Harkness Fellowship Report

Sarah Box[1]
2024 New Zealand Harkness Fellow

Introduction

My Harkness Fellowship tenure in the United States spanned three months from mid-September to mid-December 2024. The primary aim of my research was to gain a deeper understanding of U.S. policy directions for trustworthy and responsible artificial intelligence (AI) and to consider how a small country like New Zealand could best learn from and align with its strategic partner.

It was a unique time to be in the U.S., with the Presidential election creating societal tensions and uncertainties. Over my stay, I built a much better appreciation of the U.S. government system and the country's history, economic pressures, and social divides. It has been an enriching experience.

There is much speculation, as commentators parse announcements from the President-elect and his team to predict likely policy choices. However, regardless of what 2025 brings, I think it has been incredibly helpful to look back at the evolution of AI policy in the U.S., to understand more about the processes, motivations and personalities behind the current approaches, and hear reactions to what is in place.

Approach

My research methodology was a mix of interviews with experts, participation in AI-oriented events, and desk-top research. Geographically I was based in Washington D.C., but also made research trips to Atlanta and San Francisco (the latter on two occasions) and held several virtual meetings. It was a privilege to tap some of the deep knowledge in the U.S. and to have time to attend events and read more of the torrent of daily AI news and analysis.

Colleagues at my host organisation, the Observer Research Foundation (ORF) America in Washington D.C., were very welcoming and I'm grateful for their willingness to take on a visiting Fellow. They made useful introductions to several other thinktanks, organised an event that brought together their ongoing work on semiconductors and our shared interest in AI, and facilitated my participation in a workshop at Georgia Tech on AI governance. The Foundation's connections to India (it is affiliated to India's Observer Research Foundation) offer interesting perspectives more generally for New Zealand and I hope this will be a useful contact in future.

[1] This report reflects the personal research, analysis and views of the author, and does not represent the position of the New Zealand Harkness Fellowships Trust, the Observer Research Foundation America, or the New Zealand Government. It is an abridged version of the end-of-tenure report provided to the Harkness Fellowships Trust Board on 14 December 2024.

In addition, I received some initial leads from former OECD colleagues, and some very kind introductions from experts and contacts in New Zealand and the U.S. Not all introductions led to meetings, but it was enough to kick off a chain of connections.

Unfortunately, my plans to visit Tennessee were derailed by Hurricane Helene, which devastated the northwestern edge of North Carolina and impacted the Knoxville corner of Tennessee where Oak Ridge National Laboratory is based.

A full list of my meetings, plus events in which I had an active role, and events that I attended either in-person or virtually as a guest, is appended. I'm extremely grateful to everyone that gave up their time to speak with me and to those who opened doors to further meetings with experts.

AI policy in the U.S. to date

Understanding the current U.S. AI policy landscape was a key pillar of my research. This gave me a crash course on U.S. government structures, since AI policy emanates not only from the Executive Branch (White House and federal agencies) but also lawmakers in Congress, and from the states. I also gained a historical view since AI policy builds on decades of attention to information technologies. Below are some of the main policy actions that fed into my thinking.

Executive actions

AI policy took off under the first Trump administration (Jan 2017 – Jan 2021) as the practical potential of the technology became clearer. The administration focused on leveraging AI for national competitiveness, building public sector capabilities, and putting markers in the ground internationally (e.g. championing the OECD AI Principles). These policies, rolled out via Executive Order (EO) and legislation, remain in place. Notably, it was during this administration that the National Institute of Standards and Technology (NIST) was mandated to develop a risk management framework for AI. This has since evolved into a major part of the policy architecture.

The Biden administration put more emphasis on civil liberties and issues such as bias, launching an AI Bill of Rights in late 2022. But like other countries, they had their "ChatGPT moment", which pivoted the second half of their term to drafting and implementing the EO on the *Safe, Secure and Trustworthy Development and Use of Artificial Intelligence* (EO 14110). The longest EO ever to be issued, it requires a swathe of actions from agencies related to their own use of AI as well as "outward facing" domestic and international AI policy, tied to very specific (and tight) deadlines.

Agencies have been delivering at a rapid pace, apparently due to both "sticks" ("the White House asked for it") and "carrots" (agencies can point to the EO to ask for resources to fulfil their tasks). A key action was creation of the U.S. AI Safety Institute (AISI) to support EO responsibilities assigned to the Department of Commerce, including developing standards for safety, security and testing of AI models and for authenticating AI-generated content, and providing test environments for researchers. The AISI is engaging widely and clearly signals a safety and security focus. (It is still awaiting legislative backing however.) Along similar lines, a recent major deliverable is a White House memo on harnessing AI in the U.S. government to support national security objectives.

The EO emerged from an OMB-OSTP-White House Deputy Chief of Staff process. I was told there was a strong tech-sceptic viewpoint flowing through the drafting, coming from a sense that Washington had "failed" on social media regulation. The EO task list built on suggestions from agencies, who were given a relatively short timeframe to contribute. This pressure led to some "approximations", notably the contested measure of 10^{26} floating point operations (a proxy for computing power, as a proxy for AI model capability) that marks the threshold for private sector compliance requirements.

