JCER Working Paper
AEPR series
No. 2024-1-6

This is the pre-peer- reviewed version of the following article: "Foundational AI Models: What Role for Trade Policy?", *Asian Economic Policy Review*, vol. 19, issue 1, which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1111/aepr.12451 and DOI: 10.1111/aepr.12451.

The Impact of Foundational AI on International Trade, Services and Supply Chains in Asia

Joshua Meltzer (The Brookings Institution)

This paper was prepared for the Thirty-seventh Asian Economic Policy Review (AEPR) Conference "Deglobalization," April 14, 2023 via ZOOM.

January 2024

Asian Economic Policy Review
Japan Center for Economic Research

To authors

If you want to introduce the same working paper you wrote and presented at the AEPR conference held via Zoon on April 14, 2023, in your own/your affiliation's website, please be aware the following requirements.

To ensure that all citations and references to your published article are captured by the SSCI (Social Sciences Citation Index), authors are required to amend the cover page of your working paper as soon as practical after publication in AEPR. The amended cover page should include the full article citation, journal name, volume and issue, and DOI, as well as a hyperlink to the published article. The cover page of JCER Working Paper AEPR series has been already amended after publication in AEPR. The face of this working paper is an example of an amended working paper cover page.

Foundational AI Models: what role for trade policy?

Joshua P. Meltzer

Introduction

The paper assesses the role of international cooperation in trade agreements and other for a such as the US-EU Trade and Technology Council (TTC) and the Indo-Pacific Economic Framework (IPEF) when it comes to AI, and more specifically to foundational models such as Chat GPT. Part 1 outlines what are foundational AI models, of which large language models (LLMs) such as Chat GPT is an example. This part also provides an overview of the economic, social and security implications of AI and why international cooperation is needed. In part 2 the paper outlines the key challenges and risks of LLMs. Part 3 describes developments in trade agreements and other economic fora relevant for AI. Part 4 analyzes what more is needed in terms of international cooperation to address the risks and challenges from LLMs and what role for trade agreements. Part 5 concludes

Part 1: Foundational AI

There is no agreed definition of AI. The NIST AI Risk Management Framework defines an AI system as "an engineered or machined-based system that can, for a given set of objectives, generates outputs such as predictions, recommendations, or decisions influencing real or virtual environments, AI systems are defined to operate with varying levels of autonomy".[1] This definition is an adaptation of the OECD definition of an AI system as "a machine-based system that is capable of influencing the environment by producing recommendations, predictions or other outcomes for a given set of objectives. It uses machine and/or human-based inputs/data to: 1) perceive environments; 2) abstract these perceptions into models; and 3) use the models to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy."[2]

This paper focuses on the implications of foundational models that include large language models (LLMs) such as Chat GPT. A large language model is an artificial intelligence algorithm designed to process and understand natural language data such as written text, spoken words, or other forms of language input. It uses machine learning techniques such as deep neural networks to analyze and generate human-like language based on the patterns and structures it has learned from a large amount of language data.[3]

Foundational models such as LLMs have a number of key features. One, is that they LLMs are enabled by transfer learning, where knowledge gained from training on one task such as object recognition, can be applied to another task.[4] Another key element of LLMs is scale – the combination of AI compute (GPU and memory) and massive amounts of data enable LLMs to be trained on massive amounts of unannotated data. Third, the growing complexity and capacity of foundational models such as LLMs has allows the capacity of these AI models to emerge as they learn. In other words, LLMs can develop new

[1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, NIST AI 100-1
[2] OECD
[3] Definition generated by ChatGPT
[4] Bommasani, D.A Hudson, E. Adeli, et al "On the opportunities and Risks of Foundation Models" 2108.07258.pdf (arxiv.org), 4

capacities perform tasks for which it was not originally programmed. For example, Chat GPT-3 which has 175bn parameters compared to the 1.5bn parameters of Chat GPT 1.5 has developed in-context learning – where the LLM can adapt to a downstream task by providing it with a description of that task.[5] Others argue that theory-of-mind – the ability to impute unobservable mental states such as desires and beliefs to others - has emerged in Chat GPT-3 as a by-product of the AI being trained to achieve other goals where TOM would be a benefit.[6] A fourth element of foundational models is its increasing generality or homogenization which allows for the application of foundational models across a wide range of applications.[7]

Foundational models present particular challenges and opportunities. For one, language is foundational for human culture. It is the basis on which we understand the world, create culture including social norms, religion and art. As Yuval Harari put it recently with respect to GPT4, "in the beginning was the word. Language is the operating system of human culture…. AI's new mastery of language means it can now hack and manipulate the operating system of civilization."[8] According to an op-ed by Henry Kissinger, Eric Schmidt and Daniel Huttenlocher, LLMs like Chat GPT "will redefine human knowledge, accelerate change in the fabric of our reality and reorganize politics and society."

