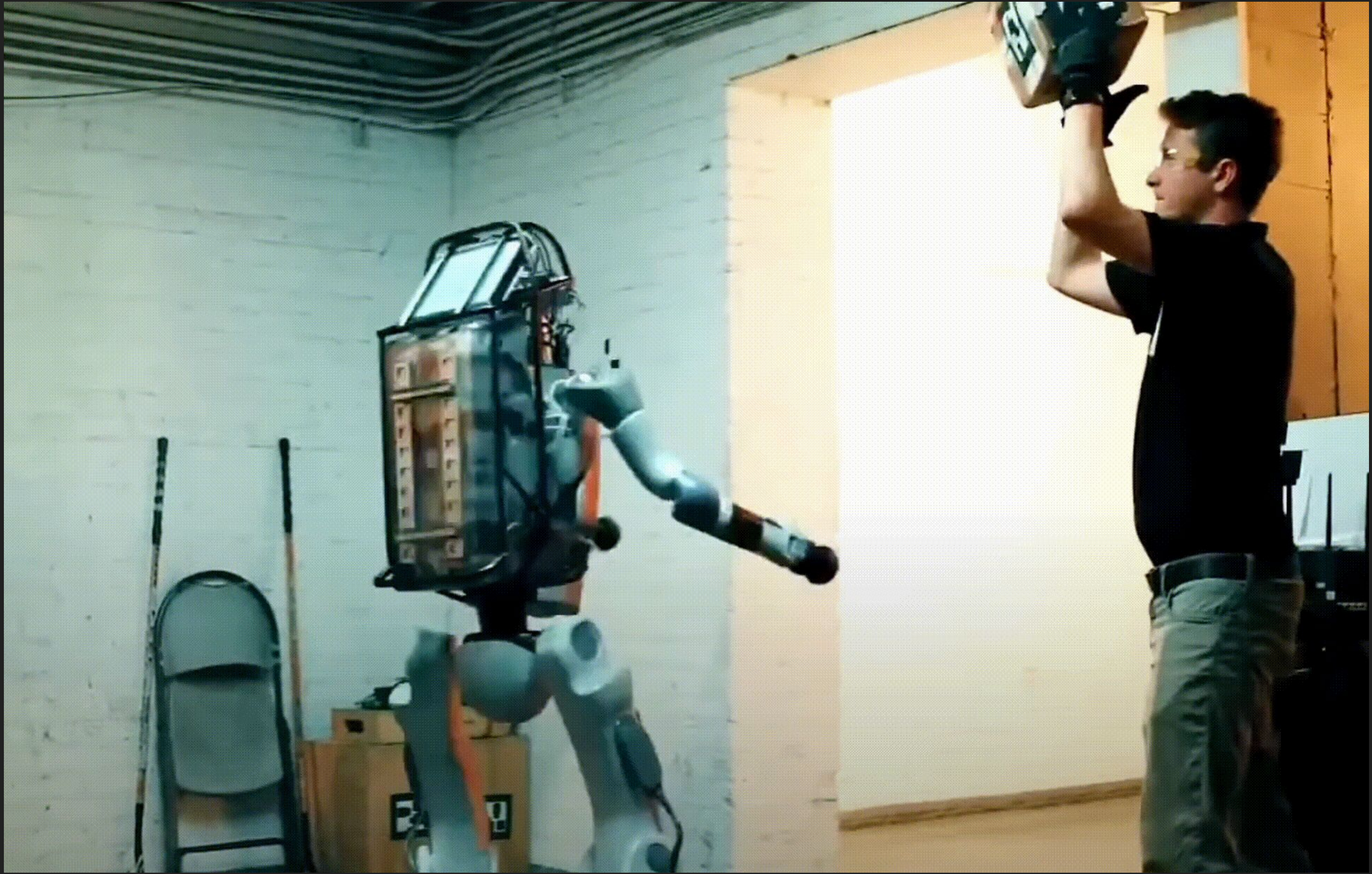
Autonomous Weapons Systems (AWS)

