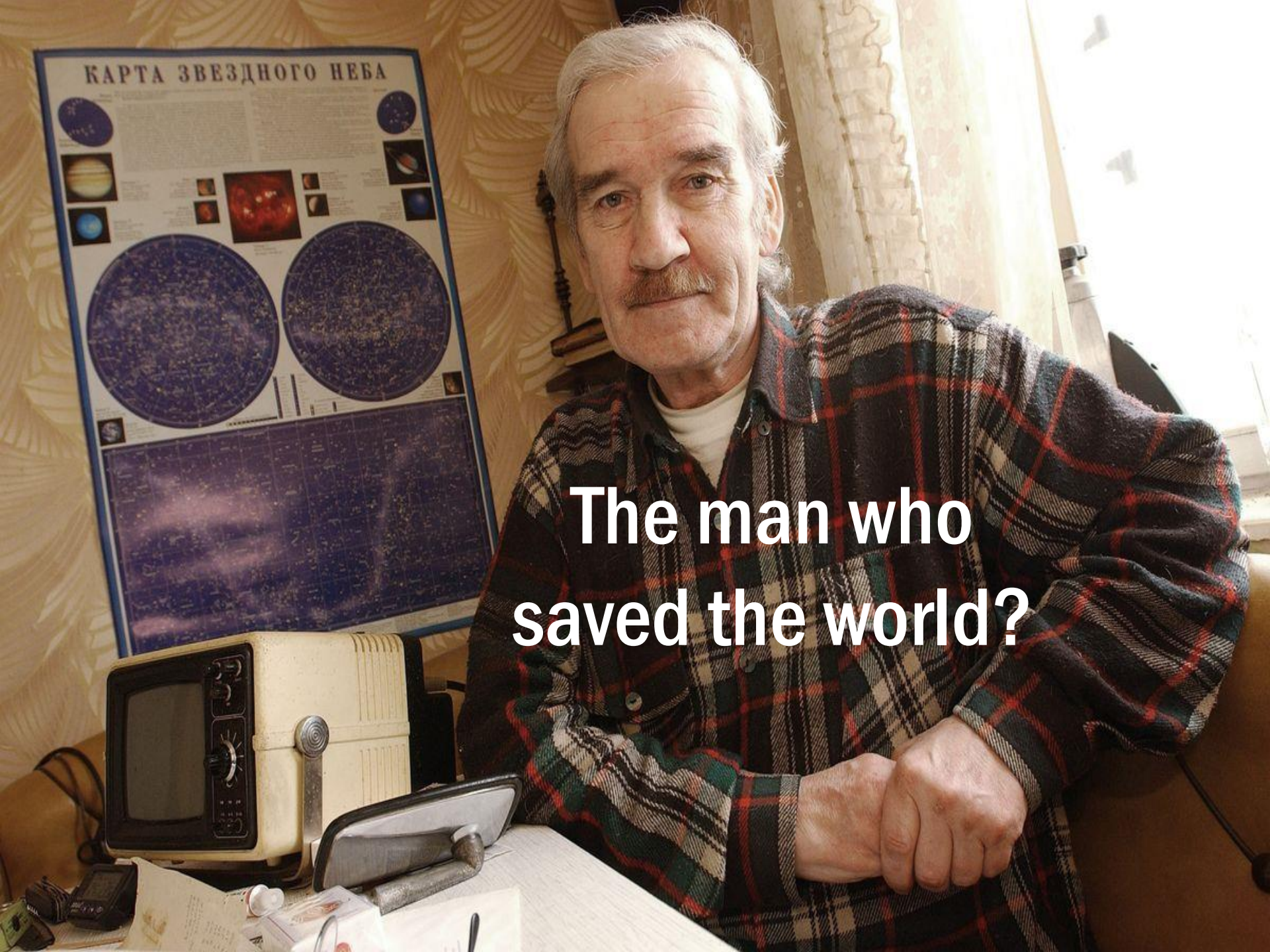
A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence

