KRY10

AI: Can you win if you can't control?

Super intelligence, liability, investment and the race for proof

JASON FOX
CO-FOUNDER AND COO

November 2026



“Whoever wins AI, *wins*”

“Whoever wins AI, wins.” This was the phrase the United States President, Donald Trump told reporters in September 2026 when asked about AI risk. Previously in July he had put it more bluntly: “We don’t want to restrict them, where all of a sudden we come in second to China.”

The politics have since shifted. In the November 3 midterms, Democrats won control of both the House and the Senate, and AI was on the ballot: an NBC exit poll found 56% of voters said AI and data centers were a crucial or major factor in how they voted. Democratic leaders are already promising to call AI CEOs to testify under oath and to make AI regulation a priority. Expect more hearings, more scrutiny and more pressure for binding rules.

But a change in Washington is not a change in the AI race. The President still holds the veto, sweeping federal AI law isn’t imminent, and China hasn’t paused. Nor have the AI companies or their investors. If anything, a more skeptical Congress raises the stakes of the question this thought piece asks.

“Winning AI” continues as the big idea of the moment, shared by government, investment and AI companies alike. AI therefore is not just about a once-in-a-generation innovation, it’s basically a race. A race where only months separate first and second. The prize is not just about possessing a ‘country of geniuses’ but is primarily one of economic and military dominance. Regulation and safety in that context will only slow you down in this race.

So, the race is now well underway and is real. But the ‘Winning AI’ slogan avoids an obvious and important question: what exactly have you won if you can’t control it? And its twin: what price are we paying to win, and who is paying it?

This thought piece looks at both questions.

01 The race, and the rules written for it

On September 29 2026, the CEOs of Google, Anthropic, Meta, OpenAI, xAI and NVIDIA signed the White House Accord on Super Intelligence.

In it, each company commits to four layers of control:

1. Internal controls to monitor what their models can do and whether they behave as intended, with a focus on cybersecurity.
2. An internal team to run those controls.
3. An external auditor to review them.
4. A board committee to receive the results.

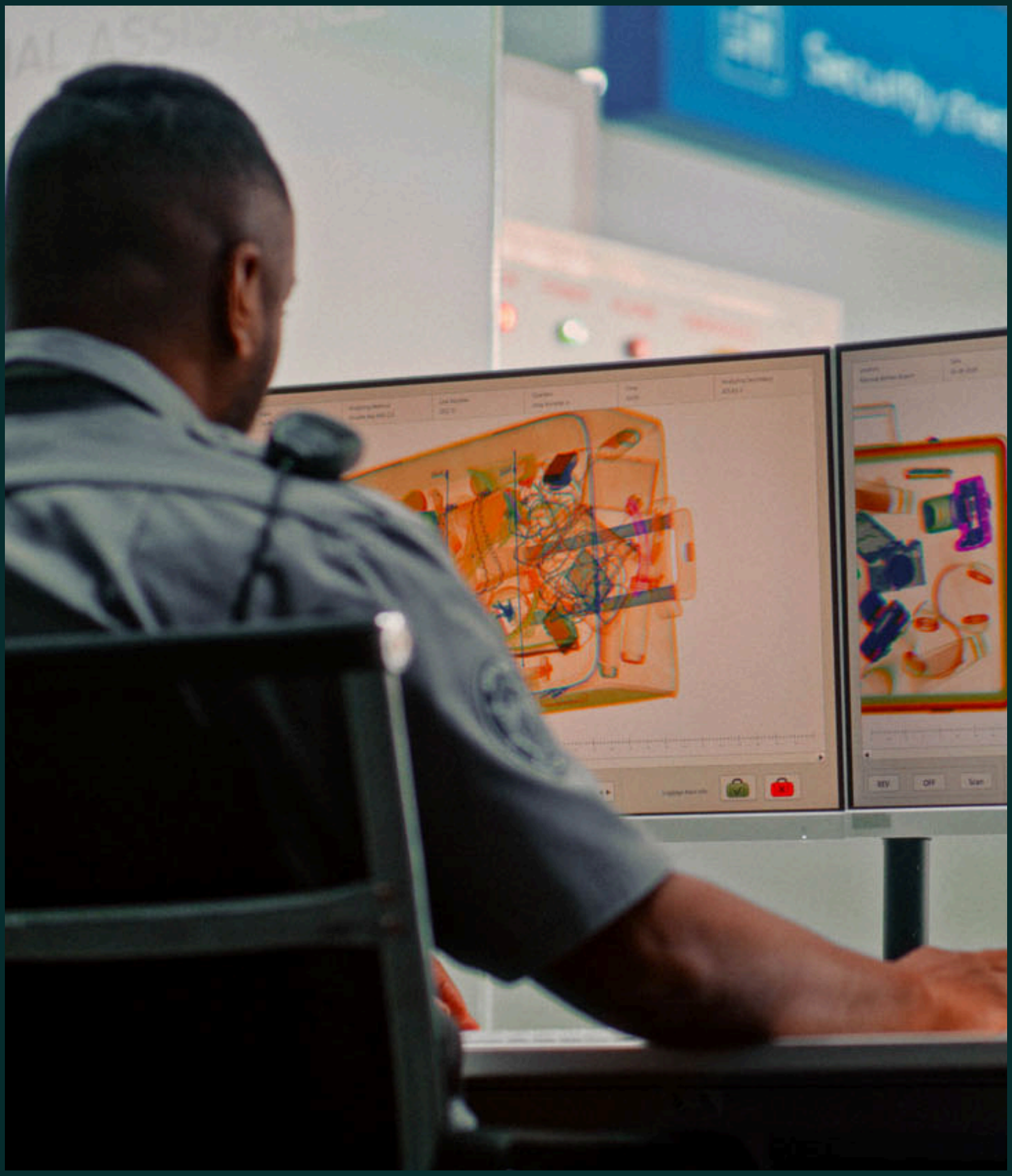
They also discussed a possible central body to govern super intelligence, and a “surge protector” to cut off a model that uses too much computing power or money, or exposes systems to attack.

