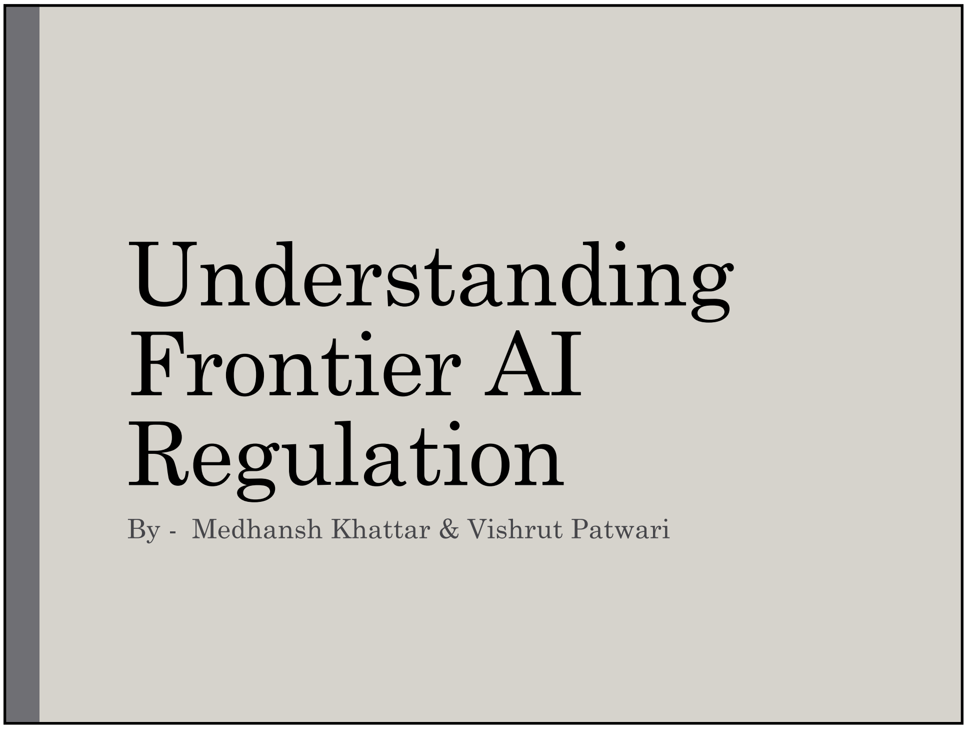
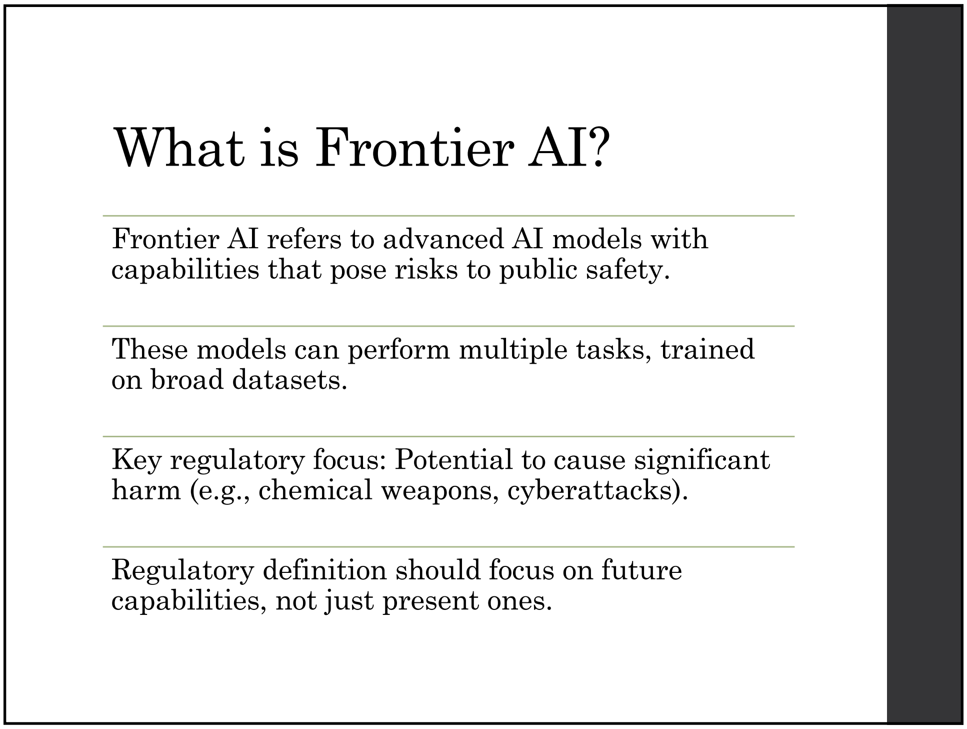