The EO implicates most if not all federal agencies and I suspect there is some jostling for policy leadership. Some people considered OMB and OSTP at the centre, with OMB holding funding levers as well as being the home agency for the Federal Chief Information Officer, who has oversight of AI in government and a voice over AI regulation, while OSTP focuses on policy drafting. Others downplayed OSTP's role, and I was interested to learn that OSTP received strong push back on its AI Bill of Rights. The Departments of Commerce (home to NIST), Energy, Defense, Justice, and Homeland Security, play very important roles.

Aside from the EO, the federal government has used "voluntary commitments" with firms as another tool to support trustworthy AI. The first were announced in mid-2023, when President Biden convened seven leading AI companies to the White House, with an agreement to ensure products are safe before introducing them to the public, build systems that put security first, and earn public trust (including developing AI-content labelling). In September this year, the White House announced new voluntary commitments to combat image-based sexual abuse.

The Partnership for Global Inclusivity on AI, announced by Secretary of State Antony Blinken at September's United Nations meetings, is probably viewed by firms as another set of voluntary commitments. OpenAI, Microsoft, IBM, Amazon, Google, Meta, Nvidia and Anthropic are investing in a public-private partnership to boost compute, skills and "context" (i.e. local datasets) in low- and middle-income countries. The government frames this initiative as part of international AI outreach, helping to spread the U.S. vision for AI governance. It is part of a growing narrative on inclusiveness, better communicating U.S. support of developing countries. Announcements of a Global AI Research Agenda and an AI in Global Development Playbook also position the U.S. to support AI ecosystems beyond its borders.

I did not delve much into the role of trade and industrial policies in the AI policy landscape, but these clearly have an influence. The U.S. imposes export controls on the sale of AI chips to China, for instance, and one imagines these tools will become a stronger policy channel under the new Trump administration. Regulatory agencies also influence AI in their specific domains. For instance, the Federal Trade Commission (FTC) is increasingly using its powers regarding unfair and deceptive conduct to bring cases related to AI. An early case was against a retailer using facial recognition to remove shoppers from stores, and in September the FTC launched five new cases against firms "using AI to supercharge deceptive or unfair conduct that harms consumers".

Congress

People say there is no federal-level AI law, but that isn't strictly true. Congress has passed several bills aimed at building U.S. AI capabilities, including the National AI Initiatives Act 2020 that established AI Research Institutes across the U.S. The House Committee on Science, Space and

Technology has passed nine bipartisan bills on AI that could progress through Congress during the lame duck period. Among these are the CREATE AI Act, which would stand up a National AI Research Resource of compute, data and testbeds for AI researchers across the U.S., as well as a bill on AI literacy for students, and an amendment to the National AI Initiatives Act to require a center to advance work on safety (I think aimed at putting the U.S. AISI on a firm footing).

What there is not, is federal-level AI regulation of the EU AI Act type, nor does there appear to be anything like this on the horizon. The Senate AI Caucus created in 2019 has apparently become less active, and the House Caucus on AI remains more focused on educating lawmakers. A more recent bipartisan Senate AI Working Group (which some say hijacked the Senate Caucus) has a roadmap setting out AI policy priorities, but these are around innovation funding, ensuring enforcement of existing laws, addressing deepfakes related to election content and looking at AI's impact on content creators and journalism, and establishing a federal privacy law, amongst others.

With the shift to a Republican majority in the Senate and House, there is perhaps a smoother path for passing legislation over the next two years. It seems unlikely there would be appetite for a broad regulatory bill affecting firms. More likely could be a bill that brings some uniformity to automated decision making in areas where civil rights loom large, where states of both colours have been putting most of their legislative efforts. This might lean on a Texas bill – more on that below.

State-level policies

States are very active in proposing laws on AI. The argument is that national lawmakers are not stepping up, so states need to fill the gap as they have previously done on privacy. Draft bills are said to number in the hundreds – no-one has a definitive figure (there are myriad trackers).

In practice, most bills target very specific issues (e.g. deepfakes in elections, setting up AI advisory boards) and many are aimed at government use of AI. Some appear tangential to AI.

Of more comprehensive bills, the Future of Privacy Forum says most are focused on regulating AI to mitigate algorithmic discrimination, in areas covered by civil rights law (such as education, housing and employment). These apparently have some consistency, e.g. using the OECD definition of AI and setting out specific responsibilities for developers vs deployers of AI. But there are big debates about how prescriptive to be about scope, whether the AI in question needs to "control" decisions (or just "facilitate" them) to be covered, and whether to prohibit algorithmic discrimination or take a softer line (e.g. impose a duty of care to prevent it).

Relatively few bills are taking a technology specific approach. California's vetoed Senate Bill (SB) 1047 – the *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act* – is perhaps the most notable. This bill was significant for the level of liability it would place on AI developers in California and for its potential to shape AI regulation more widely in the U.S. and globally due to the headquartering of many big AI firms in the state. The bill focused on large scale AI models and was heavily debated by academics, lawmakers and industry players. In the end, Californian Governor Gavin Newsom sided with those that considered it was overly focused on hypothetical risks and required technical solutions that do not yet exist. He has requested further work – so this bill remains something to watch.

When you look at legislation in force, the sample shrinks. The IAPP (another privacy body) identifies just four enacted laws that impact private sector organisations and only one of these (Utah) is in force, with the other three taking effect in early 2026.

