SPAR Project 2

MENTOR: Zhamilia Klycheva, Ph.D.

PROJECT TITLE: From Near Miss to Pause: Historical Warning Shots and AI Governance

PROJECT DESCRIPTION

This project investigates how historical "warning shots" — near-miss events in high-risk domains such as nuclear command-and-control, aerospace, biosafety, and industrial safety — have triggered (or failed to trigger) institutional change. Drawing inspiration from Guest (2023) and Barnett, Scher (2025), and the organizational-learning literature, the project aims to identify the structural factors that determine whether a warning shot produces meaningful reform, escalation, or complacency. The goal is to extract a generalizable framework for understanding institutional learning under risk.

Building on 2–3 carefully selected historical warning shots (e.g., Petrov nuclear false alarm, Challenger O-ring precursor warnings, biosafety lab containment failures), mentees will analyze the clarity of the signal, the incentives of key actors, the political context, and the mechanisms that shaped the response. The project then applies these insights to contemporary AI governance — especially US–China frontier-model competition — to explore under what conditions a modern AI warning shot would be sufficient for states or labs to halt, slow, or reform their development trajectory.

The project is primarily qualitative and conceptual. It relies on structured case studies, factor analysis, and scenario-based reasoning rather than formal game theory or quantitative modeling. The final deliverables include a cross-case comparison matrix, a theory-informed "Warning Shot Response Framework," AI-relevant scenario applications, and a short policy brief outlining mechanisms to make institutional learning more reliable in the age of advanced AI.

Key Research Questions

(Not all must be answered — mentees may prioritize based on interest and feasibility.)

1. Definition & Classification: What constitutes a meaningful “warning shot” in high-risk domains, and how should different types of near-misses be categorized?
2. Success vs Failure: Why do some warning shots lead to safety reforms, while others are ignored or misinterpreted?
3. Causal Mechanisms: What organizational, political, and psychological factors predict whether institutions learn from near-miss events?
4. Transferability: Which lessons from historical warning shots can meaningfully apply to modern AI governance?
5. Forecasting Impact: Under what conditions would a major power (e.g., the US or China) consider an AI incident dangerous enough to slow, halt, or restructure AI development?
6. Governance Design: What mechanisms (incident reporting mandates, international hotlines, safety audits, shared investigation bodies) increase the likelihood that AI-related warning shots lead to stability rather than escalation?

Tentative Project Structure

1. Warning Shot Typology (Core Conceptual Output)
    A. Define 4–6 categories of warning shots (e.g., technical failure, misinterpretation incident, governance breakdown, procedural lapse).
    B. Establish classification dimensions:
        a. signal clarity
        b. severity and potential impact
        c. attribution certainty
        d. visibility and media attention
        e. competing political incentives
    C. Produce a short typology section.
2. Historical Case Studies (2–3 cases, structured analysis)
    A. For each selected warning shot:
        a. concise narrative (what happened)
        b. actors involved
        c. clarity of the warning signal
        d. institutional incentives (aligned / misaligned)
        e. political, economic, or bureaucratic pressures
        f. decision-making structure
        g. resulting actions (reform, escalation, neglect)
        h. long-term impact
    D. Cases may include:
        a. 1983 Petrov false alarm (restraint under ambiguity)
        b. Challenger O-ring precursor warnings (ignored signals)
        c. BSL-3 lab containment failures (partial learning)
        d. Three Mile Island precursors (regulatory learning)

The mentor will help mentees select appropriate cases.

3. Cross-Case Factor Analysis
    A. Identify recurring variables across cases (e.g., ambiguity, incentives, institutional slack, centralization, reputational risk).
    B. Build a comparison matrix scoring each case along these factors.
    C. Extract general patterns and “necessary + sufficient conditions” for meaningful learning.

4. Development of a “Warning Shot Response Framework”
    A. Convert the cross-case insights into a conceptual model describing how institutions interpret, process, and react to near-miss events.

B. Distinguish enabling conditions for reform from blocking conditions such as bureaucratic inertia, political liability concerns, or ambiguity discounting.
C. Formulate the general mechanisms that explain why some warning shots trigger reform while others do not.

5. Application to Contemporary AI Governance
   A. Identify 2–3 plausible AI warning-shot scenarios such as misaligned model behavior, an AI-enabled cyber incident, or a major model-weight leak.
   B. Apply the Warning Shot Response Framework to assess how frontier labs, the United States, China, and other actors might interpret and respond to each scenario.
   C. Evaluate which conditions would make a modern warning shot strong enough to alter the pace of AI development or adjust competitive dynamics.
   D. Analyze how strategic pressures and misaligned incentives in US–China competition might suppress institutional learning.

6. Governance and Policy Implications
   A. Identify interventions that could improve institutional learning from AI-related near-misses, such as standardized incident reporting, international safety hotlines, and shared investigation procedures.
   B. Evaluate how these mechanisms perform across the AI warning-shot scenarios.
   C. Produce a final synthesis with recommendations for governments, frontier labs, and multilateral bodies.

7. Final Deliverables

A. Warning Shot Typology
B. Two Historical Case Studies
C. Cross-Case Factor Matrix
D. Warning Shot Response Framework
E. AI-relevant scenario applications
F. 2–3 page policy brief
G. 10–15 slide presentation deck
H. OPTIONAL: 10–20 page working paper (time permitting)

12-Week Proposed Timeline

Weeks 1–3:
• Select historical cases
• Draft warning shot typology

- Conduct initial literature review
- Build preliminary factor list & research plan

Weeks 4–6:
- Complete two historical case studies
- Begin cross-case comparison matrix
- Identify emerging causal patterns

Weeks 7–9:
- Finalize factor analysis
- Develop Warning Shot Response Framework
- Apply framework to 2–3 contemporary AI warning-shot scenarios

Weeks 10–12:
- Draft policy brief
- Build slide deck
- Final synthesis & presentation preparation