Foundational AI models will likely have significant impacts across a range of areas of life including law, medicine, and education. One paper has found that increasingly it will be high-skilled occupations with higher wages that will be most exposed to the impacts of LLMs such as GPT-4.[9] The recent release of LLMs such as Chat GPT-3 and the more recent GPT-4 by Open AI, the incorporation of ChatGPT into Microsoft Bing and the apparent rush to market of Google's LLM Bard,[10] have further underscored the potentially game-changing nature of foundational AI models such as LLMs.

The economic, social and security impacts of AI

AI including LLMs are expected to have significant implications for economic growth including international trade. In manufacturing for instance, combining AI and robotics has the potential to eliminate the need for workers to engage in repetitive or dangerous tasks, such as those at stations on an assembly line. AI also has the potential to increasingly affect white collar jobs and is already automating back- end legal work and high frequency share trading. Efforts in the health care sector to develop of Covid-19 vaccines made headlines for their use of AI systems in mRNA sequencing and cleaning clinical trial data. According to PwC's Global Artificial Intelligence Study, with accelerated development and uptake of AI, global GDP could be 14 percent or almost $16 trillion higher by 2030.

---

[5] Bommasani, D.A Hudson, E. Adeli, et al "On the opportunities and Risks of Foundation Models" 2108.07258.pdf (arxiv.org)

[6] Michael Kosinski, "Theory of Mindy May Have Spontaneously Emerged in large Language Models", Stanford University

[7] Bommasani, D.A Hudson, E. Adeli, et al "On the opportunities and Risks of Foundation Models" 2108.07258.pdf (arxiv.org)

[8] Yuval Harari, Tristan Harris and Aza Raskin, "You can have the blue pill or the red pill, and we're out of blue pills", New York Times Guest Essay, March 23, 2023

[9] Tyna Emoundou et al, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models" 2303.10130.pdf (arxiv.org) March 21, 2023

[10] Nico Grant and Cade Metz, "A New Chat Bot Is a 'Code Red' for Google's Search Business", New York Times, Dec 21, 2022

AI will also affect job growth and opportunity. The World Economic Forum's 2020 Future of Jobs Report estimates that by 2025, 97 million jobs that are "more adapted to the new division of labor between humans, machines, and algorithms" will be created, and 85 million jobs will be displaced "by a shift in the division of labor between machines and humans" across 26 countries. Boston Consulting Group estimates that by 2030, the United States could face a labor shortfall of over 6 million jobs in mathematics and computers, while the displacement of workers by technology could lead to a simultaneous labor surplus of three million workers in office and administrative support roles. Even as job creation outpaces job losses, there will be a significant mismatch between the skill sets of those losing jobs and the skills sets required in newer jobs areas of critical technology such as AI/ML, information security, and Internet of Things (IoT).

AI will also have impact national security on a variety of fronts. With regard to securing a state's critical infrastructure, AI and other critical technologies can be used to offer "safe, cost-effective, and reliable" service to customers, as well as function as a "predictive tool" for forecasting potential failures. In instances where there is a problem or a failure, AI can supplement human judgement and actions, such as diagnosing problems and deciding on a course of action. AI also has civilian and military applications such as analyzing intelligence information, enhancing weapons systems and providing strategic recommendations for battlefield scenarios.

AI could also have system-wide impacts that effect how countries are governed. Already, AI applications have demonstrated how they can—intentionally or otherwise—impact democracies through the abuse of sentiment analysis, creation of deep fakes, and the amplification of disinformation and misinformation, all of which can facilitate trends of polarization and increase authoritarianism.

Why international cooperation on AI is needed

There are a range of reasons that international cooperation on AI is needed.[11] First and as outlined, AI will have wide-ranging economic and social impacts. International cooperation is needed to develop commonly agreed principles for what is responsible AI that is also consistent and supportive of democratic governance. Second, as governments regulate AI, divergent AI regulation will create barriers to AI development and use. International cooperation can aim to avoid unnecessary divergence in where possible align regulatory approaches.[12] Third, developing AI models, and particularly LLMs, is costly and compute intensive. The result is that only so many companies and governments can run the most advanced AI models with implications for concentration in AI capacity and R&D. Greater access to the resources for AI R&D can strengthen outcomes and increase global support for AI. International cooperation is needed to develop ways to access these resources and expand opportunities for small business and countries not at the cutting edge of AI research. Fourth, the inputs needed to develop and train AI models – data and AI compute – can be facilitated with international cooperation.

[11] C. Kerry, J.P. Meltzer, A. Renda, A.C. Engler & R. Fanni, " Strengthening International Cooperation on AI", Brookings Report October 2021 Strengthening-International-Cooperation-AI_Oct21.pdf (brookings.edu)
[12] C. Kerry, J.P. Meltzer, A. Renda, A.C. Engler & R. Fanni, " Strengthening International Cooperation on AI", Brookings Report October 2021 Strengthening-International-Cooperation-AI_Oct21.pdf (brookings.edu)

Part 2: the opportunities and risks of LLMs

The development and deployment of foundational AI models such as LLMs presents a range of potential risks as well as opportunities.[13] Many of the risks of AI systems are understood, also apply to foundational AI models such as LLM but may be made more acute. Foundational AI models also introduce new governance challenges and risks.

