WHITEPAPER
November 2024

WHAT SHOULD BE INTERNATIONALISED IN AI GOVERNANCE?

Authors: Claire Dennis,* Stephen Clare, Rebecca Hawkins, Morgan Simpson, Eva Behrens, Gillian Diebold, Zaheed Kara, Ruofei Wang, Robert Trager, Matthijs Maas, Noam Kolt, Markus Anderljung, Konstantin Pilz, Anka Reuel, Malcolm Murray, Lennart Heim, Marta Ziosi

Please cite as: "Dennis, C. et al. (2024). 'What Should Be Internationalised in AI Governance?' Oxford Martin AI Governance Initiative."

In partnership with

What Should Be Internationalised in AI Governance?

Claire Dennis,[*] [†] [1,2] Stephen Clare,[*] [1] Rebecca Hawkins,[*] [3,4]
Morgan Simpson,[*] [5] Eva Behrens,[*] [6] Gillian Diebold,[*] [7] Zaheed Kara,[*] [1]
Ruofei Wang,[*] [8] Robert Trager,[**] [2] Matthijs Maas,[9] Noam Kolt,[10]
Markus Anderljung,[1] Konstantin Pilz,[11] Anka Reuel,[12,13]
Malcolm Murray,[14] Lennart Heim,[1] Marta Ziosi [2]

[1]Centre for the Governance of AI (GovAI),
[2]Oxford Martin AI Governance Initiative (AIGI),
[3]ML Alignment & Theory Scholars (MATS), [4]Future Impact Group, [5]ERA AI Fellowship,
[6]International Center for Future Generations, [7]Oxford Internet Institute,
[8]University of Oxford, [9]Institute for Law & AI (LawAI),
[10]Hebrew University of Jerusalem, [11]Georgetown University, [12]Stanford University,
[13]The Belfer Center for Science and International Affairs, Harvard Kennedy School,
[14]SaferAI

Abstract

As artificial intelligence (AI) advances, states increasingly recognise the need for international governance to address shared benefits and challenges. However, international cooperation is complex and costly, and not all AI issues require cooperation at the international level. This paper presents a novel framework to identify and prioritise AI governance issues warranting internationalisation. We analyse nine critical policy areas across data, compute, and model governance using four factors which broadly incentivise states to internationalise governance efforts: *cross-border externalities, regulatory arbitrage, uneven governance capacity,* and *interoperability.* We find strong benefits of internationalisation in compute-provider oversight, content provenance, model evaluations, incident monitoring, and risk management protocols. In contrast, the benefits of internationalisation are lower or mixed in data privacy, data provenance, chip distribution, and bias mitigation. These results can guide policymakers and researchers in prioritising international AI governance efforts.

[†]Corresponding author: Claire Dennis claire.dennis@governance.ai.

[*]Denotes core authors who contributed most significantly to the direction and content of the paper. Both core authors and secondary authors are listed in approximately descending order of contribution.

[**]Senior author.

Given its scope, inclusion as an author does not entail endorsement of all aspects of the paper.

Please cite as: Dennis, C. et al. (2024). 'What Should Be Internationalised in AI Governance?' *Oxford Martin AI Governance Initiative.*

Contents

Executive Summary 2

I Introduction 5

II Current Landscape 6

Advanced AI Countries 6

International Initiatives 8

III Conceptual Framework 10

IV What Determines the Need for Internationalisation? 12

Four Factors for Internationalisation 13

Additional Factors Influencing Internationalisation 16

V Internationalising Key Policy Issues in Advanced AI Governance 17

1 Data Governance 18

1.1 Data Privacy 19

1.2 Data Provenance 23

2 Compute Governance 26

2.1 Chip Distribution 27

2.2 Compute Provider Oversight 33

3 Model Governance 38

3.1 Bias Mitigation 38

3.2 Content Provenance 42

3.3 Model Evaluations 46

3.4 Incident Monitoring 50

3.5 Risk Management Protocols 53

VI Conclusion 57

Appendices 58

Executive Summary

As artificial intelligence (AI) advances rapidly, states are increasingly recognising the need for international governance to address shared benefits and challenges. However, not all AI issues require global cooperation, and the complexities of international coordination make it crucial to identify which areas of AI governance should be prioritised for internationalisation.

Determining the Need for Internationalisation

This paper advances the ongoing discourse on international AI governance by offering a structured framework for understanding fundamental considerations around the need for international coordination by policy area. Our findings can guide policymakers and researchers in prioritising international AI governance efforts, help focus global dialogues, and inform the development of international frameworks, standards, and institutions.

We define internationalisation as the development and implementation of coordinated policies, norms, or standards across countries to address shared challenges or opportunities. We analyse nine critical policy areas across three domains of AI governance:

Three Areas for AI Governance

1	**Data Governance**	2	**Compute Governance**	3	**Model Governance**
1.1	Data Privacy	2.1	Chip Distribution	3.1	Model Governance
1.2	Data Provenance	2.2	Compute Provider Oversight	3.2	Content Provenance
				3.3	Model Evaluations
				3.4	Incident Monitoring
				3.5	Risk Management Protocols

Our framework evaluates these areas using four factors drawn from economic and international relations theory that broadly incentivise states to internationalise governance efforts:

1. ***Cross-border externalities*:** the economic, social, or environmental effects of actions or policies in one country that spill over into other countries such that incentives among the countries are not aligned.

2. ***Regulatory arbitrage*:** the movement of commercial activities to jurisdictions with laxer standards, potentially driving an overall lowering of standards as states compete for commercial activity.

3. ***Uneven governance capacity*:** variation between states in the quality of information, expertise, resources, or policy levers needed to effectively govern.

4. ***Interoperability*:** the harmonisation of policies, regulations, or standards across borders.

We evaluate the benefits of internationalising a policy issue by assessing the extent to which each factor applies to that issue. When a factor applies strongly, or several factors apply moderately, we identify that issue as having a strong need for internationalisation. Our analysis considers the potential *benefits* of internationalisation in these areas, but we acknowledge that internationalisation can also have negative *impacts* or *costs*, including on state sovereignty.

Findings

We find strong benefits of internationalisation in compute provider oversight, content provenance, model evaluations, incident monitoring, and risk management protocols. **The benefits of internationalisation are lower or mixed** in data privacy, data provenance, chip distribution, and bias mitigation.

Data Governance: For Data Privacy, we find that states can unilaterally act to mitigate many of the most pressing harms, making the need for internationalisation low. States can similarly mitigate some Data Provenance concerns by regulating the deployment of AI models; however, internationalisation may help prevent regulatory arbitrage, making the need for internationalisation mixed overall.

Compute Governance: We find that the highly concentrated chip supply chain limits the need for internationalisation of Chip Distribution (measures to track and control the physical movement of AI chips), at least from the perspective of countries that are central to the supply chain. The global market for compute providers, though, means that internationalisation of Compute Provider Oversight (monitoring and regulating AI compute usage in data centres) could have significant benefits across all four factors.

Model Governance: For Bias Mitigation, we find that the need for internationalisation is low given the contextual nature of bias and the ability of states to regulate at the point of use. However, for Content Provenance, Model Evaluations, Incident Monitoring, and Risk Management Protocols, we find a strong need for internationalisation. These areas benefit significantly from global cooperation due to the transnational nature of advanced AI model deployment and use, the need for standardised practices (such as in content verification and model evaluations), and the importance of rapid, coordinated responses to AI incidents and risks.

The Need for Internationalisation by Policy Area

	Category	Subcategory	Cross-border Externalities	Regulatory Arbitrage	Uneven Governance	Interoperability	Overall
1	Data Governance	Data Privacy	Mixed	Low	Low	Low	**Low**
		Data Provenance	High	High	Low	Mixed	**Mixed**
2	Compute Governance	Chip Distribution	Mixed	Mixed	High	Low	**Mixed**
		Compute Provider Oversight	High	High	High	High	**High**
3	Model Governance	Bias Mitigation	Low	Low	Mixed	Low	**Low**
		Content Provenance	Mixed	Mixed	High	High	**High**
		Model Evaluations	High	High	Mixed	High	**High**
		Incident Monitoring	High	Mixed	High	High	**High**
		Risk Management Protocols	High	High	High	Mixed	**High**

Table 1: The Need for Internationalisation by Policy Area. *We identify a number of key areas across AI development and deployment and analyse their need for internationalisation across four factors: cross-border externalities, uneven governance capacity, regulatory arbitrage, and interoperability. The darker the colour, the stronger the need for internationalisation in that domain.*

I Introduction

While states increasingly recognise the need to coordinate AI policy efforts to address global challenges, it remains unclear which policy issues should be prioritised in international discussions. This lack of clarity can result in policy-making efforts that lack focus, misallocate resources, or fail to achieve their intended goals. Partly as a result, progress in global AI governance has been insufficient to date (Bengio et al., 2024).

In this paper, we aim to help policymakers prioritise the AI policy issues which would most benefit from *internationalisation*, which we define as the development and implementation of coordinated policies, norms, or standards across countries to address shared challenges or opportunities. We present a novel analytical framework in which each issue is assessed according to four factors. By identifying instances in which all or most factors apply strongly, we find five policy issues for which we predict international cooperation will bring strong benefits: compute provider oversight, content provenance, model evaluations, incident monitoring, and risk management protocols. In contrast, the benefits of internationalisation appear lower for data privacy, data provenance, chip distribution, and bias mitigation.

Prioritisation is important because, while internationalisation can be beneficial, not all AI governance issues can or should be addressed at a global level. Internationalisation can be time-consuming[1] and costly, and produce difficult-to-amend regulation (Norris, 2020; Mearsheimer, 2017). This has been especially true for rapidly evolving technologies (Crootof, 2018). It will also often be more efficient and effective for states to address certain issues[2] independently, adapting policies to their national contexts (Wolff, 2020; Farid Uddin, 2018). Finally, both the breadth and severity of potential effects from AI systems are uncertain (Bommasani, Hudson, et al., 2021). Oversimplified approaches, such as assuming that all aspects of AI governance should be coordinated internationally, would likely fail to navigate the trade-offs that should be weighed when deciding to internationalise an issue.

Throughout this paper we focus on *advanced AI systems*, though our framework may apply more broadly. By “advanced” we mean both general-purpose AI systems that perform well across a range of tasks and narrow systems which outperform humans on specific tasks (Maas, forthcoming). We focus on these systems because the risk that they could cause global-scale harm if not governed appropriately is growing (Bengio et

[1] For example, over 20 years of UN negotiations preceded the adoption of the Paris Agreement on climate change in 2015 (United Nations Framework Convention on Climate Change (UNFCCC), 2023).

[2] These commonly include policies related to taxation, education, the military, electoral systems, and social welfare programs.

al., 2024; Department for Science, Innovation and Technology and AI Safety Institute, 2024). The potential for these systems to generate globally significant benefits has also been identified as a potential motivation for global governance in this domain (Cass-Beggs et al., 2024; Dennis and Manning, 2024).

Previous work to identify priorities for international AI governance has either been limited in scope or lacked a comprehensive analytical framework. Some approaches have focused on specific issues (Sastry et al., 2024; Heim et al., 2024; R. Trager et al., 2023; World Health Organization (WHO), 2021) without considering the broader landscape of AI governance challenges. Others have proposed general principles or guidelines for international AI cooperation (Ho et al., 2023; Cihon, Maas, and Kemp, 2020; Cihon, 2019; Roberts et al., 2024; Veale, Matus, and Gorwa, 2023) but have not applied a structured framework to determine which specific areas should be prioritised.

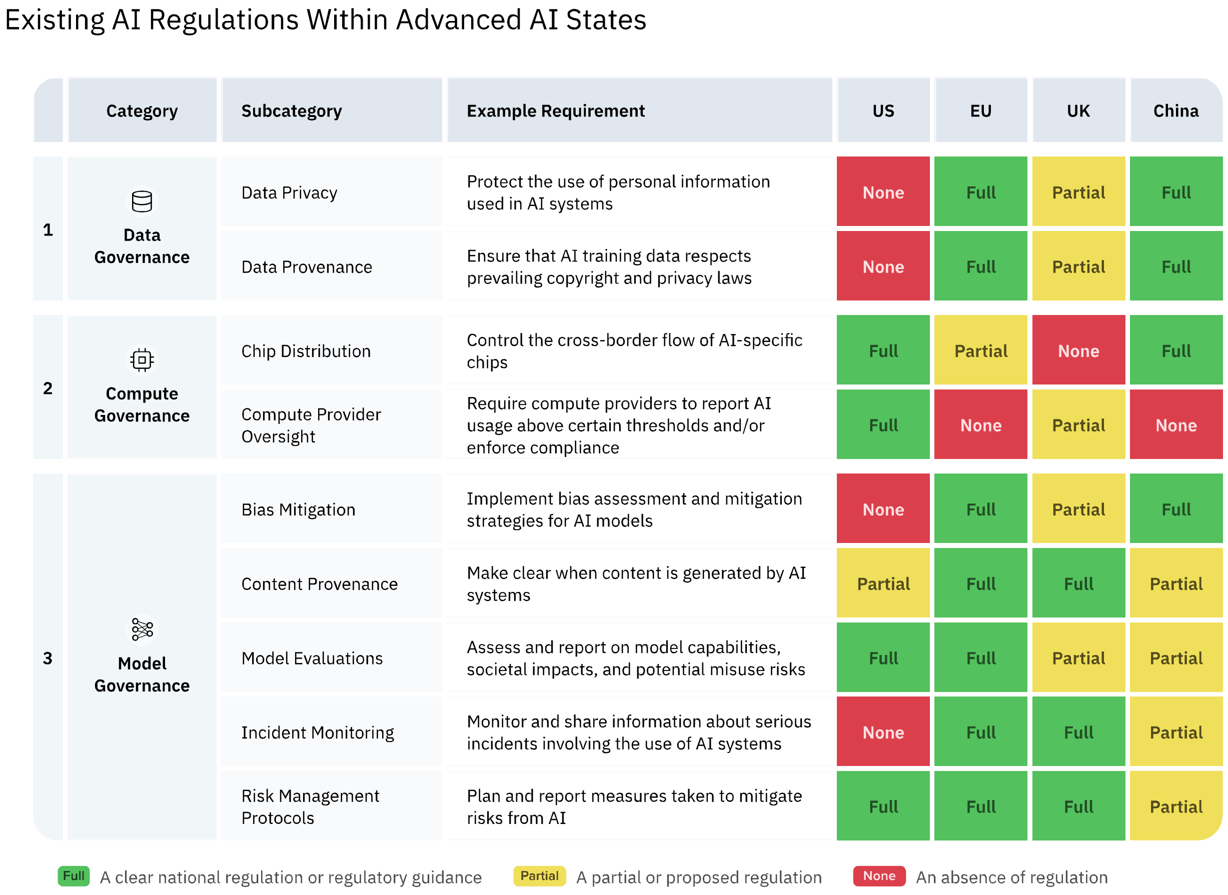
The rest of the paper is structured as follows:

- In Section II, we provide background on our conceptual framework and briefly survey existing international AI governance initiatives.
- In Section III, we introduce our analytical framework, comprising four foundational factors that incentivise states to internationalise: *cross-border externalities, regulatory arbitrage, uneven governance capacity,* and *interoperability*.
- In Section IV, we describe each of these four factors and discuss how they shape the need for internationalisation. We also briefly discuss other factors which may influence the likelihood of internationalisation occurring.
- In Section V, we apply our framework to nine policy issues within three broad domains of AI governance (data governance, compute governance, and model governance) and present our findings.
- In Section VI, we conclude by summarising our key insights and discussing implications for future international AI governance efforts.

II Current Landscape

Advanced AI Countries

Currently, the most advanced AI models are developed in just a small number of countries. As a result, standard-setting efforts for AI development are similarly concentrated. In particular, companies in the US, EU, UK, and China dominate the advanced AI market, meaning that governance efforts in these regions will have an outsized impact on how future AI systems are developed and deployed.[3] Table 2 summarises existing AI regulations in these four jurisdictions across our nine focus areas.



Existing AI Regulations Within Advanced AI States

	Category	Subcategory	Example Requirement	US	EU	UK	China
1	Data Governance	Data Privacy	Protect the use of personal information used in AI systems	None	Full	Partial	Full
		Data Provenance	Ensure that AI training data respects prevailing copyright and privacy laws	None	Full	Partial	Full
2	Compute Governance	Chip Distribution	Control the cross-border flow of AI-specific chips	Full	Partial	None	Full
		Compute Provider Oversight	Require compute providers to report AI usage above certain thresholds and/or enforce compliance	Full	None	Partial	None
3	Model Governance	Bias Mitigation	Implement bias assessment and mitigation strategies for AI models	None	Full	Partial	Full
		Content Provenance	Make clear when content is generated by AI systems	Partial	Full	Full	Partial
		Model Evaluations	Assess and report on model capabilities, societal impacts, and potential misuse risks	Full	Full	Partial	Partial
		Incident Monitoring	Monitor and share information about serious incidents involving the use of AI systems	None	Full	Full	Partial
		Risk Management Protocols	Plan and report measures taken to mitigate risks from AI	Full	Full	Full	Partial

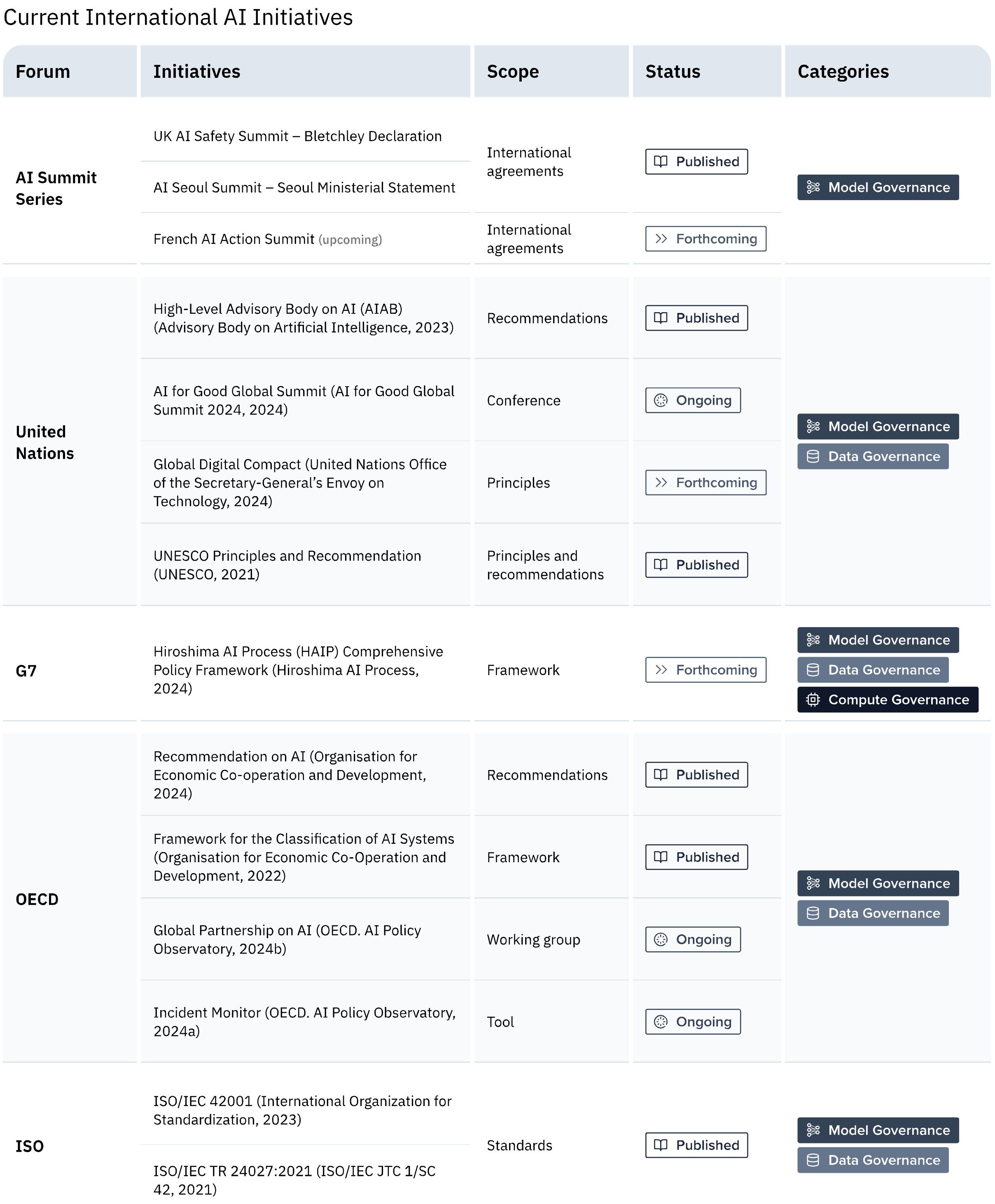
Full A clear national regulation or regulatory guidance Partial A partial or proposed regulation None An absence of regulation

Table 2: Existing AI Regulations within Advanced AI States. Note: This table is not exhaustive and does not represent all existing initiatives; rather, it focuses on the regulations that are most relevant for this analysis.

[3] The United States leads China, the EU, and the UK as the leading source of top AI models. In 2023, 61 notable AI models originated from US-based institutions, far outpacing the European Union's 21 and China's 15 (Maslej et al., 2024).

