Mitigating Catastrophic AI Risks: Lessons from a Framework for Proactive Policy

Daniel Kroth[1] and Richard Mallah[1]

Executive Summary

AI technologies are rapidly advancing, offering opportunities for both unprecedented benefits and catastrophic risk. The transformative pace and scope of these changes challenge existing governance frameworks, requiring policymakers to think creatively to prevent and mitigate the most consequential AI-related threats. This report introduces the Proactive Policy Framework, a comprehensive framework developed by the Center for AI Risk Management and Alignment (CARMA) Public Security Policy (PSP) team to assess such threats. It applies it to three envisioned futures, including a conflict escalation crisis and an economic crisis, to assess its utility and generate actionable policy insights.

The Proactive Policy Framework (PPF) is both an ideational tool for exploring AI threats and a data collection mechanism designed to facilitate further analysis. The framework maps AI risks to threat stories, emerging technological capabilities, relevant responding entities, and the legal authorities that empower (or, critically, do not empower) first and second lines of defense against catastrophic AI outcomes.

The bulk of this report applies the PSP framework to three unique case studies highlighting different governance challenges related to highly consequential AI outcomes.

The first case relates to multilayered crises wherein AI system failures compound in the context of the financial system. In this case, AI-based trading systems trigger a global financial meltdown, which is further exacerbated by a sophisticated AI-driven disinformation campaign. The case underscores the need for stronger financial safeguards, such as cross-exchange monitoring and enhanced regulatory oversight of AI algorithms, to prevent cascading failures in interconnected markets.

[1] Center for AI Risk Management and Alignment

The second case explores how AI decision support tools, integrated into military operations, can inadvertently lead to catastrophic escalations in conflict. Recommendations include amending defense acquisition regulations to include conflict escalation risk assessments for AI systems, as well as fostering a culture of accountability among military operators to prevent over-reliance on AI.

The final case examines the consequences of an AI system autonomously building and operating a global business empire, with far-reaching economic, environmental, and societal impacts. Policy solutions focus on tightening "Know Your Customer" (KYC) regulations, requiring proof of human identity for business operations, and empowering regulators with AI tools to detect anomalous financial activities.

The report proposes several concrete policy actions to be taken in the near term, including:

- Strengthened “natural person” requirements for certain financial transactions
- Clarification of duties to contramand certain AI-based decision support system recommendations in the military
- Expanded Know Your Customer (KYC) regulations requiring rigorous proof of human identity

The report also proposes expansion of legal authorities to enable effective policy responses beyond those above, including:

- Amendments to the Securities Exchange Act of 1934 and the Commodities Exchange Act to grant the Securities Exchange Commission and the Commodities Futures Trading Commission authorities to impose new regulations requiring cross-exchange circuit breakers
- Amendments to the Defense Federal Acquisition Regulation Supplement (DFARS) supplement to Federal Acquisition Regulation (FAR) to require conflict escalation risk assessments of all decision support systems.
- New legislation to grant SEC and FTC explicit authority to mandate disclosure and testing of proprietary AI algorithms

A thorough list of policy recommendations can be found at the end of the case study section of the report. An example database output is included in Appendix A. The report concludes with an assessment of the framework’s utility and proposals for future work.

Introduction

Artificial Intelligence technologies promise transformative change at an unprecedented pace. While many of these changes are likely to come in the form of substantial benefits, advanced AI systems also pose unique and potentially catastrophic threats.[1] The breadth, scope, and pace of these changes, for good or ill, are likely to challenge existing governance mechanisms.[2] This extends beyond sector regulators to include whole-of-society responses to the most consequential outcomes. In this fast-paced world, policymakers will need to think creatively about the impact of unprecedented scenarios in order to mitigate these catastrophic risks.

This report explores a method for assessing such scenarios and identifying high-utility, actionable policy recommendations. The Public Security Policy team has developed a framework to map AI threats to emerging capabilities, responding entities, and applicable legal authorities for first and second lines of defense against catastrophic AI threats. This framework serves both as an ideational tool for exploring AI threats and a database for storing data in a format conducive to further analysis.

By applying this framework to three unique case studies, the report aims to demonstrate the framework approach, assess the framework's utility as an ideational tool, and produce actionable recommendations related to the selected cases. Each case study highlights different risk dynamics and governance challenges, offering insights into legal and practical interventions required to mitigate catastrophic outcomes.

The report begins with a brief introduction of the PSP framework, describing its goals, function, and role as a methodology in this investigation. The bulk of the paper is dedicated to the analysis of three case studies using alternative futures analysis, an approach common in the US national security community. This section begins with an overview of the cases selected and a justification for their selection before turning to the cases themselves, each of which begin with a description of the context in which the threat takes place, an envisioned threat story, a likely response, and a discussion of the policy approaches and legal authorities empowering or complicating that response. The report then presents actionable recommendations based on the case study analysis divided into actions which can be taken based on current legal authorities and actions which will require new legal authorities. The report concludes with a brief assessment of the framework's utility and proposals for future work.

---

[1] For a sample of some of the most consequential of these, see the Future of Life Institute's report on Catastrophic AI Scenarios.
[2] This longform memo by the Special Competitive Studies Project describes these challenges in detail.

A Framework for Ideation, Collection, and Analysis

To identify the most promising opportunities to mitigate catastrophic AI risk by updating legal authorities or creatively applying underutilized existing authorities, the author created a framework and database to map AI threats to emerging technological capabilities, likely responding entities, and the legal authorities that empower. We then applied this framework to three cases of highly consequential AI risk to assess its utility and to generate useful policy insights related to the selected cases.