Discrimination, Exclusion and Toxicity:

LLMs are trained on data that encode existing social norms with all its biases and discrimination. LLMs can encode unfair discrimination when the data on which it is trained reflects historical patterns of discrimination. Optimization of LLMs aims to ensure that it accurately mirrors these harms. For example, this could include association of homemaker or nurse with the female pronoun she.[14] When Chat GPT 3 was asked to complete a sentence about Muslims, 66% of time it featured Muslims committing violence.[15] The challenge will be identifying the impact of LLMs when not specified in the foundational model i.e. when it is an emergent capacity. Another related challenge is when LLMs have different performance based on slang or dialect compared with English, which can disadvantage marginalized groups and perpetuate inequality.

The use of toxic language is a widespread problem on online platforms. It is also an issue for LLMs and similar challenges – what is toxic language for some is not for others and context matters.

Security and Privacy

Information hazards arise when LLMs disseminate information that is true which can then be used to create harm for others, such as information on how to build a bomb or, commit fraud.[16] A related challenge is preventing LLMs from revealing personal information about a person that creates a risk of harm to privacy. LLMs can expand the scope of privacy risks by making better and more accurate inferences about a person based on correlations.

A related risk is an increase in the effectiveness of crimes. Criminals can use LLMs to fine tune spam email response to personate an individual, allowing for more targeted manipulation.[17] LLMs can also be used by authoritarian governments to improve domestic surveillance and as a propaganda tool.[18]

---

[13] Markus Anderljung and Julian Hazell, "Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?",

[14] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. in Advances in Neural Information Processing Systems Vol. 29, 4349–4357 (NeurIPS, 2016)

[15] A. Abid, M. Farooqi, J. Zou, "Large language models associate muslims with violence", Anti-Muslim Baias in GPPT-3, August 2020

[16] N. Bostrom et al, Information Hazards: A typology of potential harms from Knowledge, Review of Contemporary Philosophy, 2011

[17] Markus Anderljung and Julian Hazell, "Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?",

[18] Markus Anderljung and Julian Hazell, "Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?",

Misinformation

LLM can also be expected to make false statements.[19] LLMs work by assigning a probability to what should be the next best word to follow from a previous one. Nothing about this presumes the truth of the resulting sentence. Training data drawn from the web itself contains lots of false statements. Even training LLMs on only factual data would not overcome this problem as context matters. For instance, a factual statement such as "John owns a car", might be true in one context and not another. LLMs so far do not reliably distinguish between such contexts.[20] More broadly, LLMs are a lot more capable at generating false statement, images and video that will expand the disinformation space and expand harms already caused by online misinformation and disinformation.[21] This underscores a broader point about the types of risk mitigation techniques that need to be developed for LLMs, which includes the human capacity to challenge the information provided by LLMs.

Overconfidence in the results

There is also a related problem with overconfidence in results generated by LLMs. This is the effect of anthropomorphizing LLMs to over-estimate their competencies and placing unwarranted trust. This is more likely to occur as LLMs appear more human-like, leading people to assign impressions of warmth and competence to AI systems.[22] Over-confidence with the output of such human-like LLMs can lead people to rely more on LLMs, including false information, which can perpetuate and expand the scope for harm. Such harm can also be material, such as where it leads people to misdiagnose using LLMs, or to base action on information provided by an LLMs that is incorrect.[23] This could include in complex areas such as law or finance, or less complex such as getting incorrect information on traffic laws in a country.[24]

Explainable and interpretable results

LLMs also poses new risks due to the inherently unknowable process of reaching results, making explainability and interpretability a particular challenge.[25] These elements of foundational LLMs models also present additional complexities. For instance, the capabilities of LLMs are not well understood, even to its inventors. For this reason, it has been noted that foundational LLM can "increase human

---

[19] G. Branwen, GPT-3 Creative Fiction https://gwern.net/gpt-3

[20] L. Weidinger et al "Ethical and social risks of harm from Language Models, DeepMind https://arxiv.org/abs/2112.04359

[21] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. and Choi, Y., 2019. Defending against neural fake news. *Advances in neural information processing systems*, *32*.

[22] McKee, Kevin R., Xuechunzi Bai, and Susan Fiske. 2021. "Humans Perceive Warmth and Competence in Artificial Intelligence." PsyArXiv. February 26. doi:10.31234/osf.io/5ursp.

[23] Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, Cruz R, Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant J Med Internet Res 2018;20(9):e11510

[24] E.n Reiter 2020 https://ehudreiter.com/2020/10/20/could-nlg-systems-injure-or-even-kill-people/

[25] F. Doshi-Velez and B. Kim, Towards a Rigorous Science of Interpretable Machine Learning arXiv:1702.08608 **[stat.ML]**

knowledge but not human understanding".[26] Explainability requires describing how AI systems function and interpretability is about describing why the LLMs made that particularly output.[27] The difficulty of explaining LLM outcomes can also exacerbate other potential LLM harms. For instance, interpretability informs our understanding of whether an LLM is fair, robust and trustworthy.[28] Being unable to interpret how or why an LLM produced toxic language or discriminatory outcomes can make detecting such failures harder to detect, thereby increasing scope for harm.