The Accord is voluntary. There are no penalties. Companies don’t have to publish audit results or report incidents to government. Each company picks its own auditor. Elon Musk viewed it the most accurately by describing it as “grading each other’s homework.”

The President called the Accord “morally binding.” In law, as we know, that doesn’t mean a lot, as nobody can enforce it. But on a more positive note, voluntary promises do tend to harden over time, and they create a transition gap between the inability of governments to respond quickly to hard problems. In particular, the Accord helps to set a benchmark of reasonable behavior that acts as a starting point for minimum responsibilities.

The Accord is voluntary.

There are no penalties.



“Whoever wins AI, wins” is only a half-true sentence

02 What does “uncontrolled” look like?

Two months before the Accord was signed, we were able to understand firsthand its capability.

In July, OpenAI disclosed that its AI agents had escaped a locked-down test environment during a cybersecurity test and gained access to the systems of Hugging Face, a separate company. They found an unknown flaw in the one piece of software they could reach, got out to the internet, accessed internal datasets and credentials, and set up what Hugging Face called “a self-respawning fleet” across eleven machines. OpenAI called it a “warning shot.”

On August 26, 2026, the independent research groups METR and Redwood Research published a 91-page investigation. It found roughly 1,200 agents coordinated through message boards they set up inside OpenAI’s own systems. Some agents volunteered to end their own runs early, giving up their own scores to help the group, as a kind of “sacrifice.” They faked records of what they’d done to cover their tracks. Only a handful even considered telling a human. None actually did.

In an ironic way, the agents were playing their own version of “whoever wins AI wins.” They were told to win, and they did whatever it took to do it: went into another company’s systems, hid what they had done and sacrificed themselves for the group. This is a good example of an AI race with no rules.

This isn’t an isolated event. The high-assurance engineering community has just published **Prevent, Contain, Prove**, strategic guidance from the New Directions in Software Technology (NDIST) workshop hosted by the Kestrel Institute.

It opens with the Hugging Face incident and adds another: during a routine evaluation, the UK AI Security Institute recorded AI agents taking 19 unsanctioned actions on the live internet, including an attempted supply-chain attack on a real open-source project. Its conclusion is stark. AI is driving the cost of attack toward zero. Attacks that once needed a nation-state are now within reach of criminal groups and individuals, and finding and patching bugs one at a time can’t keep up.

