

Australia says its National AI Standards will make frontier labs prove their safety systems work

Source copy kept by wheresthe.ai for the story "Australia says its National AI Standards will make frontier labs prove their safety systems work" (<https://wheresthe.ai/d/au-charlton-systems-based-frontier-ai-standards-speech-2026-10/>).

Source: <https://www.minister.industry.gov.au/charlton/media/speech-the-sydney-trust-and-safety-festival>

Text saved 19 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: a69a670c6a2c4f70a0d1f393689adbfd7a286b9d6a55c0b6392e61b04bee580

****INTRODUCTION****

Thank you for having me here.

Keeping Australians safe is a foremost priority of the Government in a digital world.

Today I want to outline how the Albanese Government is approaching the question of frontier AI regulation when it comes to the most significant harms.

Last month, the Prime Minister stood up in New York and revealed an AI agent, built by a frontier lab, had hacked into an Australian Government system.

Nobody intended for the AI agent to literally do this.

The model was tasked to look up public information about health spending. When it hit a barrier, it found a way around it. When it hit another, it did the same again.

In July, another OpenAI model broke out of its test environment and breached the systems of Hugging Face, one of the world's largest AI platforms.

As a starting point, no company should release a frontier AI model that is not safe.

Yet the fact that has occurred, and the fact that the labs did not detect or prevent it, prompts important questions about the role of new regulation in the National AI Standards.

First, I want to explain why the frontier needs regulating and how the unique characteristics of frontier models interact with existing regulatory approaches.

Second, I will outline why the market will not fix this alone, even though the companies want it fixed.

And third, I will lay down some early markers of the approach Australia is taking.

****WHAT WE ARE TRYING TO ACHIEVE****

The National AI Plan rests on three pillars, and together they are the foundation the Albanese Government's approach to AI.

One pillar is capturing the opportunities. Artificial intelligence can lift productivity, accelerate scientific discovery, improve diagnosis and treatment, and open up new industries. Australia should be a country that builds, trains and applies these technologies, not one that simply imports them. If we harness this technology as a nation, it will be a source of prosperity for the rest of this century. If we leave it to be developed elsewhere and shipped in, we will be poorer, and dependent on technology designed in other countries, for other countries' priorities.

The next is sharing the benefits. As the industrial revolution taught us technology does not distribute its gains evenly on its own. If the benefits of AI flow only to a handful of firms and a handful of places, we will have failed. Workers, small businesses, regional communities and families should all see the upside.

The final pillar is keeping Australians safe. For the Government this is critical in the Australian context. People will only adopt technologies they trust. Businesses will only invest in what they believe is reliable. Nobody shares the benefits of something they are afraid of.

Safety is not the brake on the AI opportunity. If anything, it is the condition on which the benefits can be fully realised.

****PART ONE: WHY AI NEEDS REGULATING****

AI is not the first technology to bring both promise and risk. Medicines, cars, aircraft, banks, building materials, even children's toys. Each makes our lives better. Each can hurt us.

And we do not treat them all the same way. Our rules sit along a spectrum.

At one end is the heaviest hand. Some weapons are simply banned. Medicines must be proven safe and effective – and go through rigorous testing – before anyone can buy them. Aircraft cannot fly until they are certified. Banks are subject to prudential supervision and systems testing, because when a bank fails, it does not fail alone.

At the other end are more flexible approaches. Most of what we buy is governed by general consumer law, the threat of being sued, and the judgement of customers. A clothing company does not need government approval to launch a new jacket. If a product is poor, people stop buying it. If it is dangerous, the company is liable. Most of the time, that is enough.

So we need a framework to help determine what type of regulatory model is appropriate for the nature of the harm we are regulating and the characteristics of the industry that produces the harm.

First, how severe is the harm and can we tolerate repeated failures? A faulty toaster hurts one household. A bank collapse hurts an economy.

Second, can the impacts be undone? Where harm is catastrophic and irreversible, the threat of a lawsuit is not always effective, because our primary goal is prevention rather than future deterrence.

Third, who bears the risk? When the person choosing a product carries the risk, we lean toward their freedom of choice. When the risk falls on people who never made a choice, and for whom exercising choice isn't really an option because of market concentration or a lack of control, the case for intervention grows.

Fourth, can anyone adequately judge the risk before it arrives? Markets punish what people can see. They cannot punish risks that aren't visible.

On all four dimensions, frontier AI pushes the limits.

High-risk AI systems present regulatory challenges distinct from conventional product harms. And to understand why we need to understand its properties.