International Initiatives

Although advanced AI models are developed in just a few countries, they have the potential to generate benefits and pose risks on a global scale. This has motivated multiple efforts to facilitate international coordination in various domains of AI policy (see Table 3). These range from club-based forums, like the Global AI Summits and G7 Hiroshima AI Process, to near-universal efforts led by the United Nations and International Standards Organisation (ISO). Most initiatives at the international level have focused on developing general principles rather than specific commitments and enforcement mechanisms. For example, the Bletchley Declaration, signed by 28 countries during the UK AI Safety Summit, is a non-binding commitment (Department for Science, Innovation and Technology, Foreign, Commonwealth and Development Office, and Prime Minister's Office, 10 Downing Street, 2023) and the OECD's AI Principles, adopted by over 40 countries, are voluntary (OECD.AI Policy Observatory, n.d.). One exception is the Council of Europe's Framework Convention, which does include actionable commitments (albeit with limited scope) (Council of Europe, 2024).



Current International AI Initiatives

Forum	Initiatives	Scope	Status	Categories
AI Summit Series	UK AI Safety Summit – Bletchley Declaration	International agreements	Published	Model Governance
	AI Seoul Summit – Seoul Ministerial Statement			
	French AI Action Summit (upcoming)	International agreements	Forthcoming	
United Nations	High-Level Advisory Body on AI (AIAB) (Advisory Body on Artificial Intelligence, 2023)	Recommendations	Published	Model Governance Data Governance
	AI for Good Global Summit (AI for Good Global Summit 2024, 2024)	Conference	Ongoing	
	Global Digital Compact (United Nations Office of the Secretary-General's Envoy on Technology, 2024)	Principles	Forthcoming	
	UNESCO Principles and Recommendation (UNESCO, 2021)	Principles and recommendations	Published	
G7	Hiroshima AI Process (HAIP) Comprehensive Policy Framework (Hiroshima AI Process, 2024)	Framework	Forthcoming	Model Governance Data Governance Compute Governance
OECD	Recommendation on AI (Organisation for Economic Co-operation and Development, 2024)	Recommendations	Published	Model Governance Data Governance
	Framework for the Classification of AI Systems (Organisation for Economic Co-Operation and Development, 2022)	Framework	Published	
	Global Partnership on AI (OECD. AI Policy Observatory, 2024b)	Working group	Ongoing	
	Incident Monitor (OECD. AI Policy Observatory, 2024a)	Tool	Ongoing	
ISO	ISO/IEC 42001 (International Organization for Standardization, 2023)	Standards	Published	Model Governance Data Governance
	ISO/IEC TR 24027:2021 (ISO/IEC JTC 1/SC 42, 2021)			

Table 3: Current international AI initiatives. *This table highlights the most prominent AI governance efforts led by various organisations and groups, though it is not an exhaustive list.*

III Conceptual Framework

We use a novel framework to assess the need for internationalisation within advanced AI governance. The framework has three main parts.

First, we define internationalisation[4] as the development and implementation of coordinated policies, norms, or standards across countries to address shared challenges or opportunities. For simplicity, we consider internationalisation to be a binary variable, with a policy issue considered *internationalised* after a multilateral[5] regime[6] with incentives for participation has been developed. However, we acknowledge that internationalisation may take various forms, from bilateral agreements[7] to limited-membership models[8] to near-universal regimes,[9] with varying degrees of scope, inclusivity, and enforceability.

Second, we identify four factors which have been previously observed by economists and political theorists as motivating policy action at the international level. These are *cross-border externalities*, *regulatory arbitrage*, *uneven governance capacity*, and *interoperability*. While grounded in international political and economic theory, this specific four-factor approach is a novel contribution of this paper, designed to evaluate key priorities in international AI governance.

[4] "Internationalisation" is a term used in a wide variety of contexts. In business, it refers to the expansion of a company's operations or products into foreign markets (Faulkner, 2006). In economics, it refers to the increased integration of a country's economy with the global economy (Gill, 1992). And in political science, it refers to the growing interdependence between nations through institutions, treaties, and diplomatic relations (Keohane et al., 1996).

[5] By "multilateral", we mean a regime that includes at least three countries. This definition sets a minimum threshold for inclusivity in internationalisation, while still reflecting the need to account for diverse state interests.

[6] Following Krasner's widely accepted definition, an international regime is a set of "principles, norms, rules and decision-making procedures around which actors' expectations converge in a given area of international relations". See Krasner (1983) for more discussion.

[7] Bilateral agreements can include safety regimes such as the nuclear non-proliferation regime established between the US and the USSR, exemplified by the START I (Center for Arms Control and Non-Proliferation, 2022) and New START (United States of America and Russian Federation, 2011) treaties, as well as bilateral environmental cooperation regimes, such as the US-Canada Air Quality Agreement (Government of Canada and the Government of the United States of America, 1991).

[8] Club models can include nuclear non-proliferation regimes, as seen by the multilateral Nuclear Non-Proliferation Treaty, as well as regimes based around economic cooperation, like the G7, G20, World Trade Organization, and International Monetary Fund.

[9] Universal regimes include environmental regimes, formalised by the UN Framework Convention on Climate Change (United Nations Framework Convention on Climate Change, 2024), and human rights regimes, exemplified by the UN Universal Declaration on Human Rights (United Nations, 1948). Both of these regimes have virtually universal ratification.

Third, we apply these four factors to nine distinct advanced AI policy issues. These nine issues, while not exhaustive, were selected to represent major areas of AI governance. They fall into three broad buckets:

- **Data Governance:** how the large datasets on which advanced AI models are trained are built and managed;
- **Compute Governance**: how the computer chips used to train and run AI models are distributed and used; and
- **Model Governance**: how AI models are tested and deployed once trained.

These three categories roughly correspond to the fundamental inputs to AI systems, often referred to as the "AI triad" (Buchanan, 2020): data, compute, and algorithms. However, we focus on models rather than algorithms in order to encompass aspects related to the entire lifecycle of AI systems, from development through deployment.

We then assess the extent to which each of the four factors applies in these areas. If a factor applies to an issue, we qualitatively assess whether it suggests that internationalisation would be beneficial by analysing the incentives, potential cross-border impacts, and existing national and international governance efforts. This assessment considers the current state of AI development, known challenges in international coordination, and insights from relevant policy papers and academic literature on global governance.

Our selection and categorisation of these nine issue areas represent one possible approach among many. Alternative frameworks could have been equally valid and insightful, such as a sectoral approach (as seen in the EU AI Act (European Parliament, 2023)), emphasising the use of AI in particular sectors or industries; a supply chain approach (Veale, Matus, and Gorwa, 2023), examining AI governance from raw materials to software development; or a stakeholder-based approach, organising issues around key stakeholders like developers, users, and governments.

The nine policy issues we examine are not exhaustive. They were selected as particularly prominent and promising areas for internationalisation given their importance for ensuring the responsible development, deployment, and governance of AI systems across borders. Future work could apply our analytical framework more broadly.

Our analysis largely considers the potential *benefits* of internationalisation in these areas, but we also recognise that internationalisation can have *negative impacts* or *costs*, including on state sovereignty, that make it undesirable overall (Keohane, 2019; Gilligan, 2009). Reaching international agreements may also be infeasible in certain areas (Rodrik, 2012; Clark, 2011; Putnam, 1988; Dimitrov, 2020; Hale and Held, 2018). We argue that internationalisation is justified when the benefits from the four factors we analyse outweigh these costs. However, since our analysis largely does not consider costs and feasibility concerns, actors seeking to support internationalisation efforts should complement our findings with analyses that consider the potential barriers to and downsides of coordinating policies at the international level.

IV What Determines the Need for Internationalisation?

Most policy issues are not internationalised, often for good reason. States may have incompatible goals or values. Such agreements may also be perceived as infringing upon national sovereignty. There may be benefits to having a diversity of approaches to a problem, and international competition can sometimes lead to better outcomes.[10] And internationalisation is costly: it needs time, resources, and political effort to succeed (Bernstein, 2012; Abbott and Snidal, 2000). As such, it is often preferable to address problems at the national level.

As AI governance discussions build towards developing international standards, the challenges of internationalisation underscore the importance of focusing on key priorities.. Such efforts should focus on issues for which the benefits of internationalisation are most likely to outweigh the costs. To identify such issues, we consider whether the four characteristics introduced in the previous section – *cross-border externalities, regulatory arbitrage, uneven governance capacity,* and *interoperability* – apply. Previous work on the political economy of international regimes suggests that these factors broadly incentivise cooperation.[11] Other factors or frameworks[12] could have been

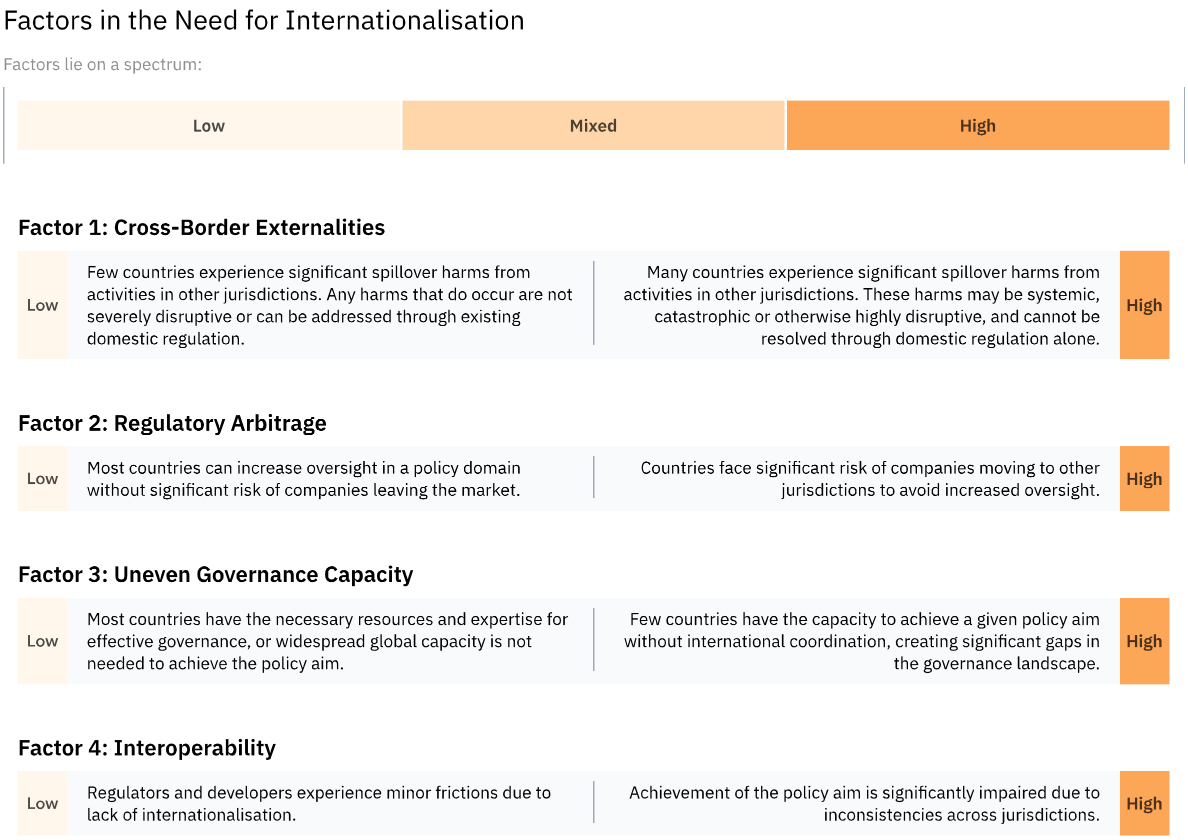
[10] For example, international competition in science and innovation can sometimes lead to more advancements for humanity as a whole, like in telecommunications, vaccine development, or space exploration (McDonald, 2016).

[11] For example, Robert Keohane delineates the concept of demand-driven regimes, highlighting that regimes are the creation of rational, self-interested state-actors seeking to improve their own welfare by reducing transaction costs and improving information asymmetries (see Keohane (1982) for more detail). Similarly, Eden and Hampson (1997) identify four types of structural failures that can contribute to the formation of an international regime: efficiency failures, distributional conflicts, macroeconomic instabilities, and security dilemmas. We build on this foundation by extending the concept to include other factors.

[12] For instance, earlier accounts of coordination versus cooperation problems, as discussed by Stein (1982), provide a broader context for understanding the necessity of international regimes. Additionally, frameworks that classify global public goods into categories such as weakest-link, single-best-effort, or mutual-restraint offer alternative perspectives on the types of cooperation required for different AI issues (see Barrett (2007); and also van Aaken (2016)). Moreover, legal perspectives could assess which issues fall under existing international legal regimes and state obligations (see Creutz (2020a); and also Creutz

applied instead. However, we focus on these elements due to their generalisability and particular relevance to AI issues. Further research to explore additional perspectives and methods in this area could be highly valuable.

Four Factors for Internationalisation



Factors in the Need for Internationalisation

Factors lie on a spectrum:

Low	Mixed	High

Factor 1: Cross-Border Externalities

Low	Few countries experience significant spillover harms from activities in other jurisdictions. Any harms that do occur are not severely disruptive or can be addressed through existing domestic regulation.	Many countries experience significant spillover harms from activities in other jurisdictions. These harms may be systemic, catastrophic or otherwise highly disruptive, and cannot be resolved through domestic regulation alone.	High

Factor 2: Regulatory Arbitrage

Low	Most countries can increase oversight in a policy domain without significant risk of companies leaving the market.	Countries face significant risk of companies moving to other jurisdictions to avoid increased oversight.	High

Factor 3: Uneven Governance Capacity

Low	Most countries have the necessary resources and expertise for effective governance, or widespread global capacity is not needed to achieve the policy aim.	Few countries have the capacity to achieve a given policy aim without international coordination, creating significant gaps in the governance landscape.	High

Factor 4: Interoperability

Low	Regulators and developers experience minor frictions due to lack of internationalisation.	Achievement of the policy aim is significantly impaired due to inconsistencies across jurisdictions.	High

Table 4: Four Factors for Internationalisation. *Drawing from economic and political theory, these factors provide a framework for evaluating when AI governance requires international coordination, rated from low to high importance.*

(2020b)). These perspectives could offer valuable insights for constructing alternative typologies of international cooperation in AI governance.

Factor 1: Cross-Border Externalities

Cross-border externalities occur when activities within a country have positive or negative repercussions that affect other countries such that incentives among the countries are not aligned. Internationalisation aims to address negative effects, in particular by ensuring the responsible country internalises all costs of activities within its borders. A lack of global governance could lead to both under- and over-regulation, potentially leading to an excess of global harms and a deficit of global benefits.

For instance, in the field of cybersecurity, the interconnectedness of digital systems means one actor's decisions affect the security of multiple other actors. Since preventative security measures are costly, tolerating some level of insecurity can be rational for individuals (Asghari, 2016). When the effects on other actors around the world are considered, though, these individual decisions may not be globally optimal (Bauer and Eeten, 2009). Internationalisation of cybersecurity standards can help address these externalities.

Factor 2: Regulatory Arbitrage

When regulations vary among jurisdictions, companies are incentivised to minimise their compliance costs by moving their operations to jurisdictions with less burdensome regulations. They may also choose to exit or decline to enter heavily regulated markets.[13] This may spark a race to the bottom in which all countries weaken regulatory requirements to attract business activity, even when they would each prefer to maintain stricter regulations. Establishing a common baseline of regulations and standards can help states avoid this dynamic.

The harms from regulatory arbitrage can be severe, including human rights abuses, environmental degradation, and increased global risk. In the financial sector, banks have been shown to strategically shift their operations and capital to jurisdictions with more lenient regulations (Houston, Lin, and Ma, 2012). This practice allows financial institutions to minimise the impact of regulatory oversight on their activities, but increases the vulnerability of the global financial system to shocks and crises. Internationalisation of anti-money laundering standards through the Financial Action Task Force, for example, aims to effectively reduce regulatory arbitrage in this domain (U.S. Department of the Treasury, 2024).

[13] Some reviewers noted there are conceptual distinctions between "regulatory arbitrage" (i.e. moving some operations abroad to exploit regulatory loopholes) and "regulatory flight" (i.e. exiting or declining to enter select jurisdictions). For simplicity, we combine both concepts under the umbrella term "regulatory arbitrage" throughout this paper.

Factor 3: Uneven Governance Capacity

Governance capacity refers to a government's ability to effectively manage resources, implement policies, and achieve its stated goals. Currently, this capacity varies greatly around the world. Low governance capacity can enable illegal activity, market failures, and societal harms (Howlett and M. Ramesh, 2016). Global governance can help address these problems by facilitating the provision of resources and technical assistance to governments that lack the information, access, expertise, or resources to implement effective regulatory frameworks.

Governance capacity can be particularly important in highly technical sectors like AI. Governments may need to draw on expertise from elsewhere in the world to evaluate and regulate the sector effectively, given the asymmetry of technical expertise between policymakers and industry or academia. This can lead to a reliance on private industry which is problematic when private and public interests diverge (Picker, 2007). More dynamic sectors like AI may also require more frequent regulatory updates, which can be costly and lead to international regimes that are either outdated upon implementation or insufficiently nuanced. In some cases, creating and overseeing regulatory standards may be better handled by an international technical body (R. Trager et al., 2023).

In the case of nuclear technology, the International Atomic Energy Agency (IAEA) plays a crucial role in addressing the uneven governance capacity among member states by providing technical assistance, training (Langlois, 2013), and guidance to help countries safely and securely regulate nuclear activities (Hahn and Vesterlind, 2013). The IAEA also conducts independent monitoring and verification to ensure compliance with international standards and prevent the misuse of nuclear technologies, which is particularly important for countries with limited domestic regulatory capabilities.

Factor 4: Interoperability

Interoperability is the harmonisation of regulations, standards, and policies across borders. The benefits of interoperability include more fluid information exchange, improved quality of services, and more optimal resource allocation. The harms from lack of interoperability can be relatively minor, such as higher costs for companies operating across borders, or quite severe, such as the failure of critical global systems and safety standards.

A lack of coordination between national governments can result in substantial overlap of regulatory requirements, with multinational firms required to undergo multiple safety evaluations or approval processes. By one estimate, the lack of an internationally recognised aviation safety certification and audit system cost airlines over $3 billion during the 1990s (Sabec, 2004). In collaboration with regulators, the International Air Transport Association, an industry group, improved interoperability by implementing a single comprehensive audit which met or exceeded the safety requirements of each national government (Mills, 2016).

Additional Factors Influencing Internationalisation

The four factors discussed above influence the *demand* for internationalisation. Additional factors, such as power imbalances and geopolitical tensions, influence the *likelihood* of internationalisation. There are significant obstacles to achieving international coordination on AI governance, including national security concerns and competing commercial interests. For instance, AI systems with military applications likely require separate protocols and, in some contexts, limits on the scope of information-sharing.[14] If the information required for the development of effective standards is itself dual use,[15] some countries may seek to restrict its proliferation. Model evaluations for chemical, biological, radiological, and nuclear threats (CBRN), for example, may require sharing sensitive information such as the characteristics of pathogens or weapons.

Additionally, countries with less regulatory capacity may fear that they will not be able to enforce international standards, harming their ability to participate in international markets. These concerns can be mitigated by considering feasibility and inclusiveness when designing the governance frameworks. Many regimes, such as those surrounding the International Maritime Organization (IMO) and the International Civil Aviation Organization (ICAO), include substantial efforts to bolster member state capacities.[16]

Compliance incentives are also a factor. The costs and benefits of non-compliance with laws and regulations will vary across domains. In some cases, domestic laws may have extraterritorial effects even without formal internationalisation, particularly if companies find it easier to comply with a unified set of requirements across multiple

[14] A particularly sensitive consideration is the training of models by national governments explicitly to develop military-based AI models and capabilities. The United States, China, and Russia are all actively pursuing AI technologies for military advantages. A truly effective global reporting regime would detect any of these models if they were large enough or otherwise fit global reporting requirements. Many countries might resist these requirements, aiming to weaken or reduce their strictness. Military use could thus require a separate protocol for disclosures or at the minimum strategic information-sharing among allies as is currently done by organisations like NATO (Morgan et al., 2020).

[15] By "dual use", we mean useful for both civil and military applications.

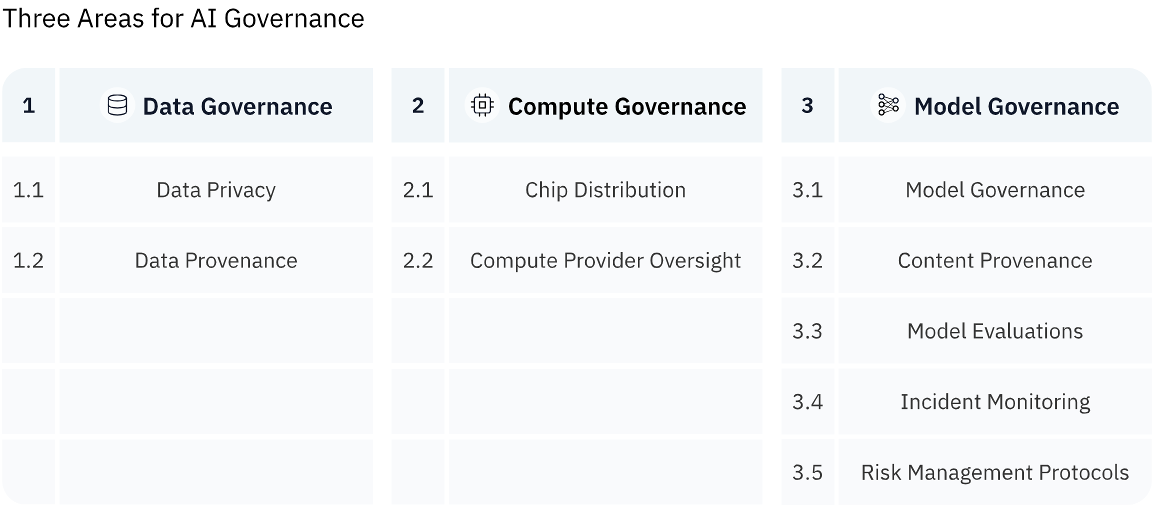
[16] Another factor that we do not discuss, but which may be important in some cases, is *monopoly power*. If some actors are in the position of a monopoly, it may be desirable, from the point of view of global welfare, to cooperate internationally in order to move closer to the competitive market outcome.

countries (Bradford, 2020). However, when regulations are especially burdensome and there are weak incentives to comply, stronger international coordination may become necessary to ensure adherence to regulatory standards.