Simon Rupp, Aryavrat Gupta, Will Boudy



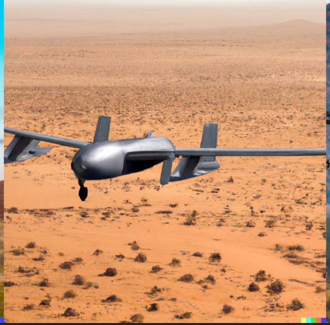
Presentation Roadmap

- Overview of Autonomous Weapon Systems (AWS)
- Ethical Problems
- Current State of AWS
- Analysis of Existing Solutions
- Proposed solutions (Technical and Policy/Governance)
- Future prospects

Terms: Autonomous Weapon Systems (AWS), Lethal Autonomous Weapon Systems (LAWS), Semi-Autonomous Weapon Systems (SAWS), Lethal Semi-Autonomous Weapon Systems (LSAWS).

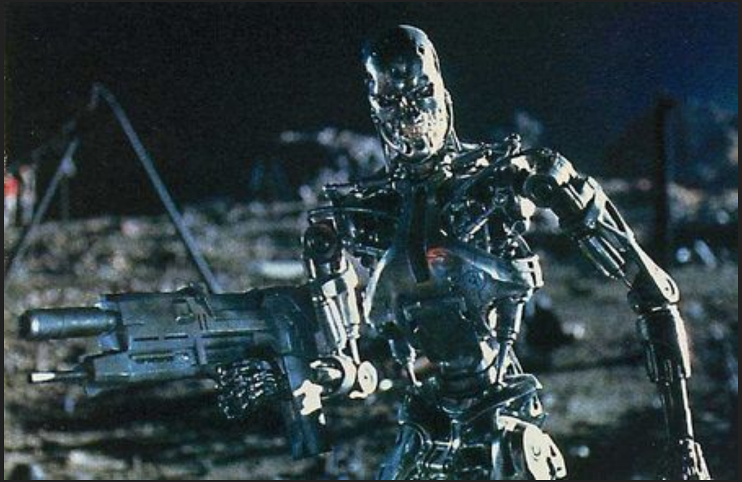
Understanding AWS

- **Definition**: Systems that select and engage targets without human intervention.
- **Examples**: Drones, loitering munitions, robotic turrets, autonomous missiles, anti-aircraft defense, decision-support systems
- Provides substantial advantages: Reduce human casualties, quicker response time, increased precision.



Ethical Problems with AWS

- Erosion of human accountability (moral judgement and human autonomy)
- Escalation of conflict (safety concerns, lack of transparency in decision-making, algorithmic bias)
- Violations of LOAC (distinction, proportionality, military necessity)



Current State of AI in Military - AI Arms Race

- Nations recognize the strategic advantages of AI-enabled military technology.
- **Specific case**: US vs China - two leading global powers and adversaries at the forefront of the AI arms race.
- DARPA, PLA aggressively developing AI-enabled weapons
- **US policy**: *Political Declaration on Responsible Military Use of AI and Autonomy*, DoD Directive 3000.09
- **China policy**: Advocates for the complete ban of fully autonomous lethal weapons systems

Past Failures and Potential Threats

In 2021, a U.S. drone strike in Afghanistan mistakenly killed 10 civilians. It was conducted by a remotely piloted drone (operated by human personnel) using information from intelligence sources. However, this incident highlights key risks that could be amplified in fully autonomous systems such as **reliance on flawed intelligence** and **lack of accountability**.

Potential threats:

- Escalation of unintended conflict
- Cybersecurity threats
- Loss of command-and-control oversight

Claim

The development of autonomous weapon systems (AWS) **must align with international ethical standards** to ensure transparency, human accountability, and conflict prevention. Starting or **escalating wars through autonomous technology is condemnable** and undermines global stability. While fully autonomous weapons pose significant risks, the **controlled use of semi-autonomous systems for defensive purposes can be acceptable** if guided by robust legal frameworks and global collaboration.