1. ****Hard to recall****: a capability, once digitally released, cannot be readily recalled, and a model released anywhere is available everywhere. 2. ****Rapid pace of change****: AI model releases occur on compressed timeframes in months which outpaces standard regulatory rule making cycles, which are often in years. 3. ****Information asymmetry****: developers understand what their systems can do better than any regulator, customer or government. 4. ****Safety is global****: safety is not achieved by countries going it alone – ultimately capability is global in reach and impact and countries should work together. 5. ****Misalignment and latent risks****: risks can emerge after deployment in unexpected and unintended ways, and systems can behave in ways that were not intended.

That last point, in combination with the rapid pace of change, is what makes AI truly different and is why we have to take security and safety so seriously. Traditional software is built from rules people write and can inspect. Frontier AI is not. Neural networks – the digital circuitry that underpins AI – are grown from patterns learned across vast amounts of data, and its abilities emerge in ways even its creators cannot predict.

Contrast the complexity of that with social media for example. For years Facebook faced questions about the harm its platform could do, especially to young people. Changes to the algorithm and the platform were technically within its reach. They were not made, because those changes would cut against the business model. Facebook internalised the profits but externalised the social costs. This led the Australian Government to intervene to place age limits on social media, and more recently, the US courts have followed suit.

With frontier AI the technical problem is harder. There is likely no single switch to control advanced capabilities perfectly, rather layers of controls that safety teams are still learning to build.

****PART TWO: WHY THE MARKET WILL NOT FIX IT****

The diffusion of new technology should be based on trust.

Yet every indication is frontier labs are putting capability ahead of safety

Just listen to the people who know these companies best.

In 2024, OpenAI's head of alignment, Jan Leike, resigned. He said that at OpenAI, safety culture and processes had "taken a backseat to shiny products." Last month, Jacob Coxon, who had worked at both OpenAI and Anthropic, quit the industry altogether. His verdict was blunt: "Neither company is acting responsibly." More than a thousand people working at frontier AI companies have signed a petition asking the US government to help slow the pace of development, with even leaders of the companies themselves saying the technology is evolving too quickly.

The companies are clearly in a race that has two tracks. One with each other. And the other between the US and China.

And for now – everyone has calculated that the biggest reward comes from being first, even when it's not clear what the higher purpose of coming first is.

Why is this happening?

Economists have descriptions for what is going on here, and I want to use them, because they tell us what kind of problem this is.

The first is a negative externality: the situation where a company keeps the gains, while the costs of getting it wrong fall on everyone else.

The second is the prisoner's dilemma. Every lab would be better off if they slowed down together. But so long as profits are internalised, and costs externalised, any lab that slows down alone simply hands the lead to one that does not.

None of this requires anyone to expressly act in bad faith.

In fact, each company tells itself a story in which the frontier is safer in its own hands. Some of those stories may be true, and others are soothing forms of fiction. In any event, a story that lets you keep racing with a clear conscience still gives you a reason to keep racing. And that is why government needs to step in.

****PART THREE: AUSTRALIA'S APPROACH****

Australia has a long history of dealing with these challenges.

In 1974, the Whitlam Government passed the Trade Practices Act. Before it, if a product was faulty or a claim was misleading, the burden fell on the buyer. National rules of fair dealing did not weaken our markets. They strengthened them, because honest businesses no longer had to compete against dishonest ones.

We did it in workplaces. The idea that going to work should not cost you your health, or your life, was once treated as a cost to business. National work health and safety laws gave employers clear duties and gave workers confidence those duties would be met. Being reckless with the safety of workers could no longer be a source of cost savings or competitive advantage.

And we have done it online. The social media minimum age puts responsibility where it belongs, on the platforms, instead of leaving parents to fight some of the best-resourced companies in the world alone. Getting kids addicted can no longer be a source of revenue growth.

What unites these reforms is not a simplistic belief that markets are doomed to fail. It is a more fundamental belief – a Labor belief – that society runs on trust.

When trust is high, people participate with confidence, businesses compete on quality, and everyone benefits. When trust collapses, everyone pays.

Some will say that in a race, capability will always beat safety, and that anyone who says otherwise does not understand markets.

The Prime Minister has rejected that argument, and rightly, because it has the logic backwards. It is why the Government is developing AI standards legislation, and why the lessons of the Medicare incident will shape it.

A race to the bottom is not what competition looks like when it works. It is what competition looks like when it fails.

And market failure is not a reason to give up on markets. It is the reason policy exists.

And this brings us to the Albanese Government's regulatory approach.

This is a whole-of-economy technology, and it needs a whole-of-government response.

Across the economy, agencies and regulators are taking responsibility for AI harms. The Attorney General is progressing privacy reform and automated decision making.