- Utah's AI Policy Act aims to clarify that existing consumer protection laws apply to generative AI. It also sets up an Office of AI Policy, which incorporates a regulatory sandbox.
- Colorado's AI Act is focused on AI for consequential decision making. It imposes a variety of duty of care, disclosure and impact assessment requirements. It is contested (especially by the start-up community) and will be amended before taking effect to address concerns it is overly broad and deters investment.
- California's AI Transparency Act requires genAI system providers to offer AI detection tools and provide users with methods to place a disclosure on AI-generated material. It also requires providers to themselves include disclosures on gen-AI material.
- California's Generative AI – Training Data Transparency Act requires developers to post documentation regarding the data used to train the genAI, including sources, data descriptions, whether any data is protected by IP, etc.

The sample is slightly wider than this if you include AI-relevant legislation pertaining to the public sector and to individuals (e.g. California passed amendments to its Labor Code to clarify situations of digital replicas and require informed consent, as well as amendments to its Penal Code related to distribution of intimate images). But the point remains that enacted legislation remains limited.

People suggested it is worth watching legislative activity in Connecticut, Texas, California, and New York, as their examples could cascade out to other states. The Colorado Act mentioned above was apparently modelled on a Connecticut bill that failed but is being "retooled" for 2025 and may become the "Democrat template" for state-level AI bills. A "Republican template" may emerge from Texas, with a draft Texas Responsible AI Governance Act looking to regulate high-risk AI systems making consequential decisions, set up a regulatory sandbox, and make grants for upskilling Texans on AI. While one might think Republican Texas would be "anti-regulation", it has tried to enforce responsible AI practices – its Attorney General recently settled with a genAI health care company that purportedly violated consumer laws by making misleading claims about its product (which summarised patient information and drafted medical notes). California may see a rerun of SB 1047, and in addition there is action by the California Privacy Protection Agency to add an article on automated decision making to existing privacy regulations, which would require disclosure and opt-out possibilities for consumers. In New York, there is potential legislation based on product liability that would hold AI firms liable if their systems "go awry".

Possible policy directions under Trump II

Intense speculation started almost the day after the election, as to what President-elect Trump may have in mind for AI policy. Not much was said on the campaign trail by either side, though I had heard Trump mention the massive energy requirements for AI, and the Republican manifesto proposed repealing the Biden EO. There is a growing wave of opinion pieces and webinars from experts with their take on the future, and most people I spoke to offered their views.

What I take away is that in many areas there is a commonality of purpose, notably national competitiveness and building capability in the public sector and research community. The

transition team is being advised on technology policy by Trump's former Chief Technology Officer, Michael Kratsios, which lends itself to continuity on those issues. It may also be a force in favour of ongoing international engagement given Kratsios' support for the OECD AI Principles. Several firms commented that Trump had good AI advisors in his last administration and on that basis, they were relatively optimistic about policy under Trump II. Shortly before my departure, Trump announced that tech investor David Sacks would be "White House AI and Crypto Czar" and would focus on making America the clear global leader in both areas.

Trump's focus on competing with China and strengthening industry in the U.S. is consistent with building capacity and is likely to see more effort placed on AI infrastructure, including energy. This would please AI firms (including OpenAI, which has already made public its recommendations in a 'Blueprint for AI Infrastructure') and investors, some of whom are close to the Trump team. New spending will be constrained though, as several trillion dollars are needed to renew the shortly expiring Trump-era tax cuts.

Stripping out "woke" policy (that which is viewed as too progressive or left leaning) will be where change is most likely. I was told that agency policies developed and finalised under the Biden EO (e.g. on impact assessments) would be difficult to roll back. However, I suspect with the appointment of Russell Vought (involved in conservative Project 2025) as OMB head, there will be a strong push to roll back regulation and effectuate Trump's priorities.

Reporting requirements on firms, as required under the EO, are likely to be cut in support of business innovation, particularly if tech voices remain close to the administration and if Elon Musk's Department of Government Efficiency is stood up. Republicans have a slightly love-hate relationship with big tech, reflected in Trump's announcement of David Sacks, saying "He will safeguard Free Speech online, and steer us away from Big Tech bias and censorship". But the common threads seem to remain a pro-innovation and pro-business stance. The replacement of FTC head Lina Khan (who has been active against big tech) likely further supports this.

I would expect there to be pressure on the U.S. AISI to focus more on supporting technical standards and issues of competitiveness and national security, less on issues of bias, transparency and content oversight, and less on evaluations of firms' models unless this is security oriented. The AISI may also come under pressure to reduce ties with counterparts abroad, if the concerns of Texan Republican Senator Ted Cruz gain momentum. He has written to the U.S. Attorney General alleging that a centre aligned with the UK AISI is acting as a foreign agent and influencing U.S. policy. He notes Congress has given no direction to engage with the EU (another AISI network member) on AI initiatives. He views the EU/UK approach as a threat to U.S. competition with China and as an effort to enable government censorship of free speech.

I also think that international initiatives that might look "aid-oriented" (such as the Global Partnership for Inclusivity on AI) will be deprioritised in favour of relationships with more immediate win-win outcomes for U.S. AI.

Stakeholder perspectives on AI regulation

Another key part of my research was to gather up views on current U.S. AI policies and where they should head. I found divergent opinions on the Biden EO and what (and how) – if anything – should

be brought under stricter government oversight. Overall, though, the general sentiment seemed to signal a relatively pragmatic and innovation-friendly ask of policy makers. This is not just a business view (I highlight their perspectives further below) but one from thinktanks, academics and experts across the board.