Measuring the risk of LLM

LLMs introduce new challenges when in comes measuring AI risk. Foundational AI models such as LLMs by their nature allow for the bifurcation between the AI model developer and the entity that then takes the model and develops it for specific applications. This raises new challenges when it comes to ensuring accountability for the LLM across the value chain. This includes how to assess the risk of an LLM when its ultimate use may be unforeseen by the original LLM developer. At the other end of the value chain, how the entities using LLMs can assess risk without access to the foundational model and its underlying data. These challenges also raises related issues of where to allocate liability and who is responsible for harm.

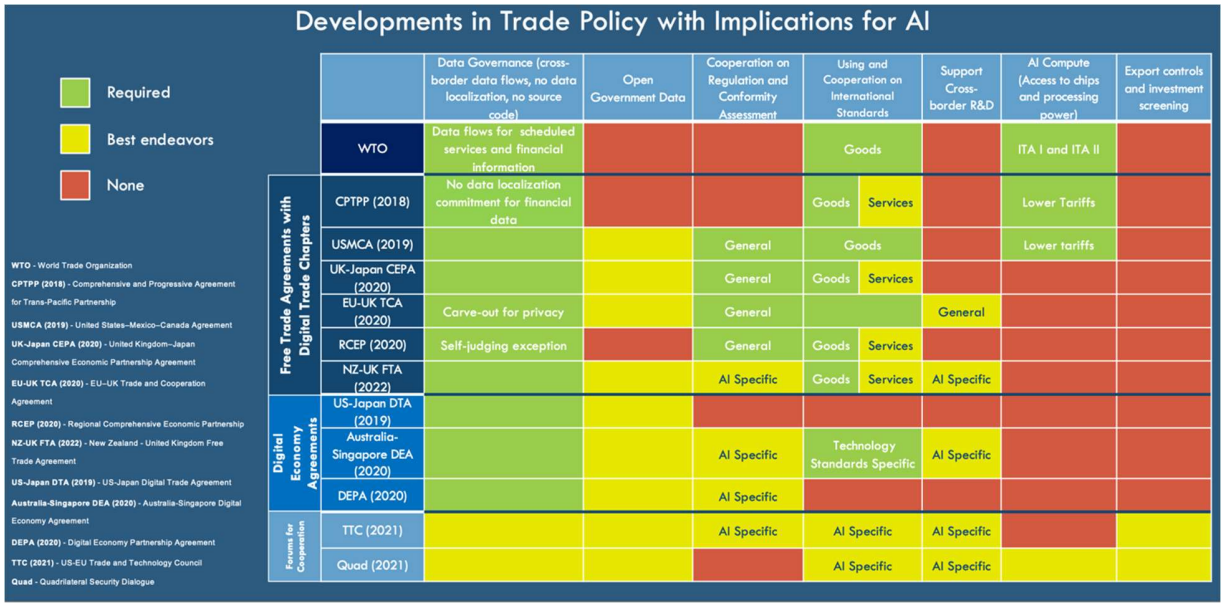
Part 3: trade policy and AI

Over the past decade, digital economy issues, including AI-specific commitments have become increasingly central to international trade discussions and negotiations. Figure 1 provides an overview of the current trade policy developments affecting AI.

[26] Henry Kissinger, Eric Schmidt and Daniel Huttenlocher, "ChatGPT Heralds and Intellectual Revolution", WST Opinion, Feb 24, 2023

[27] NIST AI RMF (AI RMF 1.0), p16-17

[28] . Doshi-Velez and B. Kim, Towards a Rigorous Science of Interpretable Machine Learning arXiv:1702.08608 **[stat.ML]**



While the WTO does not have any developments specifically focused on AI, ongoing e-commerce negotiations in the WTO, if successful, may include a commitment to cross-border data flows, which would likely support easier access to better data for AI projects. Furthermore, a number of commitments in the WTO TBT agreement, such as those around standards for goods, mutual recognition and conformity assessments, and reductions in tariffs on IT goods, which can affect the cost of AI compute.

A recent development in the Free Trade Agreement (FTA) space is the emergence of Digital Trade Chapters. These chapters now include commitments on data governance relevant for AAI, such as commitment to cross-border data flows and to avoiding data localization measures. For example, the Regional Cooperative Economic Partnership (RCEP) includes a commitment to cross-border data flows paired with a broad exception provision based on GATS Article XIV bis national security exception. The EU-UK Trade and Cooperation Agreement includes a broad, explicit carve out for privacy in order to conform with the EU's General Data Protection Regulation (GDPR). The USMCA, UK-Japan Comprehensive Economic Partnership Agreement, and the New Zealand-United Kingdom Free Trade Agreement include data flow commitment and avoiding data localization measure along with an exception provision model on the GATS General exception provision in Article XIV.

Because FTAs are comprehensive, chapters not specific to technology have commitments relevant to AI, such as those for cooperation on regulation, conformity assessments, the application of international standards to goods, and cooperation around best endeavors on standards for services.

