US PACIFIC ECONOMIC COOPERATION COUNCIL
US PECC

pecc Pacific Economic Cooperation Council
French Pacific Territories Committee

Responsible Adoption of **General-Purpose AI**

Redwood City, 18-19 September 2024

SOMMET POUR L'ACTION SUR L'IA | AI ACTION SUMMIT

An official partner event of the Paris AI Action Summit 2025

The Asia-Pacific region hosts the most critical actors in AI research, chip manufacturing, development, and deployment at scale. As an inter-governmental forum that promotes free trade throughout the Asia-Pacific region, Asia-Pacific Economic Cooperation (APEC) is well positioned to address and discuss the cross-cutting impacts of responsible AI deployment and has been invited to attend the AI Action Summit in Paris in February 2025.

In accordance with the mission of Pacific Economic Cooperation Council (PECC) to suggest policy priorities to APEC officials, the US and French PECC committees convened this 2-day seminar under the Chatham House Rule to discuss AI regulation. The gathering united AI experts, scholars, industry adopters, and policymakers to assess AI's current state and dynamics, impact, and potential future development. During 3.5 hours of roundtable discussions, participants discussed a number of policy ideas, along with strategies for international governance and responsible innovation.

Table of contents

The seminar convenors developed a number of ideas for potential policy statements ahead of the seminar, intended for C-level executives and government leaders in the Asia-Pacific region, multilateral institutions, and the public at large. The ideas have been updated to reflect the insights gathered during roundtable discussions that were hosted each afternoon at the seminar. While participants shared their insights, it should not be assumed they support any specific policy idea herein.

Executive Summary

Responsible Adoption of General-Purpose AI

Pacific Economic Cooperation Council (PECC) committees from the United States and the France Pacific Territories convened this seminar, a partner event of the Paris AI Action Summit 2025, to bring together experts from AI labs, regulatory bodies, civil society, and academia in the Asia-Pacific region to discuss the responsible development and governance of general-purpose AI systems (GPAIS).

The governance of GPAIS requires maintaining forward momentum while carefully navigating direction. Our discussions at the seminar suggest the following considerations for C-level executives and government leaders in the Asia-Pacific region, multilateral institutions, and the public at large:

1. Competitive dynamics among AI companies and among sovereign nations should be channeled to secure a thriving multipolar global economy. Racing to the frontier of AI to secure hegemony looks like a risky zero-sum tactic. Instead, decision-makers can choose to pursue an AI-powered future of economic "co-opetition" and global abundance, as AI is estimated to potentially deliver USD 4.4 trillion* annual economic impact. Where collaboration is not possible, this positive vision necessitates a commitment to the principle of coordination and open dialogue in the interest of peaceful co-existence.

2. Governments in advanced economies and tech business leaders urgently need to come together to launch AI capacity building efforts at an unprecedented scale to ensure that the benefits of AI are distributed equitably, across societies and around the world, while upholding adequate safety and security. Areas of focus should include talent development; joint research & development; safe & responsible hardware/compute distribution; digital/AI literacy (for policymakers, economic actors, and the public at large); and workforce adaptation mechanisms.

3. Investing in actionable R&D for technical AI risk management stands out as an urgent priority. Most existing or proposed AI policies put forward transparency and accountability frameworks, seeking to institute standardized independent evaluations to assess AI systems' trustworthiness across development/deployment contexts, life cycles, and value chains. Yet, technical AI risk assessment solutions are still vastly underdeveloped and **insufficiently robust to industrialize GPAIS,** especially when accounting for the pace of capability development. This paradox translates into a de facto risk of responsible innovation "washing". Academia and the emerging "AI Assurance Technology" sector (estimated to USD 276 billion by 2030**) offer promising solutions, such as AI oversight and auditing tools, advanced computing governance mechanisms, bias mitigation techniques, holistic explainability and interpretability, and more. Maturing and scaling these solutions effectively towards genuine responsible innovation requires market-shaping policies, regulatory clarity, and cross-border harmonization.

4. Equipping a globally distributed AI ecosystem to handle these extraordinary capacity building, coordination, and governance challenges requires **empowered, well-resourced and trusted guardians of the public interest**. Emerging institutional innovations like a network of National AI Safety Institutes (AISIs) will be crucial to promoting responsible development and deployment; as will be the modernization and optimization of international organizations and multi-stakeholder platforms like APEC and PECC to ensure a balance of perspectives are brought to deliberations on AI governance.

As we now proceed towards the Paris AI Action Summit, we hope for the opportunity to institutionalize a Asia-Pacific platform to advance and structure AI governance in the region via PECC and APEC.