General views

People with a human rights lens were the most appreciative of the EO, noting its actions to strengthen agencies' AI practices in areas with civil rights implications. Many people seemed ambivalent – "it's ok" – while several criticised it for overreaching and diverting agencies from other important work. Several highlighted the heavy burden on NIST and the lack of commensurate resourcing. Those with experience in government were concerned that some new governance structures currently added more compliance cost than value, especially the new Chief AI Officer role, which overlaps with existing roles (such as information or digital officers), has unclear mandates, and is very early in building a collective culture. Some were disappointed that the EO did not provide help for firms grappling with the emerging state-level patchwork of AI regulation and other policies, a task that the underused National AI Initiative Office could perform.

Very few people called for a federal AI regulator and AI Act. Those that did suggested that there are regulators for finance, transport etc, so why not AI? One thought the regulator's goal could be to seek transparency and accountability and set up systems to achieve this, rather than starting from a "harms perspective". Some thought an AI Act could set a useful common baseline on, e.g. disclosure and transparency in a federal law. They noted the FTC is becoming more active in AI cases, and states are beginning to lean on consumer protection actions to pursue AI firms when things go wrong. More often though, people suggested movement on a federal privacy law would be more useful than an AI law, saying treatment of personal data is behind many AI issues.

In most cases, people advocated a cautious approach to formal AI regulation and governance. Those with a long memory in technology policy considered that AI is just further progress of the digital ecosystem, that the technology is neutral ("don't make AI the crime"), and that poorly designed AI regulation could risk rolling back the democratisation of digital technologies. Many pointed to the large body of law already in existence, noted there didn't seem to be any bespoke AI harms that need new laws (for instance, the Department of Justice considers it has the federal tools needed to tackle AI obscenity and child sexual abuse material), and suggested existing laws can be iterated as harms emerge or if enforcement needs to be adapted. One observation was that norms coalesce faster than laws and hasty regulation could prevent valuable learning from ongoing discussions between standards bodies and industry.

Caution does not mean zero regulation, however. Most people agreed that in high-risk areas, such as health and the military, it makes sense to have rules and laws. The idea of "precision regulation" was popular, with most saying this should be aimed at use cases not the technology per se, given that risks are typically shaped by context and occur at different points of the value chain. Detail is difficult though. People highlighted the need to clarify what are the red flags for risk (one option is the "rights impacting" approach in the EO), answer the fundamental question of "what is harm?" so that guardrails can be designed, and specify when intervention occurs – pre-deployment, post-deployment or throughout a system's lifecycle.

One lively debate is who takes responsibility for what risks. State-level bills have been distinguishing between developers and deployers, where the latter take responsibility for context-specific risks. However, some noted there is a third layer emerging, of entities making "wrappers" for AI systems and on-selling to deployers, which could upset carefully crafted legislative language. There was general agreement that post-market monitoring (perhaps learning from pharmaceutical post-market surveillance) plus adequate resources for enforcement are needed.

Given the emerging patchwork of state laws, people stressed that interoperability is critical. There was praise for the Colorado Act on this point, as it would grant full compliance to entities that have already adhered to relevant rules under other federal or state law, been approved by a federal agency, or achieved equivalence in other similar situations. This may become more important if the pace of "general AI legislation" at state level slows due to difficulty in crafting and getting agreement on a wide set of issues, which some thought was likely. A focus on specifics, e.g. AI in privacy, AI in health, etc, may be easier to craft, but it will add up to a busy regulatory landscape.

An issue that came up multiple times is the need to look more at governance of AI in the physical world, i.e. cases where AI (generative or otherwise) is effectuating action in the real world. Robotics is one way this can occur, with use cases in health and hospitality; another is AI systems linked to critical infrastructure. Agentic AI models, which are getting increasing coverage, also trigger this concern. This ties in with an interest (and unease) with systemic AI, where multiple AI systems are working at once and could collectively trigger negative outcomes.

Finally, on AI safety, many people agreed the term is too broad and that you need to pick a lens. There is also general agreement that safety is not only a technical issue, but one that is influenced by people, culture, infrastructure and other factors, which calls for a "sociotechnical" approach to policy. Some considered the U.S. was very exposed to the safety debate and work of international AISIs, as targeted firms are mostly U.S. based, and their AI development work could be curtailed. However, the interest of countries like Brazil, India, Japan, Korea and Singapore, may keep AI safety work from stifling innovation. It was observed that as countries are thinking more about AI, there is increasing reticence to establish AI governance regulations that infringe on competitiveness. This may also mean AI governance at the international level may favour measures that don't run counter to economic interests.

What do businesses think?

Coming from MBIE, I was particularly keen to draw out how businesses view the U.S. environment for AI innovation. Industry doesn't speak with one voice, and consumer-facing firms face different challenges to firms supplying AI to enterprises. People pointed out that there are also many internal tensions within firms, with techies pushing out new products while AI and privacy staff try to keep up with guardrails. However, some consistent messages came through.

They are actively addressing AI governance

While some observers are cynical about the motives, AI firms and large tech-driven firms are taking AI governance seriously and have incentives to do so. There is a mix of firm-level and collective action, as firms identify challenges with the technology, seek to build trust with stakeholders, and leverage AI governance tools for meeting wider risk management and legal oversight goals.