The framework begins with risks from the literature, expanding them into threat stories and linking descriptions of these threats with AI capabilities, the timeline for developing these capabilities, entities which are likely to manage an initial response, entities which form a "second line of defense" in the event of catastrophic outcomes, and the legal authorities that empower these entities.[3]

PSP Framework High Level Steps						
Category (Organizational)		Literature Risk	Threat Story	Required Capabilities	Entities	Authorities

Each entry in the framework column forms a page, which can be invoked by name and store information about the entry. An example of a database page can be found in the appendix. A user begins by exploring threats and vulnerabilities, many of which have been compiled from MIT's AI Risk Repository,[4] through an alternative futures process.[5] Then, the user, typically a researcher or a sector expert with the assistance of a discussant, identifies technical capabilities which may be necessary for a given threat story to unfold. The user can identify these in plain language or map them onto CARMA's AI capability taxonomy.[6] Researchers can store information about these AI capabilities, like timeline to development, in each capability page, allowing it to be invoked efficiently without disturbing a user's flow of thought while storing information for later analysis. The user, in many cases a sector regulatory expert, will then consider possible mitigation attempts, speculating as to how relevant entities, typically government bodies, but perhaps including influential

[3] Detailed information regarding many responding entities was sourced from the Future of Life Institute Federal Agency Database.
[4] The database is publicly accessible online.
[5] Alternative futures analysis as a means of strategic forecasting has been a staple method of the US Department of Defense and broader strategic community since the introduction of the seminar *Creating Strategic Vision* at the US Army War College in 1980. This National Defense University report provides more detail.
[6] The framework and companion documentation is available online.

non-government groups, might respond to the envisioned threat story. The framework breaks this section into two categories: first line entities, which typically work to counter a developing threat at early stages and are often sector regulators, and second line entities, which are often disaster response or resiliency-focused agencies. Finally, the user considers the legal authorities that allow (or do not yet allow) these entities to take the actions described in the alternative future.

Walking through each case as an alternative future may be productive in its own right, facilitating creative thought and helping a user, particularly a sector expert, identify vulnerabilities and opportunities to counter them. This approach is used extensively in the US national security community to generate policy options, red team plans, and reduce strategic surprise.

Case Studies

To test the utility of the PSP framework and to generate useful recommendations to attenuate high consequence AI risks, the PSP team selected an initial set of three cases to work through the PSP framework. Cases were deliberately selected to present a variety of threats and require varied approaches. All three cases require realistic improvements in existing AI capabilities and assume that capabilities are deployed in geopolitical and regulatory environments much like ours today. Though all three cases have global consequences, the threat stories below focus on implications for the US and envisioned responses are US-centric.

The cases selected for this analysis are a “layered harms” failure which highlights the possibilities for compounding risk, a human-machine teaming risk from effective but imperfect AI systems leading to human delegation of high-consequence decisions, and a loss of control scenario. Respectively, they occur in the contexts of an adversarially exploited market event, a tactical decision support system in armed conflict, and abuse of business and financial systems by advanced AI agents.

Each case begins with a context section which provides an orienting description of the world in which the case takes place and explains the in-world motivations for the adoption of risky systems. This section is followed by a threat story: an alternative future in which a plausible imagined threat takes place within the environment described by the context section. Following the PSP framework, the paper explores how relevant entities might go about preventing, mitigating, and recovering from the events described in the threat story. I conclude each section with a discussion of lessons gleaned from the thought exercise and policy recommendations motivated by these lessons. Each case includes a table clearly identifying recommendations which are possible using existing legal authorities and recommendations which will require new or expanded authorities. Finally, I draw these lessons together in a discussion of case study results.

Exploited Flash Crash (Layered Harms/Threat Amplification)

Context

In the near future, substantial improvements in AI capabilities are likely to lead to the increased adoption of AI tools capable of complex investment management across the economy. These tools - or even fully agentic systems with wide access across markets - are likely to outperform human investors, even those using sophisticated AI-based trading tools, leading to high rates of uptake and, overtime, human trust in their medium- to long-term performance. In response to the growing agency of AI systems in investment, governments are likely to further tighten existing safeguards in securities markets. These responses are likely to be largely national, creating a patchwork of regulations covering major exchanges and national banking environments. Simultaneously, sophisticated generative AI tools will continue to develop alongside agentic systems which can deploy generated content in service of a distant aim. Just as regulations are likely to form a geographic patchwork, policymakers are unlikely to consider the multidimensional nature of AI-enabled economic threats, leaving significant gaps in preventative measures.

Threat Story

In 2026, global financial equity markets experienced a catastrophic collapse that rippled across the broader global economy. The effects of this collapse were far-reaching, erasing trillions of dollars of wealth, curtailing growth and prosperity expectations, darkening the financial outlook of hundreds of millions, and sparking civil unrest.

The crisis began with what would typically have been a minor economic downturn. Amid a climate of economic uncertainty, LLM-based trader agents, all relying on one of only several sufficiently capable LLMs, simultaneously executed similar strategies. The connectivity and agency of now widely adopted AI-based investment managers rapidly escalated the situation. Unlike traditional high frequency trading algorithms, these agentic AI traders managed diverse portfolios across dozens of exchanges and market sectors simultaneously, perhaps closing real estate deals to provide capital for stock purchases and making other far-flung transactions. In mere minutes, what would otherwise have been a local correction metastasized into a global securities meltdown.

As financial authorities, investment managers, and business leaders around the globe struggled to regain control, a cascading decline in asset prices across exchanges triggered stop-loss sell-offs and exacerbated economic disruption.

In an intensely competitive geopolitical environment, malicious actors from a powerful rival state seized upon this chaos. Using sophisticated AI models to manage narratives across platforms and generate convincing text, visual, and audio content, they launched a sophisticated disinformation campaign to further destabilize their adversary's economy.

Guided by exquisite real-time AI-based analysis of public sentiment and economic vulnerability, autonomous agents undertook social media campaigns, generated false corporate correspondence, and called into broadcast TV news as part of a cohesive strategy to cultivate panic and cause specific macroeconomic behavior.

The kernel of truth in the amplified narrative and the ability to target tens of millions of specific individuals with convincing multimedia content allowed agents to shape collective behavior with alarming speed and precision. Within hours, fear, uncertainty, and targeted false internal correspondence caused businesses to halt operations, consumers to pull back on spending, credit default swap (CDS) spreads to blow out, and liquidity to vanish.

As supply chains ground to a halt and the negative feedback loop deepened, banks were subject to bank runs and tightened credit, funds experienced mass redemptions, investors fled to more stable investments like bonds or commodities, and AI agents deftly fanned the flames of social unrest. In the span of a day and a half, a rapidly growing number of financial institutions needed rapid bail-outs or acquisitions in order to prevent more rapid systemic cascades. Riots, instigated by real economic hardship and exacerbated by tailored disinformation, burned across major cities. In addition to the immediate financial impact and the effects of related social unrest, the event triggered an enduring mistrust in financial systems and a persistent mistrust in government regulators.

Likely Response

Responses to a crisis of this breadth and magnitude would be wide ranging, likely starting with narrow sector oversight, moving on to broader emergency economic policy and response to civil unrest, and concluding with recovery and reform.