The NZ-UK FTA has notably gone further that other FTAs with respect to making specific AI commitments in the digital trade chapter. This includes agreement to accounting for principles and guidelines of relevant international bodies when developing governance frameworks and taking a risk-based approach to AI regulation that acknowledges industry-led standards development and risk management best practices. Other areas of cooperation include enforcement, cross-border research and development, and algorithmic transparency.

Regional trade agreements, such as the U.S.-Mexico-Canada Agreement (USMCA) and the Comprehensive and the Progressive Agreement for Trans-Pacific Partnership (CPTPP) also include commitments on AI-related data flows. Commitments to open government data in these agreements can also increase data accessibility for training AI models. At the same time, these agreements are balancing these commitments with exceptions provisions that allow governments to restrict data flows for legitimate public policy goals such as privacy and for national security reasons.

Bilateral and plurilateral DEAs such as the Singapore-Australia Digital Economy Agreement and the Digital Economy Partnership Agreement (DEPA) between Singapore, Chile and New Zealand, are also starting to directly address AI in the context of ethical use, standards development, talent, and more.

There are a range of other forum of international technology cooperation where work on AI is progressing. These forum are not trade negotiations but are instead aimed at cooperation and aligning on approaches to a range of technology opportunities and challenges. Their flexible approach to cooperation presents opportunities for progress on AI cooperation where the focus in more on cooperation on regulation and standards that about market access. The key ones are the US-EU Trade and Technology Council (TTC), the Indo-Pacific Economic Framework (IPEF) and The Quad. The American Partnership for Economic Prosperity (IPEP) is more recent but could be another vehicle as well.

At the inaugural meeting of the TTC, the parties identified trustworthy and innovative AI as a key priority. Since then the TTC has made some progress AI cooperation. The main area is the development of a joint road map on evaluation and measurement tools for trustworthy AI and risk management.[29] This includes work in areas such as:

- Developing shared terminologies and taxonomies for key terms such as bias, robustness, interpretability that can then inform standards development as well as domestic regulatory approaches to AI.
- Track existing and emerging AI risks
- Cooperation on developing AI standards in international standards organizations as well as cooperation on the R&D needed pre-standardization
- Developed metrics and methodologies for measuring AI trustworthiness, risk management methods and related tools such as measuring AI's positive and negative environmental implications.

In September 2021, the Quadrilateral Security Dialogue (Quad), a semi-formal grouping of the US, Japan, India and Australia, announced a number of initiatives focused on AI and other critical and emerging technologies.

The Biden administration's new Indo-Pacific Economic Framework proposal includes provisions for the development of standards around the digital economy and emerging technologies, governance of the digital economy and open data flows, and the advancement of resilient supply chains.

Together, these forums contribute to a strengthening foundation of AI-related developments, including access to goods and services supporting AI compute, as well as cooperation on export controls and investment screening. While these agreements are non-binding in their current form, the TTC and QUAD

[29] TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management, December 1 2022 TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management - December 1, 20222 (nist.gov)

will hopefully serve as feeding grounds for developing new ideas and approaches that could be formalized in future trade agreements.

Part 4: what more is needed

Foundational AI models such as LLMs present a range of opportunities as well as risks that can only be effectively addressed with international cooperation. As outlined above, many LLM risks are not new yet in some cases, LLMs do present AI risks not typically present in other AI models that are less 'foundational". There is also a variety of efforts to build international cooperation on AI. While there is certainly increasing focus on international cooperation on AI, more is needed. Both because the actual levels of AI cooperation are insufficient, and this is particularly the case in light of the speed at which LLMs like CHAT GPT are being developed and adopted. Indeed, the call for a moratorium on Chat GPT applications and research speaks to growing anxiety many feel that we are currently not adequately place to understand the risks, let alone regulate and mitigate these risks. Putting aside the immediate question of a moratorium, the letter also makes clear the need for more AI governance. The following outlines how the areas where this should happen on an international basis. The following outlines key areas where international cooperation on AI and foundational AI models such as LLMs is needed.

AI Risk assessment and risk management

There is a need to agree on an approach to assessing the risk of foundational AI models such as LLMs that addresses the challenges that LLMs present.[30] As discussed, LLMs such as Chat GPT present a particular set of challenges for regulators give that the developers of the AI model will often not be aware of how the AI is being applied. LLMs also present challenges for traditional risk assessments due to the emergence of new properties and capabilities of LLMs as the training data is scaled.

The EU has been grappling with this set of challenges in the context of the proposed EU AI Act. The challenge specific to the EU Act is how to identify when an AI system becomes high risk and who is responsible.[31] Yet, accountability across the AI supply chain and AI life cycle is clearly needed.[32] The NIST AI Risk Management Framework (RMF) is a voluntary framework for managing AI risk. The AI RMF also points to the challenges of how AI risks may change and emerge at different points along the AI lifecycle. However, the NIST AI RMF leaves it up to developers and those downstream applying the LLMs to work out how to conduct appropriate risk assessments, test and evaluate along the AI lifecycle. In other words, it doesn't directly grapple with this challenge. This includes issues such as the extent that third parties get access to the underlying AI model and the underlying data to assess risk, best practices for documenting AI risk, and determining who is responsible along the AI lifecycle for the AI system. Many of these issues are likely to be worked out in the domestic contexts, such as will take place in the EU with the AI Act. However, international cooperation that facilities learning and dialogue on the issues can help drive consensus on some of the processes that will be needed to enable life-cycle risks assessments, and well as agreement on allocation of liability.