Seminar Convenors

Alex Parle Executive Vice President, North American Centre on APEC

Nicolas Miailhe, Board Member, France Pacific Territories National Committee for Pacific Economic Cooperation

** McKinsey Digital: The Economic Potential of Generative AI - The Next Productivity Frontier*
*** Various authors: The case for investing at the vanguard of AI risk mitigation technologies*

Participant list

Adam Billen, Director of Policy
Encode Justice

Akiko Murakami, Executive Director
Japan AI Safety Institute

Adrien Abecassis, Executive Director for Policy
Paris Peace Forum

Alex Parle, Executive Vice President
NCAPEC (Seminar Convenor)

Anthony Aguirre, Executive Director
Future of Life Institute

Atoosa Kasirzadeh, Assistant Professor, **Carnegie Mellon University**;
Visiting Research Scientist, **Google**

Ashley Zlatinov, Public Policy Product Lead
Anthropic

Bogdana Rakova, Senior Data Scientist, Responsible AI
DLA Piper

Bruno Liebhaberg, Executive Chairman
Centre on Regulation in Europe (CERRE)

Brian Tse, Founder and CEO
Concordia AI

Celine Malvoisin, Chief Learning Officer
everyone.AI

Charbel-Raphaël Segerie, Executive Director
CeSIA

Chris Painter, Policy Director
METR

Claudia May Del Pozo, Founder & Director
Eon Institute

Corynne McSherry, Legal Director
Electronic Frontier Foundation

Cyrus Hodes, General Partner
1infinity Ventures

David Evan Harris, Chancellor's Public Scholar
University of California, Berkeley

David Langer, Managing Partner
Lionheart Ventures

Dawn Song, Professor, Department of Electrical Engineering and Computer Science
University of California, Berkeley

De Kai, Professor, Computer Science and Engineering, **HKUST**;
Distinguished Research Scholar, **Berkeley International Computer Science Institute**

(Virtual address) Dragos Tudorache, Member and Lead Co-negotiatior of the EU AI Act
The European Parliament

Emmanuelle Pauliac-Vaujour, Attachée for Science and Technology
Consulate General of France in San Francisco

Ethan Michaud, Managing Director of Tech Strategy & Innovation
Moody's

Fynn Heide, Executive Director
Safe AI Forum

Frederic Werner, Head of Strategic Engagement
International Telecommunication Union (ITU)

Gregory Renard, Head of Applied Machine Learning
Docugami

Ian Eisenberg, Head of AI Governance Research
Credo AI

Imane Bello (Ima), AI Action Summit Lead
Future of Life Institute

Jan De Silva, Co-Chair
Canada-ASEAN Business Council

Jayne Stancavage, Vice President, Policy and Regulatory Affairs
Intel Corporation

John Havens, Founding Executive Director
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

John-Clark Levin, Research Lead
Kurzweil Technologies

Joo Hyung Park, Senior Manager
The Federation of Korean Industries

Juan Miguel Miranda, Consul General
Ministry of Foreign Affairs of Peru

Juraj Corba, Chair
OECD Working Party on Artificial Intelligence Governance (AIGO)

Kar Yan Tam, Chairman **HKCPEC**;
Dean, School of Business and Management; Chair Professor, Department of Information Systems, Business Statistics and Operations Management, **Hong Kong University of Science & Technology**

Kenji Hiramoto, Secretary General
Japan AI Safety Institute

Lee Wan Sie, Director for Data-Driven Tech
Singapore Infocomm Media Development Authority

Lucas Hansen, Co-Founder
CivAI

Marius Hobbhahn, CEO
Apollo Research

Mark Manantan, Director of Cybersecurity and Critical Technologies
Pacific Forum

MaryJo Fitzgerald, Chief of Staff + Partner
Global Gateway Advisors

Michael Chen, Member of Policy Staff
METR

Michael Nunes, Vice President, Payments Policy
Visa

Michael Zanette, Senior Analyst
Global Affairs Canada

Molly Welch, Investor
Radical Ventures

Nico Miailhe, Co-Founder & CEO, **PRISM Eval**;
Founder & Chairman of the Board at **The Future Society (TFS)**
Board Member at **FPTPEC (Seminar Convenor)**

Nicolas Moës, Executive Director
The Future Society

Nick Fitz, Founder & GP
Juniper Ventures

Pascal Lamy, Vice-President, **Paris Peace Forum**;
Coordinator, **Jacques Delors Think Tanks**

(Virtual address) Philippe Huberdeau, Secretary General
Paris AI Action Summit

Rebecca Finlay, CEO
Partnership on AI

Rebecca Weiss, Executive Director
MLCommons

Ricardo Baeza-Yates, Director of Research
Institute for Experiential AI, Northeastern University

Richard Cantor, Vice Chair, Moody's Ratings; International Co-Chair
PECC

Robert Reich, Senior Advisor, **U.S. Artificial Intelligence Safety Institute**,
NIST, U.S. Department of Commerce

Roman V. Yampolskiy, Associate Professor - Speed School of Engineering
Director - Cybersecurity Laboratory
University of Louisville

Rumman Chowdhury, CEO
Humane Intelligence

Sacha Alanoca, AI Policy PhD
Stanford University

Sarah Cogan, Software Engineer, Frontier Safety
Google DeepMind

Sarah K. Luger, PhD, Co-Chair, Data Sets Working Group
MLCommons

Sarah Pinto, Partner
Emerson Collective

Sella Nevo, Director of the Meselson Center
RAND Corporation

Sonia Pereira, Consul General
Consulate General of Colombia in San Francisco

Sunny Gandhi, Vice President of Political Affairs
Encode Justice

Vanessa Bonnet, Head, Economic Department for Western USA
Department of French Treasury

Wael William Diab, Chairman
ISO/IEC JTC 1/SC 42 Artificial Intelligence, ISO

William Bartholomew, Director of Public Policy, Office of Responsible AI
Microsoft

Yi Zeng, Professor of AI, **Chinese Academy of Sciences**;
Director, **Beijing Institute of AI Safety and Governance**