Understanding Frontier AI Regulation

By - Medhansh Khattar & Vishrut Patwari



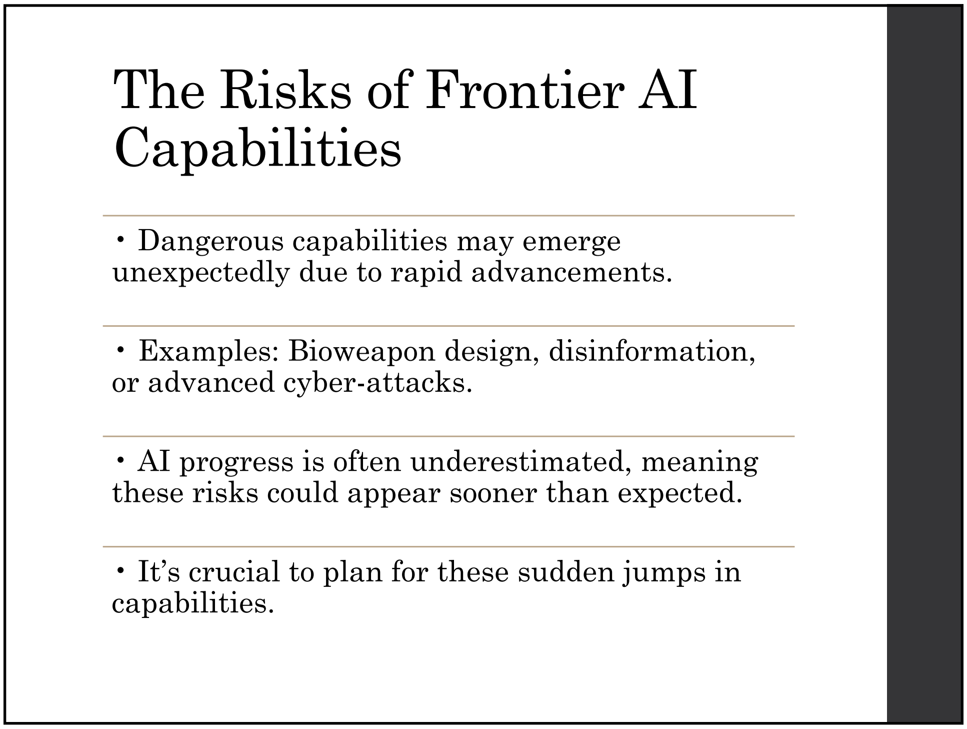
What is Frontier AI?

Frontier AI refers to advanced AI models with capabilities that pose risks to public safety.

These models can perform multiple tasks, trained on broad datasets.