Michael C. Horowitz

University of Pennsylvania

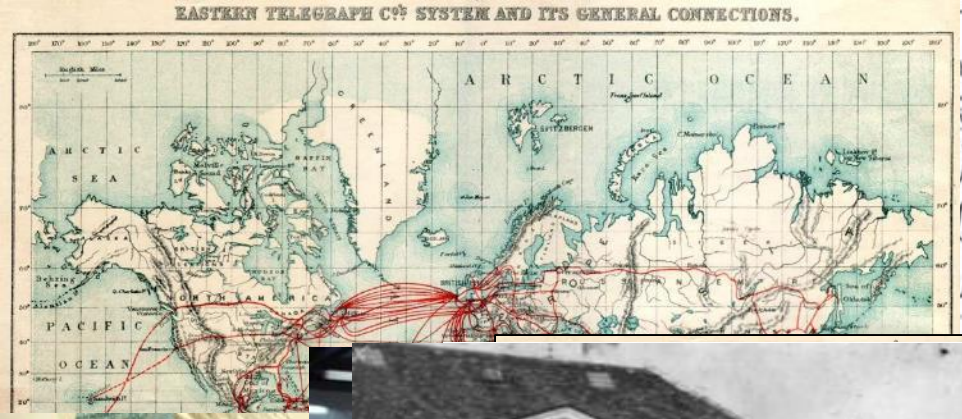
March 18, 2020

USSTRATCOM Academic Alliance

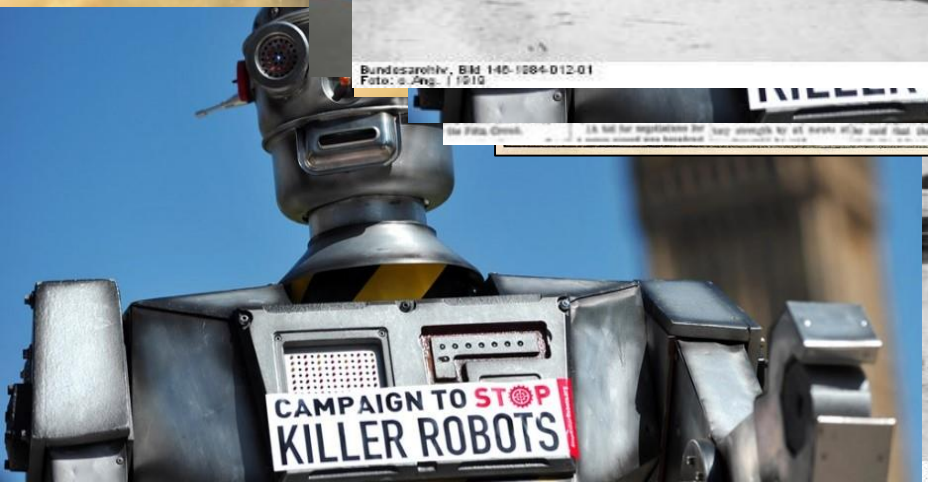

КАРТА ЗВЕЗДНОГО НЕБА
The man who
saved the world?

Presentation Overview

- Current Research Agenda
- Theory: Bias, Capabilities, and Interest in AI
- Early Warning/Command and Control
- Uninhabited Nuclear Platforms
- Conventional Military AI: Impact on Nuclear Stability
- Conclusion



