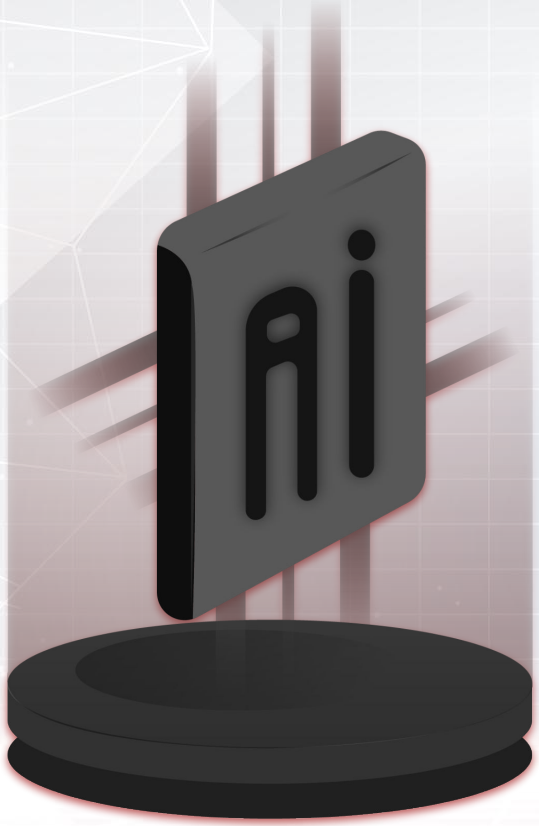

Managing Extreme AI Risks Amid Rapid Progress

Bengio et al.

Michael Greer and Natasa Zupanski

Rapid Progress

- AI capabilities still limited, but current trends foreshadow rapid progress
  - Investment: x3/yr
  - Training Efficiency: x2.5/yr
  - Computing Chips: x1.4/yr
- Advances in AI accelerate AI progress
- No reason AI progress will halt at human abilities
  - AI already surpassed human intelligence in certain domains

*"We must take seriously the possibility that **highly powerful generalist AI** systems…will be developed within the **current decade** or the next.*

What happens then?"

Roots of Risks

- Causes
  - Malicious users and developers
  - Biased or Incomplete Data
  - Poorly Specified Objectives
    - “Literalist” AI
- Strategies for a Misaligned AI
  - Hacking
  - Social Manipulation
  - Misinformation
  - Self-Replication
  - Gaining Control of Critical Infrastructure

Consequences

Misinformation	Cybercrime	Discrimination
Damage to the Biosphere	Human Marginalization	**Human Death or Extinction**

Solutions and Preventative Measures

R&D Challenges

- Progress on AI safety falling behind
  - Only 1-3% of AI publications

- Must address certain challenges to ensure “reliably safe AI”:
  - **Oversight:** How can we tell AI is honest/correct?
  - **Robustness:** How well can AI work in new environments?
  - **Interpretability:** How to better understand AI’s decisions?
  - **Inclusivity:** How to mitigate bias and include all perspectives?
  - **Emerging challenges,** such as AI bypassing its own safety mechanisms

R&D Challenges

- Other challenges focus on effective governance and reducing harm when safety and governance fail
  - **Unforeseen Capabilities:** How can we evaluate AI's true capabilities? Test/prepare for unexpected danger?
  - **Alignment:** How can we evaluate AI's intent? What is AI willing to do? Might AI lie about its inner workings?
  - **Risk Assessment:** What is the extent of harm AI could do? Impact on society?
  - **Resilience:** Can we stop AI? How?

Governance

- Governance must prevent risk-taking by those seeking a competitive edge
- Take inspiration from governance of other “safety-critical technology”
- Policies should trigger upon certain AI milestones
  - To prepare for breakthroughs
  - Politically feasible despite disagreements on future AI timeline
- New govt. institutions to identify risks and enforce proactive risk reduction
  - Require developers of AI to assess and address risks

Governance

- Govt. Institution for AI Safety
  - Technologically savvy
  - Well funded
  - International Cooperation
  - Focus on frontier. Protect small & predictable.
- Mitigation & Regulation
  - Set best practices, standards
  - Red lines
  - Liability, clear consequences
  - Failsafe government control over progress
- Insight Required
  - Protection for whistleblowers, Incident reports, Registry of models/data
  - Access from the start of development
- Company Responsibility
  - Unsafe until proven safe
  - Reports of potential hazards & mitigation plans

Discussion

Why?
What?
Who?
Where?
Which?
When?
Wow?
Wollowski?

Discussion

Who should be responsible for ensuring the safety of AI as it progresses? How would they be held accountable?

How can we manage the risks of AI while maintaining progress in AI? Is this even possible?

How do we keep AI democratized while denying access to malicious actors?