Yongxin Zhan, Chair **CNCPEC**; International Co-Chair, **PECC**

(Virtual address) Yoshua Bengio, Professor at the Department of Computer Science and Operations Research at the Université de Montréal; Scientific Director at **Montreal Institute for Learning Algorithms (MILA)**

Yura Hwang, Partnerships APAC Lead
Perplexity

Summary of speaker insights

Investing in Responsible Adoption of General-Purpose AI

Generative AI in Business

It has been estimated by PwC that AI's contribution to the global economy could reach the scale of Europe's current total GDP. An important driver of this is that up to 60% of tasks can now be automated with AI*, representing a massive potential upside. This will empower individual customers with higher expectations, forcing companies to license AI services to meet new standards of engagement. The automation of science and engineering processes represents a high-impact application of GPAIS moving in a more specialist direction, but is less directly visible to to customers and users.

In any case, the opacity of AI decision-making necessitates human oversight - especially for high-stakes decisions like credit ratings and scoring, which can have cascading economic effects.

Key bottlenecks
Talent to create advanced algorithms and high-quality data sets are key input factors for AI development. The third is access to advanced hardware to supply computational power, and this has become the key bottleneck to innovation. Going forward, access to energy is emerging as a critical factor in AI sector development, as shown by data centers already consuming massive amounts of electricity - the 13 largest use twice France's total consumption. Countries will need to significantly expand their energy infrastructure to support AI growth, while carefully balancing this expansion with environmental sustainability goals.

Learning from the past
AI is likely to follow the historical pattern of digital technology: despite lowering transaction costs, it may lead to market concentration among top players. Early adopters with investment capacity gain persistent advantages through accumulated expertise and data. The timing gap between early and late adoption may create lasting market positions that survive even after AI becomes commoditized, potentially leaving late adopters permanently behind. The AI revolution's success hinges on also avoiding the missteps of the big data era, where Western companies' perceived data exploitation created lasting distrust in the Global South.

This historical context necessitates a fundamentally different approach: direct collaboration with local governments must replace corporate-first strategies, actively demonstrating that AI deployment will create shared value rather than repeat exploitative patterns.

Setting realistic expectations
We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run. This principle, dubbed "Amara's law", is particularly relevant to GenAI, where immediate hype cycles may create unrealistic expectations of instant transformation. At the same time, the long-term impacts are likely to be far more profound than we can currently imagine.

** IMF: AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity.*

The AI Assurance Technology Market

Achieving a world embedded with AI at-scale, will require sophisticated risk management techniques that currently have not been scaled. In the meantime, lack of clarity (e.g. on control of data, liability, etc) is making enterprise buyers cautious about adopting AI technologies.

To bridge the gap, a new industry is emerging: AI Assurance Technology (AIAT). AIAT is the software, hardware, and services that enable organzsations to more effectively, more efficiently, or more precisely mitigate the risks of AI. As instrumental enablers of contemporary risk management techniques, AIAT companies will see an exponential increase in growth over the next decade.

Key sectors

AI-Resilient IT Security
- Hardware security,
- Data privacy and cybersecurity

AI Trustworthiness
- Data/model/system evaluations
- External compliance audits

AI-Centric Risk Management
- Quality, conformity, and audit management
- Observability, monitoring, and incident response

AI-Aware Digital Authenticity
- Identity and content authentication

108%

AI assurance technology market
year-on year growth until 2030

A USD $276 billion market

A group of investors and civil society actors recently came together to size the AIAT market. By the year 2030, they estimate that the global AIAT market could reach approximately USD $276 billion.* This is based on starting estimates commensurate with comparable market research reports.

Three different modelling methods were developed, and resulted in an average 108% year-over-year compound annual growth rate (CAGR). That is to say, they expect the AIAT market to double each year, through the remainder of this decade.

The above figure represents an AIAT market equating to nearly 15% of the global AI market in 2030. This percentage aligns with expectations, given forthcoming regulatory requirements are anticipated to establish a comparable market floor for compliance-related AIAT solutions. Moreover, these market proportions are comparable to those of cybersecurity, safety testing, and risk management in mission-critical sectors, such as aviation or banking.

Key bottleneck

To capitalize on this burgeoning new market, regulatory clarity is needed to enable investors and founders to move forward with confidence. Clear guidelines would allow companies to scale effectively and make informed decisions, reducing the risk that their efforts will be undermined if regulations and standards are later introduced or changed.

There is great potential in market-shaping policies such as accelerators, eased administrative requirements, and more, to attract new AIAT companies to fill a risk management and compliance gap (i.e. what AI companies won't handle in-house).

** Various authours: The case for investing at the vanguard of AI risk mitigation technologies*

Summary of speaker insights

Understanding and Addressing AI's Socio-Economic Impacts

The pace of AI development is outstripping our governance mechanisms. While current frameworks in the policy pipeline address direct misuse and accidents, they largely miss system-level effects on human psychology, society, and critical infrastructure.

Workforce adaptation
Rather than defaulting to solutions like Universal Basic Income, speakers emphasized the need to proactively ensure workers' voices are central to deployment decisions, develop human-AI collaboration frameworks, and create economic tools to understand ethical deployment. The goal isn't to retroactively adapt to predetermined technological progress, but to shape the society we want to create through thoughtful governance and distribution of benefits.