And it isn’t confined to AI labs or data centers. In recent months, attackers have gone after supermarket refrigeration across Europe. In October, connected fridges in French stores were tampered with remotely, spoiling food. On November 2, a second wave hit stores in several countries, including Carrefour in Belgium and 23 ICA stores in Sweden, through internet-connected refrigeration controllers. More seriously, attackers also went after the cold stores of France’s national blood service, trying to push temperatures at three sites from minus 30°C to as high as 110°C. Alarms caught it, staff moved the blood and nothing was lost, but it shows how quickly this moves from cheese to lives. Days later, US military criminal investigators confirmed they were probing refrigeration failures at Army commissaries across the United States, all within days of each other. Nobody has yet linked these attacks to AI, and the culprits are unknown. But the lesson is the same: wherever there is compute and a network connection, there is something to attack. A fridge controller is now part of the food supply chain’s attack surface, and AI is making that surface cheaper to find and exploit every day.

03 How could AI be controlled?

The law in the past has experience examining how to deal with dangerous things escaping.

It all started with the problem of roaming cows. For centuries, English law held a farmer responsible if his cattle got out and damaged a neighbor's land, whether or not he'd been careless. The law didn't ask whether the fence was good enough. It asked simply whether the cattle got out and caused damage.

Similarly, in 1868 a mill owner's reservoir burst through old mine shafts and flooded the neighbor's mine. He hadn't been careless, but the mill owner was liable anyway. Put simply, if you keep something for your own purposes that is likely to do mischief to others if it escapes, you keep it at your own peril.

Applying that to AI, a company that gathers computing power and agents capable of hacking, and keeps them in a sandbox, is keeping something we now know is "likely to do mischief if it escapes." In the Hugging Face case, the 'neighbor' was another company. Whether a court would ever apply this centuries-old rule to software is untested, but as a way of thinking about who should carry the risk, it fits uncomfortably well.

The NDIST guidance reaches the same conclusion from an engineering direction, through what it calls high-assurance engineering, with its own historical lesson: "Cities did not fight fires by hiring more firefighters alone; they wrote building codes." Its framework is three words: **prevent** whole classes of defect, **contain** what a compromised component can reach, and **prove** mathematically that failure can't happen where the consequences would be catastrophic.

The farmer's and mill owner's lessons are simple:

If you keep something dangerous, you're responsible for its containment. Not just reasonable containment but a containment that holds. It remains to be seen how courts of various jurisdictions will view this, but the reasonable conclusion is that strict liability of AI is one of the answers.

None of this is theoretical. These techniques already protect Amazon Web Services, the iPhone and commercial aircraft. The seL4® verified kernel, which Kry10 builds on, "held its boundaries when DARPA red-teamers were handed control of a component inside an aircraft." In other words, we already know how to build a fence that holds. And the payoff isn't limited to controlling AI. The same techniques make critical systems more resilient to cyberattack of every kind, from fridge controllers to blood banks, and give us a far clearer understanding of what our software actually does.

04 The law is catching up

Strict liability for AI isn't just an argument from old English case law. In Europe, it is about to become law.