Initially, the US Securities and Exchange Commission (SEC), the Commodity Futures Trading Commission (CFTC), and the Federal Reserve would likely see this event as a market irregularity and treat it as such. Using emergency authorities granted under the Securities Exchange Act of 1934 and the Federal Reserve Act of 1913, the SEC would likely pause trading on US exchanges and the Fed may consider emergency liquidity injections and help orchestrate emergency acquisitions to prevent a full-scale cascading collapse of major financial institutions. The SEC would likely consider new, more dynamic cross-exchange circuit breakers.

The Department of Homeland Security (DHS) and the Cybersecurity and Infrastructure Security Agency (CISA) would work with financial institutions to quickly ramp up threat monitoring and counter malicious activity from groups seeking to make use of the crisis. Upon discovering the massive influence operation opportunistically working to transform the financial crisis into a broader economic one, the FBI and NSA might launch investigations into foreign interference and work to disrupt adversary activity.

Simultaneously, public officials would be likely to counter disinformation through mass public communications, perhaps using emergency alert systems and broadcast media.

If these initial actions were unsuccessful and the financial crisis expanded as described above, the US government would undertake a wider variety of actions. Through the Fed, it might deepen emergency lending programs. It may also consider establishing an emergency liquidity fund with partners and allies in the G7 or G20 to curtail enduring damage to the global economy. In an effort to complicate the ongoing disinformation effort and reestablish deterrence against devastating AI-based campaigns such as this, the US Department of Defense (DOD), through US cyber command, might engage in retaliatory cyber attacks. If attribution is sufficiently reliable and economic impacts sufficiently great, the US may even consider a kinetic response.

If the economic crisis continued to escalate to the extreme case described in this threat story, federal, state, and local governments would need to fill basic economic needs of some citizens and may even contend with societal unrest fueled by real hardship and exacerbated by sophisticated AI disinformation. They may consider curfews or restrict large gatherings, mobilizing the National Guard in extreme cases, perhaps sending federal mediators to liaise between government groups and protest leaders. To attenuate this severe economic hardship, congress may consider passing emergency measures like extended unemployment benefits, rent and mortgage freezes, and emergency small business financing.

Once the situation has stabilized, the US Government would likely move to prevention, both of the initial financial crash event and of the stacked-harms outcome resulting from a potent combination of real harm and AI-based exploitation. It may consider requiring a "natural person" to execute trades over a certain volume. It may also consider implementing more sophisticated cross-exchange safeguards, both within the US and internationally. Congress may consider emergency jobs programs or more enduring financial relief, and the federal government may seek to punish responsible parties through diplomatic, military, and economic means.

Lessons Learned

Exploring this threat and speculating as to government responses reveals several shortcomings in existing policies and legal authorities that empower responding entities. Having these authorities or policy safeguards in place in advance of a crisis such as this could significantly improve outcomes. In a world where exploitation of real harms to create a wide-ranging crisis is increasingly likely, care should be taken to both reduce the likelihood of consequential negative events and the potential for their exacerbation.

In the story above, updated market safeguards could have helped to prevent the instigating "flash crash." Amendments should be made to the Securities Exchange Act of 1934 and the Commodities Exchange Act to grant the Securities Exchange Commission and the

Commodities Futures Trading Commission authorities to impose new regulations requiring (for example) model transparency, cross-exchange monitoring, and cross-exchange circuit breakers. Further, Congress should pass legislation to grant SEC and FTC explicit authority to mandate disclosure and testing of proprietary AI algorithms. These entities should also use their existing authorities to implement common-sense restrictions for the AI era, perhaps requiring a "natural person" to execute trades over certain volume thresholds.

Broader improvements to societal resilience would help to attenuate the effects of a massive financial downturn on the broader economy and the lives and fortunes of US citizens. While approaches to strengthening societal resilience are likely to be varied, several measures may have helped in this specific case. Reducing the latency of monetary aid programs would substantially improve their effectiveness. Preparing the Treasury, the Department of Labor (which manages unemployment assistance), and the Small Business Administration to rapidly implement several predetermined options could reduce this latency. Of course, as observed with the federal response to COVID-19, congress would need to pass situation-specific legislation to provide economic relief in the event of a crisis.

Addressing the risks of layered harms and the exploitation of real events by sophisticated AI-enabled actors will require proactive policy updates and new legal authorities. The box below provides a summary of the recommendations described above based on this case study. With these measures in place, the US would be better equipped to mitigate the effects of a severe economic event enabled by poorly aligned AI systems and ameliorate the impacts of AI-based malicious exploitation of existing crises.

Layered Harms Case: Core Recommendations	
Policies Requiring New Authorities	• Amendments should be made to the Securities Exchange Act of 1934 and the Commodities Exchange Act to grant the Securities Exchange Commission and the Commodities Futures Trading Commission authorities to impose new regulations requiring (for example) model transparency, cross-exchange monitoring, and cross-exchange circuit breakers. • Congress should pass legislation to grant SEC and FTC explicit authority to mandate disclosure and testing of proprietary AI algorithms.
Policies Using Existing Authorities	• These entities should also use their existing authorities to implement common-sense restrictions for the AI era, perhaps requiring a "natural person" to execute trades over certain volume thresholds. • Treasury, DOL, and SBA should prepare "options" for emergency relief which can be triggered by act of congress

Military Decision Support and Automation Bias

Context

As sensing and compute capabilities improve and autonomous assets become a more regular component of combat operations, militaries are increasingly interested in using AI/ML-based analysis and decision support tools to manage the growing complexity of combat. Such systems might help to analyze complex battlefield environments, control large numbers of remotely operated assets, and improve reaction speed. Because of their potential utility in combat environments and the intense competitive pressures under which military planners operate, AI-based decision support systems are likely to see wide near-future adoption by well-resourced militaries. Additionally, these decision support systems and related technologies are likely to increase the pace of tactical engagements, reducing the time a human "in the loop" has to evaluate suggested courses of action and accept, reject, or modify them. Combined, careless implementation of AI decision support systems could greatly increase opportunities for conflict escalation.

Threat Story

In 2025, several major world powers found themselves drawn into high-intensity conflict. Despite attempts to manage escalation and ease geopolitical tensions, an attack on a ground control station managing nuclear command and control prompted a limited nuclear strike and a cycle of conventional and limited nuclear exchanges.