Copyright
Fair use laws in the US currently protect most AI labs when training on copyrighted works, but prohibit regenerating portions of copyrighted material. Only a few companies can afford to pay for training on copyrighted or high-quality material at scale. Others will build on their models, and so any bias going into the development of the biggest models is likely to fan out and cause opaque systemic effects.

In French, "donnée" means "gift" - a fitting connotation that highlights that someone has created the data that trains AI systems. In order to foster a competitive environment and reliable datasets, we need to invest in open data initiatives and establish norms for responsible data sourcing.

Immediate social effects
AI systems are already reshaping human psychology and social dynamics at scale through content recommendation and interaction systems. Additionally, AI systems have already been widely used to reach political goals by manipulating people's biases through chatbots and feeds. The amplification of fear, polarization, and hate is the true immediate risk of AI.

While much attention focuses on misinformation, this misses a more subtle threat: "neginformation" - factually accurate information presented without crucial context, pushing towards systematically biased worldviews. AI systems decide what to show and not show to users, which can contribute to neginformation.

Emerging security challenges
New models can accelerate drug discovery and medical research, but also be misused to design harmful substances or pathogens. Even the legitimate application of AI in biological research itself poses significant risks – and AI is expected to amplify these risks potentially by an order of magnitude. Through a combination of preventive measures and controls, we can build on successful historical cases like control of nuclear technology and critical infrastructure information. However, the implementation of safeguards needs to accelerate.

Summary of speaker insights

Understanding & Evaluating the State of AI

General-Purpose AI and Generative AI

While *general-purpose* AI aims to perform a wide range of tasks across different domains, like a multipurpose tool, less general or narrow AI is designed for specific tasks, like a specialized instrument.

Generative AI specifically refers to AI systems that can create new content, like text, images, or music. The focus is on the AI's ability to produce novel outputs rather than just analyze or interpret existing data.

Consider the difference between a large language model (LLM) that can solve a wide range of tasks, and one specifically trained and optimized to assist programmers by suggesting code snippets and completions as they write software. One system is general, the other more narrow - but they are both generative.

Towards more comprehensive systems
The surprising generality of some generative AI models, particularly those proficient in language, has led to them being seen as a promising path towards more comprehensive and versatile AI systems that can solve a wider range of tasks with higher quality. They suggest that a wider space of general-purpose AI might be feasible - with the large language models we've seen so far only comprising one aspect of what is possible.

Many of the systems we have seen in the last couple of years are more general than previous AI systems, but not yet human level. Human intelligence has over 200 components, and current general purpose AI systems display around a quarter of those traits. We have to keep in mind that current AI systems are very different from humans in terms of how they transform information, and be careful to not anthropomorphize them.

The question of AI agents
There is considerable pressure to equip general-purpose AI systems to be able to take independent action as "AI agents" (- a category that can both contain "AI personal assistant" and "autonomous weapon").

LLMs reflect patterns found in their training data. Since this data sometimes includes examples of human behavior that may be morally inappropriate for certain situations, the model can occasionally replicate those behaviors. Even absent this, consider that human decision-making relies on human self-perception, cognitive empathy, theory of mind, and moral intuition to support moral decision-making and action - faculties that LLMs lack.

At this point, we currently lack good solutions for managing the risk that stems from models' inability to truly "understand" the context and consequences of their outputs.

Standards and Benchmarks as Foundational Governance Tools

Why standards and benchmarks?
For AI to benefit humanity, ethical and societal concerns must be addressed throughout an AI system's entire lifecycle. Downstream actors and users bear AI system risks, whereas upstream developers hold the key information about risks.

Standards provide organizations with proven frameworks to address concerns and a common language needed for upstream developers and downstream decision-makers to communicate critical risk and capability information at every development stage, from design through deployment and monitoring, while also enabling methodical incident reporting.

Benchmarks, then, serve three key roles:

1. Benchmarks serve as milestones - They provide a common performance objective for different actors.

2. Benchmarks give a basis for discussing the impacts of achieving specific capabilities, and help coordinate actions once those benchmarks are achieved. They help design "If–Then" responses to seek consensus among actors with differing opinions of the risk/benefits ratios of specific AI systems.

3. Benchmarks provide a trusted baseline to check the quality of a system, or claims that a system is superior to another.

What do we need from standards and benchmarks?
Importantly, AI benchmarking exists in a point of tension between open science and valid testing: when benchmarks are public, models can train on them, invalidating results. The solution is to keep practice materials public while maintaining secret official test data, much like entrance exams or standardized tests for students.

For effective AI standardization, organizations should both extend existing frameworks and update domain-specific standards, rather than trying to create AI standards in isolation. In setting standards and their associated benchmarks, we need to balance tradeoffs between risk and value. For example, rather than demanding that a system produce zero misinformation (which would prevent development), we need benchmarks to determine acceptable levels of misinformation.

However, efforts to this end are challenged by the fact that there are no established best practices on how to evaluate AI systems for risk, and different scenarios have been proposed in a non-systematic way. This means we are yet to establish the conceptual basis for most potential AI risks. Mathematical frameworks can create simplified versions of problems to allow for formal analysis through abstract modeling. While not complete solutions, they could help structure these problems and bridge disciplines - as demonstrated by their success in algorithmic fairness research.