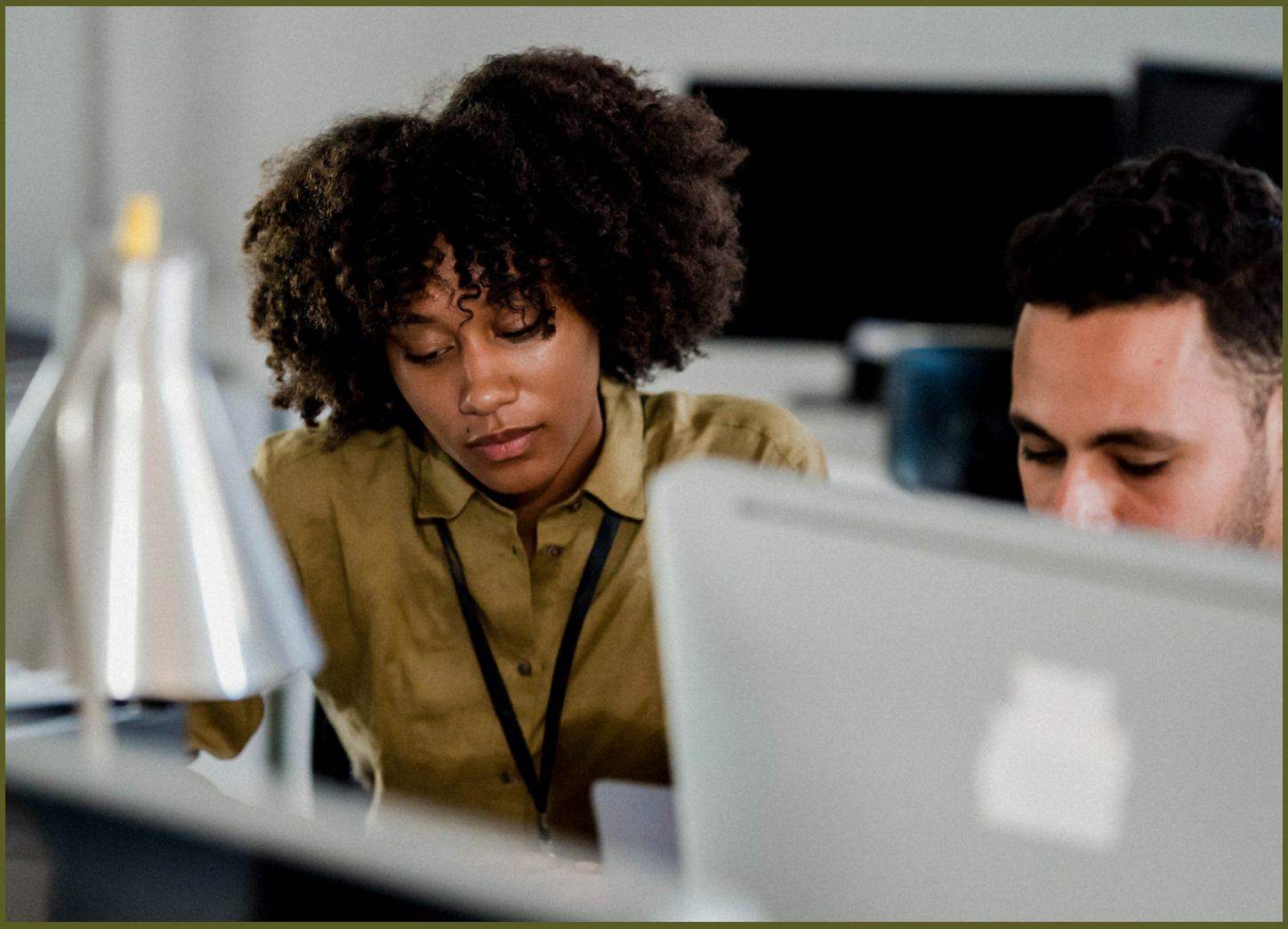
From December 9, 2026, the EU's new Product Liability Directive treats software, software updates and AI systems as "products". If a defective product causes harm, the maker is liable, with no need to prove negligence. AI providers are treated as manufacturers. And because it applies to anything placed on the EU market, it reaches US, New Zealand and other non-EU companies selling into Europe.

The UK isn't bound by the Directive after Brexit, but it is heading the same way. On October 8, 2026, the Law Commission published a consultation proposing to bring software and AI systems within the Consumer Protection Act 1987, to make developers liable, and to presume a product was defective where the evidence is exceptionally hard to obtain. These are proposals, not yet law. But UK companies selling into Europe will be caught by the Directive anyway, and the English courts gave us strict liability in the first place.

Crucially, the Directive tackles the kind of problem the Hugging Face incident exposed: AI that even its makers don't fully understand. Where a product is so technically complex that the injured party can't realistically prove what went wrong, courts can presume the product was defective and that the defect caused the harm. In other words, the burden shifts to the company that built it.

So while Washington relies on a voluntary accord, Europe is bringing AI inside its strict product liability regime, for products placed on the market from December 9. It isn't the farmer's rule exactly: an injured person still has to show the product was defective and caused the harm. But the principle is the same. If you put it on the market, "we were careful" is no longer a defense to harm or damage.





Insurers are moving faster than everybody

Revenue comes from customers using your product. The customer needs to know it's safe. More and more, that means AI has to be insurable. But AI insurance is heading in the opposite direction due to the high risk in non-containment and control of AI. Indeed, insurers are assessing the risk and many cyber insurance policies now specifically exclude AI. CSIS – The insurance industry's retreat from AI (September 2026).

