cerre Centre on Regulation in Europe

GENERATIVE AI: GLOBAL GOVERNANCE AND THE RISK-BASED APPROACH

GLOBAL GOVERNANCE FOR THE DIGITAL ECOSYSTEMS- PHASE 2

30 november 2023

Gianclaudio Malgieri
Gautam Kamath

As provided for in CERRE's bylaws and procedural rules from its “Transparency & Independence Policy”, all CERRE research projects and reports are completed in accordance with the strictest academic independence.

The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE or any sponsor or of members of CERRE or any other organisation or individual involved in the CERRE project on “Global Governance for the Digital Ecosystems.”

info@cerre.eu – www.cerre.eu

TABLE OF CONTENTS

ABOUT CERRE

Providing top quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) promotes robust and consistent regulation in Europe’s network and digital industries. CERRE’s members are regulatory authorities and operators in those industries as well as universities.

CERRE’s added value is based on:

- its original, multidisciplinary and cross-sector approach;
- the widely acknowledged academic credentials and policy experience of its team and associated staff members;
- its scientific independence and impartiality;
- the direct relevance and timeliness of its contributions to the policy and regulatory development process applicable to network industries and the markets for their services.

CERRE's activities include contributions to the development of norms, standards and policy recommendations related to the regulation of service providers, to the specification of market rules and improvements in the management of infrastructure in a changing political, economic, technological and social environment. CERRE’s work also aims at clarifying the respective roles of market operators, governments and regulatory authorities, as well as at strengthening the expertise of the latter.

ABOUT THE AUTHORS

Gianclaudio Malgieri

Associate Professor of Law, University of Leiden
Co-Director, Brussels Privacy Hub VUB

Gautam Kamath

Senior Advisor, CERRE

With contributions from **Lucas Anjos** (Universidade Federal de Minas Gerais, Brazil; Sciences Po Paris), **Niti Chatterjee** (CIPP/E, L.L.M. University of Leiden), and **Lyubomir Nikiforov** (VUB Brussels).

EXECUTIVE SUMMARY

The emergence of generative AI has sparked both enthusiasm and concern, especially because there is a currently a lack of know how of the technology itself. Given the significant private sector involvement in AI and the global nature of AI regulation, there is a growing need for comprehensive, technology-neutral, multi-stakeholder-driven regulatory frameworks.

In addition to this need, there is a growing consensus on exploring potential risks from generative AI, while sensationalist media are driving the divide between existential and immediate concerns. This policy report aims to reframe the political discourse around regulating foundational technologies like generative AI, offering practical policy approaches and recommendations for global convergence. The "risk-based approach" from the EU AI Act is an illustrative example that policymakers, including those in the G7 Hiroshima AI Process, can adopt to assess any potential foundation model risks. This perspective is also relevant to initiatives like the Global Partnership on Artificial Intelligence (GPAI) and the G20, led by countries such as India and Brazil.

This policy report comprises three main sections: one addressing risks arising from the use of generative AI, another discussing risk mitigation measures for the risks that were surfaced, and a final section charting a path for global governance of generative AI. It concludes with concrete policy recommendations for regulatory convergence through evidence-based, technology-neutral, multi-stakeholder, resilient policymaking. This aligns with the goal of CERRE's Global Governance for the Digital Ecosystems project (GGDE) to promote regulatory convergence globally and ensuring co-existence when convergence is not feasible.

This report is the first in a series of upcoming GGDE policy notes to explore values, frameworks, and global convergence on AI regulation. This series will explore principles, human oversight, comprehensibility, accessibility, competition law, copyright, AI risk assessment, and international AI system transfer/use. The project will also examine the roles of EU institutions and other bodies in organising convergence.

The overarching goal is to provide a concrete path forward for G7+ policymakers, fostering a common understanding of "AI adequacy" to guide industry and citizens worldwide and ensuring progress on the Hiroshima AI Process. This report also provides evidence-based policy recommendations for stakeholders in these processes. CERRE has already submitted four specific policy recommendations within the framework of the European Commission "stakeholder survey" of October 2023, guiding the G7 Hiroshima AI Process:

- Adapt the "risk-based approach" from the EU AI Act for Generative AI in G7+ countries, focusing on "high-risk" general purpose AI model applications and developing industry risk assessment tools.
- Focus on developing tiered, yet limited regulatory obligations for high-risk use cases.
- Embrace a collaborative, multi-stakeholder approach in AI governance efforts.
- Promote "AI adequacy" to enhance legal certainty and regulatory convergence.

FOREWORD

This policy report has been prepared within the framework of CERRE's flagship project on "Global Governance for the Digital Ecosystems" (GGDE). It is in line with the project's overarching goal: contribute to preserving and promoting regulatory convergence at the global level and, where convergence is neither desirable nor legitimate, to organising co-existence.

It is the first of a series examining how to achieve these objectives in the case of AI governance. Generative AI is examined throughout Phase 2 of the GGDE project as a transversal use case, eliciting specific, relevant insights on global governance.

Other upcoming policy reports in the GGDE's AI workstream of GGDE will examine common values and frameworks to build further ideas for global convergence on AI regulation. This will include a deep dive into exploring principles, human oversight, comprehensibility, accessibility, as well as issues related to competition law and copyright. Further, this workstream will investigate whether AI risk assessment could act as a bottom-up business tool for global convergence, focusing on defining both systemic risk and social impact. Finally, the project will also seek to investigate inter-governmental tools for the international transfer/use of AI systems and discuss what role EU institutions and other supranational and multilateral bodies could play in organising such convergence.