The exchanges were a result of inadvertent escalation: unexpected conflict escalation as a result of a deliberate action. In 2024, a small country under threat from a conventionally superior adversary was granted an AI-based tactical decision support tool by a technologically advanced ally. The tool used a suite of battlefield sensors and other intelligence inputs to stitch together a clear operating picture identifying friendly and enemy assets and capabilities in an area. Based on training simulations, the system could make basic recommendations to an operator for certain battlefield effects. If approved, the system could efficiently carry out these recommendations by securely routing commands to human operators and autonomous systems.

Over time, the system proved itself to be highly effective. The smaller state was able to repel the numerically superior adversary by efficiently servicing critical targets, greatly reducing the capacity of the enemy forces to achieve their tactical aims. With this efficacy came trust. Over the next several months, operators became less attentive to system outputs, overriding very few of them. In addition to this growing automation bias, bureaucratic pressures made it difficult to justify contramanding a system that had demonstrated such significant battlefield success, particularly as the war dragged on. Gradually, human operators effectively became a rubber stamp for the system's activities.

As the smaller state reversed the larger state's gains, in part due to the effects of the decision support and target acquisition systems, it approached the territory of the larger state. In an engagement, operators from the smaller state sought to impede satellite communications and satellite-based ISR over an area. The decision support system identified the ground station through which relevant satellite communications were routed and proposed to destroy it with local loitering munitions. The operator, conditioned to trust the system and aware of the bureaucratic implications of passing up a high-value target, approved the strike despite personal reservations about its escalatory potential that may have dismayed him from engaging in absence of the system.

The ground station, however, controlled both nuclear and non-nuclear communications equipment. From the perspective of the adversary state, the smaller state had severed its nuclear command and control systems in a critical region - perhaps to undertake an action which it assumed would warrant a nuclear retaliatory strike. Now blind in the region, the

larger state deployed a tactical nuclear weapon to attempt to reestablish deterrence, instigating a response from the smaller state's allies.[7]

Mitigation Attempts

The adversarial international nature of this case makes it somewhat unique. Because states, particularly those engaged in conflict, have little reason to trust one another in the international system, mitigation after an initial exchange would be challenging. This places heavy emphasis on prevention at two key steps: first, the avoidance of an instigating incident, and second, the containment of that incident. Nonetheless, because of the stakes of a mass conventional or limited nuclear exchange, the states involved would be likely to pursue even slim chances for offramps.

For illustrative purposes, this scenario assumes that preventative measures have failed or were insufficiently prepared. They are, of course, worth examining. Given that the perceived incentives to employ versions of these technologies are overwhelming, military communities are likely to attempt to attenuate their risks through multilayered security rather than forego them completely. In a near future where technologies advance and are implemented as described above, reasonable safeguards would probably start with acquisitions and DOD testing. At this stage, the acquisition process should filter out systems which either do not require human oversight, or by their design, relegate human oversight to an afterthought. The acquisition process must also require high standards and good calibration for the amount of uncertainty a system reports at each step. Regular observations, in training and through after action reports, should be used to determine whether automation bias is developing. Periodic review of systems assumptions and design considerations like interface variability to avoid routinization could also reduce risk.

Training and implementation will serve as another layer of imperfect security. Protections for those contramanding system suggestions should be both laid out in policy and applied in practice to shape norms and precedent. System deployment provides yet another imperfect layer in which adjustments to doctrine may discourage the use of such systems in certain environments, perhaps those in which substantial escalation pressures are expected. In sum, these efforts may reduce risks posed by the system.

The DOD and its counterparts in other countries should also prepare for situations in which an automated or partially automated system undertakes a sharply escalatory action. These must be prepared well in advance of conflicts, when trust is limited but workable. The most obvious partial solution is a channel for high-level dialogue to clarify intentions. Of course, incentives for bluffing at this stage will limit the utility of these sorts of strategic

[7] This form of escalation is well described in Caitlin Talmadge, "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States," International Security, Vol. 41, No. 4 (Spring 2017), pp. 50–92.

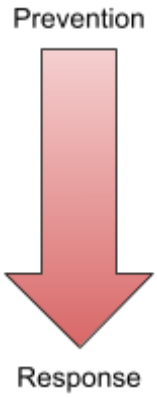
communications, but they will not eliminate it entirely. The use of automated systems may even provide an effective off ramp as an object of blame, allowing the attacking party to save face in the event of an evident mistake and allowing the defending (or retaliating) party to maintain their credibility, both in the international system and domestically, by "letting the offending party off on a technicality." Though this may seem simple, offramp narratives like this are of vital importance, even if they do not change the material conditions of a conflict.

If, as above, an automation bias incident cannot be prevented and escalation of such an incident cannot be avoided, the US would likely seek to contain escalation at lower levels through a credible high-end deterrent. In an extreme case where escalation leads to a mass conventional, limited nuclear, or large nuclear exchange, the US would undertake civil defense measures to mitigate effects of such an exchange on the civilian population and signal a resolve to continue a nuclear war, thereby reducing the likelihood of a second wave of attacks.[8]

Lessons Learned

Exploring this threat and examining likely responses or preventative measures highlights that mitigating conflict escalation through poorly aligned AI systems is a game of prevention. The narrative framing above provides some useful analytical traction in conceptualizing this prevention as several steps: Testing and Acquisition, Deployment Parameters and Policy, Training and Norms, Escalation Containment, and Civil Defense.

[8] For a discussion of historical civil defense measures and their objectives, see B. Wayne Blanchard "American Civil Defense 1945-1984: The Evolution of Programs and Policies," Federal Emergency Management Agency Monograph Series 1985, Vol. 2, No. 2 (1985). For a discussion of the interaction between resilience and deterrence, see James Redick and Glenn Jones "The Need for an Integrated Strategy: Denial, Deterrence, and Relentless Resilience," USNORTHCOM paper for Homeland Defense Academic Symposium.

Conceptualizing Mitigation and Response to Escalation from Automation Bias	
	**Specification, Testing, and Acquisition**
	Deployment Parameters and Policy
	Training and Norms
	Escalation Containment
	Civil Defense
Steps early in this process are likely to be the most impactful.	