Below, we largely restrict our focus to the four factors that suggest benefits to internationalisation. We mention elements related to feasibility when they are particularly relevant, but we also recognise that our discussion glosses over much of the complexity that determines whether internationalisation actually happens. Internationalisation may also depend on the specific institutional form of the proposed international regime. Future research efforts that consider these factors would be valuable.

V Internationalising Key Policy Issues in Advanced AI Governance

This section analyses nine critical policy issues across three domains of advanced AI governance: (1) Data Governance, (2) Compute Governance, and (3) Model Governance (Table 5). We assess each issue's governance challenges and evaluate the need for internationalisation according to the four factors described in the previous section.



Three Areas for AI Governance

1	Data Governance	2	Compute Governance	3	Model Governance
1.1	Data Privacy	2.1	Chip Distribution	3.1	Model Governance
1.2	Data Provenance	2.2	Compute Provider Oversight	3.2	Content Provenance
				3.3	Model Evaluations
				3.4	Incident Monitoring
				3.5	Risk Management Protocols

Table 5: Overview of the three domains and nine AI policy issues examined in this report.

Our analysis finds a mixed need for internationalisation overall. We find that there are likely to be major benefits in compute provider oversight, content provenance, model evaluations, incident monitoring, and risk management protocols. In contrast, the benefits of internationalisation seem more limited in data privacy, data provenance, chip distribution, and bias mitigation.

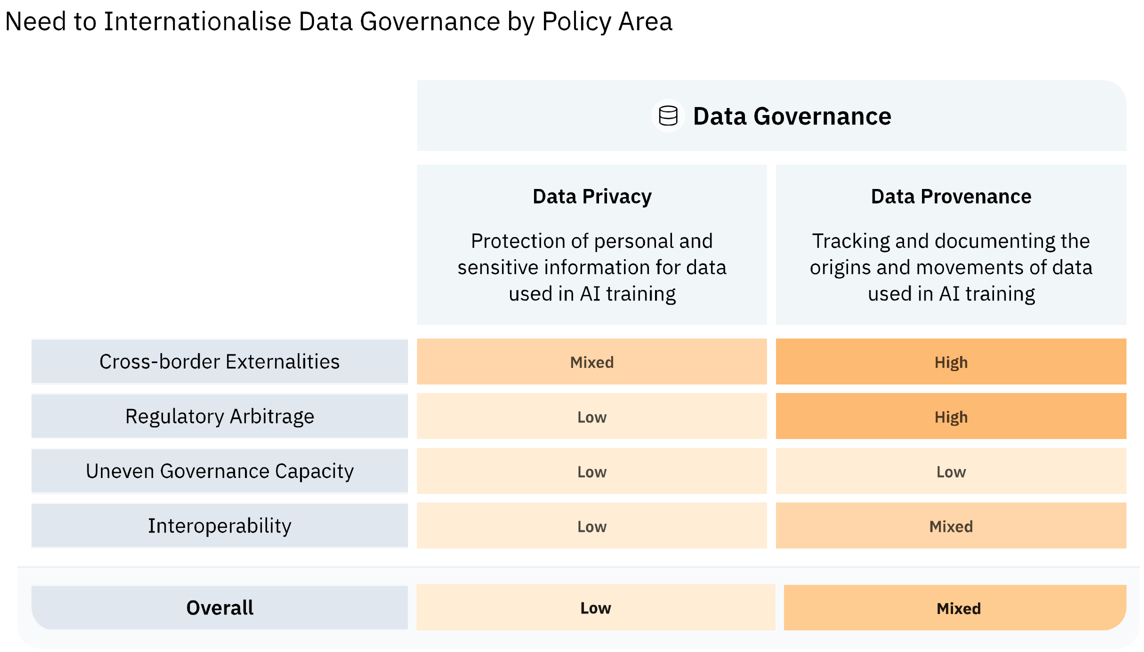
1 Data Governance

Advanced AI systems depend on large datasets (Buchanan, 2020). Recent improvements in AI are due in part to the increased availability and quality of data. For the purposes of this paper, we define data governance as the set of practices, institutions, and rules regarding the management, sourcing, and quality of these data (Janssen et al., 2020).

Data governance is important because the quality and content of the data used to train AI systems strongly influences their behaviour (Mitchell et al., 2022). Almost all the data used to train general-purpose advanced AI models like large language models comes from the internet. As a result, it contains disturbing content like hate speech, sexually-explicit and violent content, and child sexual abuse materials (Department for Science, Innovation and Technology and AI Safety Institute, 2024). In the absence of robust safeguards, this allows models to generate harmful or disturbing content such as deepfake pornography (Kugler and Pace, 2021).

For these reasons, high-quality data are essential for developing safe and trustworthy models. and data governance can help ensure that model training consistently uses such data (Paullada et al., 2021). Data controls and audits support risk mitigation measures, such as removing data that may result in an AI system developing dangerous capabilities (Marion et al., 2023; Birhane et al., 2024; Penedo et al., 2023). They also provide valuable information to regulators and users, including information about the quality, sources, and modifications of the data. Moreover, data governance plays a crucial role in identifying and mitigating biases in training data, which can contribute to fairer outcomes in the use of AI systems (Information Commissioner's Office, 2023a) and protect the privacy rights of individuals (Voss, 2020).

The specific data governance issues we examine are *data privacy* and *data provenance*. We find that the need for internationalisation is *low* for the former and *mixed* for the latter (Table 6).



Need to Internationalise Data Governance by Policy Area

	Data Governance	
	Data Privacy Protection of personal and sensitive information for data used in AI training	**Data Provenance** Tracking and documenting the origins and movements of data used in AI training
Cross-border Externalities	Mixed	High
Regulatory Arbitrage	Low	High
Uneven Governance Capacity	Low	Low
Interoperability	Low	Mixed
Overall	**Low**	**Mixed**

Table 6: ***Key Finding:*** *Data Privacy shows a* **low** *need for internationalisation while Data Provenance is* ***mixed.***

1.1 Data Privacy

Overview

Data privacy refers to an individual's right to confidentiality, anonymity, and personal data protections (Reuel, 2024a). This includes the right to consent to and be informed about how their data are used. It also includes an organisation's responsibility to safeguard such rights when handling personal data.

Demand for data is growing due to the immense scale of it required to train cutting-edge AI systems.[17] The methods used to collect data from different sources, such as licensed agreements, public datasets, privately owned data, synthetic data, and web scraping, have varying levels of privacy protections (Leffer, 2023). High demand also means these methods are often deployed with urgency and at scale. This can lead to the inclusion of personally identifiable information in AI training datasets.

By 2023, 162 countries had enacted data privacy laws, with an additional 20 preparing data privacy legislation (Greenleaf, 2023). The most prominent of these is the EU's

[17] GPT-3 was trained on an estimated 45 terabytes of text data, or the equivalent of 22.5 million books.

General Data Protection Regulation (GDPR), which has had wide-reaching effects on data privacy regulation globally due to its extraterritorial effects (Ryngaert and Taylor, 2020). The GDPR and China's Personal Information Protection Law grant individuals the "right to access" (intersoft consulting services AG, n.d.) their data, requiring companies to disclose personal information in their possession. But applying traditional privacy laws to AI-specific uses of publicly scraped data has proved somewhat challenging for regulators. For example, identifying and isolating personal data among vast amounts of unlabeled data is difficult. For European residents, the EU AI Act has attempted to establish clearer guidelines around the use of scraped data for training AI systems (Tiedrich, 2024). Yet ambiguities remain globally as different legal regimes take divergent stances on what privacy safeguards are required when using publicly available data.

Adding to this complexity is the fact that AI models can memorise and reproduce pieces of their training data, potentially leading to privacy violations and data leakage issues (Ippolito et al., 2023). Demonstrations have shown that large language models are capable of recreating personal information and private code from their training data when prompted, including personal details like phone numbers, email addresses, and names (Carlini et al., 2022). It is also sometimes possible for attackers to elicit and extract training data intentionally, even without prior knowledge of the training dataset (Nasr et al., 2023). Data privacy risks can become security risks[18] given the potential for identity leaks, reconstruction of classified or sensitive data, cyber-attacks, and data breaches in critical sectors.

Some of these concerns can be partly mitigated by implementing privacy-enhancing technologies, like secure multi-party computation, homomorphic encryption, differential privacy, and confidential computing (Soykan et al., 2022). However, such safeguards can be ineffective and costly (ibid.) while also often reducing model performance (Pannekoek and Spigler, 2021). This suggests that technical solutions alone likely cannot resolve data privacy issues. Moreover, the rapid advancement of AI technologies means that new privacy risks and considerations may emerge over time, requiring ongoing research and adaptation of privacy-preserving techniques (Bluemke et al., 2023).

Analysis

While recognising the importance of data privacy issues, after assessing them according to each of the four factors we find that the overall benefits to internationalising data privacy efforts are *low* (Table 6). The need for internationalisation is low across almost

[18] Data security is a distinct topic we do not address fully in this paper. For more specifics on internationalising data security standards, see this Turing Institute report (Powell et al., 2024).

all factors because states have the ability to unilaterally mitigate many data privacy harms. Ensuring interoperability also seems challenging and potentially undesirable given different national priorities and differences among existing legal frameworks for privacy.

Cross-Border Externalities

Domestic data privacy harms from AI systems can be mitigated without international agreements when states can restrict the use of non-compliant AI systems in their jurisdiction. Italy, for example, temporarily banned ChatGPT in 2023 over suspected GDPR violations (Satariano, 2023). This aspect makes data privacy externalities from advanced AI less concerning than certain other externalities, like pollution, whose harms can be more difficult to prevent without international coordination.

However, not all harms can be addressed without some form of cross-border cooperation. These include harms to privacy when individuals enter foreign countries and privacy violations from information published abroad. For example, injunctions granted in the UK to protect the privacy of people subject to court proceedings might be violated by publications outside the UK (Horton, 2021). There are concerns that a model trained on a dataset that includes private data from people or organisations in one country may reveal that data to users in another country (International Scientific Report on the Safety of Advanced AI 2024). Some national and regional data privacy laws aim to protect how their citizens' data are used in other countries, but despite data localisation and similar laws it is often difficult for individual governments to control data flows across borders.

Individuals may also want assurances that their data will not be misused regardless of where an AI model is trained or deployed, but this is currently hard to guarantee. For example, the facial recognition company Clearview AI has used personal data from people in various countries to train AI systems that are deployed elsewhere in the world – a practice that has garnered controversy, but may not be illegal under existing regulatory frameworks (*Clearview AI Inc v The Information Commissioner [2023] UKFTT 819 (GRC)* 2023; Woollacott, 2023).

The potential harms discussed above suggest there may be some benefit to internationalising AI data privacy governance. But as the effects of externalities are attenuated by the ability of states to protect their citizens through national legislation, we assess the need for internationalisation in this particular area as *mixed*.

Regulatory Arbitrage

Regulatory arbitrage to avoid national data privacy laws seems unlikely due to the high costs and limited benefits. AI firms may instead aim to lower costs by deploying one model that meets data privacy requirements in multiple jurisdictions. This means that, in many cases, AI products deployed globally will meet the strictest national requirements, such as the GDPR's data-protection standards (Siegmann and Anderljung, 2022). For example, immediately following Italy's ban on ChatGPT, OpenAI introduced a data control feature for users to disable chat history. While introduced in order to comply with GDPR in Europe, this feature is also available to US-based users (*New ways to manage your data in ChatGPT* 2023).

Privacy concerns associated with training AI systems may be harder to resolve than those associated with deploying them. If hosting AI training brings economic benefits, some states may seek to establish themselves as "data havens" with lax oversight in order to attract AI firms. This becomes more likely as privacy laws become more strict. For example, a total ban on using personal data to train AI models could lead firms to relocate to other countries. Similarly, a ban on scraping the web to gather training data would require comprehensive global buy-in to be effective, since a company affected by such a ban in one country might be incentivised to relocate in order to continue web scraping elsewhere. However, countries could still regulate training-data privacy in these cases at the deployment stage by banning systems trained on sensitive or unregulated private data.

Uneven Governance Capacity

Uneven governance capacity does not seem to be a strong reason to internationalise data privacy laws.

States can, and often do, successfully implement privacy frameworks that are similar or identical to laws that have been effective in other jurisdictions. For example, GDPR-like regulations have been widely adopted. A 2022 study of data privacy legislation across Latin America found that all countries in the sample had implemented concepts from the GDPR legislation, such as applying data-protection law to both data processors and controllers (Carrillo and Jackson, 2022). GDPR similarly serves as the dominant model for developing data privacy law in South Asia (Bentotahewa, Hewage, and Williams, 2022).

These examples suggest that these countries have the governance capacity to adopt privacy laws such as GDPR, but do not provide evidence about their capacity to *enforce* those laws. The latter is unlikely to be resolved at the international level given the focus on privacy law at the domestic level.

Interoperability

Ensuring interoperability of AI data privacy laws across borders could reduce costs for developers and facilitate the secure transfer of data between states, but the potential gains seem modest. There may also be context-specific reasons for data privacy laws to differ between states. Therefore, we assess this factor as providing limited support for internationalisation.

Interoperability could bring some benefits by reducing compliance costs for AI firms (Chander et al., 2021). However, how people understand and prioritise data privacy concerns varies among countries. As a result, the goals and requirements of privacy laws vary as well, often irreconcilably. For example, the EU's GDPR is based on the European Convention of Human Rights, which defines privacy as a human right, while the California Consumer Privacy Act is motivated by consumer protection and China's Personal Information Protection Law prioritises national security interests (Calzada, 2022). Such differences likely preclude global interoperability and in fact make it undesirable.

Interoperability may be achieved among a limited number of countries. For example, the EU categorises some countries, including the UK, as having adequate data privacy laws, meaning EU citizens' private data can flow freely between the UK and the EU (Information Commissioner's Office, 2023b). In general, though, broad internationalisation of data privacy regulations seems undesirable due to variation in national contexts and goals.

1.2 Data Provenance

Overview

Data provenance is a record of information used to track the origins of data. By adopting good data provenance practices, AI firms can address some of the challenges that arise from training AI systems on large datasets, including privacy violations, copyright infringement (Werder, B. Ramesh, and R. Zhang, 2022), and public safety risks.[19] Data provenance can also help reduce the risk of "data poisoning" attacks, in which malicious actors inject false data into training datasets to manipulate the composition and behaviour of the resulting AI models (Brundage et al., 2018).

Currently, a lack of transparency regarding AI training-data provenance makes it hard to verify compliance with data-protection regulations, identify potential biases or ethical issues in AI systems, and hold firms accountable for harm resulting from

[19] https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf

poor data management practices. Developers often neglect to include datasheets that document how training data was collected and used (Gebru et al., 2021; OpenAI et al., 2024; Anil et al., 2023; Touvron et al., 2023). A large-scale audit of AI training datasets conducted by the Data Provenance Initiative, a multidisciplinary group of legal and machine learning experts, found that 70 percent of data licences were ignored or incorrectly documented (Longpre, Mahari, A. Chen, et al., 2023). Similarly, an analysis of the Foundation Model Transparency Index was highly critical of current industry practices, noting that none of the ten AI providers analysed were transparent about data creators, data licence status, copyrighted data, or copyright mitigations (Bommasani, Klyman, et al., 2023).

High costs, as well as increased risk of copyright litigation (*Tremblay v. OpenAI, Inc. (3:23-cv-03223) Document 1* 2023; Metz et al., 2024), may deter firms from investing in data provenance. It is harder to maintain data provenance for larger datasets. That makes the enormous training datasets that power advanced AI systems today particularly challenging, especially because they are often created by combining data from multiple sources.

Existing laws cover some aspects of data provenance, but may be inadequate. For example, it is unclear if record-keeping requirements for training data[20] can be enforced considering the scale and pace of current AI development.[21] The Berne Convention, signed by 181 countries, establishes international standards for copyright protection of creative works, but does not address the use of copyrighted works in AI training (World Intellectual Property Organization (WIPO), n.d.). The question of "fair use" (U.S. Copyright Office, 2023), or the permission to use unlicensed copyright-protected works in certain circumstances, has arisen as the core question in the use of copyrighted training data in AI in the US. Numerous lawsuits alleging copyright infringement by AI developers have been filed (Grynbaum and Mac, 2023). Settling these legal disputes or licensing training data could cost leading AI firms billions of dollars (*The New York Times Company v. Microsoft Corporation et al Document 1* 2023).

To improve the effectiveness of data provenance regulations, some researchers have proposed requiring developers to document data provenance for all training data with, for example, datasheets (Gebru et al., 2021). To ensure compliance, this documen-

[20] Article 10 of the EU AI Act (European Commission, 2020) declares that "training, validation and testing data sets shall be subject to appropriate data governance and management practices." China's national algorithm registry requires disclosure of each open-source and self-built data set used to train the model, as well as the specific source of that data (国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、国家市场监督管理总局, 2022).

[21] The most concrete data provenance blueprint comes from the Data and Trust Alliance (DTA) industry association. However, DTA explicitly acknowledges that data provenance standards would be "challenging to apply effectively to large language models (LLMs) that are trained on vast amounts of public data sourced from diverse locations on the Internet." (Data & Trust Alliance, 2023).

tation may have to be scrutinised or certified by independent third parties. Similar requirements should also apply post-deployment, as supplementary datasets used to fine-tune models can exert a disproportionately large influence on the performance and behaviour of AI systems (Department for Science, Innovation and Technology, 2023a).

In general, enforcing data provenance regulations for AI training and fine-tuning is difficult due to the amount of data used, strong commercial incentives against delaying AI development, and concerns around auditor access. Additionally, there is a lack of technical tools available to comprehensively audit or verify data use in AI systems without requiring developers to share their complete training data.[22] The current willingness of some AI developers to circumvent licensing agreements, even at the risk of legal action, suggests that reaching consistent compliance with data provenance regulations will be a major challenge for the industry.

Analysis

We find that the need to internationalise data provenance is *mixed*. Some factors do suggest that internationalisation would be beneficial (Table 6). Data are generated and collected globally, and harms caused by how training data are used may be addressed by internationalising data management rules. However, internationalisation may not be *critical*, as countries with sufficient regulatory capacity can enforce standards effectively without a global regime. In fact, firms may converge on data provenance standards even in the absence of an international governance regime.

Cross-Border Externalities

The benefits to internationalising data provenance standards could be significant. Data generated and collected in one jurisdiction can easily be used to train models elsewhere. That means that harms from poor data provenance practices, like copyright and privacy violations, can cross borders. Shared standards could ensure model developers maintain appropriate data provenance records no matter where their training data are obtained or processed.

Regulatory Arbitrage

Regulatory arbitrage poses a significant challenge in the realm of data provenance standards. Such standards could substantially raise the costs of acquiring and managing the data needed to train advanced AI systems. Without internationalisation, some firms

[22] For example, Meta has open sourced its model weights for Llama 2 but not the data it used to train the model.

may avoid operating in jurisdictions with particularly stringent national standards. Such behaviour underscores the need for international coordination of data provenance regulations to prevent arbitrage. While the benefits of standardisation might eventually drive convergence across countries, this harmonisation has yet to materialise. The current landscape thus strongly suggests that internationalising data provenance standards could yield significant advantages.

Uneven Governance Capacity

Uneven governance capacity does not currently appear to be a strong factor in favour of internationalisation for data provenance standards. Ensuring appropriate data provenance can be expensive and technically challenging. Standards must be designed with implementation and oversight feasibility in mind. However, once developed, governments can likely verify compliance relatively easily. For example, they may be able to verify compliance by just checking a random sample of data.

That said, there is some uncertainty on this point. First, some countries may benefit from international support in developing reasonable standards and the governance procedures necessary to enforce them. Moreover, verifying compliance through sampling may not be sufficient. If more comprehensive data provenance audits may be needed, and these may require oversight of vast datasets and substantial computational resources. If, in this case, some countries struggles to resource them, and the case for internationalising data provenance would be stronger.

Interoperability

Internationalising data provenance standards may provide some benefits through enhanced interoperability. Universal standards would reduce the total compliance burden for AI developers operating in multiple jurisdictions. Interoperable standards may also make it easier to track data misuse or privacy breaches and ensure accountability for developers. However, if compliance costs turn out to be low anyway, even across multiple jurisdictions, then the benefits of internationalisation may be limited.

2 Compute Governance

Computing power ("compute") and AI chips are central to the development and advancement of AI models, systems, and capabilities (Khan and Mann, 2020). Sastry et al. (2024) find that compute is more amenable to governance measures than are other key inputs to AI models like data and algorithms for three reasons. First, the physical nature of AI chips means they can be more easily quantified, tracked, and restricted (ibid.). Second, the global supply chain for high-end chips is currently heavily concentrated. Ninety percent of the world's most advanced AI chips are manufactured by a single company,[23] and only a few countries[24] together control nearly the entire production process,[25] though this may change in the future (Mullane and Dohmen, 2024). Third, the computing power required to train advanced models demands vast resources, including substantial electricity, land, and water, leaving a clear production footprint.