People from AWS, Google Cloud, IBM, and Workday spoke to me about the extensive AI governance processes and people they have in place. Firms outside the "big names" are also putting structures in place. A Future of Privacy Forum report found many firms are implementing AI impact assessments. It is still early days though and one challenge they identified for this group is getting sufficient / useful information from 3rd party model providers; another is predicting unintended use cases. Resources are also a challenge for compliance teams.

One interesting collective effort amongst business is the Frontier Model Forum (FMF), which brings together experts from member companies to share information and experiences about public safety and security of large general purpose AI models. They focus on chemical, biological, radiological and nuclear (CBRN) threats and advanced cyber risks, i.e. physical harm to people and property. The FMF was set up as a trade association with a narrow focus to allow technical experts from competing firms to come together. It has parallels to aviation, where engineers focus on technical issues to advance safety. The FMF is like asking "will the plane fall out of the sky?".

People suggested this kind of technical work, bitten off in small chunks, helps build foresight that could help avoid tech-driven failures in future. They agreed that there are heightened risks where AI capabilities control actions in the real world – "if there is a failure, it's not just a funny thing ChatGPT said, but rather actual physical harm".

I was interested to hear there is fatigue with the federal government's use of voluntary actions, which are ceasing to resonate and come with heavy reporting requirements. The commitments reflect what big firms are already doing (and has led to some competition between firms as to whose internal procedures get most airtime) and are likely to be difficult for smaller firms to meet. They are also seen as a step backwards to principles, when firms have already shifted to implementing processes such as AI boards and tracking.

They like some aspects of current governance approaches

Firms are very complimentary of NIST's RMF, which is influencing both private sector and government agency development and use of AI and is now accompanied by additional guidance on genAI. Some suggested that the RMF could form a "safe harbour" in legislation, meaning compliance with the RMF would be considered regulatory compliance overall. Apparently, a Connecticut Congressman has endorsed this idea, which may be influential given that state's legislation is picked as one to watch.

One firm commented that if every business adhered to the RMF, "AI governance would be in a different place". The RMF is considered high level enough to be applicable across industries and can be approached in a modular way – i.e. firms with more limited compliance resources can start with a few actions (notably governance aspects such as teams and norms) and build up over time. I was told Google has already audited itself against the NIST RMF and ISO 42001 (an international standard on AI management systems), anticipating future compliance requirements.

There was also praise for OMB's 2024 memo on *Advancing Governance Innovation and Risk Management for Agency Use of AI* (M-24-10), which sets up requirements for agencies including a Chief AI Officer, enterprise strategies, and impact assessment and testing of safety- or rights-impacting AI. Several firms suggested the approach of the memo could equally be applied to firms. At the same time, they commented the memo is hard for agencies to implement, as it implies

significant investment (especially in IT) with no resources behind it – a criticism that would probably also apply if the memo's actions were rolled out to smaller firms.

They don't expect fast federal action on AI regulation

Several firms pointed to the U.S. experience with privacy legislation and suggested it could take ten years to arrive at a federal approach to AI regulation. This view may also reflect awareness of the general bipartisan agreement not to stifle big U.S. firms, if that risks the U.S. falling behind China.

They observed a state-by-state approach is costly for smaller firms – while large firms can afford the compliance burden, start-ups cannot. On the wave of state AI bills, they suggested lawmakers face a tension between being practical and using existing legislation vs needing to brand themselves and being seen to act. They are watching key states, including Texas. They noted they appreciated state bills that had clear – and narrow – scoping on AI systems for consequential decisions.

They counsel against onerous regulation

Business stakeholders favour a pragmatic, cautious approach to AI regulation. This is mainly because the technology is still evolving quickly, and they are concerned not to choke the ecosystem. One firm commented that when you don't know the full capabilities of a technology, a regulator is "throwing darts in the dark", which is why the U.S. usually starts with a light-handed approach. Another remarked that it is impossible to align AI governance to unrealised risks – there are no metrics for firms to measure against. Debate over the vetoed Californian bill (SB 1047) apparently became "turbocharged" when venture capitalists (VCs) worried the precautionary approach in the AI bill could be rolled out more widely in the tech arena. I was told that while VCs are used to living with uncertainty framed around traditional liability regimes, in a precautionary world "there's no way a VC would write a check", thereby breaking the VC ecosystem for AI and beyond.

Like other stakeholders, firms were more favourable to precision regulation, particularly to maintain incentives for innovation that support a growing ecosystem. One called for "surgical interventions" focused on specific concerns or market failures. The same firm underlined that if formal guidance or rules are proposed, these must be crystal clear, since legal teams must interpret them as broadly as possible for fear of litigation. Firms also noted that their customers are often already operating in highly regulated sectors, such as banking, and are highly attuned to risk management and maintaining customer trust. In this context, several firms pointed to Trump I-era OMB guidance to agencies on how to regulate AI and suggested it be revived.

Firms advocate strongly for aligning responsibilities to spheres of control. Their main worry is not putting liability on developers for use cases that they cannot predict and have no control over – "you have to let models run in the wild to see how they function". California's SB 1047 was considered to have gone too far in that direction. Joint and several liability, where any defendant is fully liable for the plaintiff's damages even if others share blame, was also considered an "innovation killer". They recognise risks arise throughout the value chain and suggest AI model testing needs to take place at multiple stages in the AI lifecycle (i.e. in the lab, pre-deployment, post-deployment).