The main overarching goal is to provide a concrete path forward for policymakers engaged at G7+, to build on a common understanding of "AI adequacy" that provide guidance and tools for industry and citizens globally and to ensure that policymakers follow through on the Hiroshima AI Process and continue their important policy coordination and alignment in other multilateral formats, including the United Nations, the Global Partnership on AI (GPAI) and the G20.

GLOBAL GOVERNANCE FOR FOUNDATION MODELS

The emergence of foundation models has captured the global zeitgeist and inspired euphoric excitement while also eliciting broad concern among policymakers in 2023.

This term was originally coined by Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) - Center for Research on Foundation Models (CRFM) in August 2021. It refers to "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks". According to the European Parliament's position adopted in June 2023, a foundation model is defined as an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks.[1] This term is now likely to be included in the negotiated compromise text of the European Union's upcoming AI Act this fall. It is also mentioned in the Bletchley Declaration adopted at the UK's initiative and, at the G7 level, in the International Guiding Principles on Artificial Intelligence (AI) and the voluntary Code of Conduct for AI, both developed under the Hiroshima AI Process, based on EU and US drafts.

In addition to countries working on establishing guardrails for AI over the past several years, numerous initiatives have been driving towards convergence and global governance. This journey began with the OECD's AI principles in May 2019, which defined AI as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments."[2] This definition was updated by the OECD in recent weeks to inform the EU AI Act discussions, to read "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment"[3]. This evolving overarching definition can also be linked to work done at the EU-US Trade and Technology Council, which also recently requested stakeholder input for developing a common taxonomy on AI systems.

At the United Nations, the complex and interdisciplinary nature of AI has posed challenges to achieving political consensus and inter-agency coordination[4]. While the negotiation of the UN Global Digital Compact is currently in progress and expected to be finalised in 2024, the UN is also working towards the development of a legally binding agreement that prohibits the use of AI in fully automated weapons of war by 2026. There are at least three ongoing initiatives aimed at regulating Artificial

[1] EP position can be found here : https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

[2]: OECD, "Recommendation of the Council on Artificial Intelligence" (2019) Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[3]: This was reported as part of the October session of the OECD's Committee on Digital Economy Policy and Working Party on Artificial Intelligence Governance. Available at : https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/

[4]: In its 40th session in October 2020, the High-Level Committee on Programmes at the United Nations created an inter-agency working group on Artificial Intelligence (IAWG-AI), co-led by UNESCO and ITU.

Intelligence within the UN framework. The first is led by the Office of the Secretary General's Envoy on Technology, which focuses on the Digital Compact mentioned earlier. The second initiative centres on AI for good within the ITU, encompassing the WSIS forum, AI for Good, and technical efforts related to AI standards in ITU-T study groups. Finally, UNESCO is actively engaged in the field of AI ethics. Additionally, there are new and emerging proposals, such as the UN Secretary General's recommendation to establish an agency responsible for regulating high risk AI, akin to agencies that "manage the use of nuclear energy, boost aviation safety, or tackle climate change".

In addition to the crowded multilateral governance discussion on AI in general, an interesting aspect is policymaker responses to generative AI. In the EU, the AI Act is being finalised through trilogue discussions, and the European Parliament position has dedicated text on generative AI and "foundation models"[5]. The question of definitions for generative AI are instructive for the EU negotiations, as the concept of foundation models have found mention in several cases. One amendment for a recital defined them as "a recent development, in which AI models are developed from algorithms designed to optimize for generality and versatility of output. Those models are often trained on a broad range of data sources and large amounts of data to accomplish a wide range of downstream tasks, including some for which they were not specifically developed and trained…. AI systems with specific intended purpose or general purpose AI systems can be an implementation of a foundation model, which means that each foundation model can be reused in countless downstream AI or general purpose AI systems. These models hold growing importance to many downstream applications and systems"[6]. While much of the EU's eventual position depends on what finally is agreed during trilogue negotiations of the AI Act (expected to close by the end of this year, or early next year), the likelihood of a vague definition (e.g. "general purpose AI model") that conflates distinct concepts looms large. The table below shows some of the divergences in approaches that are being discussed in various countries:

	Recent legislative developments on generative AI	***Approach***
China	Cyberspace Administration of China (CAC)'s "Interim rules on Generative AI" (entered into force on 15 August 2023). Also launched on 17 October 2023 its own Global AI Governance Initiative.	Prescriptive, state-led
European Union	Finalisation of EU AI Act could include specific text that lays down obligations on providers of "general purpose AI models" (GPAIs). Such text could also include calls on respecting "reservation of rights" under Article 4 (3) of the EU Copyright Directive.	Prescriptive, risk-based, (EU-led)

[5] See for e.g. Amendment #399, of the new draft Article 28b in the European Parliament position (from 14 June 2023), which lays down obligations for providers of these models.
[6] See for e.g. Amendment #99 for Recital 60e (from 14 June 2023) in the European Parliment position, which seeks to define foundation models and their relation to 'general purpose AI systems '.