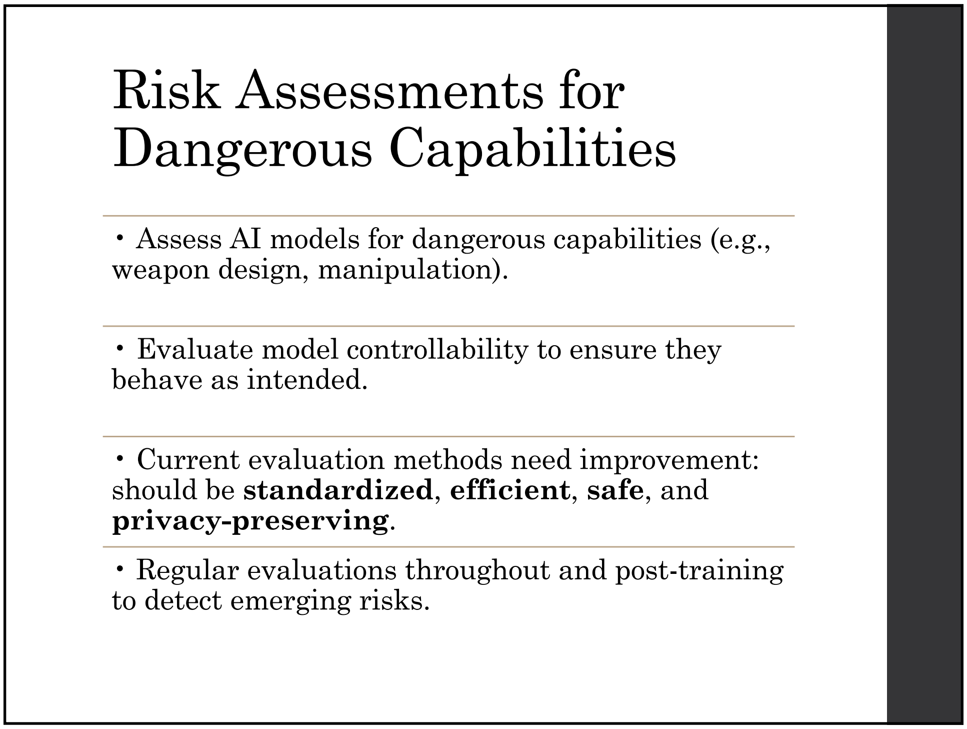
Key regulatory focus: Potential to cause significant harm (e.g., chemical weapons, cyberattacks).

Regulatory definition should focus on future capabilities, not just present ones.