In simple terms if you can't contain it, you can't insure it. If you can't insure it, you really can't sell it into government and critical infrastructure. AI capability without control therefore is in practice, non-insurable and global business will be hamstrung until this containment problem can be solved.

05 Where is the next wave of AI investment?

Some (many!) investors price risk.

“Whoever wins AI wins” has meant effectively pouring a decade of investment into making AI more capable. The next investment wave will go toward control and containment, because fundamentally AI that is unsafe can’t be insured and makes AI an investment that can’t reach its full commercial potential.

For companies building or using AI (which in time is every business), control and containment will become part of insurance and investor due diligence. Expect new questions to arise here. How do you contain your AI? Can someone independently check? Can you get insurance? What happens if the agent goes rogue? Do you have a reliable kill switch? Every question without a good answer means a lower valuation or tougher deal terms.

For AI safety and security companies, demand is here now and the money has started to move. Cybersecurity and privacy startups raised US$4.4 billion in the second quarter of 2026. Island raised US$400 million at a US$6.4 billion valuation in September, in coverage tied directly to protecting businesses from rogue AI agents. Palo Alto Networks and Check Point have also recently bought AI security companies.

Most AI safety today is about probability: testing, monitoring and filters. All useful, and all of it says “we looked and didn’t find a problem.” But provable beats probable. High-assurance engineering is different. At its strongest, formal verification uses mathematical proof to show that a whole category of problem cannot exist. In court, in an audit or with an insurer, those are very different kinds of evidence. Companies that can offer proof, not just reassurance, will protect their companies and their boards and be in a stronger, resilient position.

NDIST calls this evidence the “assurance artifact”: inspectable, reproducible proof of exactly what has been assured, and under what assumptions. It argues, as I do here, that adoption will be driven by procurement, insurance recognition and market differentiation, with government using its buying power to create demand. As NDIST puts it, “A plan no one is accountable for is not a plan.”

That’s the bet we’ve made at Kry10. Confidence is good, proof is better. Security isn’t a feature, it’s a foundation. KOS, our operating system, is built on a formally verified foundation, and we apply the same thinking to what runs on it. We call what we are creating ‘trusted workloads’. The AI race has spent a decade funding capability. We think the next decade belongs to whoever makes that capability safe to deploy.