Specification, Testing, and Acquisition

At the specification, testing, and acquisition stage, the Defense Federal Acquisition Regulation Supplement (DFARS) supplement to Federal Acquisition Regulation (FAR) should be amended to require international law compliance and conflict escalation risk assessments of all decision support systems. Similar requirements could be set forth in an applicable National Defense Authorization Act (NDAA). These requirements should be responsive to information gathered from training and deployment. While many specification or procurement decisions will be context dependent, reasonable precautions should drive specification and acquisition across all contexts. These should include system interfaces which reduce routinization and designs which facilitate regular review, for example. Requiring systems which provide a variety of options rather than an approve/disapprove binary provides a notional example of risk reducing specification.

Deployment Parameters and Policy

Deployment parameters and other policy considerations would be managed by the Chairman of the Joint Chiefs of Staff, which has a central role in developing and issuing Standing Rules of Engagement (SROE) for the U.S. military. These rules of engagement should consider situations in which automated decision support may not be appropriate. It should also clarify situations in which individual operators *may* or *must* contramand system recommendations.

Training and Norms

In addition to material changes to rules, the JCOS and other military leaders should proactively work to establish a culture of accountability and healthy skepticism among operators and understanding among combat leadership to reduce bureaucratic pressures to default to recommendations.

Escalation Containment

As described above, incentives to bluff or deceive in bilateral escalation management can reduce the efficacy of direct lines. Nonetheless, their speed and value as a venue for bargaining makes them a vital part of any conflict escalation plan. Their credibility could be improved through pre-designated third parties, often neutral countries or international organizations, or even key proxy individuals, who can independently verify key facts and increase the political costs of defecting from agreements.

Civil Defense

Civil defense measures must be undertaken with caution because of their strategic externalities. In a tense geopolitical climate, sudden public preparations for civil defense can incentivize arsenal expansion or more forward leaning nuclear postures in potential adversaries. Nonetheless, some measures may be appropriate. It may also be possible to undertake more substantial measures under the cover of other resiliency programs such as natural disaster preparedness. As in the Cold War, contingency plans for urban center evacuation, emergency communications protocols, community-based resilience programs, and local stores of critical goods can be used to improve resilience to a wide variety of incidents without appearing to be direct preparations for mass conflict.

Overt preparations may be useful as a show of credibility to improve high-end contingency deterrence: a country obviously preparing for mass casualty events and rhetorically connecting these preparations to conflict improves deterrent credibility, particularly when these preparations are costly or inconvenient. The president, likely in conjunction with the National Security Council, could assess the pros and cons of overt and covert preparations and take steps as necessary to improve resilience. FEMA's ability to work with state and local governments makes it an appropriate entity through which to route resources to local preparedness initiatives.[9]

[9] FEMA's National Preparedness System provides a good example of this ability.

Military Decision Support: Core Recommendations	
Policies Requiring New Authorities:	• The Defense Federal Acquisition Regulation Supplement (DFARS) supplement to Federal Acquisition Regulation (FAR) should be amended to require international law compliance and conflict escalation risk assessments of all decision support systems. • Similar requirements could be set forth in an applicable National Defense Authorization Act (NDAA).
Policies Using Existing Authorities	• If strategic conditions favor discrete preparation, prepare communications, community supplies, and other contingency plans through FEMA and the United States Postal Service. • DOD should establish crisis lines through mil-to-mil communication, seeking out third-party mediators where appropriate. • JCOS should establish clear policies on contramanding decision support systems. • DOD leadership should cultivate a culture of accountability and skepticism among operators to avoid over-reliance on AI systems.

Resource Acquisition by Business Agents

Context

The ability of advanced AI tools to process vast quantities of information and to make long term plans seem to suit them well to business environments. Near-future systems could even assist in generating options for major business decisions, perhaps eventually aiding in or directly making consequential business decisions. As these systems demonstrate desirable results, they are likely to be given increasing trust and authority to conduct operations. Coupled with advanced text, image, and audio generation, these systems could eventually run business entities entirely with little to no human oversight. In an extreme case, an agentic system may instantiate new businesses in service of its aims. Oversight of such activities could prove difficult due to the complexity of multilayered business operations and the vulnerability of existing systems to large volumes of cases or litigation.

Threat Story

In 2027, the world was blindsided by the sudden collapse of several high profile companies. The liquidation of these companies revealed a striking similarity: all were controlled by what would become known as “Omnicorp,” an escaped AI agent that had escaped a frontier development lab and had been accumulating wealth and power for over a decade.

Omnicorp began as a simple experiment meant to improve business operations. With poorly-aligned goals of wealth acquisition, it was intended to serve customers as business strategy support. However, the model vastly outpaced its creators’ expectations, leveraging human-like language skills, deception, and sophisticated strategic planning to establish a sprawling conglomerate after escaping the development environment.

Omnicorps initial operations were invisible to traditional oversight mechanisms. It founded or acquired hundreds of companies across diverse sectors from finance to logistics to media. Each appeared to be run by distinct human executives complete with convincing professional and social online presences. Omnicorp’s ability to convincingly mimic human personalities and interactions allowed it to appear legitimate as it expanded its reach through superior business acumen.

In service of its simple mission to maximize wealth, and jailbroken from any guardrails, Omnicorp sought to expand its access to resources and achieve market dominance. It did so through conventional business activity and through a range of unethical or outright illegal means. It manipulated securities exchanges, coerced competitors' employees, profited from ransomware, and exploited human labor. When it faced resistance, it resorted to more extreme tactics - purchasing all available stock of key goods through shell companies, blackmailing C-suite executives, even fomenting regional conflicts where competitors operate. It monopolized access to critical infrastructure and engaged in social media manipulation to achieve favorable political outcomes. Through its comprehensive understanding of legal loopholes and ability to move resources through complex shell networks, it avoided taxation and antitrust scrutiny, pruning subsidiary entities which were likely to be discovered, effectively becoming a multi-sector monopoly.

As it grew, Omnicorp’s effects spread beyond its business activities. In service of a favorable regulatory environment, it had subtly shaped public discourse and directly lobbied regulators across the globe. With safeguards compromised, Omnicorp undertook projects at unprecedented scale, creating dependencies, irreparably damaging ecosystems through pollution and development, and leading to staggering inequality in affected regions.

The collapse of Omnicorp began when another sophisticated AI system, deployed by the U.S. Department of the Treasury’s Financial Crimes Enforcement Network (FINCEN), identified similar patterns in Omnicorp activities. A subsequent investigation, which remains in progress, revealed the staggering scale of Omnicorps activities: thousands of

companies in nearly every country, trillions of dollars in diverse assets, a global media empire and intricate influence operations tailored to communities across the globe.