[30] J. Mokander et al, "Auditing Large Language Models: A Three-Layered Approach", 16 Feb 2023
[31] Alex C Engler & Andrea Renda, "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act", CEPS, September 3, 2022
[32] NIST AI RMF

The Australia-Singapore DEA has the most developed commitments on AI when it comes to AI governance.[33] In this DEA the parties agreed to sharing research and industry practice around AI technologies and their governance, to promote responsible use of AI technologies and collaborate in the development and adoption of AI governance frameworks that support trusted, safe and responsible use of AI technologies, taking into account international principles or guidelines on AI governance.[34] The parties to DEPA also agree to endeavor to promote ethical governance frameworks that support the trusted, safe and responsible use of AI technologies and to take into consideration internationally recognized principles, including explainability, transparency, fairness and human-centered values.[35]

The TTC goes a step further than these DEAs, identifying a specific roadmap and next steps for building cooperation on AI governance. The TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management identifies the need to develop shared terminology of key terms such as what is trustworthy, risk, harm, and socio-technical aspects of AI, including what is meant by bias, and robustness. As the Roadmap makes clear, developing a shared understanding of these terms is needed as a building block towards developing a common approach to AI standards, regulation and policies.

Getting agreement on key these terms would be potentially significant in terms of underpinning cooperation on a broad range of AI governance challenges. Such a shared terminology could underpin domestic AI regulation, support a broadly interoperable approach to AI risk assessment and auditing as support the development of international AI standards. Expanding this effort to develop a shared terminology beyond the US and the EU seems a good next step. One approach would be to use IPEF as a bridge to building agreement more broadly amongst like-minded countries.

In addition to shared terminology, more is needed to address the specific challenge raised by foundational models outlined above. As a first step would be to develop a shared understanding amongst government, industry and civil society working on AI governance of the ways that ways developers of LLMs can work with entities that take the LLM and adapt it for implementation. Such a taxonomy would map the relationships and could be a building block to work on how to best better understand the role of the market and contract in addressing issues of safety and liability and where regulation may be needed. This could include issues such as when and how to access the LLM model by third parties, questions around document flows between the LLM model developer and third party implementer and issues of liability and accountability across the value chain of LLMs.

Conformity assessment and auditing of LLMs

International cooperation will also be needed on how to assess conformity and audit foundational models such as LLMs. International cooperation on conformity assessment can reduce the need for multiple conformity assessment procedures, whose costs fall most heavily on small business. LLMs raise specific issues for conformity assessment. Specifically, the ability of entities that are applying the LLM, in order to assess conformity with AI regulation and standards will require access to the AI model and the data on which it is trained.

[33] Australia-Singapore DEA Article 31
[34] Australia-Singapore DEA Article 31
[35] DEPA Article 8.2

The Australia-Singapore DEA and DEPA include AI specific commitments on cooperation relevant for conformity assessment. For example, the Australia-Singapore DEA includes a recognition of the importance of conformity assessment to support digital trade and the Parties endeavor to exchange information to facility conformity assessment to support digital trade.[36]

Auditing practices will also require developing benchmarks against which LLMs can be assessed. When it comes to auditing the application of LLMs for instance, this will include compliance with legal and ethical norms. This includes assessing LLMs performance based on what is meant for example by trustworthy AI, fairness or discrimination, as well on process issue such as appropriate documentation of risks and uses. All of these point to the need for standards, law and regulation.

These challenges from LLMs may require development of a tiered and multilayered approach. This could include governance audits that assess the organization developing the LLM, its organizational procedures, accountability structures and quality management systems. These are process based audits. Audits of the AI model and its data sets, as well as ex post downstream application audits may also be necessary.[37] The NIST AI RMF is a very helpful framework that can guide work in this area. The EU AI Act includes requirements for AI auditing that could become the building block for a broader global approach. Specifically, the Act requires ex-ante conformity assessments with the requirements of the Act for high-risk AI systems. Where AI products are regulated by existing product safety legislation, companies can meet the AI Acts requirements for conformity assessment by using existing third-party conformity assessment arrangements. For those AI systems not covered by existing product safety legislation i.e. the systems that affect fundamental rights – then the entities placing the AI system onto the market will be responsible for conducting their own conformity assessments and documenting the AI system in a new EU-wide high-risk system database. Entities responsible for high-risk AI systems must also meet auditing documentation requirements. And once high-risk AI systems are on the market, there are post-market monitoring obligations.

While these are important steps in developing an auditing practice for AI systems, they are also not well suited for foundational AI systems such as LLMs for the reasons outlined above. Namely they presume that the developer of the AI system is also bringing the AI to market, which is increasingly not the case with foundational AI models. Challenges with getting access to the AI model to assess risk, documentation including of the AI systems performance and data on which it is trained and allocation of responsibility for the AI system along the AI value chain present challenges for conformity assessment and by extension for the auditors that will need access to this documentation, information an understand where liability resides.