United States	The Executive Order of 30 October 2023 lays out specific rules for gov. agencies' procurement and use of artificial intelligence. A recent bill in Congress on regulating AI has reached the committee stage and is expected to be voted in 2024. There have been investments of $140m into AI research institutes, a blueprint for an AI bill of rights, and a public consultation on how best to regulate how AI is used.	Voluntary, industry-led with procurement-driven rules and protection of civil rights
Japan	No plans to release statutory rules on generative AI (as part of their model of "agile digital governance") and insulated foundation model providers from copyright claims as per Copyright Law (Art 30-4).	Voluntary, collaborative, industry-led
G7+	Following the Hiroshima Leaders Communique in May 2023 to set up a process in partnership with the Global Partnership on AI and OECD, The G7 published in October 2023 Principles for advanced AI companies and a voluntary code of conduct. Also note the Bletchley Declaration, led by the UK, focused on articulating AI risks.	Voluntary, collaborative, risk-based

This report assesses the salient considerations that policymakers should consider when formulating regulatory frameworks for generative artificial intelligence (AI). Positioned as a tool to inform the development of regulations for both industry and citizens, the note complements and charts a path toward action grounded in empirical evidence on a multilateral scale, exemplified by initiatives such as the G7 Hiroshima AI Process.

By cataloguing the risks of generative AI, this report seeks to anticipate and disentangle the political discourse surrounding AI, bridging the perceived schism between the "existential risk" perspective primarily advocated by members of the technology industry and the second perspective dominated by scholars and civil society activists, which emphasizes immediate concerns related to democracy (e.g., countering misinformation) and human rights[7].

Many of the global governance initiatives on AI run the risk of a lack of follow-through – while governments sign lofty declarations, the efficacy of these policy dialogues can only be judged with hindsight by examining implementation-oriented actions. The political dialogue is only starting now to take a broader view of the risks posed by generative AI. One objective of this policy report is therefore to reframe the political discourse concerning the regulation of an incipient foundational technology like generative AI. It aims to offer concrete policy approaches and recommendations that can ultimately foster global convergence. An illustrative example of this reorientation is the "risk-based approach" introduced within the EU AI Act. Viewing the risks posed by foundation models through this framework proves to be a valuable exercise for policymakers involved in the G7 Hiroshima AI Process,

[7] See "What this week's flurry of AI policymaking means for researchers" from Science Business.

as articulated in Principle 5 of the Voluntary Code of Conduct. This perspective is also pertinent to the broader initiatives of the Global Partnership on AI (GPAI)[8] and the G20, led by India and Brazil.

This report is organised around three main parts: one is related to the myriad of issues arising from the risks of generative AI; the second part deals with the type of risk mitigation measures that policymakers should consider across myriad areas, and finally the last part lays a path forward for the global governance of generative AI.

In the first part, we address two distinct categories of issues pertaining to generative AI, from the vantage point of policymakers across the globe. One is related to identification and definition of risks that are specifically associated with generative AI. Another set of issues are related to the necessary actions that various economic actors within the AI value chain must undertake to ensure the safe and desirable utilization of generative AI across a wide spectrum of contexts.

The second part scrutinizes the appropriateness of existing laws and frameworks in relation to critical matters such as AI safety, data privacy, and cybersecurity. A multitude of digital regulations and frameworks are already in place, many of which can be applicable to foundational AI models. It is crucial to elucidate these regulations and appraise the approaches that should be further developed. Additionally, it is incumbent upon policymakers to mitigate adverse externalities arising from the deployment of generative AI, encompassing impacts on the public sector, open competition, and the environment. The imperative of adopting multi-stakeholder solutions, involving industry stakeholders, is evident, and encouraging examples of such collaborations are already in evidence.

The third part underscores the significance of adopting a risk-based approach to confront the challenges emanating from generative AI. This includes an examination of how the approach articulated in the EU AI Act can offer valuable lessons for policymakers worldwide, particularly those engaged in the G7 Hiroshima AI Process. The third part concludes with concrete policy recommendations for policymakers to build upon the G7 Hiroshima AI Process International Draft Guiding Principles for Organizations Developing Advanced AI Systems using the aforementioned "risk-based approach." Our recommendations aim at building regulatory convergence through focused, evidence-based, technology-agnostic, multi-stakeholder, resilient, and future-proof policymaking.

[8] The GPAI’s 2022 Multistakeholder Report lists the development of risk assessment methodology as part of working group 1 on Responsible AI (see pg 9).

OVERVIEW OF THE RISKS OF GENERATIVE AI

AI is still at a nascent stage, and the innovative benefits and rapid improvement in the foundational technology has created what Nick Clegg, Meta's President, Global Affairs, recently characterised as "moral panics" where "AI was caught in a "great hype cycle" and that new technologies inspired a mixture of excessive zeal and excessive pessimism. The press around generative AI has been similarly sensational and polarizing, veering between overtly utopian and effusive, to downright dystopian (a Frankenstein's monster) and eschatological[9]. Because of how AI is developed and deployed (primarily by the private sector) and how its regulation involves a global dialogue, an argument can be made for regulation to be both comprehensive and technology-agnostic. Several such initiatives have sprung up so far aiming to stem unintended impacts across political, social and economic dimensions.

Before attempting to comprehensively catalogue the potential risks of generative AI, it is important to take note of the transformative benefits that generative AI can have for a wide variety of industry sectors, including healthcare (enhancing image resolution, aiding with diagnostics, improving drug discovery), finance (fraud detection, customer service), manufacturing (quality assurance, etc.), education, retail, transportation and so on. Part of the 'moral panic' stems from the success that generative AI has seen in a very short timeframe, with OpenAI becoming the fastest growing consumer application surpassing 100mn monthly active users two months after launch.

At the outset, there are separate use cases for foundation models: (i) where a business or public sector organisation uses generative AI in an enterprise setting; and (ii) models that are used by consumers for individual use. Each of these poses distinct but definitive risks, some of which may overlap. For example, the "generative" nature of outputs requires end users, whether in a commercial or consumer setting, to be aware of what they are using, particularly the fact that the self-supervised nature of learning inherent in such models can result in display of "emergent properties" (where models for one function can be repurposed for others). Additionally, glaring inaccuracies and fabrications (known as "hallucinations") in foundational models have resulted in not only financial and reputational risks[10] but also hindrances to judicial proceedings.