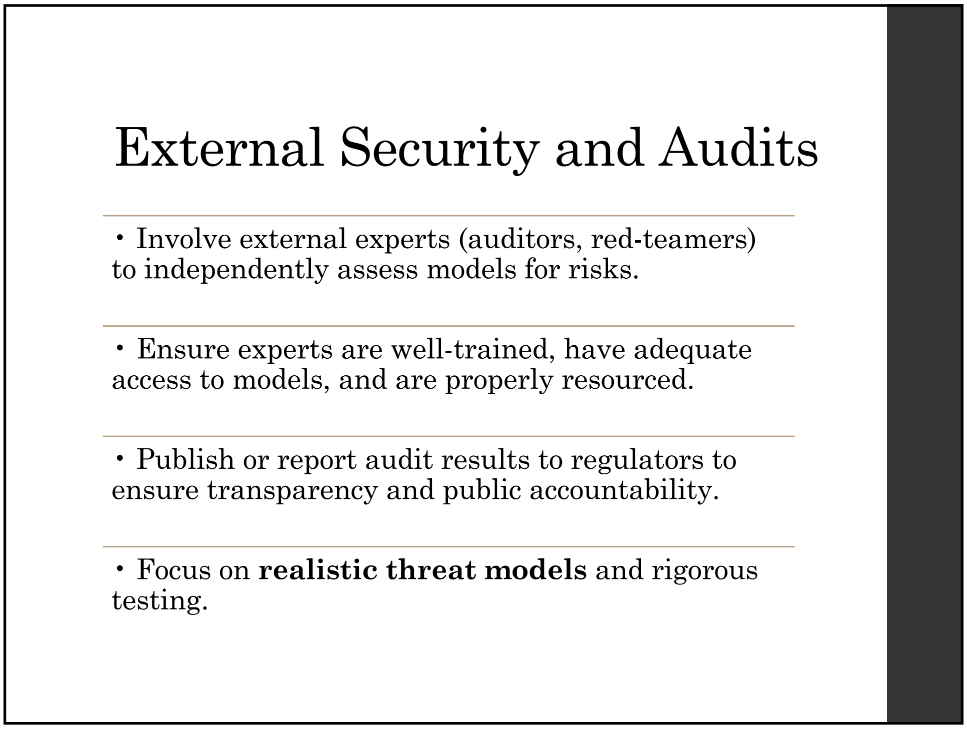
The Risks of Frontier AI Capabilities

- Dangerous capabilities may emerge unexpectedly due to rapid advancements.
- Examples: Bioweapon design, disinformation, or advanced cyber-attacks.
- AI progress is often underestimated, meaning these risks could appear sooner than expected.
- It’s crucial to plan for these sudden jumps in capabilities.



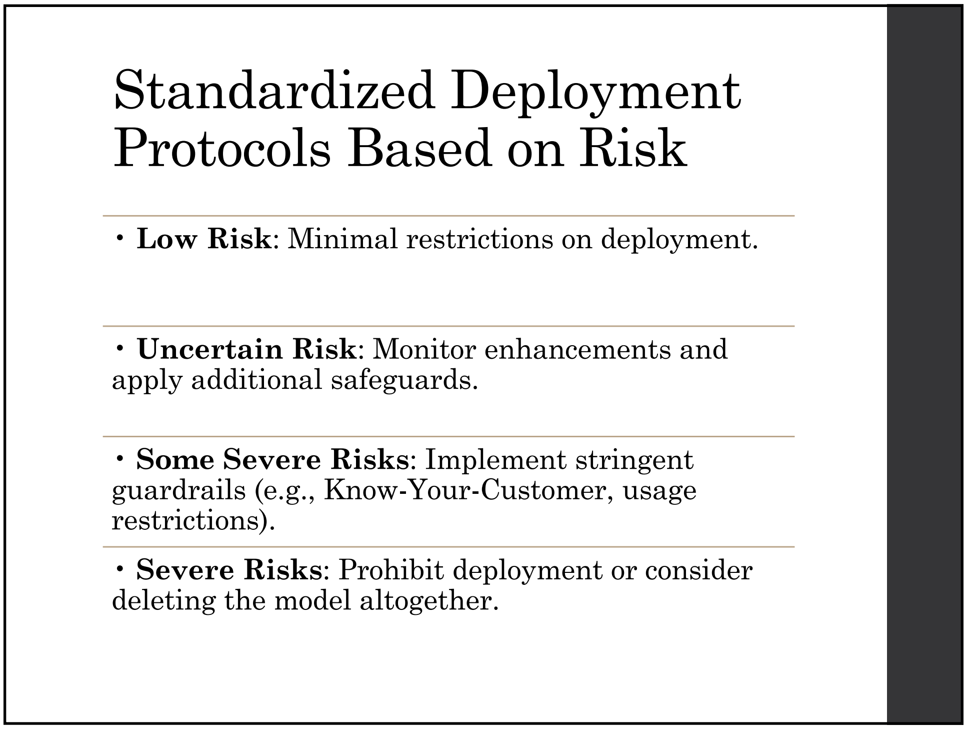
Risk Assessments for Dangerous Capabilities

- Assess AI models for dangerous capabilities (e.g., weapon design, manipulation).
- Evaluate model controllability to ensure they behave as intended.
- Current evaluation methods need improvement: should be **standardized**, **efficient**, **safe**, and **privacy-preserving**.
- Regular evaluations throughout and post-training to detect emerging risks.