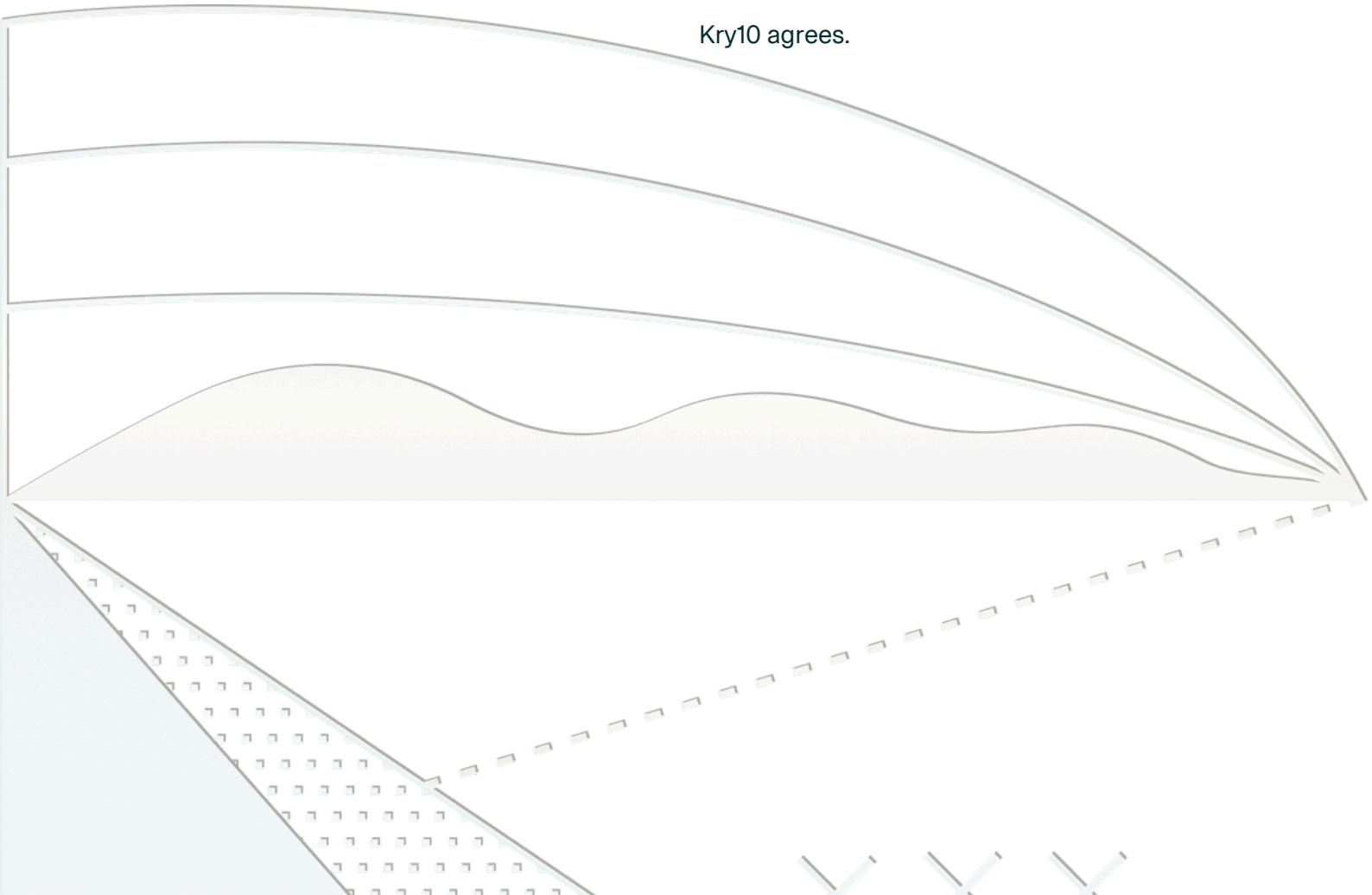
06 Our support for Prevent, Contain, Prove