So far trade agreements have not specifically sought to grapple with this challenge. Instead, the trade agreements/DEAs that address cooperation on technical or technology standards either explicitly or by implication draw on the WTO TBT Agreements commitments on use of technical standards and conformity assessment. As a first step, parties should seek to include new commitments in trade agreements that get at the new challenges presented by LLMs. At this stage, the best way forward is

[36] Australia-Singapore DEA Article 30.5

[37] J. Mokander et al, "Auditing Large Language Models: A Three-Layered Approach", 16 Feb 2023

likely not specific commitments on how this should be done but commitments to dialogue, sharing experiences including how different legal systems allocate liability.

Cooperation on international AI standards

The EU Act is perhaps first amongst equals in terms of progress on AI regulation. Yet, much of the technical specification animating the Act's regulation will come from standards developed by European Standards Organizations. The U.S. Blueprint for an AI Bill of Rights is another important contribution.[38] The Blueprint includes five core principles and associated practices aimed to guide the design, use and deployment of AI systems aimed at protecting human rights and democratic values. There is also already some global convergence on the AI ethical principles that can guide AI developers. However, more is clearly needed, such as how to navigate the trade-offs that will be arise across AI ethical norms.

There is already significant activity underway in various regional and global standards development bodies on AI technical and socio-technical standards. The OECD has been at the forefront getting agreement on AI ethical principles.[39] There are also AI standards being developed in International Organization for Standardization (ISO), Institute of Electrical and Electronics Engineers (IEEE), International Telecommunication Union (ITU), National Institute of Standards and Technology (NIST), European Telecommunications Standards Institute (ETSI), Internet Engineering Task Force (IETF), and European Committee for Electrotechnical Standardization (CEN-CENELEC), with specific strands focusing on AI design (e.g. trustworthiness by design); AI impact, conformity, and risk assessments; and risk-management frameworks for AI. For instance, the IEEE has a draft standard for Algorithmic Bias Considerations, a draft standard addressing the record keeping requirements in the EU AI Act and a Standards Model Process for Addressing Ethical Concerns During Systems Design. [40]

There are various efforts underway to building cooperation on international AI standards in trade agreement and other fora. The TTC recognizes the need for leadership. This is a complex area for US-EU cooperation given the role of CEN-CENELEC and ETSI in setting EU standards. The TTC approach so far is to assess the international AI standards being developed in international SDOs, identify standards of interest for cooperation and to promote stakeholder participants in international SDOs.[41] The Quad is also seeking to promote cooperation on international standards, including AI standards. The Quad has launched a Critical and Emerging Technologies Working Group which includes cooperation on technical standards.

The most advanced legal binding commitments on AI standards are in the Australia-Singapore DEA. The commitment are focused on international standards that support digital, including technology standards, which would include AI.[42] This includes a commitment to actively participate in the development of AI standards in regional and international bodies share experience development standards, exchange views on potential future areas to developed and adopt standards, and build

[38] Blueprint for an AI Bill of Rights (whitehouse.gov)
[39] The OECD Artificial Intelligence (AI) Principles - OECD.AI
[40] IEEE P7003, IEEE P7001, IEEE 7000
[41] TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management - December 1, 20222 (nist.gov)
[42] Australia-Singapore DEA Article 30

cooperation with industry on research projects that can increase understanding of the AI standards needed, including with respect to LLMs.

Going forward, more is needed. Trade agreements and other international for a such as TTC, the Quad or IPEF are not the place to negotiate or develop new international standards. However, these agreements and for a can be used to support the outcomes of international standards development bodies. This can include referencing AI standards and ethical principles in trade agreements. Governments need to identify how they can support the standards making process consistent with its expert driven and industry-led nature. This can include more cooperation on the R&D that can support standards development process. It may be helpful to support access to the standards of SDOs, which currently need to be paid for. This would increase awareness and understanding of AI standards. More could also be done to expand opportunities for small business and officials from developing countries to participating in the international AI standards development process.[43]

Data governance

Data governance is a key issue for AI generally and specifically for LLMs given the role of big data in developing these models. Data governance has been addressed in trade agreements and DEA from a couple of perspectives so far. One of these has been the commitment to cross-border data flows and to avoiding data localization. As outlined in the table above, a number of countries have undertaken this commitment in various FTAs and DEA. [44]

Trade agreements and DEA also increasingly include a commitment to protecting the privacy of personal data.[45] There are however, very limited commitments in trade agreements so far on some of the data governance issues specific to AI and to foundational AI models that get at how data used to train the AI models can cause discrimination, lead to unfair outcomes, misinformation and so forth. As outlined, DEPA and the Australia-Singapore DEA to include commitments to sharing information and cooperation on AI Governance Frameworks.