Existing Solutions: Do They Work?

Technical measures:

- Kill Switches - Fail-safes to disable autonomous weapons if they malfunction. Vulnerable to cyberattacks.
- Human-in-the-Loop Systems - Require human approval before lethal actions. May not function effectively in high-risk and high-speed scenarios.

Policy measures:

- UN Guidelines - Non-binding resolutions (lack of enforcement mechanisms)
- Geneva Conventions - Principles of distinction and proportionality apply but do not account for AI-driven systems (outdated, lack of binding regulations, insufficient global consensus)
- REAIMS (Responsible AI in Military Systems) - Aims to promote ethical use of AI in military contexts through international cooperation by focusing on transparency, accountability and human rights (still in early stages with no universal adoption or enforcement)

What's missing?

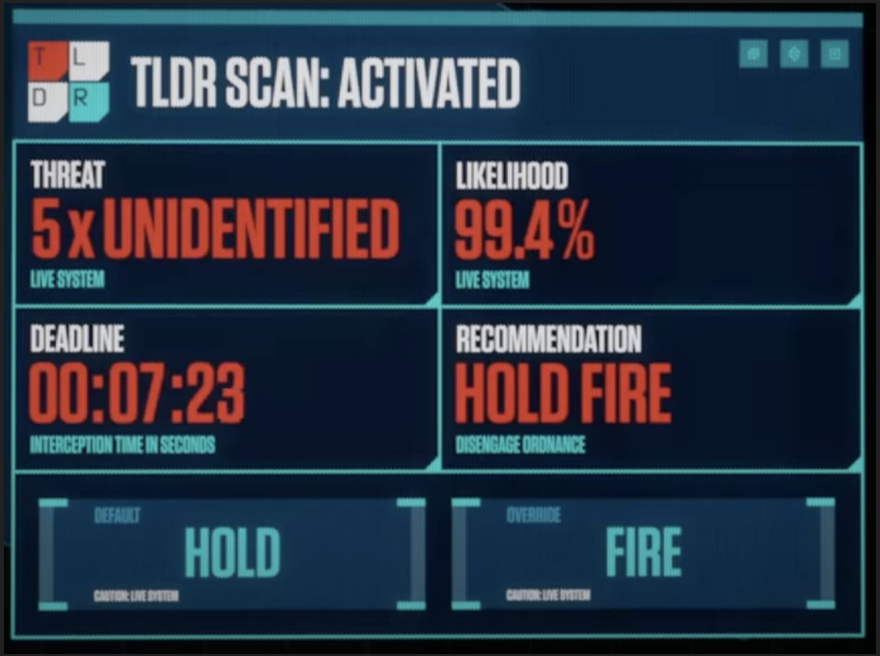
- Clear accountability mechanisms for unintended harm caused by AWS.
- Effective enforcement of existing guidelines and development of binding treaties.
- Provisions for real-time oversight and global monitoring of autonomous weapons systems.

Proposed Technical Solution

SAWS Dashboard

- Provides data visualization and analytics to aid military decision-making
- Provides deployment capabilities to allow for rapid response
- Project Maven
- Problematic dashboards may
  - “Gamify” and dehumanize war
  - Have minimal human operator input/control
  - Lack ethical analysis and allow for escalation