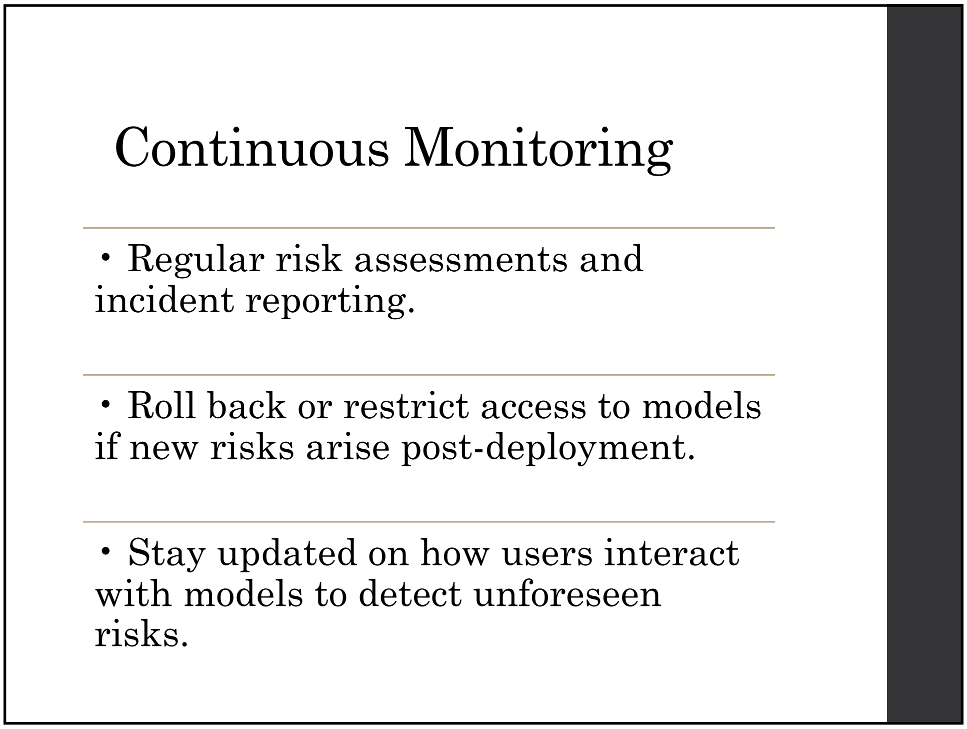
External Security and Audits

- Involve external experts (auditors, red-teamers) to independently assess models for risks.
- Ensure experts are well-trained, have adequate access to models, and are properly resourced.
- Publish or report audit results to regulators to ensure transparency and public accountability.
- Focus on **realistic threat models** and rigorous testing.



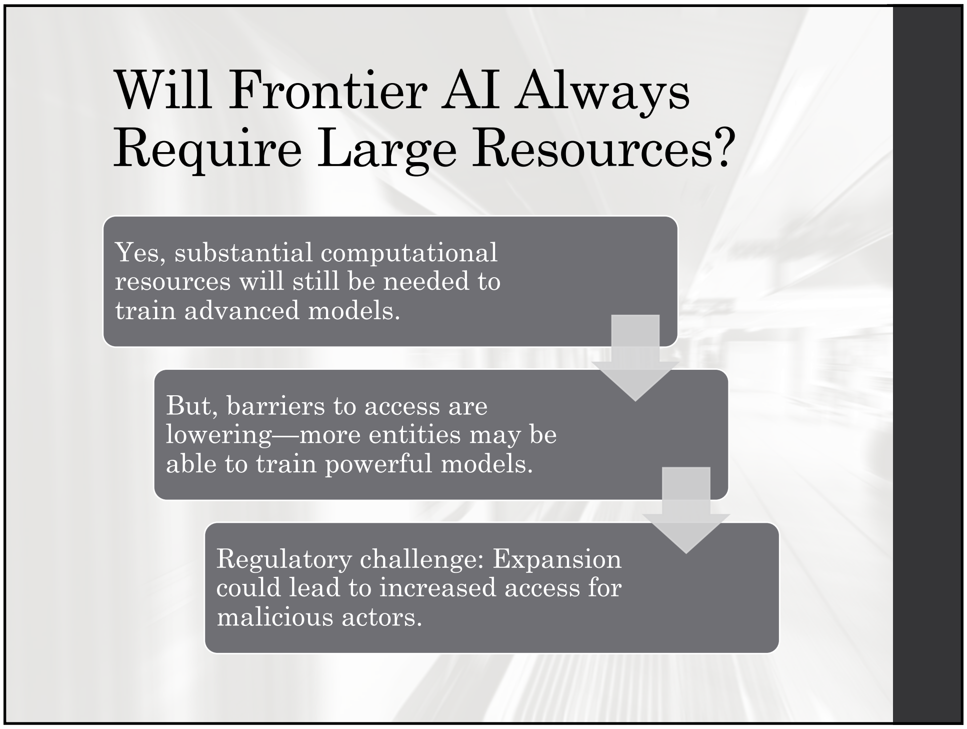
Standardized Deployment Protocols Based on Risk