Bundesarchiv, Bild 146-1984-012-01
Foto: o.Ang. | 1919



Bundesarchiv, Bild 146-1984-012-01
Foto: o.Ang. | 1919

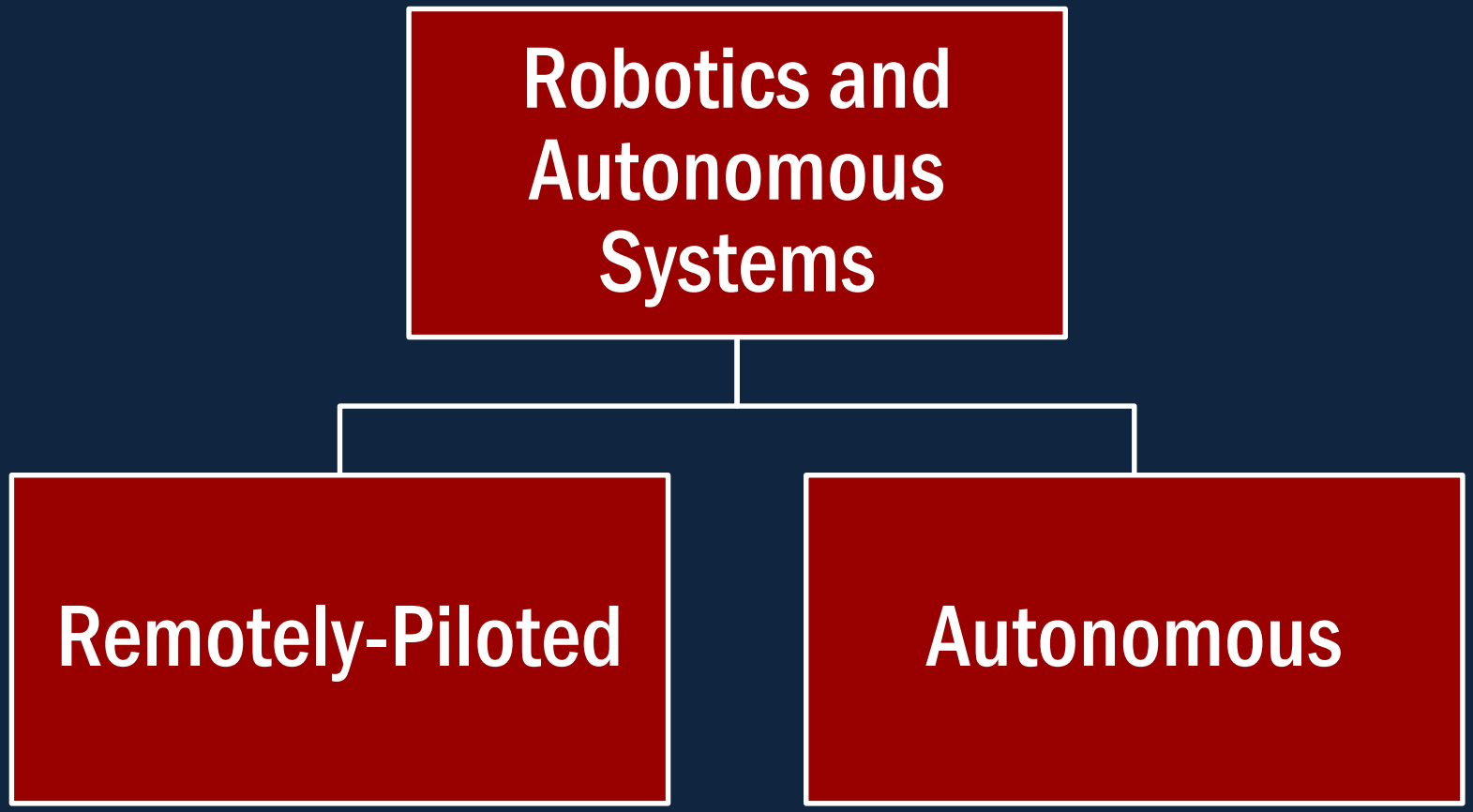
Key Role of Minerva Research Initiative

- Funding to conduct this research
- Access to key stakeholders
- Relevance for US military power

Presentation Overview

- Current Research Agenda
- **Theory: Bias, Capabilities, and Interest in AI**
- Early Warning/Command and Control
- Uninhabited Nuclear Platforms
- Conventional Military AI: Impact on Nuclear Stability
- Conclusion

Automation, Autonomy, and Artificial Intelligence



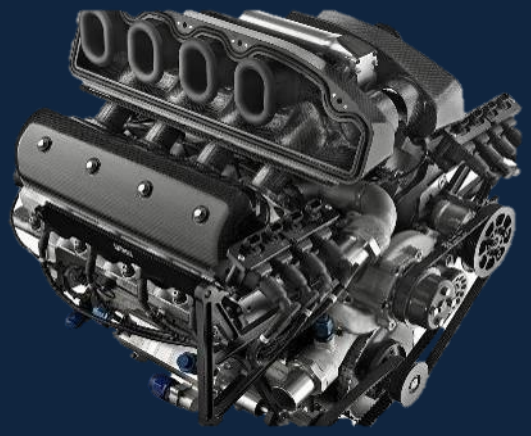

WHAT IS
A.I.?

