AI
5. Human-AI Interaction
6. Autonomous Systems
7. Machine Ethics
8. Moral Agents
9. Singularity

Asimov's POV and Previous Discussions

- **Paper:** *Model Explanations with Differential Privacy*
  - **Idea:** An AI model should be transparent in its decision process, but an attacker shouldn't be able to use the model's transparency to circumvent privacy
- **Paper:** *Sleeper Agents: Training Deceptive LLMs that Persists Through Safety Training*
  - **Idea:** Companies can build malicious LLMs that appear benign. Current AI Safety methods may not apply to LLMs.

1. Privacy and Surveillance
2. Manipulation of Behavior
3. Opacity of AI systems
4. Bias in AI
5. Human-AI Interaction
6. Autonomous Systems
7. Machine Ethics
8. Moral Agents
9. Singularity