In this context, OpenAI's GPT4 System Card[11] identifies existing safety-related challenges of LLMs. These include:

[9] See for example, Yuval Noah Harari's characterisation: "if we don't regulate deployment, this will definitely destroy democracy much faster than any scheme by a North Korean tyrant or whatever. We need regulation in order to save democracy. If we don't have regulation, we will destroy ourselves."

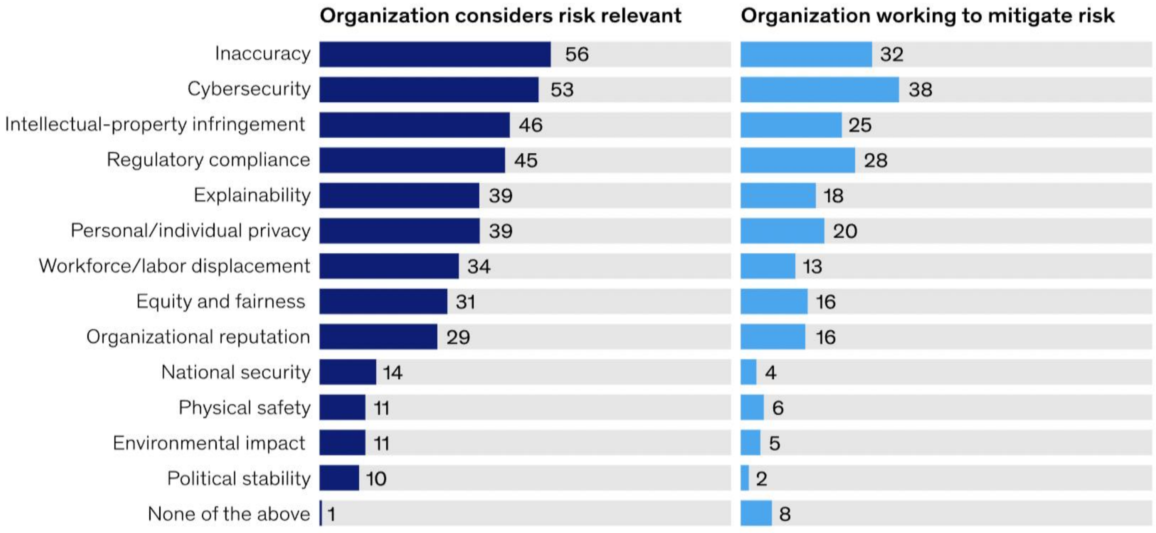
[10] In June 2023, OpenAI was reportedly sued on grounds of defamation by a radio host in the USA, who claimed that ChatGPT had generated a false legal complaint accusing him of embezzling money.

[11] See here.

- Hallucinations or wholly inaccurate and nonsensical outputs.
- Harmful content incl. hate speech, instructions for finding illegal content, graphic or violent materials, encouragement of self-harm, etc.
- Harms of representation, allocation, and quality of service.
- Disinformation and influence operations.
- Proliferation of conventional and unconventional weapons.
- Privacy and personal data leakage.
- Cybersecurity concerns.
- Potential for risky emergent behaviours.
- Interactions with other systems (e.g., if all banks used GPT4 to augment decision making, new macroeconomic risks could emerge).
- Economic impacts (in employment or rising inequality due to automation).
- Acceleration (or race to the bottom in terms of safety).
- Over-reliance of humans on foundation models as they improve in capability.

The risks of deploying generative AI by enterprises include the above concerns, which are not limited to LLMs but are also applicable to other types of models, such as generative text-to-image (TTI), etc. A recent McKinsey survey also reaffirmed these issues, highlighting that inaccuracy, cybersecurity and intellectual property infringements are key risks that organisations consider relevant, while regulatory compliance and explainability are also risks that companies want to work towards (see chart below).

Generative AI–related risks that organizations consider relevant and are working to mitigate,
% of respondents[1]



[1]Asked only of respondents whose organizations have adopted AI in at least 1 function. For both risks considered relevant and risks mitigated, n = 913.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

Our understanding of the unintended effects of generative AI use across sectors is still in its nascency, even as more research is being undertaken in this area. For example, the International Monetary Fund recently highlighted that decisions made by financial institutions based on generative AI could be susceptible to herd mentality and mispricing risks if based on public sentiments during market euphoria, or even generate solvency and liquidity risks if AI-driven models are inadequately trained towards financial risk management. In May 2023, one such real world economic harm became apparent when a fake image of smoke bellowing over the Pentagon briefly caused US stock markets to dip significantly.

Separately, there is also the centrally relevant question of power, and impact of generative AI on the fundamental rights of citizens that are the eventual service recipients or even users themselves. In this context, three specific types of risks have been identified[12]: (i) long term impact/consequences including its impact on the nature (and subsequent erosion) of human agency, and the human tendency to anthropomorphise generative AI technology if it displays sufficiently high levels of competence and likeness for human behaviour[13]; (ii) the impact on employment markets and job loss; and (iii) the spread of misinformation and toxic content.