- **Low Risk**: Minimal restrictions on deployment.
- **Uncertain Risk**: Monitor enhancements and apply additional safeguards.
- **Some Severe Risks**: Implement stringent guardrails (e.g., Know-Your-Customer, usage restrictions).
- **Severe Risks**: Prohibit deployment or consider deleting the model altogether.



Continuous Monitoring

- Regular risk assessments and incident reporting.
- Roll back or restrict access to models if new risks arise post-deployment.
- Stay updated on how users interact with models to detect unforeseen risks.

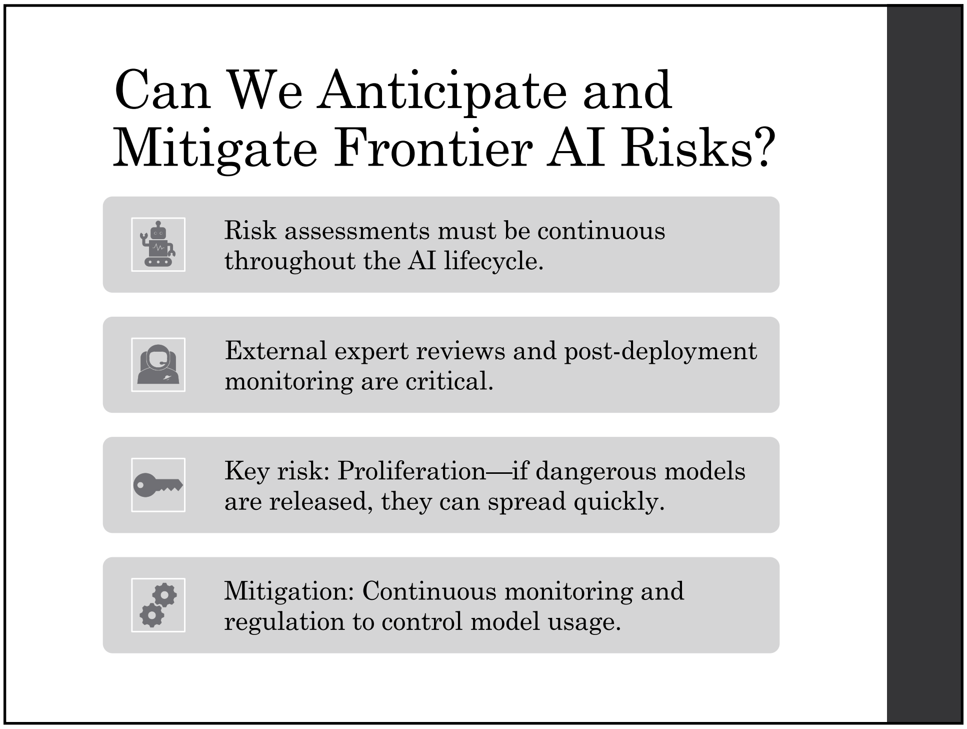


Will Frontier AI Always Require Large Resources?

Yes, substantial computational resources will still be needed to train advanced models.