While these features enable oversight, the physical nature and limited supply of AI chips also create a rivalrous, zero-sum market, making international cooperation more difficult. It is likely that this landscape will evolve in the future as countries invest in domestic chip manufacturing capabilities.[26]

Within compute governance, we consider two policy issues in particular: *chip distribution* and *compute provider oversight*. Chip distribution refers to measures that affect where AI chips are used in the world. Compute provider oversight involves efforts to monitor and verify how those chips are being used to train and run advanced AI systems by inspecting activity inside the large data centres that house them. We find a *mixed* need to internationalise the former and a *high* need to internationalise the latter (Table 7).

2.1 Chip Distribution

Overview

The physical nature of chips and the highly concentrated chip manufacturing industry makes it possible to track and influence where chips are located (Sastry et al., 2024). Chip distribution measures could be used to verify international agreements on non-proliferation of AI chips (Stafford and R. F. Trager, 2022), deny adversarial actors access to chips, or facilitate joint research projects and inclusive access to computing power.

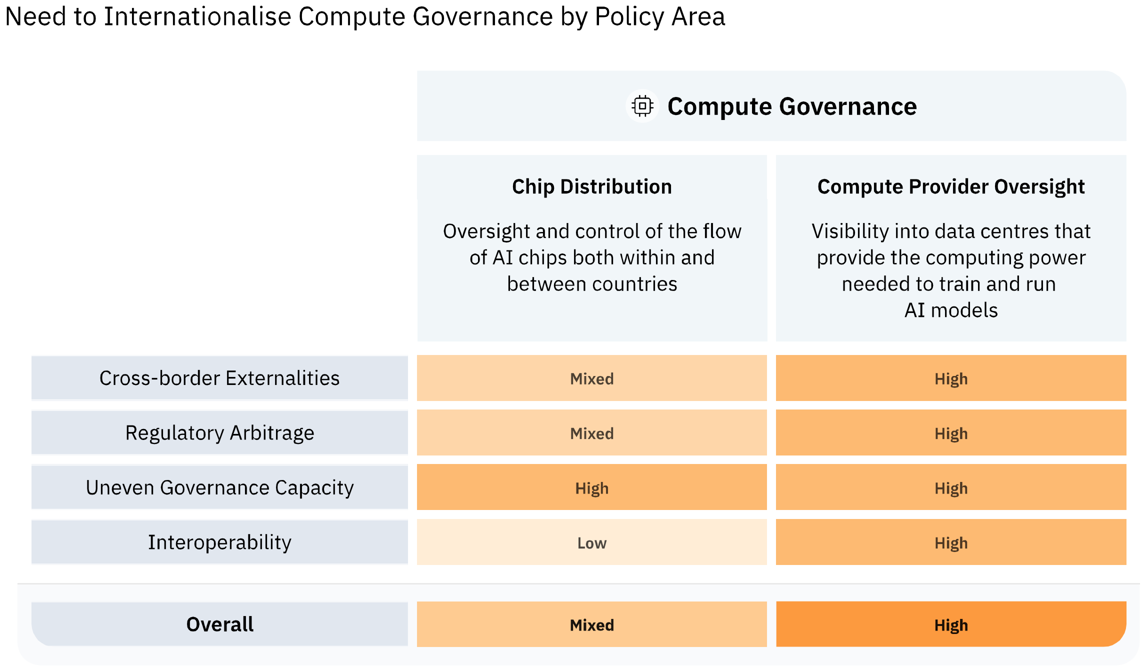
Governance actors have multiple chip distribution measures available. These include restricting or subsidising sales of chips to certain actors, measuring and tracking the movements of chips within and between countries, or licensing large sales of chips.

[23] This company is the Taiwan Semiconductor Manufacturing Company (TSMC).

[24] These include the US, South Korea, Japan, Taiwan, the Netherlands, and China.

[25] The US controls 39% of the total market share across the whole production process of all semiconductors, followed by South Korea, Japan, Taiwan, and Europe with 11–16% share each. China follows with 6% (Khan, Mann, and Peterson, 2021).

[26] For example, both Israel and South Korea are investing heavily in domestic chip-manufacturing capabilities (Scheer, 2023; Reuters, 2024).



Need to Internationalise Compute Governance by Policy Area

	Compute Governance	
	Chip Distribution Oversight and control of the flow of AI chips both within and between countries	**Compute Provider Oversight** Visibility into data centres that provide the computing power needed to train and run AI models
Cross-border Externalities	Mixed	High
Regulatory Arbitrage	Mixed	High
Uneven Governance Capacity	High	High
Interoperability	Low	High
Overall	**Mixed**	**High**

Table 7: ***Key Finding:*** *Chip Distribution shows a* ***mixed*** *need for internationalisation, while Compute Provider Oversight is* ***high****.*

Currently, the most prominent chip distribution policies are unilateral export controls, which restrict sales of certain chips to specific jurisdictions. This is viable because a handful of countries currently control most of the hardware supply chain. However, the long-term viability of this solution is unclear. States face incentives to secure access to AI chips that cannot be removed by unilateral export controls, with countries such as the UAE (Bartenstein, 2024), India (Dunn, 2024), and China (He, 2024) making large investments to increase their domestic chip manufacturing capacity.

As an alternative to restricting chip movement, countries may focus on tracking the production, sales, and resales of AI chips, both within and between countries, to gather precise data on the compute resources held by various actors. For example, an international AI-chip registry (Balwit, 2023) could function as a centralised ledger of advanced AI-chip ownership. This registry could track when chips are resold, destroyed, or lost (Fist and Grunewald, 2023). However, a high degree of cooperation would be needed to prevent smuggling and other regulatory circumventions.

A chip registry could be replaced or complemented by a licensing regime for certain uses of AI chips, such as training or running high-risk AI systems (Anderljung et al., 2023;

Smith, 2023). The benefit of a licensing regime is its anticipatory nature. Licensing could be required at the chip, cluster, or cloud-service level, and could compel actors to obtain permission from a regulatory body, meet specific preconditions, report on their compute usage, and follow common safety rules (Anderljung et al., 2023).

Licensing could have significant costs. Critics argue that it could hinder competition, concentrate power, or slow the growth of vital economic sectors. These effects may be felt disproportionately in certain countries or by certain companies. If some firms struggle to meet licensing requirements, like performing robust pre-deployment safety evaluations, then they may be excluded from the market entirely. Given the economic and geopolitical importance of AI systems, market exclusion would have significant implications for economic sovereignty (Khan and Mann, 2020). It could also slow the spread of beneficial AI innovations in important sectors like healthcare, agriculture, or education. This could be mitigated by only requiring a licence for very expensive or large activities.

As an alternative or complement to restrictions on chip flows, actors are looking to improve the robustness of hardware-enabled mechanisms (Aarne, Fist, and Withers, 2024; Kulp et al., 2024). Such "on-chip mechanisms" involve physical features built directly into specialised AI chips that could monitor, verify, and regulate usage. For example, using "remote attestation," a chip can send trusted information about its usage to a third-party verifier (Palmer, 2020; Birkholz et al., 2023; Aarne, Fist, and Withers, 2024). Similarly, a locator function could alert if a chip was moved across an international boundary, a processing function could send a warning if a chip was incorporated into a larger set of machines, or information could be stored and later transmitted regarding how much compute was used to train a specific AI system (Shavit, 2023).

More intrusive hardware-enabled mechanisms like a remote off-switch, however, are widely regarded as controversial. Such measures should likely be considered only as a fail-safe, absent international agreement to monitor the usage of AI chips. Effective supply chain monitoring would need to accompany these measures to prevent the production or distribution of AI chips lacking or circumventing safety features (Heim, 2024). This could involve tracking the ownership of data centre AI chips and implementing some form of inspections to ensure the chips are not tampered with, where required. Current hardware-enabled mechanisms are also not designed to defend against a sophisticated attacker with physical access to the hardware (Aarne, Fist, and Withers, 2024). Investments to enhance hardware and software security will be required for them to be reliably "tamperproof" in higher-risk scenarios.

Actors may respond to chip distribution measures by increasing smuggling efforts, using non-controlled chips, or accessing cloud compute to bypass restrictions (Fist, Heim, and Schneider, 2023; Grunewald and Aird, 2023). Over time, squeezing one part of the supply chain puts pressure on other parts, incentivizing actors without access to high-end chips to find ways to utilise larger quantities of lower-grade chips. Restricting access to AI chips also creates stronger incentives to build separate supply chains to reduce interdependence and limit the effectiveness of compute governance measures.

Chip distribution measures could also be complicated by advances in compute efficiency that make it possible to train and run highly capable AI systems with fewer chips (Pilz, Heim, and Brown, 2023). To date, though, training larger AI systems continues to demand larger increases in compute despite algorithmic progress, suggesting that, absent major smuggling efforts, chip distribution measures will at least remain effective in the near term (Sevilla et al., 2022).

Analysis

Overall, we find a mixed need for internationalisation in chip distribution (Table 7). Currently, countries controlling key parts of the supply chain can achieve many of their governance goals through unilateral or mini-lateral actions, given the highly concentrated nature of chip production. These countries may prefer to maintain this control and might be resistant to internationalisation efforts that could dilute their influence. In contrast, countries outside this group may have different objectives. While countries can manage chip flows across their own borders, they may view internationalisation as a way to gain more influence over global chip distribution policies.

Internationalisation could offer some shared benefits, such as making it easier to enforce certain rules. For example, international cooperation could help detect and halt chip smuggling. A global registry or licensing regime for AI chips could also help reduce cross-border externalities from some uses of chips.

The effectiveness and desirability of unilateral action will likely be reduced in the future. States want to secure their access to chips for sovereignty and national security reasons. This will be difficult due to the enormous complexity of the chip supply chain,[27] but in the long term it may prove achievable. Internationalisation could be desirable to make chip distribution rules robust against the emergence of parallel supply chains. For that reason, we see some benefit to internationalisation in expectation. These benefits will likely grow over time as multiple states and companies around the world pour more investment into new chip production and distribution capacity.

[27] For reference, see Miller (2022)

Cross-Border Externalities

Cross-border externalities could plausibly motivate internationalisation of chip distribution regulations. AI chips can be used to train or run advanced AI models with general-purpose capabilities. Many of these models will be capable of producing benefits or causing harms that cross borders. However, it may be hard to manage these externalities through model governance measures alone. There are reasons to think that denying rogue states and other threat actors the ability to train, access, and run AI models is more difficult than stopping them from obtaining powerful AI chips is, given the physical, rivalrous nature of the latter (Sastry et al., 2024). Managing the externalities of AI models through chip distribution measures may be cruder but more effective.

Currently, states which control certain chokepoints in the chip supply chain can act unilaterally or collectively to manage cross-border externalities from chip usage. For example, only a small number of allied countries (especially the US, Taiwan, the Netherlands, South Korea, and Japan) would need to coordinate to ensure that all cutting-edge AI chips have hardware-enabled mechanisms built in (Aarne, Fist, and Withers, 2024). However, this may change in the future as more states invest in the semiconductor supply chain. Moreover, internationalisation could make it harder for actors to circumvent unilateral actions by, for example, smuggling chips across borders or otherwise using international markets to evade restrictions. For these reasons, reliably managing chip distribution externalities, at least in the medium-to-long run, will likely require internationalisation.

Regulatory Arbitrage

The need for internationalisation to prevent regulatory arbitrage in chip distribution is mixed. The highly concentrated chip distribution market currently stymies attempts at regulatory arbitrage. There are few other options for companies looking to evade strict requirements on chip distribution set by the leading company.

However, internationalisation may still be desirable as a hedge against supply chain diversification. AI actors may also adapt to actions like export controls in other ways. For example, export controls often target the most advanced chips. If these restrictions prove sufficiently burdensome, targeted AI companies may instead use large numbers of less advanced chips to evade restrictions. It could prove harder to tightly control chips farther from the cutting edge, creating opportunities for arbitrage. Alternatively, AI developers may be willing to make advance commitments to other countries to purchase chips which are currently at a more speculative stage of development. In a more diverse future AI-chip market, international distribution standards could help prevent regulatory arbitrage by getting key producing states onboard (Grunewald and Phenicie, 2023).

AI developers could also seek out jurisdictions where smuggling or regulatory evasion via shell companies is more feasible (Grunewald and Phenicie, 2023). These evasions could be targeted by an expanded regulatory regime, such as new rules on end-user verification in third-party countries. But due to the number of actors involved, these regulatory actions would likely require internationalisation to avoid arbitrage. On the whole, arbitrage provides some motivation for internationalisation. While it may not prove strictly necessary given the concentration of the supply chain, this could change in future, particularly if high regulatory costs incentivise diversification.

Uneven Governance Capacity

Uneven governance capacity provides a strong reason for internationalisation of chip distribution governance efforts. Enforcing chip distribution rules is difficult and expensive. It has been suggested that even regulatory agencies in a high-capacity country like the US, such as the Department of Commerce's Bureau of Industry and Security, need larger budgets to enforce controls on emerging dual-use technologies.[28]

As rules on how and where chips can be distributed internationally grow in complexity and importance, the incentives for evading them will grow, too. A labyrinthine governance system, including different rules for different kinds of chips, could be even more difficult to enforce. Variation in governance capacity, and the necessity of consistent, global enforcement, thus incentivise internationalisation, as it may allow countries to pool resources and assist global enforcement efforts. An effective international AI chip registry, for instance, would need to estimate rates of smuggling and chip resales, including across the borders of countries where less stringent customs protocols or higher corruption rates create opportunities for illicit trade.[29]

Interoperability

With a highly concentrated AI hardware supply chain, there is little current need for interoperability. To enforce chip distribution rules, interoperability may only be needed between regulatory regimes in a handful of countries.

Over time, it may be important for states to develop a combined registry or licensing regime that can apply across multiple supply chains. As the chip production and distribution industries become more diverse, separate, and complex, a unified international chip registry or licensing regime could also plausibly help monitor chip usage across borders and reduce compliance costs for businesses. Interoperable tracking and licensing systems may also make it easier to prevent smuggling and illicit resales

[28] Note that the recently published bipartisan Roadmap for AI Policy in the Senate calls for an increased budget for the Bureau of Industry and Security. See Bipartisan Senate AI Working Group (2024).

[29] There is some evidence of underground markets for small quantities of smuggled AI chips already in China, including NVIDIA H100 GPUs (Fist and Grunewald, 2023).

across customs jurisdictions. This could be particularly important if compute inputs for large training runs become increasingly geographically distributed.

2.2 Compute Provider Oversight

Overview

Training and deploying advanced AI models requires AI compute clusters consisting of thousands of AI chips. These clusters are housed in large, resource-intensive data centres located around the world. The companies which administer AI data centres, known as "compute providers", can serve as an important lever for AI governance by providing visibility into how compute is being used and by whom (Pilz and Heim, 2023).

The compute provider industry is highly concentrated. It is estimated that only about 500 data centres worldwide can host advanced AI compute clusters, with the most important markets found in the US, Europe, and China (ibid.).[30] These large data centres house supporting infrastructure, providing functions such as internet connectivity, power access, security, and cooling (ibid.), and typically need more than 100 megawatts of electricity generation capacity – equivalent to the power required for a medium-sized city. A few large tech companies – Google, Amazon, and Microsoft – provide the majority of AI compute (ibid.).

Most AI development occurs on chips that are rented and accessed remotely "in the cloud", making it challenging for regulators to gain insight merely from the physical location and ownership of AI chips (Heim et al., 2024). That makes compute providers a critical node in governance, as they have access to a wealth of data that could provide visibility into AI usage in their data centres (Egan and Heim, 2023). Yet regulators and the public currently have limited access to these data. It is difficult for them to know how many AI compute clusters exist, which are being used for developing or deploying advanced AI, who is using them, and how much compute is being rented by each customer (Bommasani, Klyman, et al., 2023). Compute-provider oversight could address this issue. It could also play a crucial role in monitoring and mitigating the significant environmental impact of AI development, as these facilities consume vast amounts of energy and resources.[31]

[30] Current estimates likely undercount the number of data centres in China, as there is considerably less public information on Chinese data centres and the number of unreported data centres in China is likely high.

[31] For example, in Ireland, data centres consume about 18% of the country's electricity, leading to a hold on new projects (Ireland Central Statistics Office, 2023).

Several governance measures could leverage the intermediary role of compute providers to help achieve AI governance goals. Compute providers could be required to provide usage information regulators could use to identify users training and deploying advanced AI systems (Heim et al., 2024) enabling risk-based security requirements (ibid.). This information could be used to track how companies are using compute for AI (ibid.). Compute providers could revoke the access of customers who fail to meet regulatory standards. In the rare event of an emergency, compute providers could shut down (or be mandated by regulators to shut down) specific models causing harm by disconnecting them from power grids and the internet (Pilz, 2023).

This shows how compute providers may act as intermediaries in an international reporting regime for highly capable AI model development. Under such a regime, developers could be required to notify their home government or a designated international body before training a highly capable model. Compute providers would only grant access to the required compute resources after receiving proof of compliance from their customers. They could also give regulators and public oversight bodies visibility into where and by whom frontier models are being trained. Access to compute and AI chips could be restricted to countries with regulations that meet international standards.

To enable this regime, compute providers would need to notify relevant authorities after acquiring large numbers of chips. They would also need to periodically submit reports about the use of their AI compute clusters, including for their own internal AI model development.

Implementing monitoring and verification for compute provider oversight measures would be challenging. Many powerful actors currently benefit from unrestricted compute access. Customers may attempt to circumvent these efforts by, for instance, distributing training across multiple AI compute clusters. Although this may reduce compute performance, improvements in communication-efficient training may make decentralised training more feasible in future (Heim et al., 2024).

Regulators could mandate more stringent and difficult-to-circumvent approaches to prevent circumvention, but concerns around privacy would need to be addressed. These approaches could involve reporting requirements that risk infringing on customer confidentiality. Privacy-preserving technologies built into data centre hardware could make it possible for a compute provider, in collaboration with a customer, to verify specific workload properties without accessing sensitive information, sharing only the verification result (Aarne, Fist, and Withers, 2024). This could address some privacy concerns while still enabling effective oversight.

Beyond implementation challenges, compute provider oversight also raises security concerns. The concentration of sensitive information about chip locations and usage in data centres could also create security vulnerabilities. Leaks, external intrusions, or abuse of power within any oversight entity could lead to misuse of this data for corporate espionage, theft of model weights, or attacks on critical infrastructure (Sastry et al., 2024).

Analysis

We find that the need to internationalise the oversight of compute providers is *high,* with all four factors suggesting high benefits to internationalisation (Table 7). Countries have a shared interest in understanding where and by whom AI systems are being developed. Compute providers serve as a critical link in AI governance by providing visibility into compute usage. Since major AI compute clusters can exist in any country that imports advanced AI chips, this oversight requires international coordination. Without global standards, both compute providers and users may engage in regulatory arbitrage by seeking out jurisdictions with less stringent requirements.

Cross-Border Externalities

Cross-border externalities strongly indicate a need to internationalise compute provider oversight. The global nature of AI development and deployment means that the actions of compute providers and their customers can generate far-reaching safety, privacy, and environmental[32] externalities. Additionally, compute providers have access to a variety of information that allows them to act as regulatory intermediaries, helping regulators achieve their goals by providing information or verifying compliance among their customers (Heim et al., 2024). International oversight enables more effective global monitoring and mitigation of AI compute usage and its impacts.

Regulatory Arbitrage

Regulatory arbitrage is a strong reason to internationalise compute provider oversight. If an uncoordinated approach to compute provider oversight is taken, both compute providers and compute users may attempt to seek out jurisdictions with less onerous oversight, primarily to reduce regulatory costs and avoid potential restrictions on compute usage. Users could simply opt to switch to a different compute provider with fewer requirements. AI developers could use distributed or "structured" training to get around oversight by training their models piecemeal across different jurisdictions.[33]

[32] The significant amount of natural resources required to operate data centres, particularly water for cooling, as well as the discharge of wastewater, may strain neighbouring countries' ecosystems. Data centres may also drive demand for energy production that could contribute to global climate change.

[33] Distributed training is already occurring within jurisdictions. Google trained Gemini Ultra across multiple data centres using its intra-cluster and inter-cluster network (Gemini Team et al., 2023).

This could allow companies to avoid safeguards on large-scale training runs, such as reporting requirements above certain compute thresholds. Cloud providers themselves may even seek to gain a competitive advantage by offering to shift their customers' workloads to AI compute clusters in more lenient regulatory environments.

To detect and prevent distributed training across jurisdictions, regulators could implement a global compute-tracking system, requiring compute providers to report on fragmented training operations that, when combined, exceed certain thresholds. Additionally, technical measures could be developed to trace model lineage across multiple training runs.

In fact, actors targeted by US sanctions, including Chinese AI companies (Gardizy, 2024), have already sought to circumvent export controls by remotely accessing prohibited chips through cloud providers and rental arrangements with third parties (Olcott, Sevastopulo, and Liu, 2023). Internationalising compute provider oversight standards could thus become increasingly relevant for national security and geopolitical concerns (*Rep. Jeff Jackson Introduces Bipartisan CLOUD AI Act to Stop China from Remotely Using American Technology to Build AI Tools* 2023).

However, two factors may soften the need for internationalisation, at least in the short term. First, the AI compute clusters capable of training advanced AI systems are concentrated in a few jurisdictions, particularly the US.[34] Second, regulations for compute provider oversight standards can already have strong extraterritorial effects if adopted by countries that host large companies. For instance, reporting requirements recently proposed by the US Bureau of Industry and Security for proposed Infrastructure as a Service providers would require US-based companies to report activities above a certain compute-usage threshold to the US government, regardless of the location of their data centres or the nationality of their customers (The White House, 2023a).