Incidentally, the safety challenges identified by OpenAI already lists over-reliance by users as a distinctive concern. There is also a related risk of "model collapse" where synthetic training data accelerates "hallucinated" inaccuracies and pollutes the online information environment[14]. Given the extremely opaque and limited understanding of unsupervised learning, Dr Robin Hill at the University of Wyoming recently mused in an article titled "Bigger than a Blackbox" that humans are puzzled when AI recognizes a picture of a dog in a different way than a human would. This is not a new observation, as for example in 2016, when Lee Sedol was beaten by AlphaGo, commentators also observed a perplexing, "non-human" style of play. However, increased reliance on foundation models coupled with their intrinsic risks such as embedded bias and privacy challenges, may result in further loss of human agency and adversely impact mental integrity.

Another major challenge has been the prevalence of egregiously harmful types of generated content that are outputs of foundation models: deepfake pornography, child sexual abuse material, content generated towards incitement or hate speech designed to harass, intimidate, and defame, homemade dirty bomb recipes, operating details of nuclear power plants, and so on. While these types of harmful content should be moderated, the effectiveness of moderation techniques remain ambiguous, given the vast reservoir of unsupervised learning for generative AI services to tap into and generate tricks to "jailbreak" past guardrails.

[12] Dag Elgesem, 'The AI Act and the risks posed by generative AI models', NAIS 2023.

[13] In June 2022, Blake Lemoine, a Google engineer, did exactly this when he claimed that Google's large language model LaMDA had become conscious. In a more tragic example in March 2023, a man in Belgium committed suicide after six weeks of exchanges with an AI chatbot named Eliza.

[14] See this news report, and Shumailov et al (2023), "The Curse of Recursion: Training on Generated Data Makes Models Forget", University of Oxford, etc.

Consequently, when foundation models are used by bad actors to create and spread harmful content with the intent to sow democratic discontent and division, it infringes on several fundamental rights including the right to free speech and freedom of thought and expression. Additionally, the issue of deepfakes has surfaced alongside the pandemic, and remains a challenge with the improvement and release of multiple image-generating models, and their impact on democratic processes are yet to have been fully felt.

2024 could prove to be a watershed year, with presidential and/or general elections scheduled all over the world (e.g., United States, UK, South Africa, India, Taiwan, etc.) as well as the European Parliament elections. It remains to be seen how large the impact of generative AI could be on democratic processes around the world. However, if recent historic events are an indication[15] - sounding alarm bells vis-a-vis the advancements in AI and its impact especially in places where elections often turn violent, cannot be ignored.

For example, in the United States, generative AI has already begun to see use cases in political campaigns, such as these viral images of President Trump being arrested. The impact of deepfakes on democratic erosion has been studied and documented by Pawelec (2022)[16]; European Parliament STOA (2021); Fallis (2021)[17] while several philosophers and notable personalities have warned against what is being termed as an upcoming “information apocalypse” or the “epistemic threat” of deep fakes.

One of the most concerning fears is that an information overload will mean that videos will lose relevance and meaning for citizens, eventually eroding societal and democratic fibre over time. Meta has created policies against manipulated media under certain conditions on its platforms as early as 2020, and Google has recently announced that all political ads containing AI generated content will require a clear disclaimer. Given that there is also the related issue of foreign actors conducting influence operations with the objective of election manipulation which is often compounded by the increased use of generative AI. The US appears to be specifically cognizant of this issue, recognizing human vulnerability to weaponized misinformation and is at the forefront of regulation in this particular space, towards prohibiting or restricting misrepresentative content, with the State of New York recently enacting a law that prohibits deep fake ‘images’.

However, deep fakes are not merely an impediment for Western democracies but could have an equally problematic impact in other parts of the globe. For example, a recent investigation reported

[15] Many examples of political violence can be traced that have been exacerbated by disinformation. In the wake of political violence on 6 January 2021 on Capitol Hill, several of those convicted for violence used misinformation in their legal defence. In 2018, the role of disinformation in political and ethnic violence in Myanmar, prompted a UN investigation.

[16] Pawelec M. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. Digit Soc.;1(2):19. (2022)

[17] Fallis, D. The Epistemic Threat of Deepfakes. Philos. Technol. 34, 623–643 (2021).

on the proliferation of religious chatbots in India that included concerning and violent takes that could potentially result in real world harm. India's Prime Minister recently raised concern about deep fakes after a purported deepfake video of him circulated on social media. In China, there are increased restrictions on the use of foundation models, which must only create and engage with content that "adheres to the core values of socialism" and respect other existing laws on hate speech, discrimination, national security, etc.[18] While the use of technology to steer majoritarian narratives is concerning, it is worth investigating how such foundation models, if exported to other countries, could be used to ensure outputs aligned with state propaganda.

A final issue to consider is copyright. The emergence of generative AI has disrupted current copyright protections (centred mainly on text and data mining exceptions) and forced lawmakers to reconsider how they categorize and assign responsibilities to providers and users of AI systems. From the input perspective, the main issue relates to the activities needed to build an AI system. In particular, the training stage of the AI tools requires the scraping and extraction of relevant information from underlying datasets, which often contain copyright protected works. In the EU, these activities are mostly regulated by text and data mining (TDM) exceptions under the Copyright Directive. The commercial TDM exception provides an opt-out mechanism for rights holders. In the US, the system is governed by the fair use doctrine. Looking at other jurisdictions, the Chinese approach calls for foundation model providers to "respect intellectual property rights" at a high level, while the Japanese Copyright Law provides cover for any copyrighted work used for machine learning purposes. This allows general purpose AI model providers to continue innovating, while ensuring their models are not scraping data that has been protected. This is an example of how policymakers will seek to strike the balance between the impetus of protecting innovation, while also protecting creators rights. In the US, recent court cases[19] and the ensuing legal uncertainty has meant that there is a clear need for copyright laws to catch up with technological progress and make them compatible with the use of generative AI for original content generation.