Nuro.ai – a case for confidence

A highlight of my engagement with business was a visit to the Nuro.ai offices in Mountain View, California. Nuro is a robotics company focused on autonomous vehicles (AVs). It has developed AI-enabled "Nuro Driver" technology (software and hardware, e.g. sensors) for delivery and passenger vehicles. The day I visited they had just announced an expanded AV deployment in Palo Alto/Mountain View, California and Houston, Texas.

The visit began with a ride-along in an AV (a kitted-out Prius), which was undergoing training and testing on the road. I also saw several generations of their delivery robot up close.

Visit to Nuro, Mountain View California, 18 November

I then had the chance to speak with Nuro's legal and policy team. Nuro uses genAI in the design of its AV software; the product itself also integrates AI (image recognition, machine learning and sophisticated analysis of sensor data). They develop their own AI models (which are not large enough to be captured by current "frontier model" regulatory efforts) and finetune models from suppliers. One of Nuro's challenges is retaining appropriate control over their data and intellectual property, which requires careful negotiation with AI model providers. They commented that smaller firms would struggle to deal with the legal complexities.

It was valuable to hear the views of people working on practical AI applications. In furtherance of their stated goal to make roads safer for all, they focus on a safe systems approach. They are very aware that a single incident can impact the public's trust in the technology.

Firms were relatively harsh about the EU AI Act – I would summarise their views as a "fail" mark. They were not arguing that you can't have both responsible AI and innovation, but rather that the EU approach is simply bad for innovation. Several noted the Act is ambiguous and a case of poor drafting. They don't expect a "Brussels effect", as they say the Act neither promotes homegrown AI firms nor provides an attractive environment for foreign direct investment. However, by necessity, they are carefully watching the Act's implementation. Big players are currently involved in EU

working groups on the code of practice that will detail rules for providers of general-purpose AI models. (In fact, many other stakeholders were also dubious about the EU Act.)

Some firms observed that Singapore is leading a "third way" on AI regulation, which is trickling into ASEAN guidelines, Indonesian regulatory circulars, and work in the Philippines (though not Australia, which is charting a separate course).

Potential considerations for New Zealand

To state the obvious, the U.S. is at an entirely different scale to NZ, with budget figures in billions and trillions. There is a deep and longstanding intertwining of technology and defence policy that contributes to AI innovation. The tech sector is vast and diverse and benefits from the clustering of talent in well-established hubs.

NZ – and many other countries – cannot replicate this. Nevertheless, the U.S. is facing similar AI policy questions, and discussions are not that different to what we have in NZ (particularly when it comes to government use of AI). Nothing I heard suggests NZ needs to U-turn on any nascent AI policy directions, though there may be some challenges to come. Below is my initial take on four related issues for NZ.

AI regulation

It might be tempting to view the volume of state- and federal-level AI legislation as an indicator that NZ is lagging far behind and that "everyone else is passing comprehensive laws on AI regulation". I don't subscribe to this. My view is that much federal-level legislation is building AI capabilities, a large share of state-level legislation will go nowhere, and the substantive AI legislation that is appearing is either aimed at putting some safeguards in government decision-making processes or tweaking existing laws to clarify AI is covered (e.g. consumer law, some crime laws). The Californian legislation was an outlier in focusing on regulating frontier AI models, and it focused on firms and AI activity that NZ does not have and is unlikely to have.

NZ has some guidelines in place for agency use of AI, including the Algorithm Charter and a recent Algorithm Impact Assessment Toolkit. I think the focus on this work should continue, including sharing information across agencies on good practices for AI in decision-making processes. The recent OMB memo has a definition of AI and of safety-impacting AI that I found helpfully targeted and useful. Despite some negative views in the U.S. about the CAIO role, it does provide a focal point, and this would likely be helpful in NZ as well, though needs to be budgeted for and clearly scoped around internal agency AI use.

There may be areas of existing law in NZ where we need to clarify AI is covered and where we might want to reflect on appropriate enforcement and penalties. In the U.S., this work seems to be driven by high profile issues, such as deepfakes in election contexts and non-consensual intimate imagery. Unfortunately, there are examples of laws and amendments being drafted in reaction to issues, but with less or no thought given to enforcement (e.g. a potential AI fraud bill in California).

NZ has limited bandwidth to embark on legislative drafting and limited resources for regulatory enforcement. I would urge that we focus on issues where there is evidence of a problem or significant ambiguity (related to both harms and beneficial uses of AI), and where there is an intent

and an ability to follow through to implement any new laws or regulations. In the U.S., OMB's 2021 memo on how agencies should regulate AI remains a useful resource. NZ has guidelines on good regulatory practice that should get us to a similar endpoint.

More generally, I consider NZ is on the right track with an agile approach to regulation, starting with voluntary guidelines. The comment that "norms coalesce faster than laws" and that you gain significantly from discussions among stakeholders and standards bodies, is one that I found appealing. There is broad acceptance of the NIST RMF in the U.S. and growing momentum for this to be a safe harbour, showing its value. To complement this, we could consider how to be more systematic about surveillance of AI outcomes in the marketplace. Internationally there is interest in incident monitors, which can help inform potential future regulatory steps. Accompanying this with purposeful and consistent engagement with AI stakeholders would be ideal, though acknowledging that competing pressures on officials can cause engagement to ebb and flow. Letting industry experts know it is ok to share experiences (as in the FMF) might also be useful.