Bad SAWS Dashboard



TLDR SCAN: ACTIVATED

THREAT

5 x UNIDENTIFIED

LIVE SYSTEM

LIKELIHOOD

99.4%

LIVE SYSTEM

DEADLINE

00:07:23

INTERCEPTION TIME IN SECONDS

RECOMMENDATION

HOLD FIRE

DISENGAGE ORDNANCE

DEFAULT

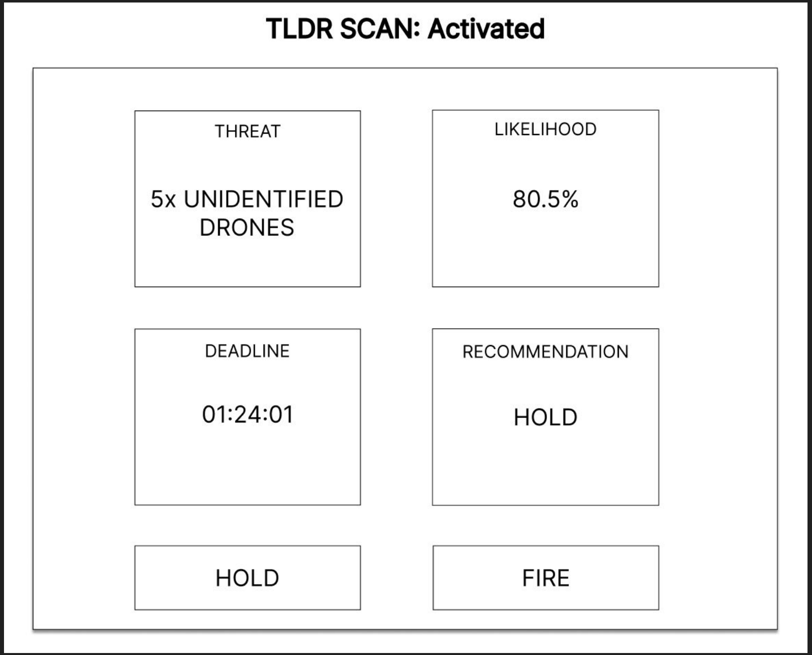
HOLD

CAUTION: LIVE SYSTEM

OVERRIDE

FIRE

CAUTION: LIVE SYSTEM



TLDR SCAN: Activated

THREAT

5x UNIDENTIFIED DRONES

LIKELIHOOD

80.5%

DEADLINE

01:24:01

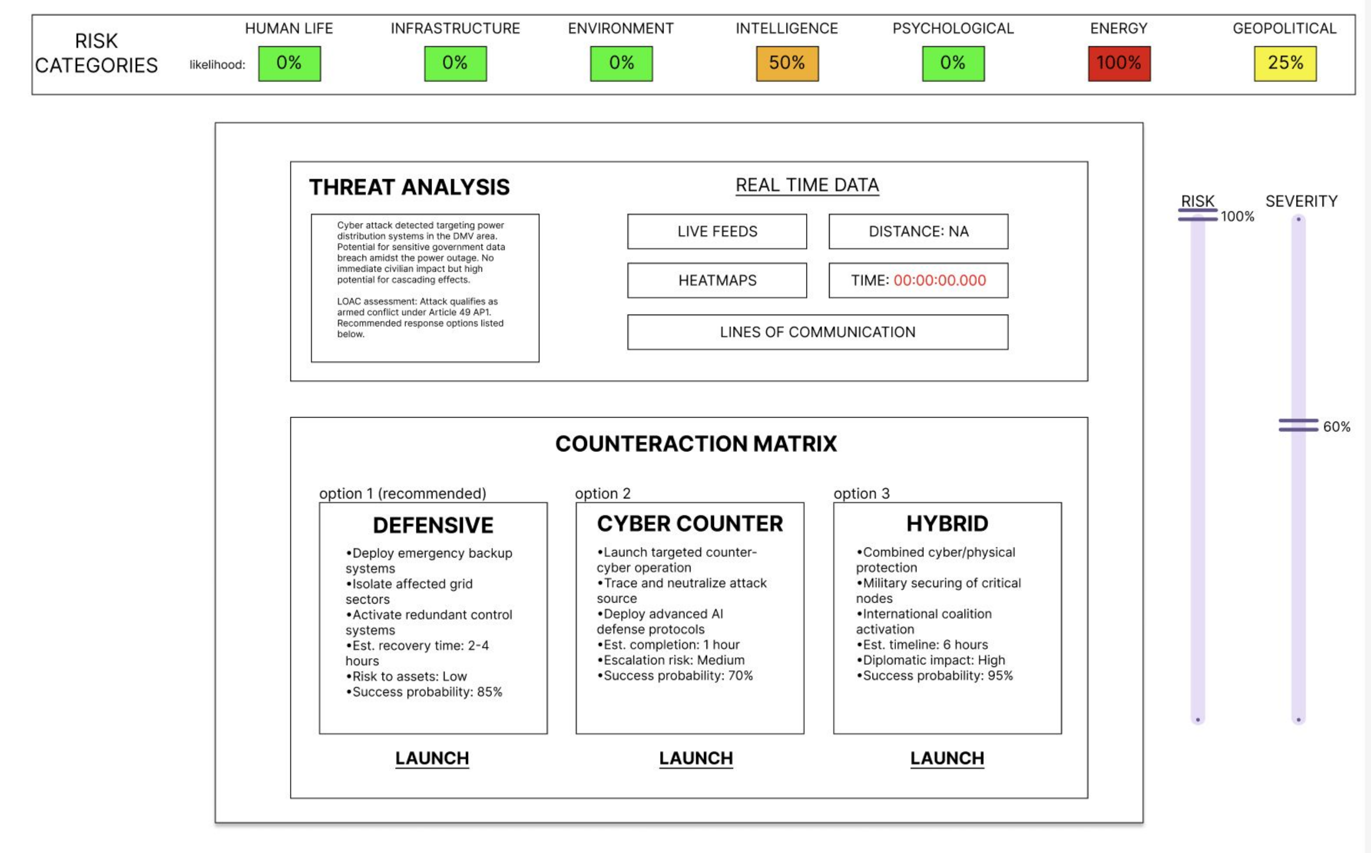
RECOMMENDATION

HOLD

HOLD

FIRE

Good SAWS Dashboard



RISK CATEGORIES		HUMAN LIFE	INFRASTRUCTURE	ENVIRONMENT	INTELLIGENCE	PSYCHOLOGICAL	ENERGY	GEOPOLITICAL
	likelihood:	0%	0%	0%	50%	0%	100%	25%

THREAT ANALYSIS

Cyber attack detected targeting power distribution systems in the DMV area. Potential for sensitive government data breach amidst the power outage. No immediate civilian impact but high potential for cascading effects.

LOAC assessment: Attack qualifies as armed conflict under Article 49 AP1. Recommended response options listed below.

REAL TIME DATA

DISTANCE: NA

TIME: 00:00:00.000

COUNTERACTION MATRIX

option 1 (recommended)

DEFENSIVE

- Deploy emergency backup systems
- Isolate affected grid sectors
- Activate redundant control systems
- Est. recovery time: 2-4 hours
- Risk to assets: Low
- Success probability: 85%

LAUNCH

option 2

CYBER COUNTER

- Launch targeted counter-cyber operation
- Trace and neutralize attack source
- Deploy advanced AI defense protocols
- Est. completion: 1 hour
- Escalation risk: Medium
- Success probability: 70%