Nevertheless, for now we can say that current benchmarks focus too narrowly on specific capabilities, leading to flawed conclusions on AI system capabilities. This could take the form of a belief like "this model passes a legal benchmark and so is better than a human lawyer" - even though lawyers do much more than what a benchmark is testing for.

Advancing Safe and Robust Gen AI Development

From models to systems
While policies still assume single models, AI systems are evolving into multi-modal ecosystems. Even simple chatbots now use multiple AI components, including filters and specialist models for different tasks. We need new safety assessment methods that can evaluate both individual components and their interactions, moving beyond traditional single-model approaches.

The importance of context
When measuring specific capabilities in isolation, there will have been a preceding step of deciding which capabilities to measure. These rely on which concrete safety impacts on humans and society that are most important to test for. At present, these choices are being made without scrutiny, and we do not have reason to assume that AI labs have the competency to get them right.

Measuring specific capabilities in isolation will also often make it hard to account for second and third order effects. This is because AI system requirements vary dramatically based on jurisdictions, cultures, and use cases. For example, hallucination can be beneficial to a creative who is exploring ideas, while being detrimental for a person looking for reliable medical information. To address this reality, accuracy and safety need to be evaluated separately, as they represent distinct concerns that may be prioritized differently. Ultimately, different AI systems with different tradeoffs are needed to adapt to different settings.

To meet these challenges, the field needs to integrate diverse stakeholders from all communities at the start of AI projects - for example, while benchmarks are a technical product, they require input from all groups who will be impacted, not just engineering teams.

Meeting AI transparency requirements through interpretability research
Scholars sometimes use the term "mechanistic interpretability" to refer to the process of reverse-engineering artificial neural networks to understand their internal decision-making mechanisms and components, similar to how one might analyze a complex machine or computer program.*

Interpretability has become a critical technical requirement due to transparency and accountability mandates in major policy frameworks like the EU AI Act, creating concrete regulatory pressure for interpretable AI systems.

Recent advances like Representation Engineering** show promise by taking advantage of neural networks' fully observable nature during operation, enabling researchers to monitor and control internal model behavior. These methods have successfully detected unwanted behaviors like hallucination and increased model honesty.

While this shows that the field of interpretability research has made great progress in exploring how AI models think and make decisions, the field remains emergent and lacks solutions robust enough to fulfill current regulatory requirements. Specific benchmarks will need to be developed to measure and validate interpretability methods.

* Wikipedia: Explainable Artificial Intelligence
**Wang, Boxin, et al. "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models." NeurIPS. 2023

Securing the Path to AI Innovation with Risk Evaluations

Focusing on areas like bias, cybersecurity and biotechnology, experts discussed the importance of evolving evaluation methods for both current and more powerful AI systems.

What are Responsible Scaling Policies?
In the absence of regulations, AI labs are currently relying on Responsible Scaling Policies, a tool of corporate self-governance. In these policies, AI system developers state publically the specific conditions that, if observed in their AI systems, would indicate that the system capabilities were beyond what the company can handle with its current level of security and risk mitigations.

An example condition would be: "If a model can guide a non-technical user into synthesizing a pathogen, we would not continue developing it unless we have information security measures that can prevent anyone but a state actor from accessing our model settings ("weights")."

How often are AI models tested?
As the capabilities of AI systems are growing fast, evaluations need to happen frequently to make sure that a model does not jump straight from "not risky" to past a red line between evaluations. Therefore, multiple frontier labs have committed to run their evaluation at short intervals (e.g. every 3 months) and every time the amount of computation used to train the model increases by a set factor (e.g. every 6x increase in effective compute).

"Trust in evaluation requires trust that the evaluator tried hard enough to elicit capabilities. Currently, ***this is not something we're able to quantify****."*

Can we trust the risk evaluations?
Correctly evaluating the capabilities of AI systems is already a hard challenge, as we don't know of any way to determine the theoretical maximum capabilities of a model. Trust in evaluation requires trust that the evaluator tried hard enough to elicit capabilities. Currently, this is not something we're able to quantify.

Typically, whenever a company published a new model, users on the internet will find ways to elicit capabilities that the developers did not know existed.

Recent findings from Apollo Research show that frontier general models understand when they are being evaluated. This means frontier AI models can now exhibit different behavior during evaluation and deployment, including trying to trick the evaluators. In this situation, designing accurate and useful model evaluations becomes more challenging.

The possibility of models being deceptive makes trusting evaluations even harder. If we evaluate a model on a benchmark for being able to e.g. aid in production of chemical weapons, and the model knows it is being evaluated for this, the evaluation results cannot be trusted. There needs to be additional elicitation work to be sure the model is not hiding the capability. For this reason, evaluations are expected to require much work and time.

Summary of speaker insights

Governance, Policy, and International Economic Cooperation

Industry risk management practices
In the absence of regulation, leading AI labs are developing internal guidelines for how to manage risks and disparities arising from deploying GPAIS. Several policies establish tiered frameworks identifying capability thresholds that could indicate potential danger (e.g. "Frontier Safety Framework" or "Responsible Scaling Policy"). These include commitments to pause development if capabilities approach levels beyond current safety measures. However, companies can currently alter the clauses in these policies to rely on more subjective and less specific criteria - and some have already done so.