The idea of precision regulation around use cases seems sensible and complements an agile approach. Health was mentioned as an example in the U.S., though I did not come across any specific regulations/legislation (and the U.S. health system is very different to NZ). The overwhelming reaction to comprehensive legislation aka the EU AI Act was negative, and I think the right approach is to watch how the EU implements its regulations and what impacts are emerging. The observation that the AI ecosystem is quickly evolving to include new groups of players – e.g. middlemen in addition to developers and deployers – to me underlines the challenges of comprehensive hard law. Bills drafted in the U.S. that explicitly focus on the categories of developers and deployers could already be out-of-date, and this is just one example.

One area to watch is where AI interacts with the physical world – an issue raised by academics, thinktanks and industry voices. It's not clear where this will go, but people were concerned about the risks in new technologies such as agentic AI, and in areas of critical infrastructure, "smart cities", health and hospitality, among others. Reinforcing information sharing across agencies about what we learn and hear from partners overseas would be helpful, as well as what we learn from NZ stakeholder engagement.

Finally, as suggested by people I spoke with, we should keep a watching brief on legislation coming out of California, Connecticut and Texas as a signal of where the U.S. is heading on AI regulation.

AI safety

Strongly related to AI regulation but worth a discussion in its own right, is the issue of AI safety. "AI safety" has become a buzz-phrase – ill-defined, as many people noted – that is dominating international AI discussions and colouring domestic policy debates. I think this issue is challenging for international diplomacy. Even if the new Trump administration deprioritises the U.S. AISI, the issue has sufficient traction internationally to remain on our radar.

My starting view was that an overfocus on "safety" could risk derailing efforts to boost basic adoption of AI technologies in countries less advanced than the U.S. I was keen to hear and learn more, and I am grateful to the State Department for facilitating my participation in the International AISI convening in San Francisco in November.

I heard from several interlocutors that AISI network countries such as Japan, Korea and Singapore are supportive of innovation and would not wish the international AI safety debate to chill AI development and use. Parts of the U.S. government are also keen to promote the benefits of AI for economic development and the SDGs.

It is early days and there is more to do to establish a solid research agenda and stake out the value-add niche of the network. For instance, model evaluation – a key safety topic discussed at the AISI meeting – is still nascent, focused on technical capabilities and not well informed by economic and social research on the effects of AI. As a policy maker, I find it is not immediately helpful to know that, for instance, a model can solve 36% of tasks at a "cybersecurity apprentice" level of competency (per the joint AISI test of Claude Sonnet 3.5). (Though, it is interesting to know that a model's built-in safeguards can be routinely circumvented, meaning a model provides answers that should be prevented – another test finding). Also, test results do not guarantee the same behaviour in real world conditions and what is "safe" in one context may not be in another.

NZ may be asked whether it will establish an AISI and join the network. I think we would need a much sharper picture of the main motivation of the network (i.e. the problem it seeks to solve), what we could add, and what we might expect.

If the AISI network keeps a focus on evaluation, it presumably would not be testing any NZ models, as these would surely not meet the size threshold[2]. Participation could help us follow the evaluation debate, but from a government perspective, it is not clear we would – at least in the foreseeable future – seek or implement model evaluations within the NZ market.

Building an ecosystem

The U.S. has a long history of investment in technology R&D, including AI. The last two administrations have both been interested in AI capability, including in the public sector, and AI infrastructure – the latter of which I expect to get more attention. These themes are bipartisan and important. They speak to building an ecosystem for AI, which supports U.S. national competitiveness and national security goals. They were a focus before discussion of AI regulation – perhaps reflecting the sentiment of venture capitalist Marc Andreessen who said in an interview with the German Marshall Fund, "you can't regulate if you don't have an industry".

NZ does not have the resources of the U.S., but our current work on an AI strategy I think is important for identifying the capabilities we do have and attempting to remove any blockages. Some remarks from my discussions and event participation stood out in this context:

- "The future is not in LLMs." Several people considered LLMs are too expensive in data, compute and energy, and that such intense training is not needed for discrete tasks in most use cases. The next step of the robot revolution needs LLMs with lower compute. And there is increased interest in "low resource" models (where there may not be large amounts of data, such as for lesser spoken languages). Such innovation seems more do-able for NZ.

[22] Jack Clark of Anthropic calculated that 10^{26} FLOPS implies a training cost (using Nvidia chips) of USD 104 million. The UK's domestic threshold of 10^{25} FLOPS would cost USD 10.4 million, which more firms might reach. (What does 10^25 versus 10^26 mean? | Import AI)