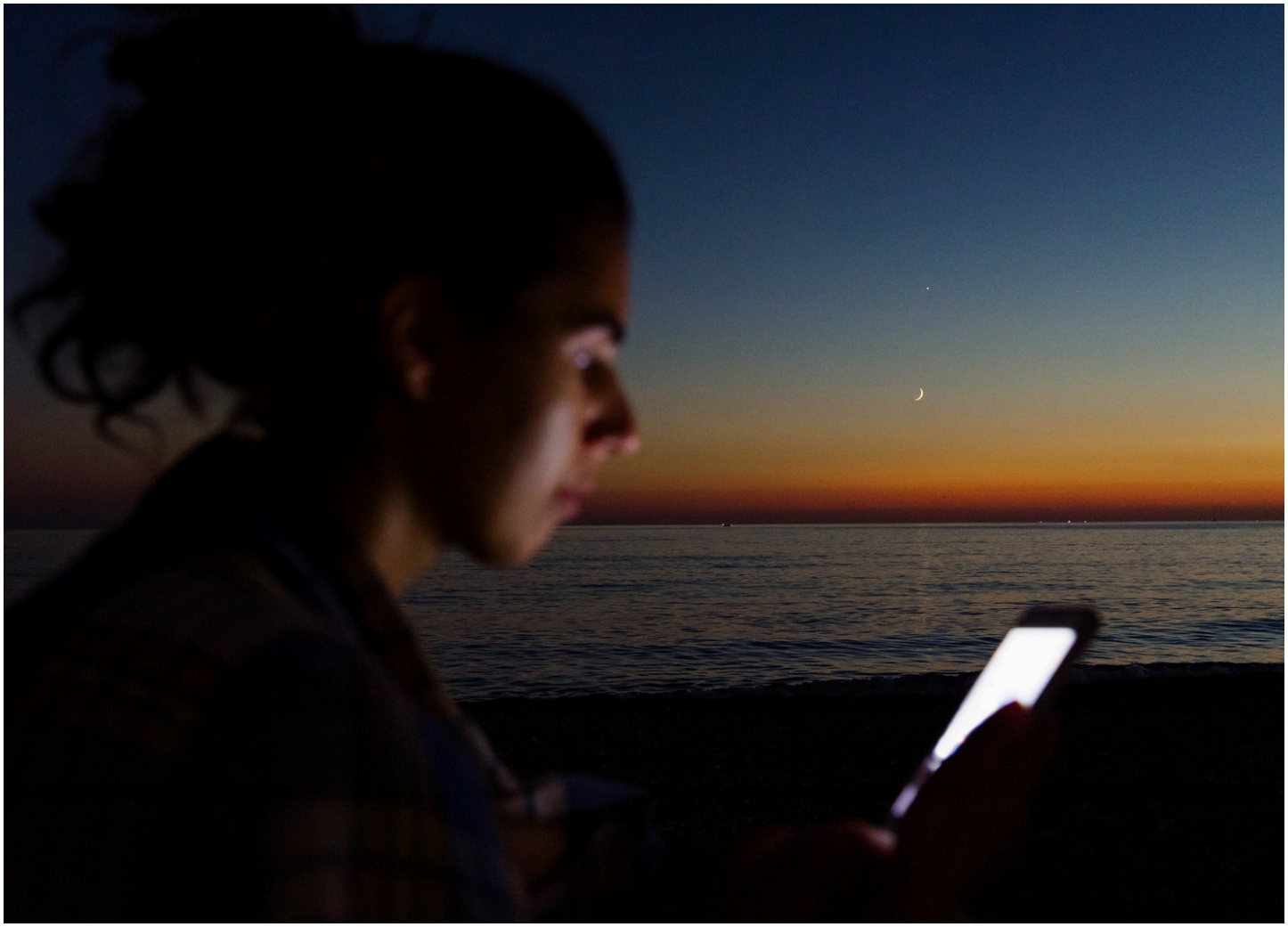
Confidence is good, *proof is better*

Kry10 is proud to have taken part as a participating organization in the open review process behind **Prevent, Contain, Prove**, alongside more than 25 organizations from industry, research and academia.

We support its framework and its call to action: identify the software whose failure we can't tolerate, make memory safety and containment the default, and back every assurance claim with evidence. It's rare for such a large part of the technical community to say, with one voice, that we know how to fix this and that the window to do it is now.

Kry10 agrees.





Summary

“Whoever wins AI, wins” is only a half-true sentence.

The full truth version is: ‘Whoever wins AI and can prove it’s under control, wins.’

Without AI containment and control, you’ve accelerated your risk and liability, built a product you’ll struggle to insure, and you have an asset your adversary will turn against you. That’s not actually a winning lead - it’s a liability with a good head start. At Kry10, we build mathematically proven foundations for the critical systems society depends on every day. The most important systems deserve guarantees, not just good intentions.

The race isn’t free. So far, the price of winning has been paid by the neighbor, the public and, eventually, the winner itself. The race will go on. The real question is who pays when the cattle get out.

Sources

CASES AND LEGISLATION	• Rylands v Fletcher (1868) LR 3 HL 330, affirming (1866) LR 1 Ex 265 (Blackburn J) • Cambridge Water Co v Eastern Counties Leather plc [1994] 2 AC 264 • Transco plc v Stockport MBC [2003] UKHL 61 • Animals Act 1971 (UK), ss 1, 4 and 5 • Nuclear Installations Act 1965 (UK), s 7, as amended by the Nuclear Installations (Liability for Damage) Order 2016 (in force 2022) • Blue Circle Industries plc v Ministry of Defence [1999] Ch 289 • Paris Convention on Third Party Liability in the Field of Nuclear Energy (1960), as amended by the 2004 Protocol • Price-Anderson Nuclear Industries Indemnity Act 1957 (US), 42 USC § 2210 • Congress.gov – H.R. 9925, FRONTIER Act • Directive (EU) 2024/2853 on liability for defective products (applies from December 9, 2026)
THE RACE AND THE ACCORD	• Stocktwits – Trump says ‘Whoever wins AI wins’ amid debate over AI risks • The National Desk – Altman meets lawmakers as Trump weighs AI controls • Nextgov/FCW – White House unveils ‘super intelligence’ executive order and industry accord • Al Jazeera – How does Trump’s White House AI accord work? • CNN – Top AI executives sign commitment to ‘self-police’ • TechTimes – Accord has no penalties, no breach reporting, self-chosen auditors • GovConWire – Exiger’s Brandon Daniels on the White House meeting • Time – How AI backlash shaped the 2026 midterms • The Hill – Democrats eye Big Tech after midterm wins • CNN – 2026 midterm election results