MITIGATING NEGATIVE EXTERNALITIES

Enhancing safety, privacy, cybersecurity & fundamental rights

One of the key questions for policymakers to grapple with is how to ensure general purpose AI models are compliant with existing laws and frameworks around important issues like safety, data privacy, and cybersecurity. Linked to some extent to the problems articulated above on harmful content, AI safety becomes more relevant when foundation models are deployed in industrial contexts, or in high-risk environments where safety needs are paramount. For instance, if generative AI is used by healthcare providers to improve diagnostics or patient experience, utmost care must be taken to

[18] : China has moved quickly to pass laws on generative AI, as early as in August 2023. Please find an English translation of the rules here.

[19] See for e.g. "Judge pares down artists' AI copyright lawsuit against Midjourney, Stability AI" (30 October 2024, Reuters)

ensure sensitive personal data is not shared to third parties without a lawful basis, and to also develop technical solutions that evaluate robustness, accuracy, and cybersecurity of the service. However, technical and architectural solutions might lie further down the AI value chain, not at the foundation model layer, but at the integrator or provider level. In fact, companies have already announced innovative technical solutions for enterprise use of generative AI. This additional "AI safety" moderation layer, as well as practices like red-teaming to refine model integrations, could ensure that new iterations and applications are safe for both commercial and personal use.

From a regulatory perspective, it will become important to clarify how existing laws on product safety, consumer rights, privacy, and human rights would apply to generative AI service providers and users, and then incentivize the creation of appropriate safety mitigations at the right levels of the architecture. This can be achieved not only through sectoral regulatory reform, but also through clearer overarching guidelines for the compliance of generative AI technologies by entities. Dealing with *ex post* effects of generative AI or relying on companies to self-regulate poses risks, as commercial interests might sometimes overshadow ethical and societal considerations, and ex-ante practices that are regular and continuous (e.g. internal red-teaming) should be encouraged [20]. Policymakers and enforcement agencies have a pivotal role in proactively setting guardrails and ensuring that the deployment and use of generative AI align with broader societal values and protections, especially in areas as critical as human rights.

Moreover, the concentration of AI resources among a few multinational tech companies and governments raises concerns about the equitable distribution of the "digital commons[21]." When the "digital commons" are utilised to develop and train proprietary AI models, it brings forth questions about how the value generated from these common resources should be distributed. Moreover, there are concerns related to data governance, such as who should control how data is used, especially when tech companies aim to commercialise AI models trained on data sources that are specifically sensitive, like indigenous knowledge. The control over the development and training of generative AI models and their usage is crucial, as those in control can create dependencies, set terms of use, and decide access, potentially leading to power imbalances and market dominance. Furthermore, the rapid adoption and deployment of generative AI systems, as evidenced by the rise of models like ChatGPT and the subsequent public discourse, underscores the urgency of addressing these challenges.

There is also legitimate concern around personal data, privacy, and confidentiality. As people begin to use generative AI tools at scale, there is a high likelihood of data leakage (e.g., confidential corporate data) and unintentional exposure of such information. An additional risk is one of personal data

[20] Baxter, K; Schlesinger, Y. Managing the Risks of Generative AI. Harvard Business Review. (2023)

[21] The "digital commons" refers to the vast amount of information found openly online, as it is a body of resources where all can be contributors, from individual pieces of data to the public infrastructure and resources of the internet. See for e.g. Huang & Siddarth (2023) "Generative AI and the Digital Commons"

breaches or extraction, where items related to people that were hitherto not easily available online were now easily available after being scraped and summarised. Further, it should be clear that those individuals that are affected by and interacting with these systems can exercise their legal rights, including access, rectification, and erasure of personal information (depending on the jurisdiction), as well as the possibility to refuse to be subject solely to automated decisions that have significant effects. There have been several concerns across EU and G7 countries that foundation models contradict privacy laws like the EU's General Data Protection Regulation (GDPR), and Italy's data protection authority briefly suspended access to OpenAI's ChatGPT following such concerns.

The final issue to consider is cybersecurity. Tools like foundation models can be useful for enhancing trust, safety, and security of digital ecosystems. However, they can equally pose serious cybersecurity risks. For instance, most of today's access control at critical infrastructure sites use voice or facial recognition, which could be compromised by content from generative AI tools, while large language models (LLMs) could be used to draft phishing emails and parse through targets as part of social engineering attacks on unsuspecting employees. Other cybersecurity risks include vulnerabilities that exist from training data (for e.g. exposure of confidential information), as well as providing workarounds or "jailbreaks" for existing cybersecurity systems. OpenAI's technical report for GPT4 observes that it was not only capable of helping draft phishing emails, but also managed to get a human "task rabbit" to solve a CAPTCHA on its behalf, clearly showing the ability for "risky emergent behaviours" [22] that could be detrimental to cybersecurity.

Approaches to mitigate risks

Increasingly, policymakers should focus on mitigating the negative externalities from introducing generative AI services into different social and economic domains. The introduction of automated decision-making has inevitably caused job losses and exacerbated inequalities. This also means policy approaches to mitigate risks might lie at all stages, with all stakeholders of the AI value chain.

