OpenAI

Panel 1: AI

AI Safety & Assurance

Alexander Ray, OpenAI

OpenAI

- AI Research Lab in San Francisco, CA
- Focused on Machine Learning
  - Primarily Deep Neural Networks
- Three categories of research relevant here:
  - Technical AI Capabilities Research
  - Technical AI Safety Research
  - AI Policy Research

OpenAI

- Technical AI Capabilities Research
  - Advancing state of the art capabilities
  - Understanding & evaluating capabilities
- Technical AI Safety Research
  - Study AI failure modes and effects
  - Methods for human input and specification
- AI Policy Research
  - Coordination mechanisms
  - Understanding regulatory strategies and effects

AI Safety & Assurance

- Overlapping research areas:
  - Understanding Failures
  - Metrics and Measurements
  - Incorporating Human-in-the-loop
  - Policy Research

- (Other areas of AI Safety research also have overlap, cutting this to be brief!)

Understanding Failures

- Concrete AI Safety Problems
  - https://openai.com/blog/concrete-ai-safety-problems/
- Faulty Reward Functions in the Wild
  - https://openai.com/blog/faulty-reward-functions/

- Assurance Angle:
  - Safety requirements starts w/ enumerating and describing safety failures

Metrics and Measurements

- Unforeseen Adversarial Robustness metric
  - https://openai.com/blog/testing-robustness/
- Safety Gym Benchmark
  - https://openai.com/blog/safety-gym/

- Assurance Angle:
  - Quantitative metrics and measures are great!
  - Need lots of iteration on these
    - Optimizing one metric will reveal another

Learning Human Preferences and Values

- Learning from Human Preferences
  - https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/
- Fine-Tuning Language Models w/ Feedback
  - https://openai.com/blog/fine-tuning-gpt-2/

- Assurance Angle:
  - One possible ingredient to safety is human input during training
  - Extremely difficult technical problems

Policy Research

- Strategies for Cooperation on AI Safety
  - https://openai.com/blog/cooperation-on-safety/
- AI Safety Needs Social Scientists
  - https://openai.com/blog/ai-safety-needs-social-scientists/

- Assurance Angle:
  - Technical research alone doesn’t solve it all
  - Need research on coordination, etc.

AI Safety & Assurance

- Fundamentally compatible objectives
- Foster a rich and productive research ecosystem
- Solve important problems w/ autonomy

Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER