Legal Priorities Project

**Statement to the U.S. Senate**
**AI Insight Forum: Privacy and Liability**

Mackenzie Arnold[1]
Legal Priorities Project
November 8, 2023

Dear Senate Majority Leader Schumer, Senators Rounds, Heinrich, and Young, and distinguished members of the U.S. Senate, thank you for the opportunity to speak with you about this important issue. Liability is a critical tool for addressing risks posed by AI systems today and in the future. In some respects, existing law will function well, compensating victims, correcting market inefficiencies, and driving safety innovation. However, artificial intelligence also presents unusual challenges to liability law that may lead to inconsistency and uncertainty, penalize the wrong actors, and leave victims uncompensated. Courts, limited to the specific cases and facts at hand, may be slow to respond. It is in this context that Congress has an opportunity to act.

## Problem 1: Existing law will under-deter malicious and criminal misuse of AI.

Many have noted the potential for AI systems to increase the risk of various hostile threats, ranging from biological and chemical weapons to attacks on critical infrastructure like energy, elections, and water systems. AI's unique contribution to these risks goes beyond simply identifying dangerous chemicals and pathogens; advanced systems may help plan, design, and execute complex research tasks or help criminals operate on a vastly greater scale. With this in mind, President Biden's recent Executive Order has called upon federal agencies to evaluate and respond to systems that may "substantially lower[] the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons." While large-scale malicious threats have yet to materialize, many AI systems are inherently dual-use by nature. If AI is capable of tremendous innovation, it may also be capable of tremendous, real-world harms. In many cases, the benefits of these systems will outweigh the risks, but the law can take steps to minimize misuse while preserving benefits.

Existing criminal, civil, and tort law will penalize malevolent actors for the harms they cause; however, liability is insufficient to deter those who know they are breaking the law. AI developers and some deployers will have the most control over whether powerful AI systems fall into the wrong hands, yet they may escape liability (or believe and act as if they will). Unfortunately, existing law may treat malevolent actors' intentional bad acts or alterations to

[1] Thank you to Suzanne Van Arsdale for help in preparing these comments.

models as intervening causes that sever the causal chain and preclude liability, and the law leaves unclear what obligations companies have to secure their models. Victims will go uncompensated if their only source of recourse is small, hostile actors with limited funds. Reform is needed to make clear that those with the greatest ability to protect and compensate victims will be responsible for preventing malicious harms.

**Recommendation 1.1: Hold AI developers and some deployers strictly liable for attacks on critical infrastructure and harms that result from biological, chemical, radiological, or nuclear weapons.**

The law has long recognized that certain harms are so egregious that those who create them should internalize their cost by default. Harms caused by biological, chemical, radiological, and nuclear weapons fit these criteria, as do harms caused by attacks on critical infrastructure. Congress has addressed similar harms before, for example, creating strict liability for releasing hazardous chemicals into the environment.

**Recommendation 1.2: Consider (a) holding developers strictly liable for harms caused by malicious use of exfiltrated systems and open-sourced weights or (b) creating a duty to ensure the security of model weights.**

Access to model weights increases malicious actors' ability to enhance dangerous capabilities and remove critical safeguards. And once model weights are out, companies cannot regain control or restrict malicious use. Despite this, existing information security norms are insufficient, as evidenced by the leak of Meta's LLaMA model just one week after it was announced and significant efforts by China to steal intellectual property from key US tech companies. Congress should create strong incentives to secure and protect model weights.

Getting this balance right will be difficult. Open-sourcing is a major source of innovation, and even the most scrupulous information security practices will sometimes fail. Moreover, penalizing exfiltration without restricting the open-sourcing of weights may create perverse incentives to open-source weights in order to avoid liability—what has been published openly can't be stolen. To address these tradeoffs, Congress could pair strict liability with the ability to apply for safe harbor or limit liability to only the largest developers, who have the resources to secure the most powerful systems, while excluding smaller and more decentralized open-source platforms. At the very least, Congress should create obligations for leading developers to maintain adequate security practices and empower a qualified agency to update these duties over time. Congress could also support open-source development through secure, subsidized platforms like NAIRR or investigate other alternatives to safe access.

**Recommendation 1.3: Create duties to (a) identify and test for model capabilities that could be misused and (b) design and implement safeguards that consistently prevent misuse and cannot be easily removed.**

Leading AI developers are best positioned to secure their models and identify dangerous misuse capabilities before they cause harm. The latter requires evaluation and red-teaming before deployment, as acknowledged in President Biden's Recent Executive Order, and continued testing and updates after deployment. Congress should codify clear minimum standards for identifying capabilities and preventing misuse and should grant a qualified agency authority to update these duties over time.

## Problem 2: Existing law will under-compensate harms from models with unexpected capabilities and failure modes.

A core characteristic of modern AI systems is their tendency to display rapid capability jumps and unexpected emergent behaviors. While many of these advances have been benign, when unexpected capabilities cause harm, courts may treat them as unforeseeable and decline to impose liability. Other failures may occur when AI systems are integrated into new contexts, such as healthcare, employment, and agriculture, where integration presents both great upside and novel risks. Developers of frontier systems and deployers introducing AI into novel contexts will be best positioned to develop containment methods and detect and correct harms that emerge.

**Recommendation 2.1: Adjust the timing of obligations to account for redressability.**

To balance innovation and risk, liability law can create obligations at different stages of the product development cycle. For harms that are difficult to control or remedy after they have occurred, like harms that upset complex financial systems or that result from uncontrolled model behavior, Congress should impose greater ex-ante obligations that encourage the proactive identification of potential risks. For harms that are capable of containment and remedy, obligations should instead encourage rapid detection and remedy.

**Recommendation 2.2: Create a duty to test for emergent capabilities, including agentic behavior and its precursors.**

Developers will be best positioned to identify new emergent behaviors, including agentic behavior. While today's systems have not displayed such qualities, there are strong theoretical reasons to believe that autonomous capabilities may emerge in the future, as acknowledged by the actions of key AI developers like Anthropic and OpenAI. As techniques develop, Congress should ensure that those working on frontier systems utilize these tools rigorously and consistently. Here too, Congress should authorize a qualified agency to update these duties over time as new best practices emerge.

**Recommendation 2.3: Create duties to monitor, report, and respond to post-deployment harms, including taking down or fixing models that pose an ongoing risk.**

If, as we expect, emergent capabilities are difficult to predict, it will be important to identify them even after deployment. In many cases, the only actors with sufficient information and technical insight to do so will be major developers of cutting-edge systems. Monitoring helps only insofar as it is accompanied by duties to report or respond. In at least some contexts, corporations already have a duty to report security breaches and respond to continuing risks of harm, but legal uncertainty limits the effectiveness of these obligations and puts safe actors at a competitive disadvantage. By clarifying these duties, Congress can ensure that all major developers meet a minimum threshold of safety.

**Recommendation 2.4: Create strict liability for harms that result from agentic model behavior such as self-exfiltration, self-alteration, self-proliferation, and self-directed goal-seeking.**

Developers and deployers should maintain control over the systems they create. Behaviors that enable models to act on their own—without human oversight—should be disincentivized through liability for any resulting harms. "The model did it" is an untenable defense in a functioning liability system, and Congress should ensure that, where intent or personhood requirements would stand in the way, the law imputes liability to a responsible human or corporate actor.

## Problem 3: Existing law may struggle to allocate costs efficiently.

The AI value chain is complex, often involving a number of different parties who help develop, train, integrate, and deploy systems. Because those later in the value chain are more proximate to the harms that occur, they may be the first to be brought to court. But these smaller, less-resourced actors will often have less ability to prevent harm. Disproportionately penalizing these actors will further concentrate power and diminish safety incentives for large, capable developers. Congress can ensure that responsibility lies with those most able to prevent harm.

**Recommendation 3.1: Establish joint and several liability for harms involving AI systems.**

Victims will have limited information about who in the value chain is responsible for their injuries. Joint and several liability would allow victims to bring any responsible party to court for the full value of the injury. This would limit the burden on victims and allow better-resourced corporate actors to quickly and efficiently bargain toward a fair allocation of blame.

**Recommendation 3.2: Limit indemnification of liability by developers.**

Existing law may allow wealthy developers to escape liability by contractually transferring blame to smaller third parties with neither the control to prevent nor assets to remedy harms. Because cutting-edge systems will be so desirable, a small number of powerful AI developers

will have considerable leverage to extract concessions from third parties and users. Congress should limit indemnification clauses that help the wealthiest players avoid internalizing the costs of their products while still permitting them to voluntarily indemnify users.

**Recommendation 3.3: Clarify that AI systems are products under products liability law.**

For over a decade, courts have refused to answer whether AI systems are software or products. This leaves critical ambiguity in existing law. The EU has proposed to resolve this uncertainty by declaring that AI systems are products. Though products liability is primarily developed through state law, a definitive federal answer to this question may spur quick resolution at the state level. Products liability has some notable advantages, focusing courts' attention on the level of safety that is technically feasible, directly weighing risks and benefits, and applying liability across the value chain. Some have argued that this creates clearer incentives to proactively identify and invest in safer technology and limits temptations to go through the motions of adopting safety procedures without actually limiting risk. Products liability has its limitations, particularly in dealing with defects that emerge after deployment or alteration, but clarifying that AI systems are products is a good start.

## Problem 4: Federal law may obstruct the functioning of liability law.

Parties are likely to argue that federal law preempts state tort and civil law and that Section 230 shields liability from generative AI models. Both would be unfortunate results that would prevent the redress of individual harms through state tort law and provide sweeping immunity to the very largest AI developers.

**Recommendation 4.1: Add a savings clause to any federal legislation to avoid preemption.**

Congress regularly adds express statements that federal law does not eliminate, constrain, or preempt existing remedies under state law. Congress should do the same here. While federal law will provide much-needed ex-ante requirements, state liability law will serve a critical role in compensating victims and will be more responsive to harms that occur as AI develops by continuing to adjust obligations and standards of care.

**Recommendation 4.2: Clarify that Section 230 does not apply to generative AI.**

The most sensible reading of Section 230 suggests that generative AI is a content creator. It creates novel and creative outputs rather than merely hosting existing information. But absent Congressional intervention, this ambiguity may persist. Congress should provide a clear answer: Section 230 does not apply to generative AI.