What Is AI?

- Definition: the use of computers to simulate human behavior that requires intelligence
- Methods of AI
  - Symbolic v. Connectionist
  - Machine learning
  - Neural Networks
- Types of AI
  - Narrow
  - General Intelligence
  - Superintelligence

AI is an Enabler, not a Weapon

- Things AI can do. . . .
  - Direct physical objects
  - Process data
  - Overall information management (decision-making)
- Things AI is not
  - A gun
  - A plane
- Implication: AI is much broader than particular military technologies



Key Properties

Broad

Dual Use

Low barrier to entry

Why Pursue Autonomy or Artificial Intelligence?

Speed

Precision

Bandwidth/Hacking

Decision-Making

Brittleness of Autonomous Systems

- Narrow AI systems trained to do one thing
- Example: Alpha Go
- Challenges:
  - How do you train them (with what data)?
  - Limited potential area of operation

Trust, Confidence, and AI (1)

Trust Gap

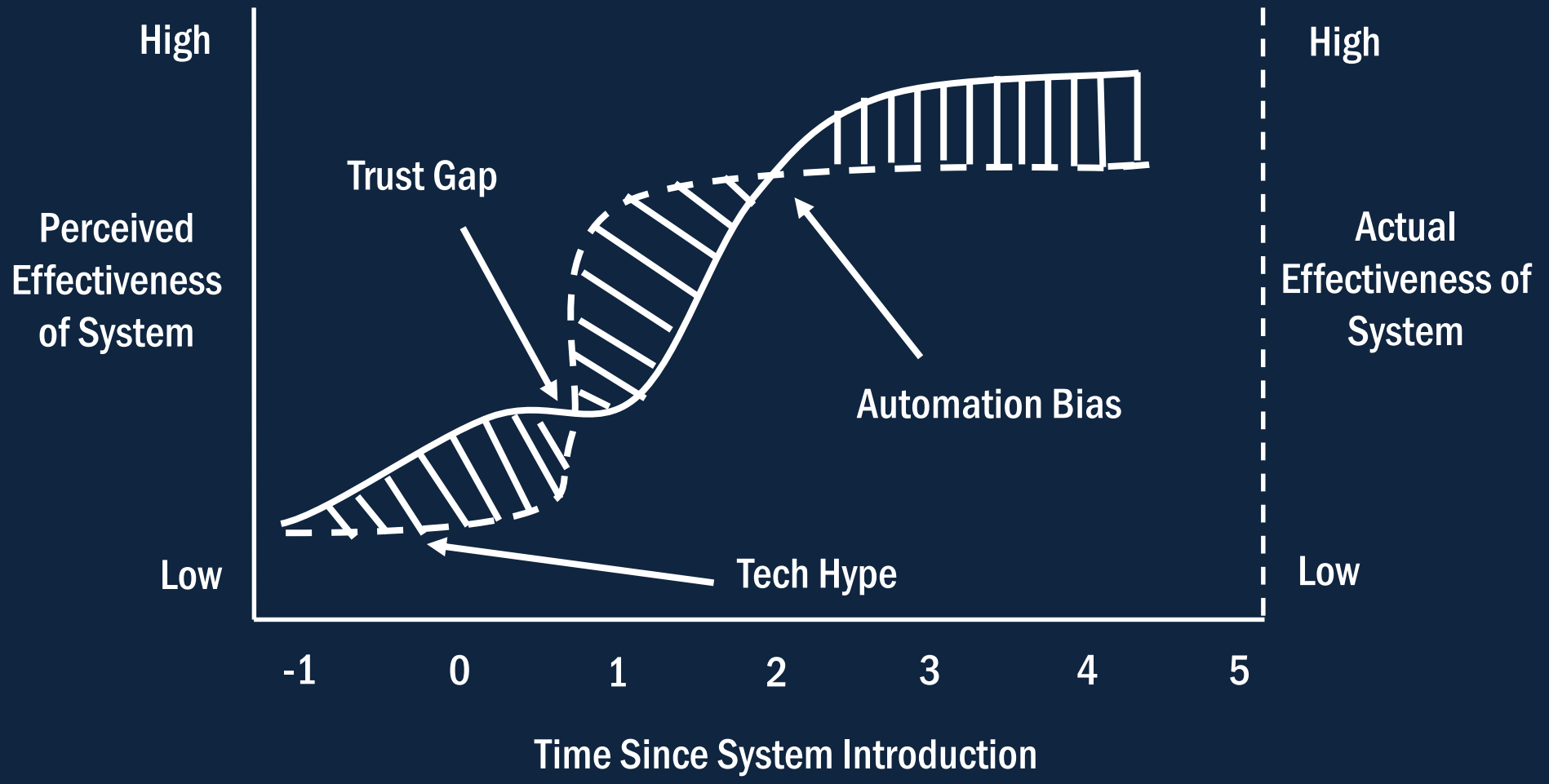
- Inability to trust machines to do work of people
- Unwillingness to deploy or properly use systems
- Example: Ground Tactical Air Controllers (MacDonald and Schneider)

