ARIZONA SENATOR
MARK KELLY

A Safety and Security Framework for Frontier AI

The more dangerous the AI, the stronger the safeguards

The same advanced AI models that could help cure deadly diseases can also help someone develop biological weapons or launch devastating cyberattacks. Yet today, companies are deciding for themselves which models are dangerous and how much protection is enough. **Too often companies act in the best interest of their bottom line, not the safety of the American people.** The result: one disclosure after another of models behaving dangerously, a string of near misses, and no system to stop them from happening again or hold anyone accountable. Whistleblowers are telling us of the risks that AI companies are taking, while experts speculate that it may only be a matter of time until there is a major incident. We can't wait until Americans are harmed to act.

This framework for AI security is based on decades of national security expertise and hard-won lessons in managing dangerous technology. It treats AI the way America already treats nuclear materials, dangerous pathogens, and weapons systems: the government defines the hazard, independent examiners verify the safeguards, and companies that hide known risks face real consequences. **We should hold AI to the same standard we have held other dangerous technologies: the more dangerous it is, the tighter the controls.**

Some will believe that the federal government can't — or shouldn't — take on this problem. They seem to have lost faith that we can do hard things. Looking at how we have handled other dangerous, rapid innovations in the past, one thing is clear: **It's not a question of if we have the power to do big, hard things and protect the American people — it's if we have the courage.**

The problem: no referee, no rulebook, no accountability

No common measuring stick. Nuclear facilities are sorted into hazard categories, and each category comes with its own required safeguards. Currently, AI has no comparable way to identify when a model becomes dangerous and no agreed set of safeguards to match each level of risk. Instead, a handful of powerful developers write their own with virtually no oversight. Without these standards, nothing else works: regulators can't set requirements, examiners can't enforce them, and courts face greater uncertainty in determining liability.

Companies oversee themselves, with outside review only after the fact. The Pentagon doesn't wait for a weapon to misfire before checking how it was built; it stations Defense Contract Management Agency staff inside contractors' manufacturing plants to oversee the work as it happens. AI developers, by contrast, police their own safety. Outside researchers typically step in only after something goes wrong, not while key decisions are being made.

A liability gap in national defense. Before a new weapon reaches the field, it must pass strict procurement standards, safety reviews, and independent testing. AI has no equivalent rules, even as the Pentagon spent nearly $200 million in 2025 to quickly bring commercial AI into military and intelligence operations. And legal protections written for traditional military weapons and equipment can shield AI companies when their models fail, even for risks they should have anticipated and addressed.



The framework: standards, oversight, accountability

Each piece builds on the last. Standards define what safeguards developers need, independent examiners make sure they follow them, and clear consequences hold them to it.

A common hazard scale — the rulebook. NIST, DOE's national laboratories, and the defense, intelligence, national security, and health agencies would work together, each bringing its own expertise and resources, to set AI hazard levels and the safeguards each requires. This is modeled on how DOE categorizes nuclear hazards and the CDC sets biosafety levels. The levels would cover weapons of mass destruction, cyberattacks, and loss of control, be tested against real incidents, and be published within 6 months with an implementation plan. The riskier the model, the stronger the protections.

Independent examiners on the inside — the referee. Modeled on federal oversight of major financial institutions and nuclear power plants, dedicated independent federal examiners team would be embedded at companies developing the most advanced AI models and review safeguards as models are built. Examiners would have full access to facilities and models to test whether safeguards match each model's risk and protect it from theft and misuse. The oversight would be strictly focused on national security, not content moderation or business decisions. Examiners identify and address national security risks before they become threats, because by the time an incident occurs, it's already too late to prevent the harm.

Weapons-grade accountability for defense AI — the consequences. AI used in high-risk military and intelligence missions should meet rigorous safety and reliability standards modeled on those already applied to existing weapons systems. As with any new weapon, AI tools and models would undergo documented risk assessments and independent testing before deployment, and defense contractors would be required to disclose known risks, testing limitations, and significant security or safety incidents to the government. Companies that knowingly conceal serious risks, recklessly disregard warning signs, or misrepresent their systems' capabilities or safety could face civil liability and penalties for resulting harm.

Proven tools for a new threat

America doesn't let companies decide for themselves how to secure nuclear material or handle deadly pathogens, and for good reason. These are technologies that impact the health and safety of all Americans, so when they were developed, our country took real steps to make them as safe as possible. Given the near misses and rapidly developing capabilities of cutting-edge AI, it's clear that frontier models deserve the same seriousness. This framework would give the country a clear hazard scale, independent examiners, and real consequences for cutting corners. It draws on decades of national security lessons. **It's the action we can take right now to protect the safety and wellbeing of Americans before a near miss becomes a disaster.**