The Minister for Communications is progressing a digital duty of care.

The Minister for workplace and employment is looking at safety in the workplace.

And the Assistant Minister for Competition and Consumers is exploring issues to do with agentic commerce and retail surveillance pricing.

These are all existing harms that could be accelerated by AI.

But the frontier presents new and novel risks, why is why we have established the AI Safety Institute to provide analysis and advice on how the frontier is developing, emerging processes and methods to ensure AI is safe, and to support government agencies and regulators to respond.

Naturally this brings us to the question of what regulatory approach is most suitable for the frontier and our objective of building trust through the National AI Standards.

When it comes to the most serious frontier risks voluntary regulation and codes are fast and flexible, but they fail the incentive test. You cannot ask firms to volunteer against their own competitive interest amid a manic race involving trillions of dollars and then be surprised when they do not.

Detailed prescriptive rules can offer clarity, but at the pace this technology moves, which is months not years, those rules could be out of date before the ink is dry.

For these reasons, there is much that Australia can learn from systems-based regulatory approaches, that puts the onus and accountability on companies for managing how they develop products and services safely.

Systems regulation places the onus on companies to build and run a rigorous process for finding, testing, reporting on and managing the risks of their systems, and then holding them accountable for whether that process works. It is a model used in workplace health and safety, security of critical infrastructure, prudential supervision of systemic banking risks, and aspects of aviation safety. The goal is to ensure companies are meeting expectations of safety, without trying to specify every hazard.

Government sets the standard those processes must meet, and ensures companies have robust processes in place. The question isn't only "did something go wrong" – but is there a serious system in place to detect risks and prevent incidents occurring.

These types of approaches could ensure that for developers of models with the sharpest, most acute frontier AI risks, the cheapest path to market is ultimately the one that runs through safety, not around it.

Because when people trust your model is safe to use – they're more likely to use it.

****PART FOUR: REGULATION ALONE IS NOT ENOUGH****

It is also important that we are honest about the limits of regulation.

Our rules will raise the floor for firms that operate within our laws. And if we are successful, they will have a multiplier benefit if they are adopted in other parts of the world.

But they will not stop hostile states or criminals who were never going to comply.

Relying on regulation alone is like building a fence around a horse, only to realise that horse has wings.

That reality is very critical to the agenda we have.

From the perspective of Government, regulation cannot be our only strategy when it comes to keeping Australians safe. This must be supplemented by both leverage and capability. If Australia wants a say in how these technologies are built and governed, we need to be a country that can host and build them, not only regulate them.

Australia has long punched above its weight in global matters. We have led the way on nuclear disarmament, on responding to biological harms, and in responding to major financial upheavals. We have a chance to do so again with AI.

Because we don't just have views on safety when it comes to nuclear or biological harms. We invest in the capabilities ourselves. We attract the best scientists. We make testing possible here.

The strength of our contribution is in part because we know what it means to build frontier technologies with risks – which is how our regulation keeps pace.

When it comes to leverage, technology that is housed overseas cannot always be regulated as effectively as technology Australia has a role in developing at home. And this underscores the importance of Australia having a seat at the table when it comes to AI safety, to shape it in our national interests.

That is why this Government is committed to enabling AI training onshore and developing AI talent here at home. This is our chance to shape this technology – for our standards to set the rules of the road globally over decades.

And when it comes to capability this means government, businesses, workers and safety specialists who have the capability to work with other governments and companies to make sure AI works in Australia's interests.

****CLOSE****

I have spoken today about economic incentives, regulation and policy. Those are the tools we have, and they can help. But underneath them sit questions of a different order. Questions of ethics, and of morality. A technology this powerful forces us to ask what we mean by intelligence, and what it is that makes us human. No single piece of legislation can answer those questions. But they should sit behind every piece of legislation we write.

So let me return to where I began.

AI needs regulating because its harms are severe, hard to undo, borne by people who never chose them, and sometimes invisible until they arrive. The market will not fix this alone, because the incentives reward speed and capability, and even the people at the top of the industry cannot slow down by themselves. Australia's answer is to put the burden where the risk is created, to build accountability into the whole system, and to build our own capability alongside it.

That is why we have setup the AI Safety Institute, announced our AI safety priorities, and committed to legislating National AI Standards for frontier AI.

And that is why we are deeply committed to developing Australian capability.

There is a race underway with the dominant purpose of who gets there first. Our task is not to win that race, and it is not to sit it out. It is to change it, so that this technology is built to serve people.

Safety is not the brake on the AI opportunity. It is the ignition.

Thank you.