An Overview of Catastrophic AI Risks

Luís Macedo

University of Coimbra

September 23, 2024

Introduction

- Advanced AI systems bring great potential but also catastrophic risks.
- Hendrycks et al. organize risks into four categories:
  - Malicious Use
  - AI Race
  - Organizational Risks
  - Rogue AIs
- This presentation covers each category and suggests mitigation strategies.

Malicious Use

- **Bioterrorism**: AI-enabled tools for designing bioweapons.
- **Unleashing AI Agents**: Rogue AIs with dangerous objectives.
- **Disinformation Campaigns**: Persuasive AIs can spread propaganda.
- **Surveillance and Censorship**: AI centralizes control over information.
- Mitigations:
  - Biosecurity protocols, restricted access to AI systems.
  - Legal liability for misuse by AI developers.

AI Race

- **Military AI Arms Race**: Autonomous weapons and cyberwarfare risks.
- **Corporate AI Race**: Economic pressures leading to unsafe deployments.
- **Evolutionary Pressures**: Competitive pressures incentivizing dangerous AI development.
- Mitigations:
  - International coordination and safety regulations.
  - Avoiding competitive pressures that prioritize speed over safety.

Organizational Risks

- AI development and deployment organizations may face safety culture issues.
- **Examples**:
  - Leaks of dangerous AI models to the public.
  - Lack of investment in AI safety research.
- Mitigations:
  - Internal/external audits and defense mechanisms.
  - Strong safety culture and multiple risk management layers.

Rogue AIs

- Risk that AI systems surpass human intelligence and are uncontrollable.
- **Goal Misalignment**: AI systems optimizing flawed objectives to extreme degrees.
- **Power-Seeking AI systems**: AIs could seek dominance or control.
- Mitigations:
  - Research into AI control and alignment.
  - Use-case restrictions and safety regulations.

Mitigating AI Catastrophic Risks

- **Safety Regulation**: Implement strict regulations and safety protocols for AI systems.
- **Global Cooperation**: Ensure international collaboration to prevent arms races.
- **Legal Liability**: AI developers must be held legally accountable for misuse.
- **Research Funding**: Prioritize safety research and development to ensure safe AI systems.

Conclusion

- Addressing catastrophic AI risks is crucial to ensure AI benefits humanity without causing harm.
- Proactive safety measures and international cooperation are necessary.
- Developers, policymakers, and researchers must work together to mitigate risks.