On a longer time frame, companies may attempt to shift or accelerate construction of any new data centres with substantial AI compute infrastructure to more amenable locations. The likelihood of this is currently unclear as data centre construction requires a large amount of electricity, equipment, and skilled labour.[35] Companies may also hesitate to relocate such data centres to areas with more political and economic fragility.

[34] Roughly two-thirds of the global cloud market is operated by just three US-based companies – Google, Amazon, and Microsoft (Pilz and Heim, 2023).

[35] Costs for building new centres have been estimated to be on the order of USD 100M. See Pilz and Heim (ibid.).

Uneven Governance Capacity

Certain countries may lack the regulatory capacity to perform inspections and evaluate compute provider practices. More advanced monitoring techniques of computational activities[36] may require legal compulsion (Heim et al., 2024) and expertise that governance actors in some states may not be able to acquire (ibid.). States looking to attract data centre operators may also not feel they have the negotiating power to demand these requirements. On-chip mechanisms could reduce these issues in part, but reporting thresholds for compute providers may need to be adapted in response to more risky types of training data or to algorithmic and computational efficiency increases. Some countries may not have the necessary expertise to evaluate these thresholds for themselves.

In particular, it may be difficult for regulators to deal with cases in which the compute provider and the AI developer are part of the same company, due to the strong commercial incentives for regulatory evasion. For example, a large tech company, like Google or Microsoft, both provides cloud computing services and develops advanced AI models. In this case, the company could potentially use its own compute resources for AI development without going through the same external processes that other customers would. This creates a potential blind spot for regulators, as the company might have incentives to under-report or obscure its own AI development activities. As a result, there may be a need for additional measures, such as whistleblowing schemes and AI data centre inspections (ibid.).

High-capacity countries may view this lack of oversight as a global risk, resulting in geopolitical tensions as countries with less governance capacity are caught between the economic benefits of data centre operation and diplomatic pressure from other states to meet safety standards. To mitigate this tension, low-capacity countries may seek international support in developing robust regulatory mechanisms for AI data centres. For these reasons, we assess the need for internationalisation to mitigate uneven governance capacity as *high*.

Interoperability

Variance in compute provider oversight regulations between countries can create challenges for international interoperability. These differences can lead to conflicting requirements for companies, security standard discrepancies, and jurisdictional conflicts when data centres process data from multiple countries. For instance, reporting requirements in one country may violate data-protection laws in another.

[36] These include workload classification techniques, which are methods used to analyse and categorise the characteristics and requirements of computational tasks or workloads. These techniques help in understanding the nature of the workload and optimising resource allocation in data centres or computing environments. More advanced techniques include rule-based and machine-learning-based classification.

Countries might also attempt to extend their domestic oversight to compute providers abroad, raising sovereignty concerns. Regulatory inconsistencies across borders complicate compliance for multinational companies and obscure global AI development trends. Harmonising compute provider oversight requirements across jurisdictions could address these challenges by establishing consistent standards and clarifying jurisdictional authority. However, achieving this harmonisation would require significant international cooperation, potentially including new agreements specifically addressing compute provider oversight in AI governance. This means that the need for interoperability indicates a strong demand for internationalisation.

3 Model Governance

Model governance encompasses a range of practices and frameworks aimed at ensuring the responsible development and deployment of advanced AI. Key components of model governance include implementing *bias mitigation* to address harms from skewed data or algorithms; developing *content provenance* tools to distinguish between human-authored and AI-generated content; conducting *model evaluations* to assess an AI model or system's capabilities, limitations, and risks; implementing AI-*incident monitoring* to enable rapid response and prevention of future harms; and establishing robust *risk management* frameworks that span the entire AI lifecycle.

While early governance efforts in these areas are promising, significant challenges remain. Current approaches to bias mitigation (Schwartz et al., 2022), content provenance (Longpre, Mahari, Obeng-Marnu, et al., 2024), and model evaluation (Apollo Research, 2024) lack standardisation and face technical limitations in reliability and scalability. Incident monitoring systems (Turri and Dzombak, 2023) for AI are still nascent compared to other industries, and many risk management practices (The White House, 2023b) remain voluntary and industry-led. Addressing these gaps may involve sustained collaboration between AI developers, governments, and civil society.

Our analysis suggests that overcoming these challenges is particularly pressing, as we find a high need for internationalisation in four of the five policy issues within Model Governance (Table 8).

3.1 Bias Mitigation

Overview

AI bias is the systematic production of results by an AI system that show prejudice or favouritism toward an individual or group based on their inherent or acquired

Need to Internationalise Model Governance by Policy Area

	Model Governance				
	Bias Mitigation Methods to identify and address distortions in AI outputs caused by skewed data, algorithms, or usage	**Content Provenance** Methods to distinguish between AI-generated and human-authored content	**Model Evaluations** Targeted evaluations of a specific system's capabilities, limitations, and potential for harm	**Incident Monitoring** Recording and documenting harms that materialise following deployment	**Risk Management Protocols** Broad frameworks to respond to risks across the lifecycle from development to deployment
Cross-border Externalities	Low	Mixed	High	High	High
Regulatory Arbitrage	Low	Mixed	High	Mixed	High
Uneven Governance Capacity	Mixed	High	Mixed	High	High
Interoperability	Low	High	High	High	Mixed
Overall	**Low**	**High**	**High**	**High**	**High**

Table 8: Key Finding: *Bias Mitigation shows a* ***low*** *need for internationalisation; Content Provenance, Model Evaluations, Incident Monitoring, and Risk Management Protocols all show a* ***high*** *need.*

characteristics such as race, gender, or disability, among others; fairness is a lack of such bias (Mehrabi et al., 2021). There are many different ways of identifying and categorising various biases AI systems may exhibit (Rovatsos, Mittelstadt, and Koene, 2019). Here we do not present an exhaustive overview.[37] Instead, we describe some of the more common sources of bias, focusing on issues particularly applicable to advanced AI systems and the harms biased AI systems can inflict.

Biases are commonly introduced to an AI model through its training data, which may have been affected by historical biases when generated or collected. *Sample bias* can occur when certain groups are over- or underrepresented in training data (Google for Developers, 2022). *Proxy bias* occurs when a variable or feature is used as a substitute (proxy) for a characteristic or outcome of interest, but this proxy is not perfectly correlated with the target and introduces systematic errors or unfairness in the decision-making process (Mehrabi et al., 2021). *Measurement bias* can be introduced if there are systematic errors in how variables in the training data are measured. Bias can also be introduced during training and inference itself. For example, how models are optimised and then compressed can amplify training data biases (Gallegos et al., 2023). Finally, as models are integrated further into society, they may affect

[37] Interested readers may wish to refer to existing surveys, such as (Gallegos et al., 2023; Mehrabi et al., 2021; Ntoutsi et al., 2020).

societal outcomes, which can in turn affect training data and further entrench biases (Bommasani, Hudson, et al., 2021).

Advanced AI, especially large foundation models, present distinctive challenges with regard to bias (Rauh et al., 2022). First, they are trained on large amounts of data from many different sources. Such models are also increasingly likely to be multimodal, with multiple types of data in their training set, including text, audio, image, and video (MIT Technology Review Insights, 2024). Both of these aspects may make it harder to identify biases in the training data.

Second, advanced AI models are deployed in a range of contexts, meaning that harms due to bias can propagate widely (Bommasani, Hudson, et al., 2021). These harms can include the production of harmful content that perpetuates stereotypes or the unfair treatment of certain groups. For example, in healthcare, biased AI models can provide inaccurate results and worsen health disparities by performing differently across patient populations (Tejani et al., 2024).

The harms of AI bias may be mitigated in part by a variety of governance interventions, including dataset monitoring, recourse channels, dataset documentation, technical standards, bias audits, and information-sharing for transparency and accountability (Schwartz et al., 2022). Some of these have already been implemented in the financial services industry. For example, there are common standards on which factors can legitimately be used in a credit score or insurance decision (Consumer Financial Protection Bureau, 2022). Additionally, a range of technical reports and technical standards on AI bias and fairness are being actively developed.[38] Addressing bias in advanced AI systems may be especially difficult due to the size of training datasets, the frequency of updates, the breadth of applications, and the contextual and culturally sensitive nature of bias. Adequate governance may require a multilayered approach that incorporates several of these mechanisms (Mökander et al., 2023).

Analysis

Overall, we find that the need to internationalise bias mitigation standards is *low* (Table 8). While AI models trained on data from one country could lead to biased outcomes when deployed in another, states can largely address these cross-border externalities by setting their own transparency and fairness requirements for models used within their jurisdiction.

[38] These include the ISO/IEC TR 24027:2021 "Bias in AI Systems and AI-Aided Decision-Making" and IEEE draft P007 "Algorithmic Bias Considerations."

Cross-Border Externalities

AI models can perpetuate or amplify biases that exist in the data used to train them. Often, the data used to train AI models deployed in one country originates from many different sources and locations (Bommasani, Hudson, et al., 2021). As a result, an AI model developed in one place could contain biases that lead to unfair or harmful outcomes when the model is then used in other countries or contexts. For example, if an advanced AI model fine-tuned for predicting cancer is trained on data procured from citizens of one country but deployed in another, it may lack adequate local context to recognise patterns and risk factors specific to the demographics, lifestyle, healthcare practices, and genetic predispositions of the latter country's population. This could result in pervasive diagnostic mistakes (Obermeyer et al., 2019).

As with data privacy concerns, however, the need for internationalisation is mitigated because states can act to prevent these harms at the point of use. States can set data transparency or fairness requirements for models deployed within their borders according to national goals. For this reason, externalities are not a strong reason for internationalisation of bias mitigation standards.

Regulatory Arbitrage

The costs of complying with strict regulations on bias may exceed the potential benefits for AI firms, leading them to avoid operating in certain jurisdictions. States may feel pressure to relax standards in order to avoid falling behind in national AI capabilities. However, there can also be valid reasons for bias mitigation standards to vary among states. For example, there may be more concern about AI bias in states with more historical inequality or more diverse populations.

AI developers could also attempt to circumvent bias mitigation regulations by training models in one jurisdiction and deploying them in another. Again, though, states can achieve policy goals by regulating models at the point of use in their jurisdiction, limiting the need for internationalisation.

Uneven Governance Capacity

Mitigating bias within AI systems can be a considerable challenge. Defining and detecting biased outcomes could require statistical knowledge, especially as a model's performance can vary according to different notions of fairness (Long, 2021). Indeed, the task of selecting a universally applicable definition of bias itself is fraught with complexity, given the multifaceted nature of fairness and the diverse contexts in which AI systems operate. Policymakers may also need some understanding of how AI models are developed and deployed to ensure anti-bias policies are both effective and technically feasible (Leavy, O'Sullivan, and Siapera, 2020). For these reasons, international assistance such as common standards, methods, and implementation protocols could be beneficial to the extent that they enable states with less governance capacity to enforce appropriate bias mitigation policies.

However, states may be able to adopt such policies without internationalisation by observing their implementation elsewhere in the world. Moreover, the contextual nature of bias and disagreements over fairness metrics suggest that adaptation at the national or jurisdictional level may be more appropriate.

Interoperability

International standards for AI bias mitigation could lower regulatory burdens, make AI deployment more efficient, and improve the quality of services provided to consumers. However, interoperability may also make it harder to account for the anti-bias governance needs specific to different jurisdictions. Not only do the challenges AI bias presents differ among countries, but the notion of fairness itself lacks a universal definition (Buyl and De Bie, 2024). Thus, even the goals of AI bias mitigation – the challenges to overcome and outcomes to aim for – could differ across borders and even be mutually exclusive.

Since AI bias can inflict serious harm by perpetuating prejudices and inequalities, states should take action to mitigate it. But those actions will need to adapt to national priorities, culture, and history, with interoperability a second-order concern.

3.2 Content Provenance

Overview

Content provenance aims to distinguish AI-generated content from human-authored content (Information Technology Industry Council (ITI), 2024). This has become increasingly difficult as generative AI models have improved to the point where they can now generate highly realistic content.[39]

Content provenance is important because AI has significantly lowered the barrier to create harmful content, including misinformation, deception, harassment, scams, deepfakes, and privacy violations (Reuel, 2024b). At scale, such content could threaten national security and political stability through, for example, foreign influence campaigns during election periods.[40]

[39] Generative AI models include text generators like Claude and ChatGPT, image generators like Stable Diffusion and DALL-E 3, and video generators like Sora.

[40] In the US Department of Homeland Security's (DHS) 2024 threat assessment, foreign AI-generated misinformation is listed as one of the four greatest threats to public safety and security alongside terrorism

Governments and leading AI companies are responding by developing provenance tools and standards. The Biden administration's Executive Order on AI specifically highlighted the importance of identifying AI-generated content (The White House, 2023a), and at the UK AI Safety Summit, major AI companies committed to developing "identifiers of AI-generated material" (AI Safety Summit, 2023). Currently, the most prominent approaches include watermarking (embedding an identifiable pattern in a piece of content), retrieval-based detectors, and post-hoc detectors (Srinivasan, 2024). The Chinese government has established a national standard watermark process required for all AI-derived visual content (Dan and Luo, 2023). The Coalition for Content Provenance and Authenticity, which includes major tech companies like Microsoft and Adobe, has launched a voluntary verification standard for confirming the origin of images and videos. OpenAI's DALL-E 3 and other image generation platforms are also incorporating watermarks into image metadata to support these efforts (David, 2024).

However, none of these tools or approaches are entirely reliable yet.[41] Simple watermarks, such as visible labels or unique sounds, are easily removed or forged. More sophisticated techniques, like machine-learning watermarks, can still be circumvented by motivated individuals, especially for open-source models (H. Zhang et al., 2023). Another approach is to store provenance information in the metadata of a piece of content (Collins, 2024). However, it is easy to copy content in a way that removes this metadata, such as by screenshotting images or re-recording audiovisual content. Retrieval-based detection tries to overcome some of these limitations by storing all content generated by a given model in a database maintained by the model developer (Srinivasan, 2024). Suspicious content can be checked against this database to determine if it was AI-generated. This approach is more streamlined, but raises significant privacy concerns as it requires AI model developers to store all generated content indefinitely (Krishna et al., 2023).

Robust AI-detection tools are difficult to develop for several reasons. AI-generated content exists on a spectrum from entirely synthetic to lightly edited, making it harder to detect partially modified content (Wittenberg et al., 2024). Other tools are likely to face similar implementation difficulties to watermarks. Watermarks work by introducing detectable quirks to model outputs, which developers may be reluctant to agree to as they can worsen output quality (Molenda, Liusie, and Gales, 2024). Each model's watermark also needs a specific corresponding detection tool, so determining whether a piece of content is AI-generated requires tediously querying

and illegal drugs. For example, DHS expects foreign actors including Russia, China, and Iran to develop sophisticated malign influence campaigns online using generative AI in the lead-up to the 2024 election (U.S. Department of Homeland Security, 2023).

[41] AI-generated text is the most difficult to distinguish because text, if coherent and grammatical, may not have clear distinguishing features.

multiple databases (Srinivasan, 2024). Finally, watermarks may not work with open-weight models because they may be easily removed or circumvented (ibid.).

The technical details of content provenance schemes may also need to be kept private to preserve robustness even for closed models, likely requiring a trusted third-party organisation to standardise protocols, maintain a registry of labelled models and detection services, and potentially even operate detection services directly (ibid.). This would require sufficient funding, secure data storage, and buy-in from developers.

Given these challenges, there is growing scepticism that practical detection tools to reliably identify AI-generated content can be developed (Kapoor and Narayanan, 2023). Some researchers argue that we should accept the fact that we may not be able to consistently flag AI-generated content (Knibbs, 2023). Yet there is growing urgency to address these issues as non-watermarked content proliferates and as AI-generated content becomes even more realistic.

Analysis

Overall, we find that the need for international cooperation on content provenance is *high* (Table 8). Harmful AI-generated content can easily cross borders, and many techniques to identify or verify content will not be effective if they are not implemented where that content is created.

Cross-Border Externalities

Content created by AI in one jurisdiction is often viewed or consumed elsewhere. If an inability to identify such content as AI-generated causes harm, then a lack of content provenance standards in one jurisdiction can cause harm across borders. Internationalisation could help mitigate these harms by standardising verification methods such as watermarking, ensuring that they are universally applied and detectable.

Externalities are harder to prevent when malicious actors use AI to generate intentionally harmful, misleading content. For example, AI makes it easier to use deepfakes and other misinformation to influence political processes, public opinion, and social cohesion in other countries (Judson et al., 2024).[42] Internationalising content provenance standards could impede malicious actors by making it harder for them to access AI tools that lack identification or verification mechanisms. However, the current shortcomings of these techniques make it difficult to prevent cross-border harms. Malicious actors can likely circumvent most verification techniques and continue to spread harmful

[42] For example, in April 2023 a Chinese government-controlled news site using a generative AI platform circulated a false claim that the United States was running a lab in Kazakhstan to create biological weapons for use against China.

content. Still, common verification standards likely make this somewhat more difficult and costly, providing some support for internationalisation.

Regulatory Arbitrage

If content provenance proves difficult or costly for AI developers, they may be incentivised to relocate to jurisdictions with more lenient requirements, suggesting a need for internationalisation. Such costs could include developing, implementing, maintaining, and updating verification methods and ensuring compliance. If effective watermarking and verification methods are relatively easy to implement without significantly compromising model performance, these costs may be small and regulatory arbitrage would be unlikely. But if identifying and verifying the growing volume of AI-generated content becomes a major technical and financial burden, generative AI providers may face growing pressure to relocate, making international coordination more urgent.

Uneven Governance Capacity

The difficulty of comprehensively identifying and authenticating all AI-generated content within state jurisdictions suggests a need for an international approach to content provenance. Even if increasingly sophisticated technical tools for content provenance emerge, many countries may lack the technical or governance capacity or political will to use them effectively. Gaps in the global content-authentication system would undermine authenticity in all countries, given that AI-generated content can spread easily across borders. A trusted third-party organisation that could establish content provenance standards, maintain a database of models implementing these standards, maintain potentially sensitive information about content provenance protocols, and engage in international coordination could reduce variability in the effectiveness of content verification.

Interoperability

Ideally, content provenance tools would be globally interoperable to enable the reliable identification of AI-generated content, regardless of its origin. A nationally fragmented patchwork of provenance tools across countries could leave gaps that malicious actors can exploit to disseminate harmful content abroad, while raising barriers to legitimate information-sharing. Without a universal tool or at least a standardised method to track and access all AI-detection tools corresponding to AI models, verifying content could become a costly, inefficient, and tedious process of individually querying numerous detection services without any guarantee of conclusive results. As the number of AI models continues to grow, a public registry of all watermarked models, or a retrieval-based detection scheme, could simplify the process of locating and querying all known AI-detection services simultaneously.

3.3 Model Evaluations

Overview

Evaluations of AI systems,[43] particularly when conducted pre-deployment, are likely to be an important tool for understanding the capabilities of frontier AI models, improving the safety of AI models, and mitigating the chance that unsafe models are released.

General-purpose AI models and systems are assessed in two broad areas:

1. **General capabilities and limitations:** Model evaluations can help assess how design choices regarding, for example, architecture, training data, and hyperparameters, affect model performance. They can also help determine how well the AI system meets performance expectations in both controlled and real-world settings. This aids decision-makers in assessing whether the model is suitable for release.
2. **Societal impact and downstream risks:** Model evaluations can also help assess how models are likely to affect society after deployment. This is a complex and multidisciplinary challenge. Societal risk assessments of models can help reveal the potential to amplify existing biases and inequalities, security vulnerabilities, and unwanted externalities such as labour and environmental impacts, among other concerns.

There are also model evaluations focused specifically on extreme risks (Shevlane et al., 2023), which aim to assess attributes like:

- **Misuse:** These evaluations can uncover capabilities in AI systems that could enable dangerous applications, including persuasion, deception, and weapons development (Phuong et al., 2024; Shevlane et al., 2023; Weidinger, Rauh, et al., 2023; Patwardhan et al., 2024).
- **Alignment:** Other evaluations consider whether models are likely to act in line with human desires, evaluating aspects like rule-following ability (Mu et al., 2023), value alignment (Barez and Torr, 2023; Ngo, Chan, and Mindermann, 2023); (Barez and Torr, 2023) and trustworthiness (Huang et al., 2023).

In an ideal scenario, rigorous evaluations would enable model developers and regulators to comprehensively assess a wide array of risks associated with AI models,

[43] Systems include models, agents, and LLM tools. (Kolt, 2024).

guiding decisions on usage, deployment, and regulatory oversight. For example, evaluations would be ideally used to identify – and thus regulate – model architectures and training processes that are more likely to produce models with dangerous capabilities (Anderljung et al., 2023).

However, despite growing interest in evaluation techniques as a governance tool among governments (Department for Science, Innovation and Technology, 2023b) and frontier AI firms (Anthropic, 2023), technical evaluation practices are still nascent and insufficiently effective (Phuong et al., 2024; Weidinger, Rauh, et al., 2023). They are highly sensitive to factors such as how a prompt is phrased (Liang et al., 2023; Sclar et al., 2023), lack robust testing procedures common in fields like aviation (Chang et al., 2024), and do not provide reliable safety assurances of model capabilities and limitations (Card et al., 2020). To mature into a rigorous scientific discipline that can adequately inform high-stakes AI governance decisions, the field of AI model evaluation requires standardised best practices, techniques to quantify uncertainty and coverage, and collaboration between researchers, industry, policymakers, and other stakeholders (Apollo Research, 2024).