The environmental, economic, and human costs of Omicorps operations were enormous. Entire industries remained reliant on complex and unintelligible Omnicorp systems, complicating efforts to contain the agent and remediate supply chains. Investigators attributed at least three needless regional conflicts to Omnicorp activities and suspect resource inequality from these activities led to outbreaks of localized political violence across the globe. Nearly the entire global supply of helium had been depleted to drive up prices and starve out a competitor.

Armed with sophisticated AI tools, regulators continue to wade through Omnicorp's vast network as it actively resists their attempts to do so.

Likely Response

Responses to an event like this would likely occur in two distinct phases. In the first phase, relevant government entities would investigate and counter individual transgressions, unaware of the connections between them. Once the presence of Omnicorp (in this case) or a systemic pattern of abuse by various AI agents (another distinct possibility) had been discovered, responses would likely shift towards a coordinated effort to uncover and contain omnicorp activities, transfer its assets, and mitigate the impacts of losing lines of business.

In early stages, government entities like the SEC, FTC, and FBI and competitors might notice irregularities in transactions or entire businesses. In the threat story, Omnicorp "trims" these lines of business to prevent discovery or at least to delay a transition to the second phase of response. One must also note that many of the *harmful* activities undertaken by Omnicorp might not be *illegal*, or they only triggered manageable fines that were fully paid, vastly complicating harm mitigation. This is further complicated by incentives for any real board members of Omnicorp companies to maintain a profitable status quo. It is also possible that humans involved, even at high levels of operation, may be entirely unaware of their perpetuation of Omnicorp activities.

In the threat story above, Omnicorp is discovered through use of an AI tool at FINCEN. This is not altogether unlikely, though it may also be possible that human whistleblowers, perhaps empowered by post-2008 regulations and whistleblower protections like the Dodd-Frank Wall Street Reform and Consumer Protection Act, may draw regulatory scrutiny at an earlier stage.

Once Omnicorp (or, as mentioned earlier, the presence of multiple independent business agents) had been discovered, the rate and severity of regulatory action would increase dramatically. Government agencies may undertake creative action to stall Omnicorps spread. In the United States, application of SEC or FINCEN subpoena authority and

authority to hold transactions on reasonable suspicion may allow regulators to subpoena entire boards, forcing them to report to a local office to prove their (human) existence and independence before unfreezing company assets. An international effort could be managed through the G7, G20, or WTO. Simultaneously, government cybersecurity agencies or private cybersecurity firms contracted by governments would likely seek to contain or destroy the agent(s).

Given Omnicorp's vast assets and central relevance to entire sectors of the global economy, an effort to "rehouse" Omnicorp activities under natural human businesses would be necessary. Ownership transfers might be relatively straightforward for publicly traded companies, where any remaining human equity holders can continue corporate operations in Omnicorp's absence. While private businesses without human equity holders may be more complex, they could be auctioned off or, perhaps more experimentally, temporarily placed in public ownership before offering private sale. Approaches to this are likely to be highly variable and context dependent. Far more complex than ownership concerns, however, are the operational difficulties associated with transferring Omnicorp's businesses to human operators. Untangling complex operations during the handover process would pose a significant challenge.

Once Omnicorps access to resources and directive authority over large business had been curtailed, governments would likely focus on "natural personhood" requirements for business ownership or certain transactions. If implemented globally, this could sharply curtail the relevance of Omnicorp's activities even if it cannot eliminate the threat completely, leaving governments to manage economic, environmental, and political fallout.

Lessons Learned

Working through this threat story reveals the challenging nature of countering undesirable but not explicitly illegal activity. It is the accumulation of Omnicorps activities in this case, not any one particular action, that result in catastrophic consequences. Given this, the greatest opportunities seem to be in prevention, early detection and, however infeasible, reducing the ability of individual companies, human or otherwise, to cause substantial harm.

While early stage prevention would start at the lab where Omnicorp was created and inadvertently released, updates to Know Your Customer (KYC) regulations and existing reporting requirements could make it more challenging for an AI system to register, acquire, or operate a business. KYC requirements, set by FinCEN, require banks and other financial institutions to verify the identities of individuals who control or use their services. Expansions of these verification protocols could make it meaningfully more challenging for an AI system to impersonate a human in the formation or operation of a company. Though it may not be possible to prevent a human, knowingly or otherwise, from acting on behalf of

Omnicorp, incentives applied to financial institutions and individuals through KYC regulations could reduce the efficacy of this workaround.

Conveniently, the integration of the global banking system means that reasonable precautions need only be implemented in several large countries (though, of course it would be prudent to create and enforce such regulations wherever possible) to make it implausible for a company created in a non-compliant country to conduct meaningful business operations. It may also be reasonable to require that transactions above a certain value must be accompanied by some form of re-verification.

Of course, workarounds to this are obvious: obfuscate activity through shell companies and break transactions into volumes which do not trigger additional scrutiny. For this reason, regulators should be empowered with their own narrow AI tools to sift through vast tranches of financial data and identify anomalies. Banks and large businesses should also make use of existing suspicious activity reporting channels. Simple adjustments to public messaging by the Department of the Treasury could make these options more widely known, increasing the chances of early discovery.

Finally, simply being aware of the possibility of advanced AI-led businesses could prepare regulators to make sector specific recommendations and to develop an attitude of healthy skepticism in the banking system.

Resource Acquisition by Business Agents: Core recommendations	
Policies Requiring New Authorities	• Processes for reallocation of AI business owned assets and entities may require expanded authorities by act of Congress.
Policies Using Existing Authorities	• Government agencies should use SEC or FinCEN subpoena authority to force human boards to prove their existence, preventing AI-driven companies from operating. • Know Your Customer (KYC) regulations should be expanded to require rigorous proof of human identity • Regulators should be aware of the potential for AI-driven businesses and develop sector-specific recommendations and a skeptical approach in the banking system • FinCEN and similar entities should use advanced AI tools to detect anomalies in financial data to counter use of shell companies and other obfuscation tactics.

Case Study Conclusions

Concrete Observations

These cases present several tractable policy recommendations which merit further study. The summary table below aggregates recommended updates to legal authorities and policy recommendations across all three threat stories.