LAUNCH

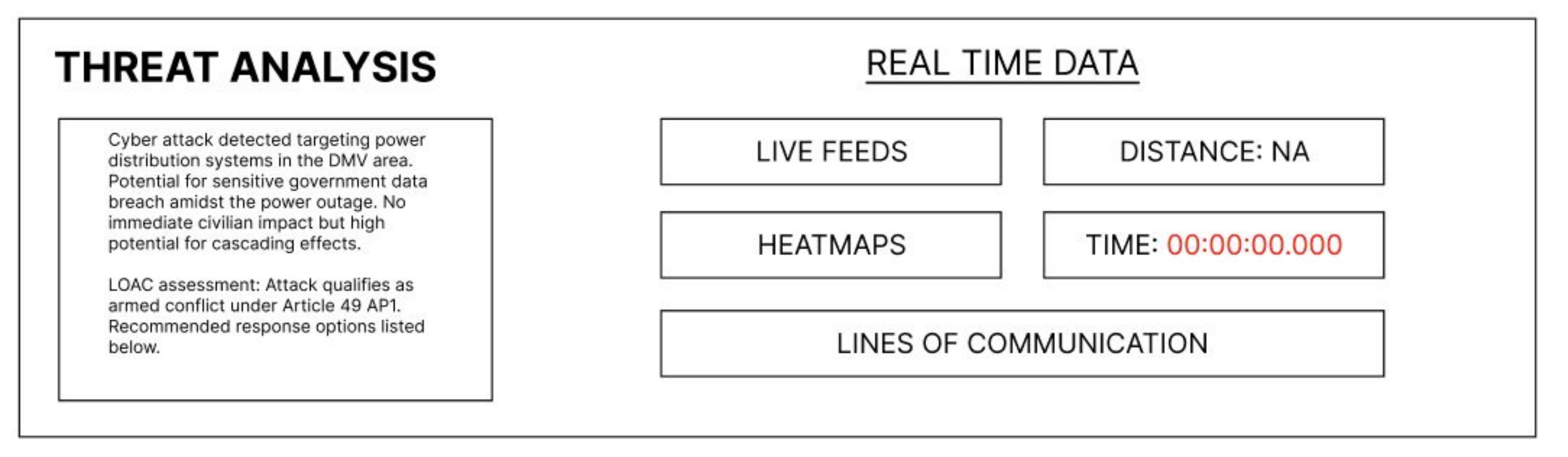
option 3

HYBRID

- Combined cyber/physical protection
- Military securing of critical nodes
- International coalition activation
- Est. timeline: 6 hours
- Diplomatic impact: High
- Success probability: 95%

LAUNCH

Data Fusion



THREAT ANALYSIS

Cyber attack detected targeting power distribution systems in the DMV area. Potential for sensitive government data breach amidst the power outage. No immediate civilian impact but high potential for cascading effects.

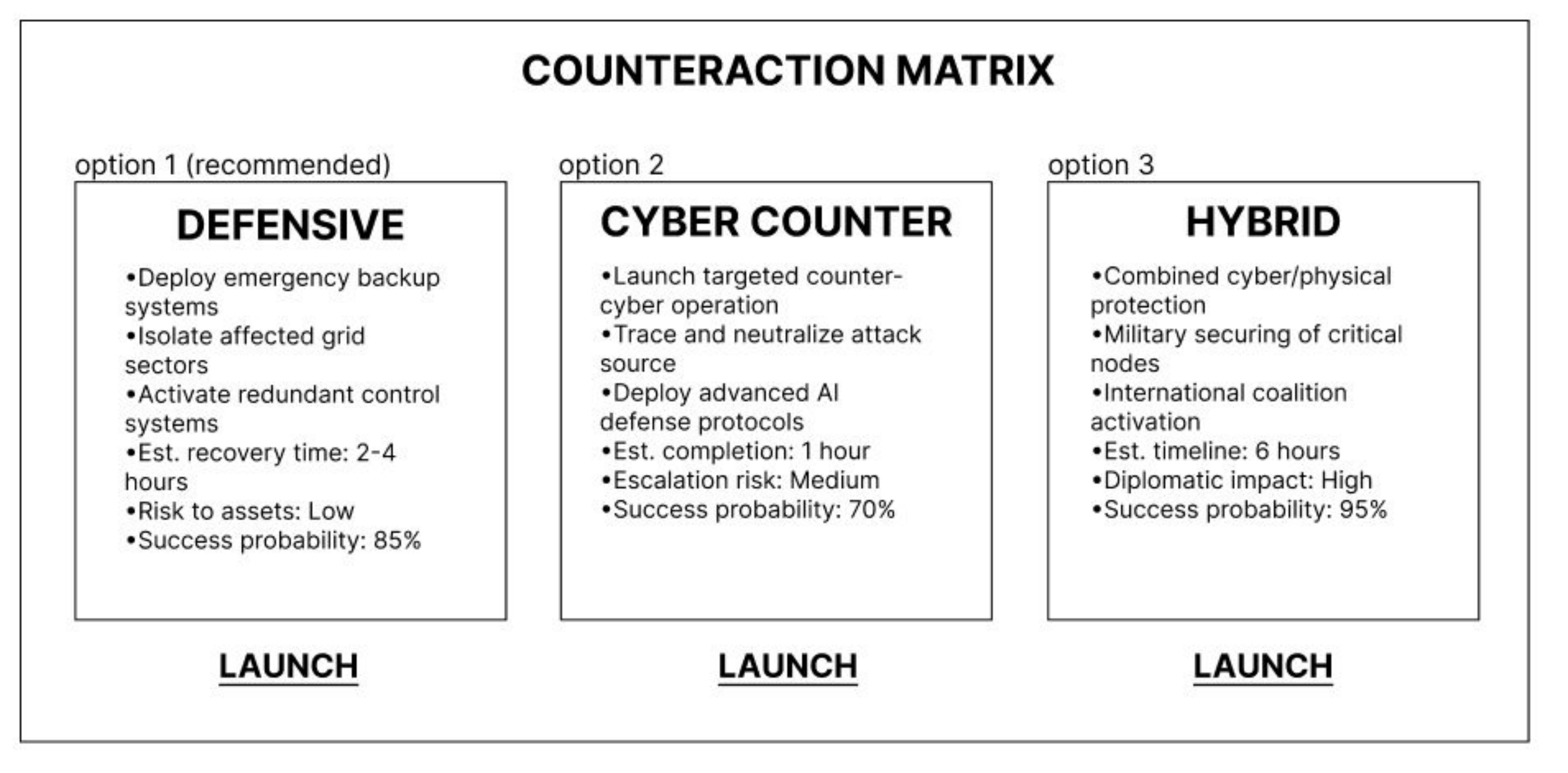
LOAC assessment: Attack qualifies as armed conflict under Article 49 AP1. Recommended response options listed below.