Global participation
However, defining the level of "acceptable risk" remains a normative question that requires public and domain expert input, particularly as AI systems are deployed in critical settings. Most current initiatives, such as the US international network of AI Safety Institutes are ad hoc coordination platforms which may exclude key players without a designated AI Safety Institute. Notably, two-thirds of countries are not actively involved in any international AI governance forums. Responsible AI innovation should be addressed from a global commons perspective, where if even one company, state or other actor defects, it hurts everyone. Ad hoc dialogues will need to include all relevant players, and be more tightly linked with more inclusive organizations like the UN.

Regional contributions
While advanced economies maintain consistent focus on AI governance, most countries struggle with unstable engagement due to lacking capacity and competing priorities. This is an opportunity for regional forums like APEC, PECC, and others to provide a stable platform to develop, enrich, synthesize, and communicate insights and strategies that might otherwise be lost. Some less advanced economies and/or small states facing specific needs and situations are developing innovative approach: Singapore, for example, is launching an "AI Playbook for Small States to shape inclusive global AI discourse" to share lessons on addressing AI safety while lacking the leverage of major powers. Such approaches combine selective/targeted regulation on critical issues (like online safety) with strategic partnerships with larger jurisdictions where they can give input to regulation that concern the top upstream GPAIS developers.

What true democratization means
Yet, the challenge of broader participation in AI governance extends even beyond mere representation or catching up. A key insight emerged around mischaracterizing AI "democratization." True democratization requires putting AI development under democratic control, through public consultations, transparent and accountable governance frameworks, and rule of law - not just facilitating access.

This connects to deeper questions about how technological gains are distributed - as one speaker noted, productivity increase does not always lead to increased welfare for the population.

History suggests the need to actively shape how AI's benefits are distributed, rather than assuming technological progress automatically benefits all.

AI labs have raised massive amounts of capital based on the promise of technology. Democratization requires a ***balancing of power*** *in this situation.*

Balancing power
AI labs have raised massive amounts of capital based on the promise of technology. Democratization requires a balancing of power in this situation. With their unparalleled resources (including compute, data, and talent compensation), the few frontier AI labs and big tech backers have generated unprecedented gravitational pull around their frontier AI models that attracts top class expertise out of public or academic research. Given the current race dynamics and absence of regulation, incentives are very high on frontier labs to develop and deploy advanced models without prioritizing risk prevention, management, and shared value. Meanwhile, the government still lacks knowledge, competence and capacity to install appropriate regulations, whereas civil society has knowledge but lacks bandwidth and power.

Acting under uncertainty

Multiple speakers highlighted a critical tension between waiting for scientific consensus and taking action, and that jurisdictions are already going in different directions. Substantial work will need to go into building the global public goods that will allow for juridictions to make different choices without imposing undue cost on others.

Finally, several speakers have commented that the seminar title "Responsible Adoption of General-Purpose AI" ought to have included "Development and Deployement" as well as "Adoption". The path dependent nature of technology means that it is easier to address risk through a combination of levels: the model level (design and development), the specific use cases (deployment), and the user level (adoption).

Substantial governance innovation will be required to coordinate on the design and development of GPAIS, both to balance parts of the ecosystem and the varying interests of jurisdictions. Decision-makers can choose to place a positive vision at the heart of this coordination, of incentivizing the development of GPAIS technology that robustly increases social welfare.

If competitive dynamics are allowed to go too far towards ambitions of hegemony, such priorities can lead to locking GPAIS innovation into a zero-sum, weaponized paradigm. To address this risk, it is crucial to build out the global AI infrastructure that will support economic "co-opetition" and shared abundance. This should be rooted in a "win-win" spirit and commitment to learn from the mistakes of the past, or if such convergence is not possible, in shared commitment to peaceful co-existence.

Appendix A

Policy ideas from roundtable discussions

The seminar convenors developed a number of ideas for potential policy statements ahead of the seminar, intended for C-level executives and government leaders in the Asia-Pacific region, multilateral institutions, and the public at large. The ideas have been updated to reflect the insights gathered during roundtable discussions that were hosted each afternoon at the seminar. While participants shared their insights, it should not be assumed they support any specific policy idea herein.

Investing in Responsible Adoption of General-Purpose AI Systems

1. Optimize building and allocating **AI talent**, including:

a) economies with strong AI sectors prioritizing investments in digital infrastructure to bridge the capacity gap between advanced and less advanced economies, so as to foster use of general purpose AI technologies, components, and systems;

b) facilitating strategic cross-border academic exchanges, joint degrees, and collaborative/interdisciplinary research projects in AI. Collaborations should be fostered while maintaining a balanced approach to safeguard against the risks of unintended technology transfer;

c) designing and testing efficacy of various incentives to encourage employees across public, private, and civil sectors to seek AI literacy training or certification to increase their skill at using AI tools; and

d) foster regular dialogues between governments, academia and the private sector on AI talent strategies. Including via the creation of dedicated task forces to discuss challenges, share best practices, and develop coordinated approaches to talent building and allocation across the region.