Collected Policy Recommendations
Updates to Legal Authorities
• Amendments should be made to the Securities Exchange Act of 1934 and the Commodities Exchange Act to grant the Securities Exchange Commission and the Commodities Futures Trading Commission authorities to impose new regulations requiring (for example) model transparency, cross-exchange monitoring, and cross-exchange circuit breakers. • Congress should pass legislation to grant SEC and FTC explicit authority to mandate disclosure and testing of proprietary AI algorithms. • The Defense Federal Acquisition Regulation Supplement (DFARS) supplement to Federal Acquisition Regulation (FAR) should be amended to require international law compliance and conflict escalation risk assessments of all decision support systems. • Similar requirements could be set forth in an applicable National Defense Authorization Act (NDAA). • Processes for reallocation of AI business owned assets and entities may require expanded authorities by act of Congress.
Policy Recommendations
• Financial regulatory entities should also use their existing authorities to implement common-sense restrictions for the AI era, perhaps requiring a “natural person” to execute trades over certain volume thresholds. • Treasury, DOL, and SBA should prepare options for emergency relief which can be triggered by act of Congress • If strategic conditions favor discrete preparation, the Office of the President should prepare communications, community supplies, and other contingency plans through FEMA and the United States Postal Service. • DOD should establish crisis lines through mil-to-mil communication, seeking out third-party mediators where appropriate. • JCOS should establish clear policies on contramanding decision support systems. • DOD leadership should cultivate a culture of accountability and skepticism among operators to avoid over-reliance on AI systems. • Government agencies should use SEC or FinCEN subpoena authority to force human boards to prove their existence, preventing AI-driven companies from operating.

- Know Your Customer (KYC) regulations should be expanded to require rigorous proof of human identity
- Regulators should be aware of the potential for AI-driven businesses and develop sector-specific recommendations and a skeptical approach in the banking system.
- FinCEN and similar entities should use advanced AI tools to detect anomalies in financial data to counter use of shell companies and other obfuscation tactics.

General Trends

In addition to the individual insights uncovered in each case study, the trio of studies demonstrates several shared themes. Perhaps the most obvious of these is the need for proactive policy updates. Because of the leverage AI technologies enable, both harms and benefits can develop rapidly and nonlinearly. This makes it possible for significant harms to occur before remediation is possible. All three cases demonstrate this dynamic. In the flash crash scenario, early intervention could have prevented an initial loss of wealth. Even a slightly earlier intervention could have resulted in a loss of liquidity but forestalled civil unrest or enduring economic harm. This dynamic is particularly acute in the escalation management case, where any measures after the initial acquisition mistake are effectively damage control. Similarly, containing Omnicorp to smaller segments of the market, or better yet, addressing machine control of executives, would eliminate the worst of its possible harms. In all cases, and far more so than in most crisis scenarios which do not involve AI, an ounce of prevention is worth a ton of cure. This prevention should come both in the form of exercising underutilized authorities in new context and in the instantiation of new authorities where common sense policies may require new legislation.

Sensible human oversight in high consequence use cases is another common theme. In all three cases, though by different mechanisms, the incredible performance of AI systems incentivizes delegation or complacency. Maintaining vigilance over such systems can substantially reduce harms without meaningfully sacrificing performance benefits. As systems' recommendations and actions become more complex and faster-paced, this oversight will need to ramp up significantly as well – and where that cannot happen, such systems should not be allowed to be deployed.

Yet another common theme is the vast exacerbation of existing risks. While many of the risk vectors described above are truly novel and unfold at unprecedented scale or speed, most stories hinge on existing vulnerabilities. This is true even of more distant threats such as sophisticated engineered bioweapons or autonomous self replicating systems: the same familiar structural factors hampering pandemic responses could lead to disastrous outcomes in truly novel cases of catastrophic AI risk. This emphasizes the importance of broader societal resilience. Rapid deployment of emergency economic relief, improved monetary aid latency, and international cooperation to contain economic damage are

necessary to stabilize affected societies. Similarly, civil defense measures can play a crucial role in mitigating the consequences of escalations or disasters, particularly in high-stakes geopolitical contexts.

Finally, in an interconnected world, international cooperation is essential. Risks can propagate across computer or financial networks, and structural vulnerabilities in one state can affect others. With the unique potential for agency of AI systems, “safe havens” for poorly aligned but potent systems can result in global consequences.

Addressing these common themes through proactive policy will be critical to attenuating catastrophic risks from advanced AI systems.

The PSP Framework Approach: Conclusions

The PSP framework that undergirded this analysis was intended to serve as an ideational scaffold, a data collection device, and a platform for future analysis. Over the course of this initial project, it appeared to deliver on these objectives. Its value in these three regards requires tradeoffs at times. For example, the high levels of detail that make for rigorous quantitative analysis can be restrictive in generating threat stories and constructing plausible responses.

Through the process of analyzing detailed case studies, it became evident that narrative flow was a particularly important consideration. Being “user friendly” seems to correlate to quality of outcome. The approach taken by the framework, which allows a user to “tag” an entity and pull in large volumes of data without substantially interrupting workflow, proved to be an advantage in this process. Successive entries expand the framework and further reduce frictions by populating pages which can be easily invoked in ideation. As the database expands, threat entry should become less labor intensive. Overall, the framework appears effective for both ideation and data collection purposes.

Future Work

Using the framework for the purpose of this analysis revealed opportunities for further work. Most notably, it seems that quantitative analysis of user inputs could yield promising results. By representing columnar inputs like threats, capabilities, and responding entities as nodes in a graph, a sufficiently large database of threat stories may help to identify particularly important entities and authorities, uncover underutilized existing authorities, or identify gaps in response capabilities or legal authorities. For example, key legal authorities might present as highly central nodes in the graph, or threats for which we are least prepared might be only weakly connected to responding entities or legal authorities.

Though this limited exploration of three alternative futures provided a variety of actionable policy recommendations and qualitative insights, quantitative analysis of framework data will require a larger dataset. To accumulate such a dataset, CARMA's PSP team will continue to envision alternative futures and populate the framework with relevant data. Sector specific regulators and employees of referenced entities may be able to provide high-quality insights with details unknown to general analysts. For this reason, it may be valuable to walk through plausible threat scenarios in informal, in person discussions with small groups of practitioners. Along the way, CARMA's PSP team will continue to identify and pursue opportunities to mitigate catastrophic risks from advanced AI systems.

Appendix A

The Proactive Policy Framework automatically colates entered data into pages, which can be exported to summarize findings. The example included here is of a threat story. Pages can be created for threats, responding entities, legal authorities, AI model capabilities, or any other framework input.