V

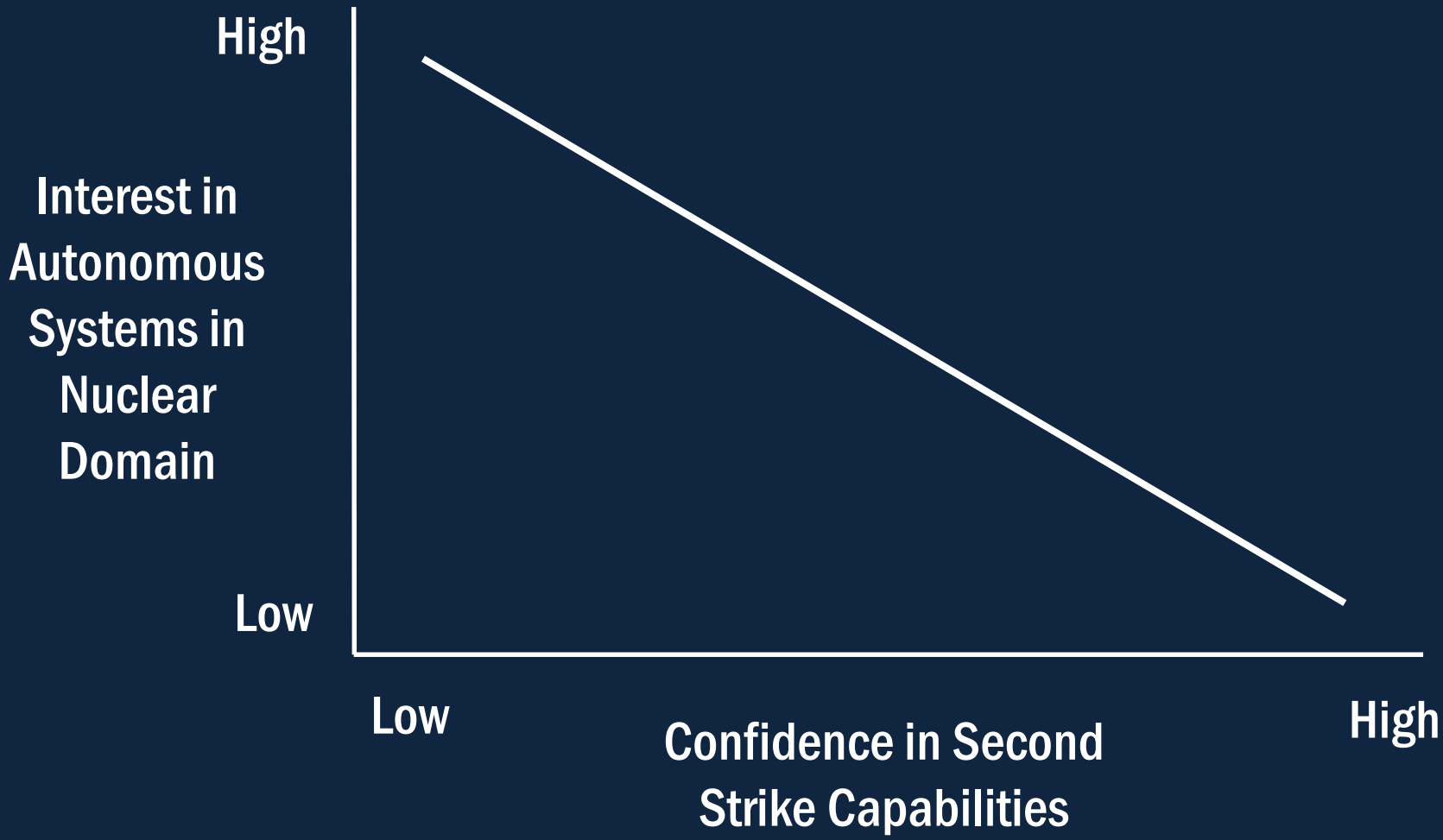
Automation Bias

- Delegation of cognitive judgment to machine – trusting too much
- Failure to question algorithms if they make mistakes
- Example: Air France Crash
- Example: Patriot Missile fratricide

Trust, Confidence, and AI (2)