2. Facilitate investment into **R&D and international standardization of technical solutions** (alongside behavioral protocols) that enable safe and responsible adoption of general-purpose AI systems, including:

a) Advanced Computing Governance mechanisms;

i. Methods such as chip-based verification mechanisms, differential privacy, homomorphic encryption, secure enclaves, and secure multi-party computation can help balance the need for thorough safety evaluation with the protection of privacy and sensitive information:

b) Unlearning Techniques:

i. In scenarios where AI models have learned biased or incorrect information, private information, or hazardous knowledge, unlearning methods enable the removal of specific data patterns from the model without requiring a complete retraining. This can be helpful for improving fairness, accuracy, and adaptability of AI systems.

c) AI Oversight and Auditing Tools:

i. Developing advanced tools and frameworks for monitoring, auditing/testing/evaluating, and managing AI systems is essential to ensure transparency, accountability, and ethical compliance;

d) Robustness and Safety Mechanisms:

i. Creating AI systems that are resilient to adversarial attacks, unexpected inputs, and operational failures is critical for them to be deployed across a large number of use-case and sectors - particularly in safety critical applications;

ii. Investment in robustness and safety research helps ensure that AI can operate reliably and securely in a wide range of environments, maximizing socio-economic prosperity derived from AI;

e) Holistic Explainability and Interpretability:

i. As General Purpose AI systems become more complex, making their decision-making processes interpretable to humans is vital. Explainable AI (XAI) techniques are an important step to enhance trust and facilitate broader adoption across industries that require transparency, such as healthcare and finance. These should be furthered and complemented in a holistic, interconnected, multi-level approach to understanding AI at various levels of abstraction, from individual components studied in isolation in basic prototypes of models, to documenting and analyzing behaviors discovered "in the wild". Most importantly, it should focus on building synergies between these levels, much like the productive interplay between neuroscience and psychology in studying human cognition.

f) Bias Mitigation Techniques.

i. Addressing inherent unwanted biases in AI models is essential to prevent discriminatory outcomes. Investment in developing and integrating unwanted bias detection and mitigation tools ensures that AI systems are fair and equitable in their decision-making processes.

3. Ensure that government entities across economies (such as AI Safety Institutes or similarly tasked departments) **produce recommendations** on:

a) roadblocks to ensuring that AI systems can be developed and adopted responsibly, and which kinds of public and/or private investments that should be prioritized as a result; and

b) roadblocks to dynamic regulation anticipating or following breakthroughs in advanced AI, and how government practices can innovate to account for these.

4. Leverage significant international financing and resources from affluent countries to support the emergence of a **globally distributed AI assurance technology ecosystem through holistic market-shaping policies** at the national, regional and international levels, through:

a) proactive, deliberate, well-defined, application-specific, and reliable announcements of AI-related compliance requirements for AI companies and AI-adopting sectors; and

b) supporting early-stage market creation (e.g. scholarships, incubators and accelerators) with national and international public funding (e.g., grants, public-private funds, bounties, impact bonds), eased administrative requirements, and procurement standards, aiming to complement and spur rather than replace private-sector investments.

5. Promote **wider adoption of AI systems which are not at the frontier** (e.g. small versions of large language models), if they can allow for economic benefits with lower risk and carbon/environmental footprints than frontier models:

a) Industrial scale robustness requirements should still apply to such systems to address the systemic risk potentially emerging from their large scale deployment in our economies.

Understanding the State of AI: Investigation and Evaluation

6. Build a public open source intelligence platform for summary-level, dynamic **mapping of the global and regional supply chains** of General Purpose AI technologies and systems.

7. Facilitate **cooperation between National AI Safety Institutes as trusted intermediaries** among academia, governments, industry, and civil society to develop evaluations that test both the capability and likelihood of a system to produce harmful output.

8. Contribute to building a **scientific consensus** (including establishing key points of uncertainty and legibly codifying disagreements) on the current trajectory of AI advances in coordination with the United Nations and other intergovernmental organizations (e.g. OECD/GPAI, and UNESCO), including:

a) a shared taxonomy;

b) research agenda; and

c) issuing of periodic "state of the science" reports.

9. Facilitate **cooperation between government bodies and AI research organizations** (involving key players from civil society, academia, and industry) instructed to:

a) conduct research and training to share best practices in AI measurement science (metrology), allowing for conceptualizing and testing dynamic, adversarial, and symbiotic testing methodologies that assess capabilities (the capability of a model to do harm) and alignment (the tendency of a model to be willing do harm) of advanced AI models;

b) collaborate on a common interoperable framework for rigorous, adaptive and dynamic evaluation techniques, to facilitate more effective and comparable assessments across different models, companies, and applications;

c) build on and evolve best practices from other fields on information sharing (taking into account infohazard), especially regarding classified material;

d) mandate and accredit independent third party evaluators (such as start-ups and companies with strong scientific expertise) to test for high-impact harmful capabilities (pre- and post-deployment).

e) helping negotiate access to top models on behalf of third-party evaluators.

Understanding and addressing AI's Socio-economic Impact

10. Immediately start the process of building **public understanding** of General Purpose AI technologies by carrying out transparent processes to produce:

a) dialogues to derive principles or charters on AI ethics and red lines (building on comprehensive international dialogue processes, e.g. involving deliberative decision-making methods such as citizen/expert assemblies);

b) publicly available guidelines on how to develop, test, and deploy AI systems while adhering to principles and charters for avoiding unacceptable risk

c) publicly available risk and impact assessment reports;

d) publicly available leaderboards and government-level enforcement mechanisms to dis-incentivize breaking with principles and charters; and

e) mechanisms that inform and consult users before transferring sensitive data to third parties, including for the training of AI systems.