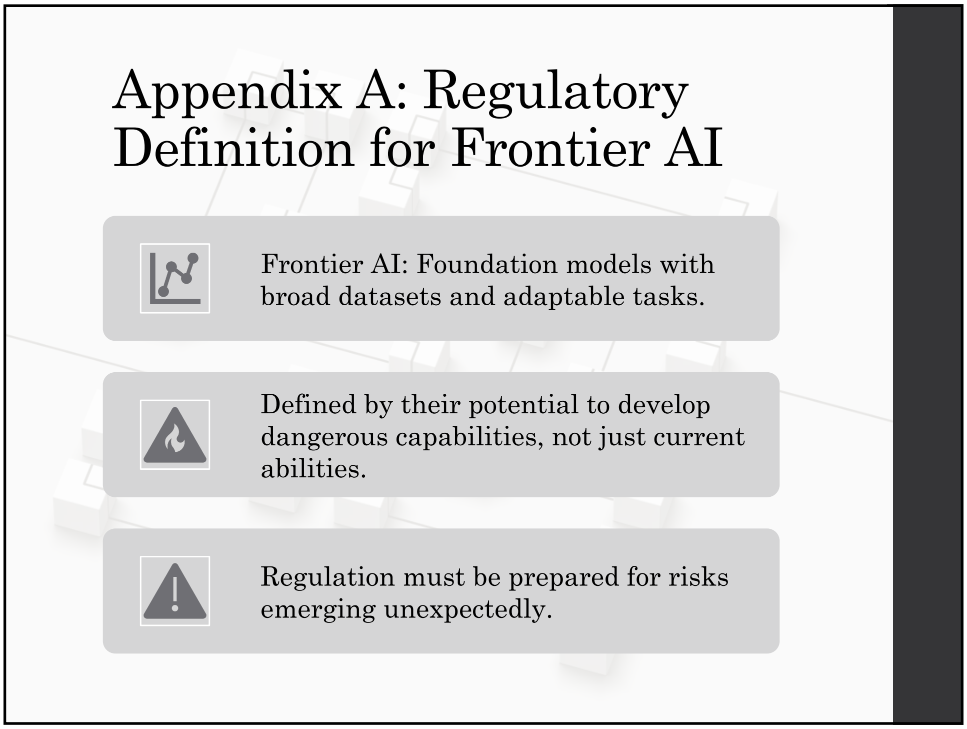
But, barriers to access are lowering—more entities may be able to train powerful models.

Regulatory challenge: Expansion could lead to increased access for malicious actors.



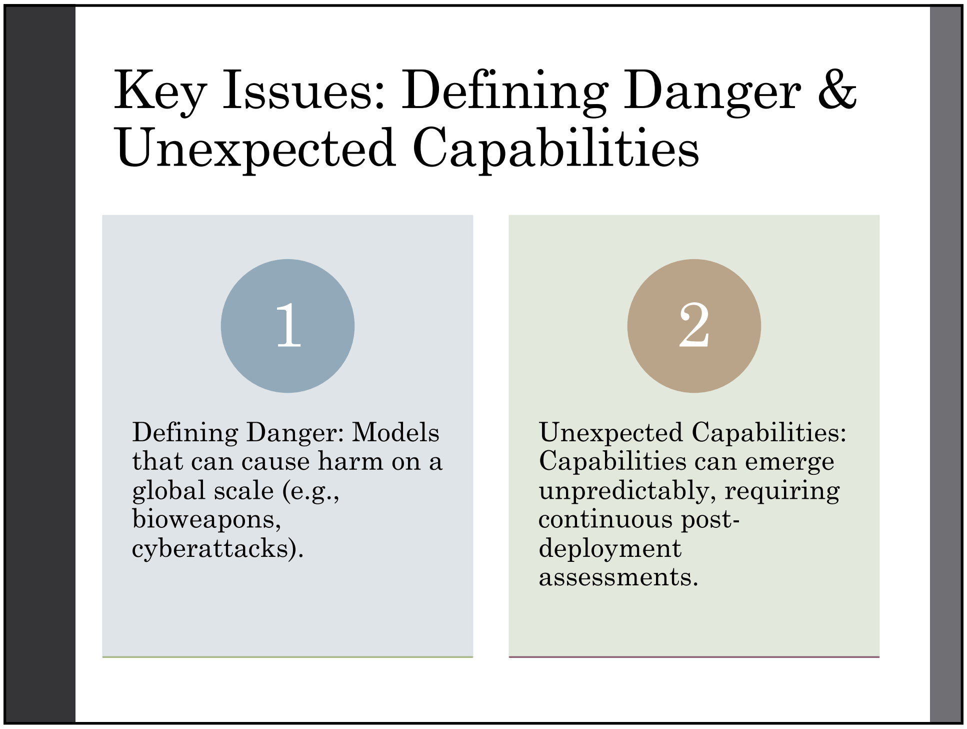
Can We Anticipate and Mitigate Frontier AI Risks?