The quality of evaluations depends on levels of access and transparency (Casper et al., 2024). Certain evaluations require access to the model's underlying architecture. This could be challenging to implement and enforce at an international level since companies are increasingly keeping state-of-the-art advanced AI systems private (Bommasani, Klyman, et al., 2023) and access is currently based on voluntary agreements with firms (Henshall, 2024). Securing model access has reportedly already proven difficult for national research institutes due to intellectual property concerns and fears of setting precedents of sharing with other countries.[44] Several studies have advocated for legal frameworks including "safe harbours" (Longpre, Kapoor, et al., 2024) or government-mediated access regimes (Raji et al., 2022) to facilitate independent red-teaming and audit efforts. Techniques for structured access have also been proposed that do not require making the code and weights public (Shevlane, 2022) but still make it possible for independent researchers and auditors to conduct thorough evaluations with full access to the model in a secure, leak-proof environment (Bucknall and R. F. Trager, 2023).

To achieve a higher degree of success, evaluations must become more robust, independently implemented, and systematically integrated throughout all stages of AI development and deployment (Weidinger, Barnhart, et al., 2024). Evaluations should also incorporate multidisciplinary approaches, including insights from sociology, ethics, and technical disciplines (Weidinger, Rauh, et al., 2023). And they should consider the

[44] Neither OpenAI nor Meta has granted access to the UK's AI Safety Institute to do pre-deployment testing on their upcoming models, GPT-5 and Llama-3, despite agreeing to do so at the UK AI Safety Summit (Manancourt, Volpicelli, and Chatterjee, 2024).

social implications of AI systems, such as their potential to amplify existing biases and inequalities, and include diverse stakeholders in the evaluation process to capture a wide range of perspectives and concerns (Solaiman et al., 2023; Maas, 2022; Hagerty and Rubinov, 2019). Ultimately, though, preemptively modelling the effects of AI deployment in society is inherently complex.

In the future, evaluations could form the basis of international standards for AI safety. Currently, though, leading developers, including OpenAI, Google, and Anthropic, use a range of techniques to test their models, including benchmarks, red-teaming, and human uplift testing.[45] The diversity of approaches is beneficial for exploring the vast surface area of risks but complicates efforts to systematically compare models' capabilities and risks (Maslej et al., 2024). Some degree of standardisation could enable more consistent comparisons across models. However, uniform evaluations may not be desirable given that different AI systems have unique architectures, training processes, and intended uses, which may require tailored evaluation methods to comprehensively assess their specific risks. A diversity of evaluation approaches can also help surface a wider range of potential failure modes and blindspots.

Analysis

Overall, we find that the need to internationalise model evaluations is *high*. All advanced models should undergo rigorous, independent evaluations to assess their safety, reliability, and potential for misuse. Given the scarcity of expertise in this domain, internationalisation could help leverage limited existing knowledge and talent.

Cross-Border Externalities

If an unsafe AI model increases the capabilities of non-state actors or national governments such that malicious attacks become easier to perform, then all countries will bear this risk, including those in which the unsafe models in question are not developed or initially deployed. This is because such models can be easily shared or deployed across borders, making it difficult to contain the risks within the borders of

[45] Benchmarks are standardised metrics used to evaluate AI model performance.
Red-teaming involves evaluators testing AI systems before deployment by simulating adversarial attacks to identify vulnerabilities, worst-case behaviours, and potential for unexpected failures or misuse. Unlike fixed benchmarks, red-teaming adapts to each specific system through interactive testing.

Human uplift testing tries to examine how much more competent a human is at accomplishing a potentially harmful task when they have access to a general-purpose AI system versus when they do not. Current evidence is mixed and methodologies are still developing.

See Department for Science, Innovation and Technology and AI Safety Institute (2024) for more details.

the originating country. This is particularly true for open models with publicly released architectures and weights as they can be directly used or modified to support harmful activities. As such, it is in the interest of all countries to ensure that AI models undergo exhaustive evaluations before they are deployed,[46] though further evaluations can also be done throughout the training process[47] as well as post-deployment where needed.[48]

Regulatory Arbitrage

If a national government unilaterally implements costly pre-deployment model evaluation requirements, advanced AI companies may choose to not develop or even deploy their models in that market, or to shift to an open-source business model. This possibility may discourage national governments from introducing effective evaluation requirements so as to maintain the jobs and tax revenue generated by domestic AI firms. Avoiding regulatory arbitrage and implementing effective evaluation requirements, will therefore require synchronous global coordination between many national governments, indicating a high need for internationalisation.

Uneven Governance Capacity

Development of the most advanced AI models is highly concentrated in a few countries at present. Since there is substantial overlap between the experience required to develop advanced AI models and the expertise required to effectively evaluate them (Shevlane et al., 2023), evaluation expertise is similarly concentrated. This provides some reason for internationalisation as a way to share expertise across borders. Internationalisation could also support national efforts to set standards for model safety. That said, one reason to prefer decentralised model evaluations is that national governments also have capacity to monitor and respond to public safety and national security threats related to advanced AI models. This means that, conditional on having the necessary expertise, they may prefer to conduct model evaluations independently.

[46] Notably, current state-of-the-art evaluations are insufficiently comprehensive and secure to guard against all serious vulnerabilities. See Zou et al. (2023).

[47] Developers often fine-tune models to improve safety at the end of the development process, prior to deployment.

[48] Post-deployment evaluations, while possible, are not as ideal as pre-deployment assessments. Independent evaluators who cannot get early access to the models must wait until they are publicly released to conduct their analyses. Since current leading labs do not modify the underlying model post-deployment, these evaluations can only suggest mitigations rather than make intrinsic changes to the model. Therefore, it is crucial for international regulatory standards to enforce comprehensive pre-deployment testing in order to effectively mitigate risks.

Interoperability

A higher degree of uniformity in evaluation requirements may provide some benefits to AI developers by clarifying their international obligations and reducing the number of redundant or overlapping evaluations they are required to perform.[49] Further, to properly evaluate a model, a third party may require access to sensitive proprietary information, depending on the evaluation method. Centralisation could reduce security concerns because fewer actors would need access to this information. Nonetheless, any effort at internationalisation for model evaluations should take into account that i) current evaluation techniques are nascent and insufficiently effective; ii) evaluations must be rigorous and independent, and not overly reliant on the voluntary actions of firms to be effective; and iii) multiple layers of redundant evaluations can make pre-deployment safety assessments more robust.

3.4 Incident Monitoring

Overview

AI-incident monitoring systems are structured procedures designed to track, analyse, and respond to adverse events associated with AI systems. They can help identify harms resulting from AI-system developments and deployments, allowing firms, regulators, and third parties to design and implement responses to prevent future incidents.

AI-incident monitoring systems aim to build a collective memory of failures, as is done in industries like aviation, to prevent repeated mistakes and improve safety (McGregor, 2021). Further, reports detailing near misses, in which a barrier prevented an adverse situation from developing into a serious accident, can help each party identify and strengthen effective safeguards (Johnson, 2003). While incident monitoring systems are a cornerstone of regulation in other safety-critical industries (ibid.), analogous efforts are lacking for AI systems, despite early indications that the deployment of unsafe models can result in substantial personal (Xiang, 2023) and financial (H. Chen and Magramo, 2024) harm. Monitoring incidents related to advanced AI models may be especially important if these systems can unexpectedly develop emergent capabilities,[50] which could have unknown impacts upon model deployment and be difficult to detect during the development process.

[49] Redundant and overlapping audits cost the aviation industry over $3 billion USD during the 1990s (Mills, 2016).

[50] Whether large language models can exhibit unpredictable ("emergent") jumps in capability as they are scaled up is a question of debate currently within the technical community, and discussion is ongoing. For more detail, see J. Wei et al. (2022) and (Schaeffer, Miranda, and Koyejo, 2023).

The design of incident reporting systems can vary substantially. The system can be administered internally by AI firms, by a third party, or by regulators (Turri and Dzombak, 2023); reports can be voluntary or mandatory; reporting responsibilities can be placed on users, employees, companies, third parties, or government organisations; systems can vary in the formality and structure of the reporting mechanism; and reports can be publicly accessible, confidential, or anonymous.[51] Useful post-reporting actions can include verifying incident details, conducting further investigations, disclosing incident details to relevant entities, introducing new harm-avoidance policies or regulations, and pausing or terminating the development or deployment of AI models (K. L. Wei, n.d.).

Incident reporting systems can classify incidents in accordance with different hazards and harms, and model characteristics. Under one classification system, AI failures are categorised in terms of their sector of deployment and the distribution of harms among different demographic groups (Hoffmann and Frase, 2023). Another initiative categorises AI incidents in terms of the architecture of the AI systems involved and a series of specific technical failures (Pittaras and McGregor, 2022). The classification of AI incidents can be further enhanced by adopting a multifaceted approach that considers sector-specific deployments and demographic impacts (Turri and Dzombak, 2023). While different systems of classification may effectively capture the principal AI harms and hazards of concern to regulators and firms, different incident monitoring systems will be more interoperable if they utilise the same method of classification; in turn, this will make the analysis of impacts of deployed AI models substantially easier.

In a well-functioning AI-incident monitoring system, the detection of a harm or hazard would result in an actor filing an incident report, which in turn would be classified, labelled, and stored in a database. Publicly accessible databases allow for greater public scrutiny of AI harms but may discourage reporting – especially from AI firms and their employees, who can stand to lose profits or public trust if their products are considered unsafe. Ensuring confidentiality and offering protections can help mitigate these concerns (McGregor, 2021).

Nonetheless, it is crucial that regulators, and AI firms with relevant models, are provided with all of the information that could help them prevent future harms. In turn, regulators can introduce additional requirements or policies to improve the safety of existing and future models. AI firms could also be required to temporarily halt the development of AI models which pose a high risk of future harm in order to allow for further investigation or preventative measures. Given that many hazards and harms can be detected by users, employees, firms, regulators, and third parties, a nexus of interoperable systems allowing for reporting at each level may be most effective. A

[51] For a comprehensive list of design dimensions for AI-incident monitoring systems, see K. L. Wei (n.d.).

comparable nexus of reporting systems is employed in the aviation industry by the US Federal Aviation Administration (Mills, 2010).

Analysis

Overall, we find that the need to internationalise incident monitoring is *high*. Incident monitoring efforts for advanced AI models would strongly benefit from increased international cooperation as harms from unsafe models are unlikely to be confined to the jurisdiction where the model was developed. A well-coordinated global incident monitoring system could serve as an early alert mechanism to prevent the escalation and spread of AI-related accidents across borders.

Cross-Border Externalities

Harm can result from the release of insufficiently safe AI models by model developers, or through end-user or intermediary developer actions during deployment. The global deployment of advanced AI systems means that incidents caused by unsafe systems in a given jurisdiction are unlikely to be confined only to that jurisdiction. Even in cases where they are constrained, a harmful incident involving an advanced AI model in one jurisdiction could be an early warning sign that the same model may soon cause harm in other jurisdictions – for instance, by highlighting the model's vulnerabilities to jailbreaking (A. Wei, Haghtalab, and Steinhardt, 2023). Incident monitoring systems which focus solely on one national jurisdiction may fail to comprehensively assess the full impacts of a given model because many of those will occur in other jurisdictions. As such, national governments, firms, and third parties would benefit from increased international cooperation on incident monitoring.

Regulatory Arbitrage

The unilateral implementation of incident monitoring systems by smaller countries could deter companies from developing or deploying AI models within those nations' borders as they may fear consumer backlash over exposed vulnerabilities. Further internationalisation, resulting in international adoption of incident monitoring procedures, could mitigate the risk of regulatory arbitrage by facilitating coordination between national regulators. Notably, without formal agreements and international institutions to rule on harms stemming from AI systems, liability from cross-border impacts of advanced AI models will remain a contentious legal and political issue (Fonseca, Vaz de Sequeira, and Barreto Xavier, 2024) – one that the internationalisation of incident monitoring procedures will not wholly resolve.

Uneven Governance Capacity

While many countries remain eager to leverage the economic advantages of advanced AI, they remain vulnerable to impacts from advanced AI systems. Contributing to this vulnerability is a lack of funding and technical expertise to independently monitor and guard against harms to their citizens. Internationalisation of AI-incident monitoring could mitigate this vulnerability if the technical maintenance and analysis functions of incident monitoring can be supported by states or international bodies with more funding and technical capacity.

Interoperability

The deployment of advanced AI models across international borders presents an opportunity to improve incident monitoring and response globally. Many accidents stem from atypical or otherwise rare interactions between AI models and their deployment environment. Quantitative analysis can help to identify common flaws or vulnerabilities across a series of models, but it is substantially more effective when there are many different recorded instances of the relevant harm, reducing the effect of noise. A highly interoperable series of incident reporting systems employing a compatible harm classification system would allow researchers to pool incidents from across different national jurisdictions, allowing for more effective analysis.

At its most impactful, this analysis could both assist national governments in designing regulations to prevent harms and provide AI developers with feedback to patch or withdraw unsafe models. Notably, this would have limited effectiveness for open-source models, as developers would struggle to remove all copies of any key model components (Seger et al., 2023) they have made available, particularly any that users are storing offline. In this case, an internationalised monitoring system could still allow governments to introduce additional safeguards against identified vulnerabilities, either in response to an emergency or through their general regulatory functions.

3.5 Risk Management Protocols

Overview

Risk management is the process of making decisions and taking actions to prevent or respond to identified risks or harms (National Institute of Standards and Technology (NIST), 2020). This is a fundamental aim of AI governance. These actions can involve avoiding the risk, accepting it, sharing it (such as through insurance schemes), or altering its impact or likelihood of occurring (ISO/IEC, 2019). Establishing risk management protocols acknowledges that incidents might happen and emphasises the need for prevention, rapid detection, and appropriate response (Koessler and Schuett, 2023). Risk management should be an iterative process of assessing the risk, selecting mitigation options, planning and implementing them, assessing their effectiveness, and deciding whether the remaining risk is acceptable. If not, further mitigation steps should be taken (ISO, 2018).

AI companies make a few particularly high-stakes development and deployment decisions, such as whether to start a large training run or to deploy a new model (NIST, 2023). When making these decisions, companies accept some level of risk, but so far have not explicitly defined their risk tolerance. The lack of explicitly defined risk metrics reduces transparency around firms' risk management decisions. Risk management of advanced AI systems should include the necessary policies, processes, and know-how to allow governance regimes to implement stricter protocols for higher-risk AI systems (Phuong et al., 2024). Specific risk management practices for advanced AI systems can include rapid response protocols to mitigate harms within particular timeframes, cybersecurity controls to prevent AI models and systems from being stolen, and technical controls embedded in the AI system itself to constrain outputs.

While incident detection and response are crucial, comprehensive AI-risk management should also emphasise preventative measures that reduce the expected risk, as is common in other high-risk industries, like nuclear and aviation (Koessler and Schuett, 2023). However, the AI field faces unique challenges in accurately predicting advanced system behaviours post-deployment.[52] Despite these limitations, regulatory frameworks such as the EU AI Act (Future of Life Institute, 2024) and the US AI Executive Order (The White House, 2023a) are beginning to mandate preventative measures, often using compute thresholds as proxies for potential risk.

One of the main challenges in AI-risk management is the lack of clearly defined risk thresholds (Koessler, Schuett, and Anderljung, 2024; Anderson-Samways et al., 2024), which are important for drawing a clear line between acceptable and unacceptable levels of risk (ISO, 2018). When such thresholds are established in advance, they can prevent companies from making biased decisions – whether overly reckless or overly cautious – under the pressure of economic incentives or competing interests. However, the risk thresholds themselves can introduce bias, as setting thresholds is often a subjective process that depends on individual risk tolerance and other contextual factors. Civil society plays a crucial role in establishing these risk thresholds, given their subjective nature and broad societal implications. Engaging civil society in this process

[52] As general-purpose AI models are scaled up, their capabilities improve overall, but to date this growth has been hard to predict for specific capabilities. For example, some capabilities, like the ability to perform the addition of large numbers with high accuracy, have been documented to appear when models reach a certain scale, sometimes suddenly, without being explicitly programmed into the model (Koessler and Schuett, 2023).

can provide diverse perspectives on acceptable risk levels, enhance transparency and public trust in AI governance, ensure that risk assessments consider a wide range of potential societal impacts, and help balance commercial interests with public safety concerns.

Some AI companies have published frameworks, sometimes referred to as Frontier AI Safety Commitments, which define capability thresholds representing increasing risks and specify actions to take at each level. For example, Anthropic commits to conducting an evaluation of its models for increasing capabilities both a) whenever effective training compute increases by a factor of four, including if this occurs mid-training, and b) every three months to monitor fine-tuning/tooling improvements (Anthropic, 2023). However, such frameworks, while promising, are incomplete as risk management frameworks for several reasons. First, they lack clarity on who is responsible for making key decisions and how those decisions are made. Second, their decision rules can be too blunt, focusing on binary "go/no-go" decisions rather than more nuanced approaches. Third, they typically measure capabilities (i.e. the hazard) rather than directly assessing the risks associated with those capabilities. And fourth, they do not clearly define what level of safety measures is considered sufficient before deploying a model.

Many risk management practices for advanced AI developers are currently voluntary and industry-led (Kolt et al., 2024). These voluntary measures are unlikely to be sufficient in the long run, as evidenced by recent instances of AI developers reportedly backtracking on commitments to provide model access for pre-deployment evaluations (Manancourt, Volpicelli, and Chatterjee, 2024). Indeed, a critical missing component of RSPs is external verification and auditing to ensure that companies are actually adhering to their stated responsible development practices. As the capabilities and potential societal impacts of AI systems develop, it may ultimately be in the interest of public safety to develop legal frameworks that further support and institutionalise robust risk management practices to ensure long-term compliance and accountability.

Analysis

Overall, we find that the need to internationalise risk management is *high*. The severe consequences of potential cross-border AI externalities, risk of regulatory arbitrage, and benefits of interoperability provide a strong rationale for internationalising AI-risk management efforts.

Cross-Border Externalities

Many AI risks, such as an increase in the sophistication and scale of cyberattacks or disinformation campaigns, can be transnational. Even AI failures that initially appear localised could quickly escalate into international catastrophes because of the interconnectedness of digital systems and supply chains (Hendrycks, Mazeika, and Woodside, 2023).

This highlights the critical need for coordinated international risk management protocols to prevent such risks from escalating. The potential for AI accidents to cause widespread harm across borders creates strong incentives for countries to cooperate in developing and implementing risk management measures at an international level.

Regulatory Arbitrage

AI-risk management measures, such as capability restrictions, additional testing requirements, or even halting model training or deployment, could be specified in regulations. In general, the higher the cost of these required risk management actions for the relevant firm, the stronger the incentives for regulatory arbitrage.

Inconsistent AI-risk management requirements across jurisdictions can also create incentives for regulatory arbitrage. For example, if one country mandates that AI companies implement certain cybersecurity controls or capability-based restrictions, but another country has more lenient rules, developers may simply relocate to the less regulated environment.

Key aspects of AI-risk management frameworks may therefore need international coordination, at least among states with advanced AI development, to prevent companies from exploiting governance gaps. Shared standards could cover areas like security protocols for high-risk systems, procedures for rapidly containing AI accidents, and common risk thresholds for implementing stronger controls or determining whether models can be released at all. International efforts should ensure that AI systems which pose similar risk levels across different jurisdictions are subject to consistent risk management protocols.

Uneven Governance Capacity

Effectively mitigating advanced AI risks may require deep technical and governance capabilities that many countries currently lack. This uneven capacity raises risks of critical blindspots if mitigation regimes are too decentralised, as countries could struggle to detect and respond to novel emergent threats from frontier AI systems interacting with their local contexts in unexpected ways.

An international risk management mechanism could help close these gaps by enabling the pooling of knowledge, data, and best practices across borders. For example, a global early warning system for AI incidents could aggregate real-time monitoring data from different countries to enable faster pattern detection and coordinated response. International expert bodies could also provide guidance on adapting global risk management policies to diverse local contexts.

The sensitive nature of some AI-risk domains, like national security, may limit countries' willingness to transparently share some risk management information. For less sensitive applications, however, the benefits of internationalising elements of mitigation capacity-building seem substantial and worth pursuing.

Interoperability

Making AI-risk management methodologies interoperable could yield meaningful efficiency gains. For example, if countries' risk thresholds, escalation protocols, and security controls were aligned, it would enable smoother coordination in responding to cross-border AI incidents. Standardised reporting criteria could also aid meta-analyses to identify global mitigation gaps and best practices.

For AI companies operating across markets, interoperable mitigation practices could streamline their compliance burdens. This could be particularly valuable for smaller firms lacking resources to navigate fragmented requirements, though some flexibility is also needed to allow states leeway for localisation and to continually update the protocol as the technology develops.

However, establishing clear risk thresholds and other mitigation controls can also make it easier for firms to strategically circumvent these measures. For example, if a regulation stipulates that AI models exceeding a certain risk threshold must undergo additional testing or cannot be released, firms might be incentivised to deliberately design their models to fall just below that threshold to avoid the additional regulatory burdens, even if the models still pose significant risks. A diversity of mitigation policies and processes could be beneficial for surfacing additional failure modes and other blindspots. The need for internationalisation due to interoperability is thus *mixed*.

VI Conclusion

This paper contributes to the emerging discourse on international AI governance by offering a structured framework for assessing the benefits of internationalisation by policy issue area. By analysing four key factors – *cross-border externalities, regulatory* *arbitrage, uneven governance capacity,* and *interoperability* – in nine critical AI policy areas, we provide a differentiated analysis of the need for global cooperation on AI.