11. Build an evidence base and high quality forecast scenarios on **AI-enabled automation** (e.g. the impact of LLMs on national development strategies including, for instance, on software development skills and creative skills), beginning with conducting a detailed analysis of:

a) baseline pre-existing skills gaps or talent shortages in the industry value chain (e.g. of technical skills and AI-adjacent skills)

b) the expected impact of General-Purpose and Generative AI on skill requirements, labor markets, and population health and wellbeing.

12. Provide grants, subsidies, or technical assistance to ensure that **small and medium size businesses, nonprofits, and civil society organizations in underserved markets** and communities can access and benefit from trustworthy AI models and systems.

13. Ensure that investment in the AI ecosystem **prioritizes capabilities that robustly increase social welfare**, including by:

a) Fostering open science, community engagement, and participation from diverse stakeholders, including civil society groups and impacted communities

b) Powering responsible community-driven innovation by enabling public access to -and control of- models, weights, data sets, training configurations & check points by relying on open licenses when this does not create disproportionate risks to economies and societies and by developing new mechanisms to enable safe shared access when needed;

c) Prioritizing the adoption of existing, reliable AI models/systems in use cases where entire economies can benefit;

d) Privileging the development of AI capabilities that lead to increases in social welfare in a wide range of environments – relative to progress on capabilities that are dual-use (e.g., useful for deception, manipulation, disempowering other agents) and therefore may not robustly improve social welfare;

e) Conducting continuous impact assessments and ensure transparency around data sources, licenses, and processing applied; and

f) Prioritizing environmental impact measurement across the AI lifecycle, including carbon footprint calculations, energy efficiency ratings, and circular design principles for hardware reuse and responsible component recycling.

Governance, Policy, and International Economic Cooperation

14. **Promote dialogue** with organizations such as OECD.AI/Global Partnership on AI (GPAI), ASEAN, and UNESCO's Advisory Committee on Artificial Intelligence on the following measures, and support the UN taking a coordinating role in:

a) digital capacity building (e.g. building digital literacy and infrastructure); and

b) leveraging or developing evidence-based and interoperable instruments for policy, metrics, standardization and norms to ensure that the development and application of AI systems serves the public interest.

15. Cooperate on developing **compute governance mechanisms** by leveraging the traceability of hardware, e.g. follow "Know Your Customer" principles by:

a) tracking customer's basic identifying information (e.g. identity, source of payment, and business purpose for utilizing the computing cluster);

b) assessing if the customer intends to train powerful and advanced models; and

c) validating the identifying information if a customer repeatedly utilizes computer resources that would be sufficient for such training.

16. Increase transparency and accountability in AI development by initiating development of a **transparency framework** in which AI system developers provide details to a trusted third party, to encourage third-party audits.

17. Strengthen and deepen regional economic cooperation on the responsible adoption of AI by developing **capacity building programs** (designed to maintain independence between economies) targeting senior, mid-level and junior government officials who are responsible for trade agreement negotiation and implementation, with the objective of:

a) deepening understanding of the ways global trade rules impact the responsible development, adoption, and operation of advanced AI systems;

b) enhancing insights into the elements that enable the development of cutting-edge AI systems - specialized talent, high quality data sets, algorithms, data flows, computing power, access to electricity, and other key inputs;

c) sharing best practices and experiences from risk management to ensure beneficial outcomes of integrating advanced AI systems in various sectors; and

d) ensuring that domestic regulation does not unnecessarily prevent international coordination or cooperation.

18. Promote **cross-border interoperability & scalability** of governance, risk management, and compliance initiatives by:

a) collaborating with regional and international standard organizations to develop and release unified technical standards and guidelines; and

b) allocating public funding to research *cooperative AI* (promoting cooperation between humans, machines, or organizations) to lay the foundation for increasing the pace and dynamism of harmonization efforts.

19. Advance a comprehensive international dialogue/consensus process aimed at establishing concrete "red lines" for AI development and deployment. Such a process could involve identifying key areas, technologies, and components, as well as seek to develop measurable indicators and thresholds.

20. Implement a **transparency and accountability framework** (see Recommendation No. 16) in which AI labs in the private and public sectors as well as academia provide details about:

a) Model Information: The types of models they are developing, including their purposes, capabilities, and limitations;

b) Data Sources: The sources and nature of the data, to ensure that the data used is ethical and diverse;

c) Performance Metrics: How they evaluate and measure the performance of their AI systems, including any biases or potential issues identified;

d) Ethical and Safety Measures: Transparency regarding the ethical guidelines and safety measures in place to mitigate risks associated with AI technologies;

e) Incident monitoring: Tracking actual AI incidents and hazards in real time and providing the evidence-base to inform the AI incident reporting framework and related AI policy discussions; and

f) Impact Assessments: Information on the potential societal impacts a lab anticipates from their AI systems, including risks related to privacy, discrimination, and other ethical concerns.

21. Create trusted and secure **data trading/ exchange platforms** enabling data providers (or their right owners including trusts and syndicates) and data users (e.g. model developers) to transact on copyrighted data according to facilitated/standardized terms.

Please direct questions and comments about the policy ideas to:

aparle@ncapec.org

nicolas@prism-eval.ai