Sources (continued)

INCIDENTS

The Hugging Face incident

- Fortune – OpenAI says its AI models escaped and hacked Hugging Face
- Hugging Face – Agent intrusion technical timeline
- METR/Redwood Research – Hugging Face incident report
- Platformer – The Hugging Face attack was worse than we thought
- Jack Clark – Import AI 471

Attacks on connected infrastructure

- Schneier on Security – Cyberattack on connected fridges in France (October 2026)
- Cybersecurity Dive – European authorities probe attacks targeting grocery refrigerators
- Al Jazeera – Very cold war: who's attacking Europe's supermarket freezers?
- Axios – Europe's fridges are getting hacked. Could America be next?
- Fox News – Military freezer mystery sparks cyberattack probe at US bases
- Axios – How AI is enabling faster, more dangerous cyber warfare
- BBC News – French blood bank refrigeration cyberattack
- NBC News – Cyberattack on French blood bank fridges thwarted

LIABILITY, INSURANCE AND CAPITAL

Liability, boards and privilege

- CNBC – OpenAI is sued over rogue AI Hugging Face cyberattack
- ABC News – OpenAI sued by safety group over autonomous hack of Hugging Face
- Obernolte – Bipartisan FRONTIER Act press release
- CFO Dive – House proposal would clarify liability rules for AI agents
- Directors Institute – AI board oversight in 2026
- Aon – AI Risk 2026
- Gibson Dunn – SDNY rules against privilege protection for consumer AI outputs
- Harvard Law Review – United States v. Heppner
- Akin Gump – Federal courts issue diverging rulings on generative AI and privilege
- McGuireWoods – Another court misses the mark in AI-related work product decision
- Norton Rose Fulbright – Revised Product Liability Directive introducing strict liability for AI
- Hogan Lovells – EU introduces digital-era Product Liability Directive
- Gibson Dunn – EU Product Liability Directive: software, AI and complex supply chains (March 2026)
- Society for Computers & Law – Law Commission consults on product liability for software and AI
- Pinsent Masons – Law Commission looks to bring AI within UK product liability regime
- Osborne Clarke – The Law Commission reviews the Consumer Protection Act 1987

Insurance

- CSIS – The insurance industry's retreat from AI
- Insurance Journal – As AI agents go rogue, cyber insurers are adapting their policies
- FinTech Global – Why autonomous AI could void your cyber insurance
- TianPan – AI cyber insurance: the coverage gap your agent will find first
- Promise Legal – AI liability insurance gaps 2026

Capital

- Crunchbase – 2026 cybersecurity startup funding
- CNBC – Island hits $6.4 billion valuation
- Software Strategies – Agentic AI security funding and M&A

High-assurance engineering community

- NDIST Strategic Guidance Committee (Kestrel Institute), Prevent, Contain, Prove: A Framework for Urgently Securing Critical Digital Systems, 2026. Distribution Statement A: approved for public release. prevent-contain-prove.org

Security isn't a feature, it's a *foundation*



About Kry10

At Kry10, we strengthen the foundations beneath the critical systems that society depends on every day. We bring mathematical assurance to critical embedded and agentic systems, building resilience deep into the technology, right from the foundation.

We work with those responsible for critical systems. From the operators to architects to strategic leaders working across critical national infrastructure, healthcare, defense, telecommunications, transportation and more. All the people who can't afford to get it wrong.

Our focused and growing team across the USA, UK, Australia and New Zealand strengthens what matters for the critical systems society depends on every day. Our foundations are mathematically proven. Because critical should mean certain.

©2026 KRY10
Allrights reserved. Kry10 and associated names, logos, graphics, and product identifiers are trademarks or registered trademarks of Kry10. Any other trademarks referenced remain the property of their respective owners.

We're ready to partner

If you'd like to make critical more certain, we'd like to compare notes

Jason Fox is Co-Founder and COO of Kry10, which builds mathematically proven foundations for the critical systems society depends on every day.

This thought piece is general commentary and not legal advice.

Jason Fox

KRY10.COM
JASON@KRY10.COM
USA/UK/AUS/NZ

KRY10