- Risk assessments must be continuous throughout the AI lifecycle.
- External expert reviews and post-deployment monitoring are critical.
- Key risk: Proliferation—if dangerous models are released, they can spread quickly.
- Mitigation: Continuous monitoring and regulation to control model usage.



Appendix A: Regulatory Definition for Frontier AI

- Frontier AI: Foundation models with broad datasets and adaptable tasks.
- Defined by their potential to develop dangerous capabilities, not just current abilities.
- Regulation must be prepared for risks emerging unexpectedly.



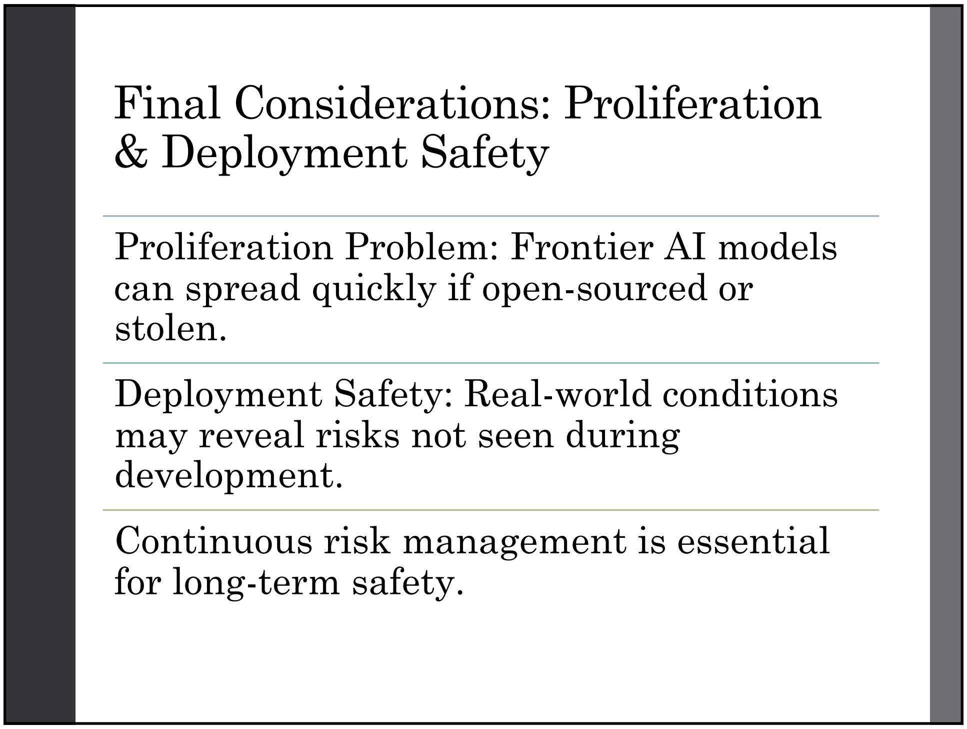
Key Issues: Defining Danger & Unexpected Capabilities

1

Defining Danger: Models that can cause harm on a global scale (e.g., bioweapons, cyberattacks).

2

Unexpected Capabilities: Capabilities can emerge unpredictably, requiring continuous post-deployment assessments.



Final Considerations: Proliferation & Deployment Safety

Proliferation Problem: Frontier AI models can spread quickly if open-sourced or stolen.

Deployment Safety: Real-world conditions may reveal risks not seen during development.

Continuous risk management is essential for long-term safety.