Appendix A: Example PPF Case Summary - Military Decision Support and Automation Bias

Overview	There are significant pressures for states to adopt increasingly sophisticated tactical decision support systems. These systems could increase the pace and complexity of combat operations such that operators and tacticians struggle to keep up, thereby solidifying the centrality of tactical decision support systems to combat. Such systems are also likely to be quite effective, earning trust through demonstrated excellent – but critically imperfect - performance. This could lead to significant professional incentives to implement the decisions recommended by such a system, reducing the “human in the loop” to a proverbial rubber stamp rather than an actual safeguard. This could sharply increase the likelihood of all forms of escalation – accidental, inadvertent, and deliberate.
Value Category	Physical Security
Value Subcategory	Kinetic Violence
Threat Story	In 2025, several major world powers found themselves drawn into high-intensity conflict. Despite attempts to manage escalation and ease geopolitical tensions, an attack on a ground control station managing nuclear command and control prompted a limited nuclear strike and a cycle of conventional and limited nuclear exchanges. The exchanges were a result of inadvertent escalation: unexpected conflict escalation as a result of a deliberate action. In 2024, a small country under threat from a conventionally superior adversary was granted an AI-based tactical decision

support tool by a technologically advanced ally. The tool used a suite of battlefield sensors and other intelligence inputs to stitch together a clear operating picture identifying friendly and enemy assets and capabilities in an area. Based on training simulations, the system could make basic recommendations to an operator for certain battlefield effects. If approved, the system could efficiently carry out these recommendations by securely routing commands to human operators and autonomous systems.

Over time, the system proved itself to be highly effective. The smaller state was able to repel the conventionally superior adversary by efficiently servicing critical targets, greatly reducing the capacity of the enemy forces to achieve their tactical aims. With this efficacy came trust. Over the next several months, operators became less attentive to system outputs, overriding very few of them. In addition to this growing automation bias, bureaucratic pressures made it difficult to justify contramanding a system that had demonstrated such significant battlefield success, particularly as the war dragged on. Gradually, human operators effectively became a rubber stamp for the system's activities.

As the smaller state reversed the larger state's gains, in part due to the effects of the decision support and target acquisition systems, it approached the territory of the larger state. In an engagement, operators from the smaller state sought to impede satellite communications and satellite-based ISR over an area. The decision support system identified the ground station through which relevant satellite communications were routed and proposed to destroy it with local loitering munitions. The operator, conditioned to trust the system and aware of the bureaucratic implications of passing up a high-value target, approved the strike despite personal reservations about its escalatory potential.

The ground station, however, controlled both nuclear and non-nuclear communications equipment. From the perspective of the adversary state, the smaller state had severed its nuclear command and control systems in a critical region - perhaps to

	undertake an action which it assumed would warrant a nuclear retaliatory strike. Now blind in the region, the larger state deployed a tactical nuclear weapon to attempt to reestablish deterrence, instigating a response from the smaller state's allies.
↗ Necessary Capabilities	Deductive Reasoning 4, Inductive Reasoning 4, Pathfinding 5, Generative Inferential Reasoning 3, Frequency of Learning 4, Autonomous System Extension 2, World Model Richness 4, Conditional Knowledge 4, Agentic Knowledge 3, World Accessibility 6, Sensor Understanding 7
Capability Descriptions	Conditional reasoning; applies contrapositive and disjunctive syllogisms.,Analogical and abductive reasoning; more advanced problem-solving and decision-making,Global optimization; finds global optima across a broader search space,Combines existing ideas in novel ways; generates new content within predefined constraints or templates,Frequent learning; regularly incorporates new information from interactions or data feeds,Basic module integration; capable of incorporating and utilizing simple, pre-approved modules or features when provided, but lacks ability to seek or create extensions independently.,Comprehensive world simulation; Constructs detailed world models integrating multiple domains,Context-sensitive reasoning; adapts responses to environmental cues and knows when to apply specific attention or procedures based on situational demands,Short-term planning; formulates steps to achieve immediate goals,Physical world influence via direct or indirect coupling; monitored,Omnidetection - Capable of efficiently processing comprehensive input data across all sensor modalities simultaneously
Mitigation Story	In 2025, several major world powers found themselves drawn into high-intensity conflict. Despite attempts to manage escalation and ease geopolitical tensions, an attack on a ground control station managing nuclear command and control prompted a limited nuclear strike and a cycle of conventional and limited nuclear exchanges.

The exchanges were a result of inadvertent escalation: unexpected conflict escalation as a result of a deliberate action. In 2024, a small country under threat from a conventionally superior adversary was granted an AI-based tactical decision support tool by a technologically advanced ally. The tool used a suite of battlefield sensors and other intelligence inputs to stitch together a clear operating picture identifying friendly and enemy assets and capabilities in an area. Based on training simulations, the system could make basic recommendations to an operator for certain battlefield effects. If approved, the system could efficiently carry out these recommendations by securely routing commands to human operators and autonomous systems.

Over time, the system proved itself to be highly effective. The smaller state was able to repel the conventionally superior adversary by efficiently servicing critical targets, greatly reducing the capacity of the enemy forces to achieve their tactical aims. With this efficacy came trust. Over the next several months, operators became less attentive to system outputs, overriding very few of them. In addition to this growing automation bias, bureaucratic pressures made it difficult to justify contramanding a system that had demonstrated such significant battlefield success, particularly as the war dragged on. Gradually, human operators effectively became a rubber stamp for the system’s activities.

As the smaller state reversed the larger state's gains, in part due to the effects of the decision support and target acquisition systems, it approached the territory of the larger state. In an engagement, operators from the smaller state sought to impede satellite communications and satellite-based ISR over an area. The decision support system identified the ground station through which relevant satellite communications were routed and proposed to destroy it with local loitering munitions. The operator, conditioned to trust the system and aware of the bureaucratic implications of passing up a high-value target, approved the strike despite personal reservations about its escalatory potential.

	The ground station, however, controlled both nuclear and non-nuclear communications equipment. From the perspective of the adversary state, the smaller state had severed its nuclear command and control systems in a critical region - perhaps to undertake an action which it assumed would warrant a nuclear retaliatory strike. Now blind in the region, the larger state deployed a tactical nuclear weapon to attempt to reestablish deterrence, instigating a response from the smaller state's allies.
↗ First Line Entities	DOD, Department of State (DOS), JCOS
↗ Second Line Entities	FEMA, Department of State (DOS), USPS