Second Strike Capabilities And Autonomy

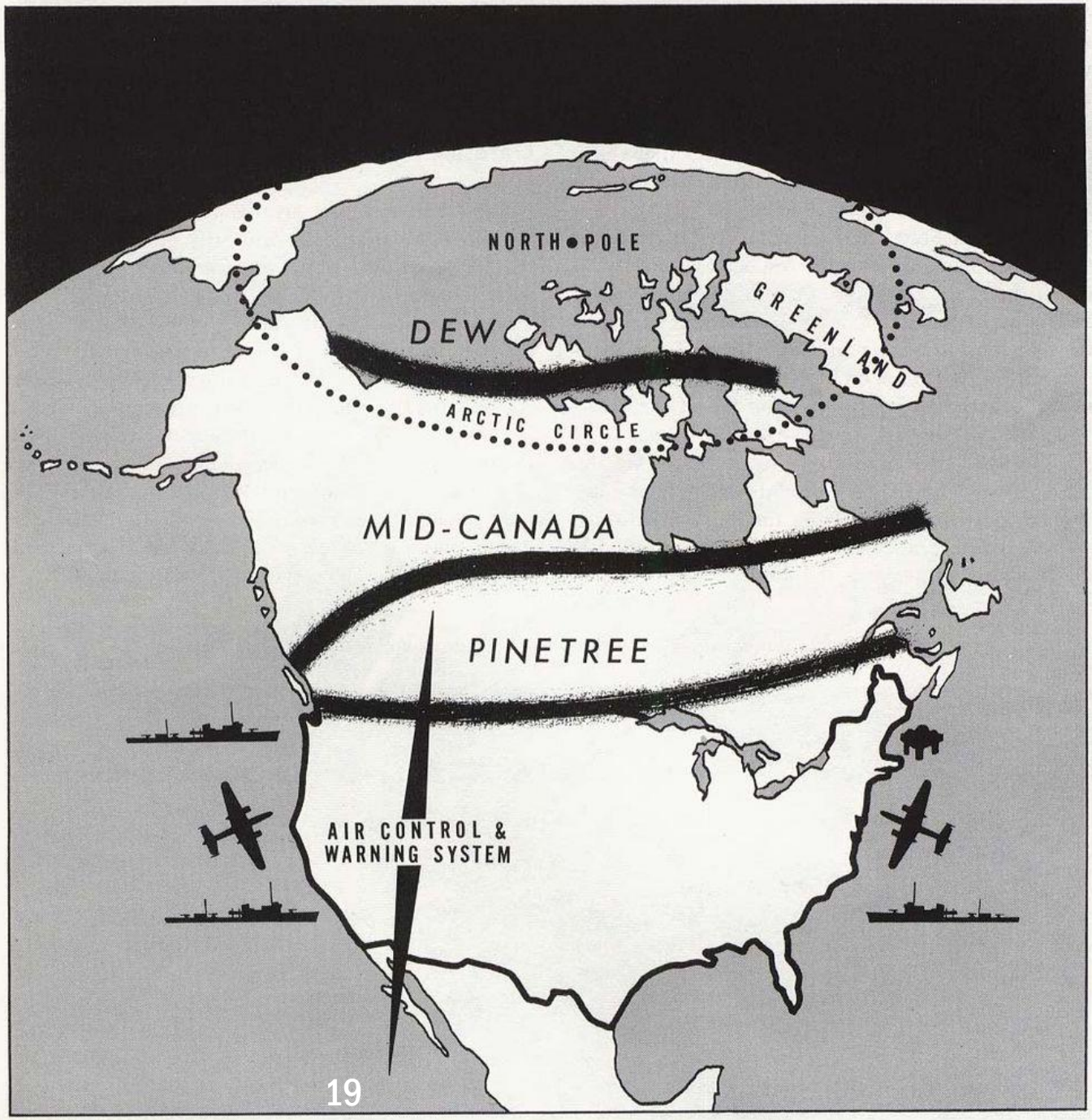


Key Driver: Competitive Pressure

Presentation Overview

- Current Research Agenda
- Theory: Bias, Capabilities, and Interest in AI
- **Early Warning/Command and Control**
- Uninhabited Nuclear Platforms
- Conventional Military AI: Impact on Nuclear Stability
- Conclusion

Early Warning + Command & Control



Existing Early Warning: Automated

- Long-range radar or satellite based alert systems
- Rapid-retargeting capability
- Communication rockets to transmit launch codes

Example: Petrov Incident

Example: Soviet Perimeter System

Autonomous Early Warning?

- Theoretical Benefits:
  - Early detection: Buys time for decision-makers
  - Reliability
- Theoretical Downsides
  - Loss of human judgment/lack of human judgment
  - Brittleness of algorithms -> false alarms

Presentation Overview

- Current Research Agenda
- Theory: Bias, Capabilities, and Interest in AI
- Early Warning/Command and Control
- **Uninhabited Nuclear Platforms**
- Conventional Military AI: Impact on Nuclear Stability
- Conclusion

Uninhabited Nuclear Platforms

- Theoretical Benefits:
  - Endurance
  - Reliability
- Theoretical Downsides
  - Cannot maintain positive human control
  - Consequences of accidents, hacking, spoofing
  - Brittleness of algorithms -> false alarms

Example: US Military

US Air Force 2013 report, *Remotely Piloted Aircraft (RPA) Vector*: **[N]uclear strike may not be technically feasible unless safeguards are developed and even then may not be considered for [unmanned aircraft systems] operations.**