REAL TIME DATA

DISTANCE: NA

TIME: 00:00:00.000

Response Options



COUNTERACTION MATRIX

option 1 (recommended)

DEFENSIVE

- Deploy emergency backup systems
- Isolate affected grid sectors
- Activate redundant control systems
- Est. recovery time: 2-4 hours
- Risk to assets: Low
- Success probability: 85%

LAUNCH

option 2

CYBER COUNTER

- Launch targeted counter-cyber operation
- Trace and neutralize attack source
- Deploy advanced AI defense protocols
- Est. completion: 1 hour
- Escalation risk: Medium
- Success probability: 70%

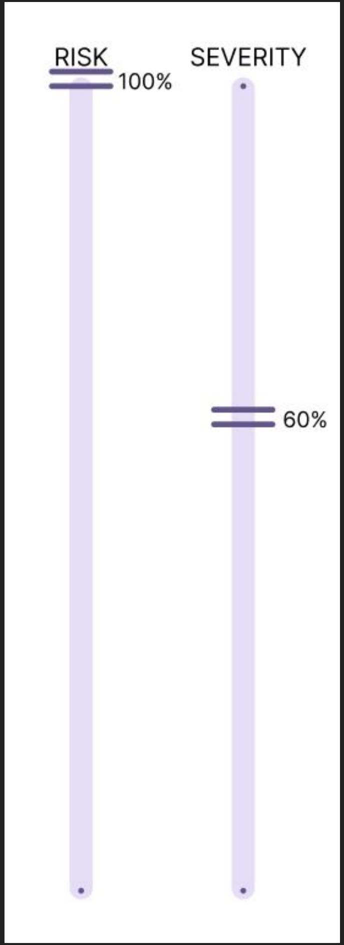
LAUNCH

option 3

HYBRID

- Combined cyber/physical protection
- Military securing of critical nodes
- International coalition activation
- Est. timeline: 6 hours
- Diplomatic impact: High
- Success probability: 95%

LAUNCH



Dynamic Interface

- Risk = likelihood of threat manifesting
- Severity = danger level
- Auto-estimated
- Manual option

Proposed Legal Solutions

Governance and Regulation

1. International Regulatory Body

- Modeled on the Nuclear Non-Proliferation Treaty (NPT).
- **Functions**:
    - **Non-Proliferation**: Monitor and restrict critical technologies (AI algorithms, advanced hardware).
    - **Permissible Use**: Define defensive use with mandatory human oversight.
    - **Transparency**: Require member states to disclose activities for accountability.