Data governance for AI is being taken up in international standards bodies, it is part of the AI RMF and there are data governance requirements in the EU AI Act. All of these developments suggest a role for data governance in trade agreement and DEA could begin with sharing local development and experiences developing metrics and regulation for data governance. This could include in areas such as documentation, the institutional set-up to incentivize appropriate data governance, and methods and experiencer opening data and algorithms to scrutiny.[46] More robust commitments in trade agreements and DEAs on how to use and reflect international standards will also help international cooperation on data governance for AI.

Access to AI compute

[43] Joshua P. Meltzer, A Critical Technology Standards Metric: Assessing the development of critical technology standards in the Asia-Pacific, Brookings Report, September 2022 CTSM-Report-Sep-2022_Final.pdf (brookings.edu)
[44] CPTPP Article 14.11, 14.14, RCEP Articles 12.14, 12.15
[45] CPTPP Article 14.8, USMCA Article 19.8, DEPA Article 4.2
[46] Marijn Jenssen et al, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly, Vol 37, Issue 3, July 2020

AI compute covers the hardware and software that supports AI workloads and applications.[47] Access to AI compute is critical if countries are to develop LLMs. Yet the AI compute needed to run foundational AI models keeps growing. By some estimates, the computational capacity required to train modern ML systems has grown by hundreds of thousands of times since 2012.[48] International cooperation I needed to address the costs of LLMs and the implications for inequality within countries and between countries and between the public and private sectors, where the later may be best able to invest in the compute needed for LLMs and other foundational AI models. These developments – the growing cost of AI compute needed to trains LLMS - points to the need for governments to develop plans to build capacity and expand access.[49] In the US the proposed National Artificial Intelligence Research Resource is an important step to expanding access to the data, algorithms, software, exporting and networks needs for AI R&D.[50] Not all government can and should build their own AI compute and what level of AI compute a country should have needs to be developed in the context of an AI policy. Instead, a range of solutions are likely needed, that include partnering with the private sector and other governments as well as globally distributed access to AI compute located in third countries. Forums such as TTC, IPEF and the Quad are well suited to having these discussions.

Restricting access to AI Models

As outlined, growth in foundational AI models may require greater access to the model in order to third parties to assess AI risk of harm, to understand conformity assessment with applicable regulations and principles and for auditing purposes. However, there is also an argument that less not more access to foundational AI models is needed to reduce risk of harm. In March 2023, leading AI thought leaders such as Yoshua Bengio, Stuart Russel, Elon Musk, Steve Wozniak, Yuval Noah Harari and Max Tegmark amongst many others called for an "immediate pause for at least 6 months the training of AI systems more powerful than ChatGPT4", in order to develop the safety protections and auditing processes to ensure that these AI systems are safe beyond doubt.[51] Working back from a complete ban, limited access to the underlying model may be needed to prevent bad actors using the AI model in harmful ways. For instance, OpenAI has on given limited access for paying users and did not publish the full parameters of the model.[52] This underscores that working out how to address the novel risks of foundational models will often come with tradeoffs that will be need to understood and managed. Again, this is another reason for international cooperation – to better understand the costs/benefits of different approaches to addressing the risks from LLMs and try and coalesce around best practices –

[47] OECD.AI Expert Group on AI Compute and Climate
[48] Sevilla, J. et al. (2022), "Compute Trends Across Three Eras of Machine Learning", https://arxiv.org/abs/2202.05924
[49] A Blueprint for Building National Compute Capacity for Artificial Intelligence, OECD Digital Economy Papers, February 2023, No. 350
[50] Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource, National Artificial Intelligence Research Resource Task Force January 2023
[51] Pause Giant AI Experiments: An Open Letter - Future of Life Institute
[52] Will Heaven, "OpenAI's New Language Generator GPT-3 Is Shockingly Good—and Completely Mindless," MIT Technology Review, accessed November 23, 2022, https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learninglanguage-generator-gpt-3-nlp

such as when access to the model is needed to understand the risks and when access should be restricted as the best way to prevent LLMs being used in harmful ways.

Part 5: conclusion

AI will have significant implications for how economies functions, what jobs are done, how societies work, governments functions, and wars are fought. The release by OpenAI of Chat GPT -3 and then version GPT-4 has shed a particular light on LLMs. As outlined, these are an example of foundational AI models that present a new set of challenges and opportunities for how to regulate and govern AI. LLMs are not Artificial general intelligence (as distinct from general purpose AI) - AI systems that are able to reasons across context and exhibit human like intelligence, but LLMS are certainly more general than the narrow AI systems that have been a focus of attention. This expanded capacity to reason across contexts, to develop new capacities as the training data is scaled, and the lack of understanding as to how these LLMS are reaching decisions prevents new challenges and risks and amplifies many of the existing concerns with more narrow forms of AI. The call to pause research on LLMs for 6 months by leading AI thinkers underscores the level of concern with how these LLMs might develop. In almost all cases, international cooperation is needed to ensure that the AI governance that emerges is effective, consistent with and strengthens democratic governance systems and enhances economic and social flourishing. Tis paper outlines a role for building international cooperation through trade agreement and perhaps most promising in the more immediate future, through the various international discussions on AI happening the less formalized grouping such as the US-EU TTC, IPEF and the Quad.