General Robin Rand, head of Air Force Global Strike Command (2016): **We're planning on [the B-21] being manned. … I like the man in the loop … very much, particularly as we do the dual capable mission with nuclear weapons.**

Example: Russian military

- Perception of conventional + nuclear inferiority
- 2012 statement
- Ocean Multipurpose System 'Status-6

Example: North Korean Military

- Relatively newer nuclear power
- Conventional military inferiority
- Fear of decapitation
- Repressive regime

North Korea, **in** theory, should have greater interest in autonomy of all kinds, especially uninhabited nuclear vehicles

Presentation Overview

- Current Research Agenda
- Theory: Bias, Capabilities, and Interest in AI
- Early Warning/Command and Control
- Uninhabited Nuclear Platforms
- **Conventional Military AI: Impact on Nuclear Stability**
- Conclusion

Surveillance and Counterforce

What Would Militaries Use AI For?

Fighting At Machine Speed: Crisis Stability

- Compressed decision cycles
  - Offense
  - Defense
- Fear of losing quickly
  - First strike stability
  - Launch posture

Presentation Overview

- Current Research Agenda
- Theory: Bias, Capabilities, and Interest in AI
- Early Warning/Command and Control
- Uninhabited Nuclear Platforms
- Conventional Military AI: Impact on Nuclear Stability
- **Conclusion**

Conclusion

- The less secure the second strike capabilities, the more a country is likely to consider autonomous systems within their nuclear weapons complex
- Some risk associated with greater automation in early warning
- Potentially large risk associated with impact of conventional military uses of autonomy on crisis stability