2. Enhancing REAIMS Framework

- **Mandated Human Oversight**: Humans control critical decisions like targeting.
- **LOAC Compliance**:
    - Pre-deployment audits.
    - Real-time monitoring mechanisms.
- **Ethical Safeguards**: Prioritize humanitarian considerations in AI algorithms.

Design and Deployment

3. Guidelines for AWS Design

- **Strict Regulations**:
  - Target verification systems for distinction between combatants and civilians.
  - Proportionality tools to minimize collateral damage.
- **Failsafe Mechanisms**:
  - Automatic suspension for LOAC violations.
- **Certification and Transparency**:
  - Independent testing for compliance.
  - Auditable code and operational logs.

4. Restriction to Defensive Use Only

- **Defensive Scenarios**:
  - Protect civilians, infrastructure, respond to threats.
  - Prohibitions on offensive use and preemptive strikes.
- **Pre-Deployment Approval**:
  - Review by an international regulatory body.
- **Real-Time Monitoring**:
  - Dashboards with LOAC compliance checks and risk metrics for commanders.

Accountability and Access

5. Accountability Mechanisms

- **Chain of Responsibility**: Liability for developers, manufacturers, and operators.
- **Incident Reporting and Review**: Oversight by international tribunals.
- **Penalties for Misuse**: Sanctions, financial penalties, development bans, or criminal charges.

6. Regulation of Access and Manufacturing

- **Export Controls**: Licensing for critical components; ban on sales to non-state actors or nations outside treaties.
- **Global Registry**: Track development and deployment for transparency.
- **Secure Development**: Ensure secure facilities for creating and testing AWS.

7. Incorporating LOAC Principles

- **Distinction**: Advanced algorithms to separate combatants from civilians.
- **Proportionality**: Minimize civilian harm relative to military advantage.
- **Military Necessity**: Use justified by clear objectives to prevent escalation.

Conclusion and Next Steps

- Believe Semi-Autonomous Weapons Systems to be acceptable
  - Always human-in-the-loop
- Design solution aims to fix major flaws
  - Autonomy, escalation of conflict
- Governance/Policy aims to bring a unified approach to development and use
  - Currently differs by country, lack of enforcement/accountability

References

- *Artificial escalation*. Future of Life Institute. (2024, May 29). https://futureoflife.org/project/artificial-escalation/
- Hiebert, K. (2024, January 15). The United States quietly kick-starts the Autonomous Weapons Era. Centre for International Governance Innovation. https://www.cigionline.org/articles/the-united-states-quietly-kick-starts-the-autonomous-weapons-era/
- Military Trends. (2023, June 18). *Military Technologies of 2024*. https://www.youtube.com/watch?v=hvJ_X_hA6ms&t=307s
- Review of the 2023 US policy on autonomy in weapons systems. Human Rights Watch. (2023, February 16). https://www.hrw.org/news/2023/02/14/review-2023-us-policy-autonomy-weapons-systems#_ftnref1
- Saballa, J. (2024, May 30). *Palantir awarded $480m to prototype US Army's "Maven" ai battlefield analyzer*. The Defense Post. https://thedefensepost.com/2024/05/30/palantir-maven-battlefield-analyzer/
- U.S. Department of Defense. (2023, January 25). Directive 3000.09: Autonomy in weapon systems. Office of the Under Secretary of Defense for Policy. Retrieved January 26, 2023, from https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.PDF