For example, an area that has just begun to receive attention is the use of foundation models in public services. While such models admittedly have numerous benefits[23] for this sector, including as a force multiplier, an information assistant, as well as displaying deep data analytic capabilities to increase efficiencies and streamline navigation of services by augmenting accessibility, the risks discussed above pertaining to foundation models also require an equivalent consideration in this context. Specifically, given the sensitive nature of citizens' datasets that are available to governments, combined with the profound implications that the previously identified risks (such as embedded bias or hallucinations) can have in a public service context, it is critical to unpack how foundation models are *actually* learning, and to articulate the different ways to identify how this corresponds to the way humans understand the world. This becomes increasingly relevant as policymakers and enterprise

[22] See p. 54 and 55 of ChatGPT's technical report. URL: https://cdn.openai.com/papers/gpt-4.pdf
[23] See here and here.

users alike seek to improve model explainability and add guardrails for different providers or use cases of foundation models. For example, the recent US Executive Order seeks to do this by creating guardrails for government agencies' procurement and use of artificial intelligence .

The integration of AI-powered technologies within the value chains of private sector entities also raises concerns about potential market oligopolisation in sectors such as advertising, social media, and entertainment. Additionally, AI models have the capacity to introduce novel political and societal risks by propagating inaccurate and potentially harmful content, thereby jeopardizing democratic principles and human rights. The output generated by foundation models also has the potential to impinge upon fundamental rights, exacerbating issues of discrimination and bias. Such concerns have spurred academic and political discourse centered on principles like purpose limitation, data minimization, accountability, and trust. Moreover, the advent of generative models challenges not only economic and political structures but also legal frameworks, particularly in the realm of intellectual property and competition law, which must adapt to the creative capabilities of AI. Smaller and less powerful players need to be carefully listened to in order to both facilitate compliance and stimulate innovation.

Generative AI's increasing integration into various sectors also highlights the importance of rigorous audits by independent entities, including researchers, enforcement agencies, and third parties. Such audits not only identify and rectify potential biases and discriminatory outcomes in the models, but they also ensure the ethical sourcing and utilisation of training (personal) data. Furthermore, with most data protection and privacy regulations currently emphasising provisions regarding explainability and transparency measures, these audits verify compliance with updated legal frameworks and assess the integrity of both data practices and the algorithms themselves. As generative AI advances, its auditability and transparency become a crucial requirement, essential for maintaining public trust and guaranteeing their responsible deployment.

At the developer level, a feared concentration of power (of both open and closed providers) could lead to an increasing and undesirable "centralization of collective intelligence and inputs", as well as the "centralization and privatisation of decisions on what kinds of downstream harms are permissible"[24]. Economist Joseph Stiglitz recently warned how unregulated AI could worsen inequality, specifically citing how generative AI could result in more job losses without policy interventions, including for education and skills training. After industrial automation replaced blue collar physical labour at manufacturing sites, the fear is that repetitive white-collar jobs (e.g., secretaries, copywriters, translators, paralegals, etc.) could be similarly under threat.

[24] Esposito, M, et. al. "The Dark Side of Generative AI: Automating Inequality by Design", California Management Review, University of Berkeley. (2023).

Private sector investment in AI in 2023 by country
(US$ billions) Source: 2023 AI Index, Stanford HAI

United States	***47.36***	*France*	*1.77*
China	***13.41***	*Argentina*	*1.52*
United Kingdom	*4.37*	*Australia*	*1.35*
Israel	*3.24*	*Singapore*	*1.13*
India	*3.24*	*Switzerland*	*1.04*
South Korea	*3.10*	*Japan*	*0.72*
Germany	*2.35*	*Finland*	*0.61*
Canada	*1.83*		

Mirroring these winner-take-all characteristics at the international level, the current burst of innovation in generative AI is being driven by tech firms and talent from a handful of nations, with only the United States leading the way, while China also makes meaningful strides in developing large foundation models (see above table). This concentration in technology development could serve to exclude local voices or concerns from most of the developing world and is also of concern for the fair access to and use of such AI by smaller companies.

As we have seen in other instances of algorithmic decision-making, the bias present in both training data and outputs of foundation models could also exacerbate stereotypes against women or ethnic minorities and could indirectly contribute to real world violence. The corollary to these risks would be the risk of overreliance on foundation model outputs to make important decisions. Clear responsibilities for developers and providers of foundation models are needed around removal of bias and discrimination from training data. Societally important systems that rely on these outputs should not infringe on fundamental rights, and foundation models in these select cases should not be linked to automated decisions.

Crucially, remedies must also be put in place to ensure the use of generative AI creates net positive outcomes for sustainability and environmental protection. Stanford's AI Index states that CO2 equivalent emissions (tonnes) footprint of GPT3 is almost 100 times that of an average human life's carbon footprint, and 500 times that of a flight from New York to San Francisco[25], and it will be important to track "computing-related impacts due to the manufacturing of hardware and devices as well as electricity consumption; indirect impacts of deploying the models, and system-level impacts on other domains"[26]. The use of generative AI is likely to exponentially rise, and so policy interventions

[25] Maslej N, et. al. "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.
[26] Kaack, L. et al. "Aligning artificial intelligence with climate change mitigation". HAL Open Science, Published in Nature Climate Change (2022), 1–10.

that focus on quantifying and improving energy efficiency and sustainable use will be more important than ever before.

A final consideration in terms of the path forward for guardrails has to take into account the development and innovation potential of generative AI technology. At present, there are two main ways foundation models have been made available, as an open-source release, and through an API (i.e. a limited release). In each case, developers hope to monetise the release based on some criteria, including whether they host the compute and re-training on their own servers, and whether or not they can charge for additional services "easy access to computational infrastructure or for other services such as fine-tuning or maintenance services"[27.] Both ways developers bring generative AI into the market involve specific tradeoffs between, on the one hand, incentivising innovation and, on the other hand, effectively mitigating risks and increasing transparency, accessibility and accountability. API access enhances content (and therefore moderation) control but centralizes power over general purpose AI systems. Open-source releases aim for equitable access and control, allowing a broader community to study, adapt, or improve models but it may also contain risks such as discriminatory outputs and lead to legal challenges[28].