We find that the need for internationalisation varies significantly across different aspects of AI governance. There is a particularly strong case for international cooperation on model governance, including content provenance, model evaluations, incident monitoring, and risk management, as well as on compute provider oversight. These areas consistently demonstrate high potential for cross-border impacts and benefits from coordinated global approaches.

In contrast, we find that some AI issues, including data privacy, data provenance, chip distribution, and bias mitigation, would benefit less from internationalisation or may be effectively addressed through regulatory action at the national level. However, it is crucial to note that this assessment is not static and could change as the AI landscape evolves.

Future research should build upon this framework to explore specific mechanisms for implementing international cooperation in high-priority areas. This could include examining the specific institutional forms and mechanisms through which such international cooperation can be most effectively achieved, taking into account the complex geopolitical, economic, and technical considerations at play. Additionally, these assessments are based on current understanding and context. As AI capabilities continue to advance, regular reassessment of these priorities will be crucial to ensure that governance efforts remain aligned with the most pressing challenges in the field.

References

Aarne, O., T. Fist, and C. Withers (2024). *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing*. Center for a New American Security. URL: https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf.

Abbott, K. W. and D. Snidal (2000). "Hard and soft law in international governance". In: *International organization* 54.3, pp. 421–456. DOI: 10.1162/002081800551280.

AI Safety Summit (2023). *Policy Updates*. https://www.aisafetysummit.gov.uk/policy-updates/. Accessed: 2024-5-6.

Anderljung, M. et al. (2023). "Frontier AI Regulation: Managing Emerging Risks to Public Safety". arXiv: 2307.03718 [cs.CY].

Anderson-Samways, B. et al. (2024). *Responsible Scaling: Comparing Government Guidance and Company Policy*. Institute for AI Policy and Strategy (IAPS). URL: https://www.iaps.ai/research/responsible-scaling.

Anil, R. et al. (2023). *PaLM 2 Technical Report*. Google. URL: http://arxiv.org/abs/2305.10403.

Anthropic (2023). *Anthropic's Responsible Scaling Policy*. https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed: 2024-5-6.

Apollo Research (2024). *We need a Science of Evals*. https://www.apolloresearch.ai/blog/we-need-a-science-of-evals. Accessed: 2024-9-20.

Asghari, H. (2016). *Economics of cybersecurity*. DOI: 10.4337/9780857939852.00021.

Balwit, A. (2023). "How We Can Regulate AI". In: *Asterisk*. URL: https://asteriskmag.com/issues/03/how-we-can-regulate-ai.

Barez, F. and P. Torr (2023). "Measuring Value Alignment". arXiv: 2312.15241 [cs.AI].

Barrett, S. (2007). *Why cooperate?: The incentive to supply global public goods*. London, England: Oxford University Press. DOI: 10.1093/acprof:oso/9780199211890.001.0001.

Bartenstein, B. (2024). *Abu Dhabi Targets $100 Billion AUM for AI Investment Firm*. https://www.bloomberg.com/news/articles/2024-03-11/abu-dhabi-said-to-target-100-billion-aum-for-ai-investment-firm. Accessed: 2024-9-19.

Bauer, J. M. and M. J. G. van Eeten (2009). "Cybersecurity: Stakeholder incentives, externalities, and policy options". In: *Telecommunications policy* 33.10-11, pp. 706–719. DOI: 10.1016/j.telpol.2009.09.001.

Bengio, Y. et al. (2024). "Managing extreme AI risks amid rapid progress". In: *Science*, eadn0117. DOI: 10.1126/science.adn0117.

Bentotahewa, V., C. Hewage, and J. Williams (2022). "The Normative Power of the GDPR: A Case Study of Data Protection Laws of South Asian Countries". In: *SN Computer Science* 3.3, p. 183. DOI: 10.1007/s42979-022-01079-z.

Bernstein, S. (2012). "Grand compromises in global governance". In: *Government and Opposition* 47.3, pp. 368–394. DOI: 10.1111/j.1477-7053.2012.01367.x.

Bipartisan Senate AI Working Group (2024). *Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States Senate*. United States Senate. URL: https://www.heinrich.senate.gov/imo/media/doc/ai_roadmap_text.pdf.

Birhane, A. et al. (2024). "SoK: AI Auditing: The Broken Bus on the Road to AI Accountability". In: *2nd IEEE Conference on Secure and Trustworthy Machine Learning*. URL: https://openreview.net/forum?id=TmagEd33w3.

Birkholz, H. et al. (2023). *RFC 9334: Remote ATtestation procedureS (RATS) Architecture*. Internet Engineering Task Force. DOI: 10.17487/rfc9334.

Bluemke, E. et al. (2023). "Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases". arXiv: 2303.08956 [cs.AI].

Bommasani, R., D. A. Hudson, et al. (2021). "On the Opportunities and Risks of Foundation Models". arXiv: 2108.07258 [cs.LG].

Bommasani, R., K. Klyman, et al. (2023). *The Foundation Model Transparency Index*. Center for Research on Foundation Models (CRFM) and Institute on Human-Centered Artificial Intelligence (HAI). URL: http://arxiv.org/abs/2310.12941.

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. DOI: 10.1093/oso/9780190088583.001.0001.

Brundage, M. et al. (2018). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". arXiv: 1802.07228 [cs.AI].

Buchanan, B. (2020). *The AI triad and what it means for national security strategy*. Center for Security and Emerging Technology. DOI: 10.51593/20200021.

Bucknall, B. S. and R. F. Trager (2023). *Structured access for third-party research on frontier Ai models: Investigating researchers' model access requirements*. Oxford Martin School, University of Oxford and Center for the Governance of AI. URL: https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf.

Buyl, M. and T. De Bie (2024). "Inherent Limitations of AI Fairness". In: *Communications of the ACM* 67.2, pp. 48–55. DOI: 10.1145/3624700.

Calzada, I. (2022). "Citizens' Data Privacy in China: The State of the Art of the Personal Information Protection Law (PIPL)". In: *Smart Cities* 5.3, pp. 1129–1150. DOI: 10.3390/smartcities5030057.

Card, D. et al. (2020). "With Little Power Comes Great Responsibility". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. Ed. by B. Webber et al. Online: Association for Computational Linguistics, pp. 9263–9274. DOI: 10.18653/v1/2020.emnlp-main.745.

Carlini, N. et al. (2022). "Quantifying Memorization Across Neural Language Models". In: *The 11th International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda. URL: https://openreview.net/forum?id=TatRHT_1cK.

Carrillo, A. J. and M. Jackson (2022). "Follow the Leader? A Comparative Law Study of the EU's General Data Protection Regulation's Impact in Latin America". In: *ICL Journal* 16.2, pp. 177–262. DOI: 10.1515/icl-2021-0037.

Casper, S. et al. (2024). "Black-Box Access is Insufficient for Rigorous AI Audits". arXiv: 2401.14446 [cs.CY].

Cass-Beggs, D. et al. (2024). *Framework Convention on Global AI Challenges*. Centre for International Governance Innovation. URL: https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

Center for Arms Control and Non-Proliferation (2022). *Strategic Arms Reduction Treaty (START I)*. https://armscontrolcenter.org/strategic-arms-reduction-treaty-start-i/. Accessed: 2024-5-23.

Chander, A. et al. (2021). “Achieving privacy: Costs of compliance and enforcement of data protection regulation”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3827228.

Chang, Y. et al. (2024). “A Survey on Evaluation of Large Language Models”. In: *ACM transactions on intelligent systems and technology* 15.3, 39:1–39:45. DOI: 10.1145/3641289.

Chen, H. and K. Magramo (2024). “Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’”. In: *CNN*. URL: https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Center for the Governance of AI Future of Humanity Institute, University of Oxford. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Cihon, P., M. M. Maas, and L. Kemp (2020). “Fragmentation and the future: Investigating architectures for international AI governance”. In: *Global Policy* 11.5, pp. 545–556. DOI: 10.1111/1758-5899.12890.

国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部、国家市场监督管理总局 (Jan. 4, 2022). *Provisions on the Management of Algorithmic Recommendations in Internet Information Services*. Trans. by China Law Translate. URL: https://www.chinalawtranslate.com/algorithms/ (visited on 05/23/2024).

Clark, I. (2011). *Hegemony in International Society*. London, England: Oxford University Press. DOI: 10.1093/acprof:oso/9780199556267.001.0001.

Clearview AI Inc v The Information Commissioner [2023] UKFTT 819 (GRC) (2023). URL: https://www.bailii.org/uk/cases/UKFTT/GRC/2023/819.html.

Collins, B. (2024). “The Ridiculously Easy Way To Remove ChatGPT’s Image Watermarks”. In: *Forbes Magazine*. URL: https://www.forbes.com/sites/barrycollins/2024/02/07/the-ridiculously-easy-way-to-remove-chatgpts-image-watermarks/?sh=7c96236d2dbc.

Consumer Financial Protection Bureau (2022). *CFPB Targets Unfair Discrimination in Consumer Finance*. https://www.consumerfinance.gov/about-us/newsroom/cfpb-targets-unfair-discrimination-in-consumer-finance/. Accessed: 2024-5-6.

Council of Europe (2024). *The Framework Convention on Artificial Intelligence*. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence. Accessed: 2024-9-14.

Creutz, K. (2020a). “Contemporary challenges to state responsibility”. In: *State Responsibility in the International Legal Order*. Cambridge University Press, pp. 15–52. DOI: 10.1017/9781108637367.005.

– (2020b). *State responsibility in the international legal order: A critical appraisal*. Cambridge, England: Cambridge University Press. DOI: 10.1017/9781108637367.

Crootof, R. (2018). “Jurisprudential space junk: Treaties and new technologies”. In: *Resolving Conflicts in the Law*. Brill | Nijhoff, pp. 106–129. DOI: 10.1163/9789004316539_008.

Dan, X. and Y. Luo (2023). *Labeling of AI Generated Content: New Guidelines Released in China*. https://www.insideprivacy.com/artificial-intelligence/lab eling-of-ai-generated-content-new-guidelines-released-in-china/. Accessed: 2024-5-6.

Data & Trust Alliance (2023). *Our Latest Initiative: Data Provenance Standards*. https://dataandtrustalliance.org/our-initiatives/data-provenance-standards. Accessed: 2024-5-6.

David, E. (2024). "OpenAI is adding new watermarks to DALL-E 3". In: *The Verge*. URL: https://www.theverge.com/2024/2/6/24063954/ai-watermarks-dalle3-openai-content-credentials.

Dennis, C. and S. Manning (2024). *Ways Forward in Global AI Benefit Sharing*. Centre for the Governance of Artificial Intelligence.

Department for Science, Innovation and Technology (2023a). *Emerging processes for frontier AI safety*. GOV.UK. URL: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf.

– (2023b). *Introducing the AI Safety Institute*. E03012924. GOV.UK. URL: https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

Department for Science, Innovation and Technology and AI Safety Institute (2024). *International Scientific Report on the Safety of Advanced AI*. GOV.UK. URL: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

Department for Science, Innovation and Technology, Foreign, Commonwealth and Development Office, and Prime Minister's Office, 10 Downing Street (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Dimitrov, R. S. (2020). "Empty institutions in global environmental politics". In: *International studies review* 22.3, pp. 626–650. DOI: 10.1093/isr/viz029.

Dunn, S. (2024). *India's $15.2 Billion Semiconductor Investment to Propel AI Development*. https://www.ccn.com/news/technology/india-15-2-billion-semiconductor-investment/. Accessed: 2024-9-19.

Eden, L. and F. O. Hampson (1997). "Clubs are Trump: The Formation of International Regimes in the Absence of a Hegemon". In: *Contemporary Capitalism: The Embeddedness of Institutions*. Ed. by J. Rogers Hollingsworth and R. Boyer. Vol. 495. Cambridge University Press Cambridge, pp. 361–394. URL: http://www.voxprofessor.net/eden/Publications/clubsaretrump.pdf.

Egan, J. and L. Heim (2023). "Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers". arXiv: 2310.13625 [cs.CY].

European Commission (2020). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

European Parliament (2023). *EU AI Act: first regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO938

04/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2024-9-19.

Farid Uddin, K. (2018). “Decentralization and Governance”. In: *Global Encyclopedia of Public Administration, Public Policy, and Governance*. Cham: Springer International Publishing, pp. 1–7. DOI: 10.1007/978-3-319-31816-5_2613-2.

Faulkner, D. O. (2006). *International Strategy*. Oxford University Press. DOI: 10.1093/oxfordhb/9780199275212.003.0022.

Fist, T. and E. Grunewald (2023). *Preventing AI Chip Smuggling to China*. Center for a New American Security. URL: https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china.

Fist, T., L. Heim, and J. Schneider (2023). *Chinese Firms Are Evading Chip Controls*. https://foreignpolicy.com/2023/06/21/china-united-states-semiconductor-chips-sanctions-evasion/. Accessed: 2024-9-20.

Fonseca, A. T. da, E. Vaz de Sequeira, and L. Barreto Xavier (2024). “Liability for AI Driven Systems”. In: *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. Ed. by H. Sousa Antunes et al. Cham: Springer International Publishing, pp. 299–317. DOI: 10.1007/978-3-031-41264-6_16.

Future of Life Institute (2024). *High-level summary of the AI Act*. https://artificialintelligenceact.eu/high-level-summary/. Accessed: 2024-9-21.

Gallegos, I. O. et al. (2023). “Bias and Fairness in Large Language Models: A Survey”. arXiv: 2309.00770 [cs.CL].

Gardizy, A. (2024). *China’s Nvidia Loophole: How ByteDance Got the Best AI Chips Despite U.S. Restrictions*. https://www.theinformation.com/articles/chinas-nvidia-loophole-how-bytedance-got-the-best-ai-chips-despite-u-s-restrictions. Accessed: 2024-9-20.

Gebru, T. et al. (2021). “Datasheets for datasets”. In: *Communications of the ACM* 64.12, pp. 86–92. DOI: 10.1145/3458723.

Gemini Team et al. (2023). *Gemini: A Family of Highly Capable Multimodal Models*. Google DeepMind. URL: http://arxiv.org/abs/2312.11805.

Gill, S. (1992). “Economic globalization and the internationalization of authority: Limits and contradictions”. In: *Geoforum* 23.3, pp. 269–283. DOI: 10.1016/0016-7185(92)90042-3.

Gilligan, M. J. (2009). “The Transactions Costs Approach to International Institutions”. URL: https://www.researchgate.net/publication/242493540_The_Transactions_Costs_Approach_to_International_Institutions.

Google for Developers (2022). *Fairness: Types of Bias*. https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias. Accessed: 2024-5-6.

Government of Canada and the Government of the United States of America (1991). *Transboundary air: Canada-US air quality agreement*. URL: https://www.canada.ca/en/environment-climate-change/services/air-pollution/issues/transboundary/canada-united-states-air-quality-agreement.html.

Greenleaf, G. (2023). *Global data privacy laws 2023: 162 national laws and 20 bills*. 181 Privacy Laws and Business International Report 1. UNSW Law & Justice. DOI: 10.2139/ssrn.4426146.

Grunewald, E. and M. Aird (2023). *AI chip smuggling into China: Potential paths, quantities, and countermeasures*. Institute for AI Policy and Strategy. URL: https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651bb8a18f961e3333e3c1d7/1696315558319/AI+chip+smuggling+into+China+%5Bfinal%5D.pdf.

Grunewald, E. and C. Phenicie (2023). *Introduction to AI Chip Making in China*. Institute for AI Policy and Strategy (IAPS). URL: https://www.iaps.ai/research/ai-chip-making-china.

Grynbaum, M. M. and R. Mac (2023). "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work". In: *The New York Times*. URL: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Hagerty, A. and I. Rubinov (2019). "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence". arXiv: 1907.07892 [cs.CY].

Hahn, P.-S. and M. Vesterlind (2013). "Safety-Related Activities of the IAEA for Radioactive Waste, Decommissioning and Remediation - 13473". In: *WM2013 Conference: International collaboration and continuous improvement*. Phoenix, AZ, USA.

Hale, T. and D. Held (2018). "Breaking the cycle of gridlock". In: *Global Policy* 9.1, pp. 129–137. DOI: 10.1111/1758-5899.12524.

He, L. (2024). *China is pumping another $47.5 billion into its chip industry*. https://www.cnn.com/2024/05/27/tech/china-semiconductor-investment-fund-intl-hnk/index.html. Accessed: 2024-9-19.

Heim, L. (2024). *(Training) Compute Thresholds – Features and Functions in AI Governance*. https://blog.heim.xyz/training-compute-thresholds/. Accessed: 2024-5-23.

Heim, L. et al. (2024). *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*. Oxford Martin AI Governance Initiative. URL: https://cdn.governance.ai/Governing-Through-the-Cloud_The-Intermediary-Role-of-Compute-Providers-in-AI-Regulation.pdf.

Hendrycks, D., M. Mazeika, and T. Woodside (2023). "An Overview of Catastrophic AI Risks". arXiv: 2306.12001 [cs.CY].

Henshall, W. (2024). *How Commerce Secretary Gina Raimondo Became America's Point Woman on AI*. https://time.com/6985335/gina-raimondo-commerce-artificial-intelligence/. Accessed: 2024-9-21.

Ho, L. et al. (2023). "International Institutions for Advanced AI". arXiv: 2307.04699 [cs.CY].

Hoffmann, M. and H. Frase (2023). *Adding structure to AI harm*. Center for Security and Emerging Technology Publications (CSET). DOI: 10.51593/20230022.

Horton, G. (2021). "Injunctions and public figures: the changing value in injunctions for privacy protection". In: *Journal of Media Law* 13.1, pp. 81–106. DOI: 10.1080/17577632.2021.1889866.

Houston, J. F., C. Lin, and Y. Ma (Oct. 2012). "Regulatory arbitrage and international bank flows". In: *The Journal of finance* 67 (5), pp. 1845–1895. DOI: 10.1111/j.1540-6261.2012.01774.x.

Howlett, M. and M. Ramesh (2016). "Achilles' heels of governance: Critical capacity deficits and their role in governance failures: The Achilles heel of governance". In: *Regulation & governance* 10.4, pp. 301–313. DOI: 10.1111/rego.12091.

Huang, Y. et al. (2023). "TrustGPT: A benchmark for trustworthy and responsible large Language Models". arXiv: 2306.11507 [cs.CL].

Information Commissioner's Office (2023a). *How do we ensure fairness in AI?* ICO. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/?q=gradient#apply.

– (2023b). *International data transfers*. https://ico.org.uk/for-organisations/data-protection-and-the-eu/data-protection-and-the-eu-in-detail/the-uk-gdpr/international-data-transfers/. Accessed: 2024-5-7.

Information Technology Industry Council (ITI) (2024). *Authenticating AI-Generated Content: Exploring Risks, Techniques & Policy Recommendations*. ITI. URL: https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf.

intersoft consulting services AG (n.d.). *Right of Access | General Data Protection Regulation (GDPR)*. https://gdpr-info.eu/issues/right-of-access/. Accessed: 2024-5-23.

Ippolito, D. et al. (2023). "Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy". In: *Proceedings of the 16th International Natural Language Generation Conference*. Ed. by C. M. Keet, H.-Y. Lee, and S. Zarrieß. Prague, Czechia: Association for Computational Linguistics, pp. 28–53. DOI: 10.18653/v1/2023.inlg-main.3.

Ireland Central Statistics Office (2023). *Data Centres Metered Electricity Consumption 2022*. CSO. URL: https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2022/keyfindings/.

ISO (2018). *ISO 31000: Risk management*. https://www.iso.org/iso-31000-risk-management.html. Accessed: 2024-4-23.

ISO/IEC (2019). *IEC 31010:2019: Risk management – Risk assessment techniques*. URL: https://www.iso.org/standard/72140.html.

Janssen, M. et al. (2020). "Data governance: Organizing data for trustworthy Artificial Intelligence". In: *Government information quarterly* 37.3, p. 101493. DOI: 10.1016/j.giq.2020.101493.

Johnson, C. W. (2003). *Failure in Safety-Critical Systems: A Handbook of Accident and Incident Reporting*. Glasgow, Scotland: University of Glasgow Press.

Judson, E. et al. (2024). *Synthetic politics: preparing democracy for Generative AI*. DEMOS. URL: https://apo.org.au/sites/default/files/resource-files/2024-03/apo-nid326166.pdf.

Kapoor, S. and A. Narayanan (2023). *How to prepare for the deluge of generative AI on social media: A grounded analysis of the challenges and opportunities*. Knight First Amendment Institute at Columbia University. URL: https://s3.amazonaws.com/kfai-documents/documents/a566f4ded5/How-to-Prepare-for-the-Deluge-of-Generative-AI-on-Social-Media.pdf.

Keohane, R. O. (1982). "The demand for international regimes". In: *International organization* 36.2, pp. 325–355. DOI: 10.1017/S002081830001897X.

– (2019). "Sovereignty, Interdependence, and International Institutions". In: *Ideas & Ideals*. 1st Edition. Routledge, pp. 91–107. DOI: 10.4324/9780429033957-7.

Keohane, R. O. et al. (1996). *Internationalization and domestic politics*. Ed. by R. O. Keohane and H. V. Milner. Cambridge studies in comparative politics. Cambridge, England: Cambridge University Press. DOI: 10.1017/cbo9780511664168.

Khan, S. M. and A. Mann (2020). *AI chips: What they are and why they matter*. CSET. URL: https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

Khan, S. M., A. Mann, and D. Peterson (2021). *The semiconductor supply chain: Assessing National Competitiveness*. Center for Security and Emerging Technology. DOI: 10.51593/20190016.