- “Simulations are powerful.” Both at a fascinating robotics conference, and at Nuro, I heard people speak about the value of AI operating in the real world, and how simulations allow testing before real world release. Simulation can allow training on unusual events that happen infrequently, and you can generate training data from simulations. I was heartened when people said, “video games have very relevant technology and skills for this”. NZ’s current investment in the game development sector takes on another dimension in this context, as does the potential of our creative sector more broadly.
- “Could y’all do something like the PIFs or USDC?” The Presidential Innovation Fellows and U.S. Digital Corps aim to bring technical talent into government, for short-term or longer-term stints. They can raise the level of general technical understanding and support policymaking, as well as boost innovation in the public sector. In the current environment, this seems difficult for NZ, but potentially worth exploring what the demand might be for different types of AI talent.
- “Scientists can’t be blamed for not including ethics in tech design.” For people involved in science diplomacy, AI governance would look different if policy makers better understood the technical landscape and technologists better understood the sociotechnical and sociopolitical system. They say part of the issue is that scientific training is most often focused on STEM and the social sciences are frequently underfunded.
- “Don’t get sidetracked into data centres.” While AI firms are distributing model training in search of cheaper options (and countries like Sweden are leaning on their clean energy credentials to supply them), there was some scepticism that NZ could sell itself as a destination for training AI models or housing data. The suggestion was to focus on AI adoption / use and getting the benefits of the technology. Data centres will be needed as the economy further digitalises, but this is broader than AI.

International engagement

AI safety is one area where NZ will likely continue its dialogue with international partners, but there are other topics and some broader considerations that may arise for us.

Taking the broad first, there is a question how strongly the new administration focuses on competition with China and national security issues as a lens for policy in general. AI is intertwined with national security, military operations, Indo-Pacific relations, supply chain resilience, and many other issues that span NZ agencies. For obvious reasons, I had limited visibility of defence-oriented activities, but for the U.S., it can be hard to draw a bright line between “national security” AI and “everyday” AI policy issues.

I think this underlines the need for good information exchange, to the extent possible, in NZ between officials working on “economic AI” and “defence AI”. There will be things that cannot be shared, but staying coordinated is vital. The good relationships built up already, including between MBIE and MFAT, are a great start, particularly if questions related to China arise.

Given the debate over AI safety and international governance of AI more generally, I would advocate staying connected with small countries who are keen to ensure their economies benefit from AI and who put weight on vibrant innovation ecosystems. I was told NZ’s voice matters when part of a broader group.

I think NZ needs to engage more with international standards work. There is momentum towards ISO standards, as a well-regarded existing fora. Just a few days before my departure, I saw Callaghan Innovation had developed ISO fact sheets to help firms conform to their AI standards and has also formed a committee to represent NZ's interests internationally in this area. Having a shared government view on this will be useful.

Finally, in my view, to stay meaningfully connected to international discussions, we must be physically present. Resources will be a problem, but we should try to identify (and budget for) a handful of key opportunities to send government policy experts abroad for meetings. The payoff flows in all directions – NZ, agencies, the individual and, in some cases, international policy outcomes.

Appendix 1: List of meetings and events

Washington D.C.

Thinktanks and civil society:

- Center for AI and Digital Policy
- Information Technology and Innovation Foundation
- Center for Data Innovation
- IBM Center for the Business of Government
- Wadhwani AI Center, Center for Strategic and International Studies (CSIS)
- Stimson Center
- Future of Privacy Forum
- Luminos Law
- RAND
- Human Rights First
- Special Competitiveness Studies Project

Industry and professional associations:

- Frontier Model Forum
- International Association of Privacy Professionals
- AWS
- IBM
- Workday
- Google Cloud
- Meta
- Institute of Electrical and Electronics Engineers (IEEE)

Government:

- State Department, Office of the Special Envoy for Critical and Emerging Technologies
- U.S. House Committee on Ways and Means
- National Institute of Standards and Technology (NIST)
- Government digital experts
    - State Department
    - Health and Human Services

- Technology Transformation Services
- General Services Administration, Cloud.gov

Academic:

- Georgetown University McDonough School of Business

Events (in-person, with role):

- Opening remarks at ORF America "Chips for Breakfast: Advanced Compute and AI" roundtable (16 October)
- Presentation on NZ AI policy to MBA students at Georgetown University McDonough School of Business "MBA Lunch and Learn" series (30 October)

Events (in-person):

- Special Competitive Studies Project (SCSP) "AI + Robotics" summit (23 October)
- Georgetown Center for Digital Ethics seminar "Why should we care about understanding AI" (30 October)
- Johns Hopkins University and German Center for Research and Innovation New York "Future Forum: Science Diplomacy in an Era of Technological Disruption" (31 October)
- CSIS "Ensuring U.S. Leadership in AI" (13 November)

Events (virtual):

- CSIS Symposium: AI in the Department of Justice (2 October)
- Center for Security and Emerging Technology (CSET) webinar "A year-end review: Recapping global AI governance efforts" (14 November)
- Center for Data Innovation webinar "Sociotechnical approaches to evaluating generative AI" with speaker from Google DeepMind (19 November)

Atlanta

Event (in person, with role):

- Panel member in session on 'Comparative Analysis of Governance Efforts' at Georgia Tech's Internet Governance Project annual workshop "Does AI Need Governance?" (17-18 October)
    -

San Francisco / Bay Area

Industry:

- Nuro.ai site visit

Academic:

- Stanford Institute for Human Centered AI (HAI)

Events (in person):

- International AI Safety Institute (AISI) convening (20 November)
- ORF America "Governance Talks" event (10 December) with panel discussion on 'A Year of Elections and Geopolitical Turmoil'.

Events (virtual):

- Stanford HAI webinar "Will copyright derail generative AI technologies?" with speaker from Berkeley Law (13 November)

Los Angeles (virtual)

Academic:

- University of California, Los Angeles (UCLA) IT Services