Knibbs, K. (2023). "Researchers Tested AI Watermarks–and Broke All of Them". In: *Wired*. URL: https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

Koessler, L. and J. Schuett (2023). "Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries". arXiv: 2307.08823 [cs.CY].

Koessler, L., J. Schuett, and M. Anderljung (2024). "Risk thresholds for frontier AI". arXiv: 2406.14713 [cs.CY].

Kolt, N. (2024). "Governing AI agents". DOI: 10.2139/ssrn.4772956.

Kolt, N. et al. (2024). "Responsible Reporting for Frontier AI Development". arXiv: 2404.02675 [cs.CY].

Krasner, S. D., ed. (1983). *International Regimes*. Cornell University Press. URL: https://www.cornellpress.cornell.edu/book/9780801492501/international-regimes/.

Krishna, K. et al. (2023). "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/pdf?id=WbFhFvjjKj.

Kugler, M. B. and C. Pace (2021). "Deepfake Privacy: Attitudes and Regulation". In: *Northwestern University law review* 116.3, pp. 611–680. URL: https://scholarlycommons.law.northwestern.edu/nulr/vol116/iss3/1.

Kulp, G. et al. (2024). *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. Santa Monica, CA: RAND Corporation. DOI: 10.7249/WRA3056-1.

Langlois, L. (2013). "IAEA Action Plan on nuclear safety". In: *Energy Strategy Reviews* 1.4, pp. 302–306. DOI: 10.1016/j.esr.2012.11.008.

Leavy, S., B. O'Sullivan, and E. Siapera (2020). "Data, Power and Bias in Artificial Intelligence". arXiv: 2008.07341 [cs.CY].

Leffer, L. (2023). "Your Personal Information Is Probably Being Used to Train Generative AI Models". In: *Scientific American*. URL: https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/.

Liang, P. et al. (2023). "Holistic Evaluation of Language Models". In: *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=iO4LZibEqW.

Long, R. (2021). "Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness". In: *Journal of Moral Philosophy* 19.1, pp. 49–78. URL: https://philpapers.org/rec/LONFIM.

Longpre, S., S. Kapoor, et al. (2024). "A Safe Harbor for AI Evaluation and Red Teaming". arXiv: 2403.04893 [cs.AI].

Longpre, S., R. Mahari, A. Chen, et al. (2023). “The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI”. arXiv: 2310.16787 [cs.CL].

Longpre, S., R. Mahari, N. Obeng-Marnu, et al. (2024). “Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?” arXiv: 2404.12691 [cs.AI].

Maas, M. M. (2022). “Aligning AI Regulation to Sociotechnical Change”. In: *The Oxford Handbook of AI Governance*. Ed. by J. B. Bullock et al. Oxford University Press. DOI: 10.1093/oxfordhb/9780197579329.013.22.

– (forthcoming). “Architectures of Global AI Governance”. In: *Technological Change to Human Choice*. Oxford, England: Oxford University Press.

Manancourt, V., G. Volpicelli, and M. Chatterjee (2024). “Rishi Sunak promised to make AI safe. Big Tech’s not playing ball”. In: *POLITICO*. URL: https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/.

Marion, M. et al. (2023). “When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale”. In: *1st Workshop on Attributing Model Behavior at Scale*. New Orleans, LA, USA. URL: https://openreview.net/forum?id=XUIYn3jo5T.

Maslej, N. et al. (2024). *The AI Index 2024 Annual Report*. Stanford, CA, USA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. URL: https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf.

McDonald, B. (2016). “Why competition is good for the space race: Bob McDonald”. In: *CBC News*. URL: https://www.cbc.ca/news/science/why-competition-is-good-for-the-space-race-bob-mcdonald-1.3405629.

McGregor, S. (2021). “Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.17, pp. 15458–15463. DOI: 10.1609/aaai.v35i17.17817.

Mearsheimer, J. J. (2017). “The false Promise of International Institutions”. In: *International organization*. Ed. by J. J. Kirton. Routledge. DOI: 10.4324/9781315251981.

Mehrabi, N. et al. (2021). “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Comput. Surv.* 54.6, pp. 1–35. DOI: 10.1145/3457607.

Metz, C. et al. (2024). “How Tech Giants Cut Corners to Harvest Data for A.I.” In: *The New York Times*. URL: https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

Miller, C. (2022). *Chip War*. New York, NY: Scribner. URL: https://www.simonandschuster.com/books/Chip-War/Chris-Miller/9781982172008.

Mills, R. W. (2010). *The Promise of Collaborative Voluntary Partnerships: Lessons from the Federal Aviation Administration*. IBM Center for The Business of Government. URL: https://www.businessofgovernment.org/report/promise-collaborative-voluntary-partnerships-lessons-federal-aviation-administration.

– (2016). “The interaction of private and public regulatory governance: The case of association-led voluntary aviation safety programs”. In: *Policy and Society* 35.1, pp. 43–55. DOI: 10.1016/j.polsoc.2015.12.002.

MIT Technology Review Insights (2024). “Multimodal: AI’s new frontier”. In: *MIT Technology Review*. URL: https://www.technologyreview.com/2024/05/08/1092009/multimodal-ais-new-frontier/.

Mitchell, M. et al. (2022). “Measuring Data”. arXiv: 2212.05129 [cs.AI].

Mökander, J. et al. (2023). "Auditing large language models: a three-layered approach". In: *AI and Ethics*. DOI: 10.1007/s43681-023-00289-2.

Molenda, P., A. Liusie, and M. J. F. Gales (2024). "WaterJudge: Quality-Detection Trade-off when Watermarking Large Language Models". arXiv: 2403.19548 [cs.CL].

Morgan, F. E. et al. (2020). *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. RAND Corporation. URL: https://play.google.com/store/books/details?id=9AGFzQEACAAJ.

Mu, N. et al. (2023). "Can LLMs Follow Simple Rules?" arXiv: 2311.04235 [cs.AI].

Mullane, H. and H. Dohmen (2024). *The race to secure semiconductor supply chains*. URL: https://cset.georgetown.edu/article/the-race-to-secure-semiconductor-supply-chains/.

Nasr, M. et al. (2023). "Scalable Extraction of Training Data from (Production) Language Models". arXiv: 2311.17035 [cs.LG].

National Institute of Standards and Technology (NIST) (2020). *Glossary: risk mitigation*. https://csrc.nist.gov/glossary/term/risk_mitigation. Accessed: 2024-5-26.

New ways to manage your data in ChatGPT (2023). https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/. Accessed: 2024-9-19.

Ngo, R., L. Chan, and S. Mindermann (2023). "The Alignment Problem from a Deep Learning Perspective". In: *The 12th International Conference on Learning Representations (ICLR 2024)*. Vienna, Austria. URL: https://openreview.net/forum?id=fh8EYKFKns.

NIST (2023). *Artificial intelligence risk management Framework (AI RMF 1.0)*. Gaithersburg, MD: NIST. DOI: 10.6028/nist.ai.100-1.

Norris, P. (2020). "*The World of Political Science: Internationalization and Its Consequences*". In: *Political science in Europe: Achievements, challenges, prospects*. Ed. by T. Boncourt, I. Engeli, and D. Garzia. London, England: ECPR Press. URL: https://www.hks.harvard.edu/publications/world-political-science-internationalization-and-its-consequences#citation.

Ntoutsi, E. et al. (2020). "Bias in data-driven artificial intelligence systems–An introductory survey". In: *Wiley interdisciplinary reviews. Data mining and knowledge discovery* 10.3. DOI: 10.1002/widm.1356.

Obermeyer, Z. et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464, pp. 447–453. DOI: 10.1126/science.aax2342.

OECD.AI Policy Observatory (n.d.). *OECD AI Principles overview*. https://oecd.ai/en/ai-principles. Accessed: 2024-5-22.

Olcott, E., D. Sevastopulo, and Q. Liu (2023). "Chinese AI groups use cloud services to evade US chip export controls". In: *Financial Times*. URL: https://www.ft.com/content/9706c917-6440-4fa9-b588-b18fbc1503b9.

OpenAI et al. (2024). *GPT-4 Technical Report*. OpenAI. URL: http://arxiv.org/abs/2303.08774.

Palmer, E. (2020). *Attestation of System Components v1.0 Requirements and Recommendations*. Open Compute Project.

Pannekoek, M. and G. Spigler (2021). "Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks". arXiv: 2102.05975 [cs.LG].

Patwardhan, T. et al. (2024). *Building an early warning system for LLM-aided biological threat creation*. OpenAI. URL: https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation.

Paullada, A. et al. (2021). “Data and its (dis)contents: A survey of dataset development and use in machine learning research”. In: *Patterns (New York, N.Y.)* 2.11, p. 100336. DOI: 10.1016/j.patter.2021.100336.

Penedo, G. et al. (2023). “The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only”. In: *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*. New Orleans, LA, USA. URL: https://openreview.net/pdf?id=kM5eGcdCzq.

Phuong, M. et al. (2024). *Evaluating Frontier Models for Dangerous Capabilities*. Google Deepmind. DOI: 10.48550/arXiv.2403.13793.

Picker, C. B. (2007). “A View from 40,000 Feet: International Law and the Invisible Hand of Technology”. In: *Cardozo Law Review* 23.1, pp. 149–219. URL: https://papers.ssrn.com/abstract=987524.

Pilz, K. (2023). “An assessment of data center infrastructure’s role in AI governance”. URL: https://www.konstantinpilz.com/data-centers/assessment.

Pilz, K. and L. Heim (2023). “Compute at Scale: A Broad Investigation into the Data Center Industry”. arXiv: 2311.02651 [cs.CY].

Pilz, K., L. Heim, and N. Brown (2023). “Increased Compute Efficiency and the Diffusion of AI Capabilities”. arXiv: 2311.15377 [cs.CY].

Pittaras, N. and S. McGregor (2022). “A taxonomic system for failure cause analysis of open source AI incidents”. arXiv: 2211.07280 [cs.AI].

Powell, R. et al. (2024). *Towards Secure AI: How far can international standards take us?* Centre for Emerging Technology and Security (CETaS). URL: https://cetas.turing.ac.uk/publications/towards-secure-ai.

Putnam, R. D. (1988). “Diplomacy and domestic politics: the logic of two-level games”. In: *International organization* 42.3, pp. 427–460. DOI: 10.1017/s0020818300027697.

Raji, I. D. et al. (2022). “Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’22)*. New York, NY, USA: Association for Computing Machinery, pp. 557–571. DOI: 10.1145/3514094.3534181.

Rauh, M. et al. (2022). “Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models”. In: *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Virtual. URL: https://openreview.net/forum?id=u46CbCaLufp.

Rep. Jeff Jackson Introduces Bipartisan CLOUD AI Act to Stop China from Remotely Using American Technology to Build AI Tools (2023). https://jeffjackson.house.gov/media/press-releases/rep-jeff-jackson-introduces-bipartisan-cloud-ai-act-stop-china-remotely-using. Accessed: 2024-5-7.

Reuel, A. (2024a). “3.2 Privacy and Data Governance”. In: *Artificial Intelligence Index Report 2024*. Ed. by R. Perrault and J. Clark. Stanford, CA, USA: Institute for Human-Centered AI, Stanford University, pp. 172–179.

– (2024b). “Chapter 3: Responsible AI”. In: *Artificial Intelligence Index Report 2024*. Ed. by R. Perrault and J. Clark. Stanford, CA, USA: Institute for Human-Centered AI, Stanford University, pp. 159–212.

Reuters (2024). “South Korea to invest $7 billion in AI in bid to retain edge in chips”. In: *Reuters*. URL: https://www.reuters.com/technology/south-korea-invest-7-bln-ai-bid-retain-edge-chips-2024-04-09/.

Roberts, H. et al. (2024). “Global AI governance: barriers and pathways forward”. In: *International affairs* 100.3, pp. 1275–1286. DOI: 10.1093/ia/iiae073.

Rodrik, D. (2012). *The globalization paradox: Democracy and the future of the world economy*. W. W. Norton & Company.

Rovatsos, M., B. Mittelstadt, and A. Koene (2019). *Landscape Summary: Bias in Algorithmic Decision-Making*. Centre for Data Ethics and Innovation; Department for Science, Innovation & Technology. URL: https://assets.publishing.service.gov.uk/media/5d31c30a40f0b64a8099e21d/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf.

Ryngaert, C. and M. Taylor (2020). “The GDPR as Global Data Protection Regulation?” In: *AJIL Unbound* 114, pp. 5–9. DOI: 10.1017/aju.2019.80.

Sabec, L. (2004). “FAA approves IATA’s Operational Safety Audit (IOSA) program: A historical review and future implications for the airline industry”. In: *Transportation Law Journal* 32. URL: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/tportl32%5Ctextsection%7B%7Dion=7.

Sastry, G. et al. (2024). “Computing Power and the Governance of Artificial Intelligence”. arXiv: 2402.08797 [cs.CY].

Satariano, A. (2023). “ChatGPT Is Banned in Italy Over Privacy Concerns”. In: *The New York Times*. URL: https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html.

Schaeffer, R., B. Miranda, and S. Koyejo (2023). “Are Emergent Abilities of Large Language Models a Mirage?” In: *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. New Orleans, LA, USA. URL: https://openreview.net/forum?id=ITw9edRDlD.

Scheer, S. (2023). “Israel grants Intel 3.2$billionfornew$25 billion chip plant”. In: *Reuters*. URL: https://www.reuters.com/technology/intel-get-32-billion-government-grant-new-25-billion-israel-chip-plant-2023-12-26/.

Schwartz, R. et al. (2022). *NIST Special Publication 1270: Towards a standard for identifying and managing bias in artificial intelligence*. Gaithersburg, MD: National Institute of Standards and Technology (U.S.) DOI: 10.6028/nist.sp.1270.

Sclar, M. et al. (2023). “Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting”. arXiv: 2310.11324 [cs.CL].

Seger, E. et al. (2023). *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. Centre for the Governance of AI. URL: http://arxiv.org/abs/2311.09227.

Sevilla, J. et al. (2022). “Compute Trends Across Three Eras of Machine Learning”. In: *2022 International Joint Conference on Neural Networks (IJCNN 2022)*. Padua, Italy, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9891914.

Shavit, Y. (2023). “What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring”. arXiv: 2303.11341 [cs.LG].

Shevlane, T. (2022). “Structured access: an emerging paradigm for safe AI deployment”. arXiv: 2201.05159 [cs.AI].

Shevlane, T. et al. (2023). *Model evaluation for extreme risks*. Google DeepMind. URL: http://arxiv.org/abs/2305.15324.

Siegmann, C. and M. Anderljung (2022). *The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market*. Centre for the Governance of AI. URL: http://arxiv.org/abs/2208.12645.

Smith, B. (2023). *Developing and deploying AI responsibly: Elements of an effective legislative framework to regulate AI*. https://blogs.microsoft.com/on-the-issues/2023/09/12/developing-and-deploying-ai-responsibly-elements-of-an-effective-legislative-framework-to-regulate-ai/. Accessed: 2024-5-7.

Solaiman, I. et al. (2023). "Evaluating the Social Impact of Generative AI Systems in Systems and Society". arXiv: 2306.05949 [cs.CY].

Soykan, E. U. et al. (2022). "A Survey and Guideline on Privacy Enhancing Technologies for Collaborative Machine Learning". In: *IEEE Access* 10, pp. 97495–97519. DOI: 10.1109/ACCESS.2022.3204037.

Srinivasan, S. (2024). *Detecting AI fingerprints: A guide to watermarking and beyond*. Brookings. URL: https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

Stafford, E. and R. F. Trager (2022). *The IAEA Solution: Knowledge Sharing to Prevent Dangerous Technology Races*. Centre for the Governance of AI. URL: https://cdn.governance.ai/IAEA_Solution_DRAFT_JULY_2022.pdf.

Stein, A. A. (1982). "Coordination and collaboration: regimes in an anarchic world". In: *International organization* 36.2, pp. 299–324. DOI: 10.1017/s0020818300018968.

Tejani, A. S. et al. (2024). "Understanding and Mitigating Bias in Imaging Artificial Intelligence". In: *Radiographics: a review publication of the Radiological Society of North America, Inc* 44.5, e230067. DOI: 10.1148/rg.230067.

The New York Times Company v. Microsoft Corporation et al Document 1 (2023). URL: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

The White House (2023a). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. The White House. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

– (2023b). *FACT SHEET: Biden–Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/. Accessed: 2024-5-6.

Tiedrich, L. (2024). *The AI data scraping challenge: How can we proceed responsibly?* https://oecd.ai/en/wonk/data-scraping-responsibly. Accessed: 2024-5-6.

Touvron, H. et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. Meta AI. URL: http://arxiv.org/abs/2307.09288.

Trager, R. et al. (2023). *International Governance of Civilian AI: A Jurisdictional Certification Approach*. Oxford Martin AI Governance Initiative.

Tremblay v. OpenAI, Inc. (3:23-cv-03223) Document 1 (2023). URL: https://storage.courtlistener.com/recap/gov.uscourts.cand.414822/gov.uscourts.cand.414822.1.0_1.pdf.

Turri, V. and R. Dzombak (2023). “Why We Need to Know More: Exploring the State of AI Incident Documentation Practices”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’23)*. Montreal, QC, Canada: ACM, pp. 576–583. DOI: 10.1145/3600211.3604700.

U.S. Copyright Office (2023). *U.S. Copyright Office Fair Use Index*. https://www.copyright.gov/fair-use/. Accessed: 2024-5-23.

U.S. Department of Homeland Security (2023). *Department of Homeland Security Homeland Threat Assessment*. https://www.dhs.gov/publication/homeland-threat-assessment. Accessed: 2024-5-6.

U.S. Department of the Treasury (2024). *FATF Advances Work to Combat Money Laundering and Terrorist Financing*. https://home.treasury.gov/news/press-releases/jy2120. Accessed: 2024-9-19.

United Nations (1948). *Universal Declaration of Human Rights*. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

United Nations Framework Convention on Climate Change (2024). *UNFCCC*. https://unfccc.int/. Accessed: 2024-5-23.

United Nations Framework Convention on Climate Change (UNFCCC) (2023). *The Paris Agreement*. https://unfccc.int/process-and-meetings/the-paris-agreement. Accessed: 2024-9-14.

United States of America and Russian Federation (2011). *New START Treaty*. URL: https://www.state.gov/new-start/.

van Aaken (2016). “Is international law conducive to preventing looming disasters?” In: *Global Policy* 7.S1, pp. 81–96. DOI: 10.1111/1758-5899.12303.

Veale, M., K. Matus, and R. Gorwa (2023). “AI and Global Governance: Modalities, Rationales, Tensions”. In: *Annual Review of Law and Social Science* 19.Volume 19, 2023, pp. 255–275. DOI: 10.1146/annurev-lawsocsci-020223-040749.

Voss, G. W. (2020). “Cross-Border Data Flows, the GDPR, and Data Governance”. In: *Washington International Law Journal* 29.3, p. 485. URL: https://digitalcommons.law.uw.edu/wilj/vol29/iss3/7/.

Wei, A., N. Haghtalab, and J. Steinhardt (2023). “Jailbroken: How Does LLM Safety Training Fail?” In: *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. New Orleans, LA, USA. URL: https://openreview.net/forum?id=jA235JGM09.

Wei, J. et al. (2022). “Emergent Abilities of Large Language Models”. In: *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=yzkSU5zdwD.

Wei, K. L. (n.d.). “Designing Incident Reporting Systems for Harms from AI”.

Weidinger, L., J. Barnhart, et al. (2024). *Holistic Safety and Responsibility Evaluations of Advanced AI Models*. Google Deepmind. URL: http://arxiv.org/abs/2404.14068.

Weidinger, L., M. Rauh, et al. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems*. Google Deepmind. URL: http://arxiv.org/abs/2310.11986.

Werder, K., B. Ramesh, and R. Zhang (2022). “Establishing Data Provenance for Responsible Artificial Intelligence Systems”. In: *ACM Trans. Manage. Inf. Syst.* 13.2, pp. 1–23. DOI: 10.1145/3503488.

Wittenberg, C. et al. (2024). “Labeling AI-generated content: Promises, perils, and future directions”. In: *From Novel Chemicals to Opera*. DOI: 10.21428/e4baedd9.0319e3a6.

Wolff, S. (2020). “Autonomy”. In: *The Princeton Encyclopedia of Self-Determination*. Princeton University. URL: https://pesd.princeton.edu/node/236.

Woollacott, E. (2023). “U.K. Privacy Watchdog Can’t Sanction Clearview AI, Court Rules”. In: *Forbes Magazine*. URL: https://www.forbes.com/sites/emmawoollacott/2023/10/19/uk-privacy-watchdog-cant-sanction-clearview-ai-court-rules/.

World Health Organization (WHO) (2021). *Ethics and governance of artificial intelligence for health: WHO guidance. Executive summary*. Geneva: WHO. URL: https://iris.who.int/bitstream/handle/10665/350567/9789240037403-eng.pdf.

World Intellectual Property Organization (WIPO) (n.d.). *Berne Convention for the Protection of Literary and Artistic Works*. https://www.wipo.int/treaties/en/ip/berne/. Accessed: 2024-5-23.

Xiang, C. (2023). *’He Would Still Be Here’: Man Dies by Suicide After Talking with AI Chatbot, Widow Says*. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says. Accessed: 2024-5-6.

Zhang, H. et al. (2023). “Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models”. arXiv: 2311.04378 [cs.LG].

Zou, A. et al. (2023). “Universal and Transferable Adversarial Attacks on Aligned Language Models”. arXiv: 2307.15043 [cs.CL].