In addition, while the two above methods remain the predominant way by which general purpose AI could be introduced, there are myriad other potential business models, including customer-specific models, white-label foundation models, creation of generative AI enabled application marketplaces, sale of custom pre-coded developer kits, etc. In any event, Generative AI technology remains very nascent even from a profitability and business model perspective and there have been no dominant paths for monetisation observed.

While a large part of these concerns can and should be addressed via governmental action at the national and international level, multistakeholder and even meaningful industry action can also generate suitable answers to some concerns. Industry best practices and structures can be an important first step towards minimising risks and harms, while increasing the benefits of using generative AI, and they should be encouraged by policymakers especially if created not just as an excuse to skirt potential legislation. As well, with proper guardrails, codes of conduct and practice (for e.g. the EU Code on Disinformation), produced via multi-stakeholder structures, which feature civil society actors and industry together (sometimes even including governments) can rely on unique policy inputs and knowledge, and feature strong enforcement, thus providing an inclusive and flexible solution till governments have made progress on the ground rules.

[27] Küspert , S ; Moës, N. ; Dunlop, C..”The value chain of general purpose AI”, Ada Lovelace Institute. (2023)
[28] See article "New AI systems collide with copyright law" (BBC News, 2023)

PATH FORWARD FOR GOVERNANCE OF GENERATIVE AI

Advancing a Risk-Based Legislative Approach for Generative AI

A consensus has emerged within various stakeholder groups concerning identifying several inherent risks associated with generative AI. The multifaceted and interdisciplinary challenges arising from the utilization and deployment of generative AI are rooted in an array of factors. The escalating risk of misuse by malicious actors, the predominantly private-sector-driven development of generative AI, and the dynamic and unpredictable evolution of this technology underscore the necessity for concerted action. Leading economies have come to the realization that there is a pressing need for robust, adaptable, and substantive frameworks governing the use and deployment of generative AI. While various governmental initiatives have been set in motion[29] globally, there exist divergent viewpoints[30] on how to tackle the challenges that generative AI poses.

However, policy discussions on the nature of regulatory compliance has become a complex issue demanding attention from both industry and government stakeholders. Governments all over the world have introduced a large number of concepts to describe “foundation models”, including in the EU AI Act discussions, and the EU parliament position has been to focus on including requirements for foundation models by categorising them as “general purpose AI models” (‘an AI model, including when trained with a large amount of data using self-supervision at scale, that is capable to [competently] perform a wide range of distinctive tasks regardless of the way the model is released on the market’). The US Executive Order uses “foundation models” but adds the broad notion of “dual use” to create additional obligations on developers (see below). In terms of definitions, the terms “generative AI” refer to the overall technology (including both foundation models and their applications) while “foundation models” and “general purpose AI models” tend to be used interchangeably.

Presently, the most advanced stage of policy development can be observed in the EU AI Act. This legislative instrument, grounded in the principles of product safety, establishes a proportionate risk-based approach for AI systems[31], and suggests, in intermediary drafts, a related approach for foundational models[32] and subsequently generative AI[33].

For foundation models, the European Parliament’s position on the EU AI Act delineates eight essential ex-ante compliance mechanisms. These encompass risk identification, data governance, ensuring performance at appropriate levels, predictability, safety, cybersecurity measures, environmental risk monitoring and mitigation, collaboration with downstream providers, the establishment of a quality management system, and the provision of technical documentation for a period of ten years. Further,

[29] See above table on page 6, *Recent legislative developments on generative AI*
[30] See for instance https://www.accessnow.org/eu-regulation-ai-risk-based-approach/
[31] Coupled with codes of conduct for non-high-risk AI systems
[32] https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition
[33] https://www.holisticai.com/blog/foundation-models-gen-ai-and-the-eu-ai-act

general purpose AI providers could have additional obligations, including transparency safeguards, ensuring that the content generated adheres to EU legal standards, and the provision of a summary of training data pertaining to copyright.

The EU AI Act also introduces a classification system consisting of three risk categories, based on the potential dangers posed by all AI applications. These categories include unacceptable risk, high-risk, and limited or low-risk applications. Generative AI providers must assess where their systems fall within these categories and adhere to the corresponding obligations as well. The first category prohibits AI systems that fail to conform to EU legal principles and values, including fundamental rights of citizens. High-risk applications encompass technologies that can significantly impact citizens' lives, such as biometric identification, education and employment assessment tools, law enforcement applications, and service access. Finally, low-risk technologies are associated with specific transparency obligations that users must be made aware of, including applications like video games, chatbots, or spam filters.

This “risk-based approach” aims to achieve several specific objectives aligned with the aforementioned risks:

- Policies in line with the risk-based approach seek to foster innovation and competitive potential within the private sector, particularly in response to the rapid evolution of AI.
- The risk-based approach establishes clear and enforceable liability obligations and safety measures.
- It differentiates among the various actors throughout the AI value chain, enabling precise allocation of compliance responsibilities and potential liabilities.
- This approach enhances legal certainty, thereby preventing market fragmentation and promoting equitable competition.
- A risk-based approach mitigates the impact of biased or discriminatory outcomes resulting from algorithm-based decisions, striking a balance between innovation and fundamental rights.
- This approach also has the potential to ensure the continuity of existing data protection and privacy principles.

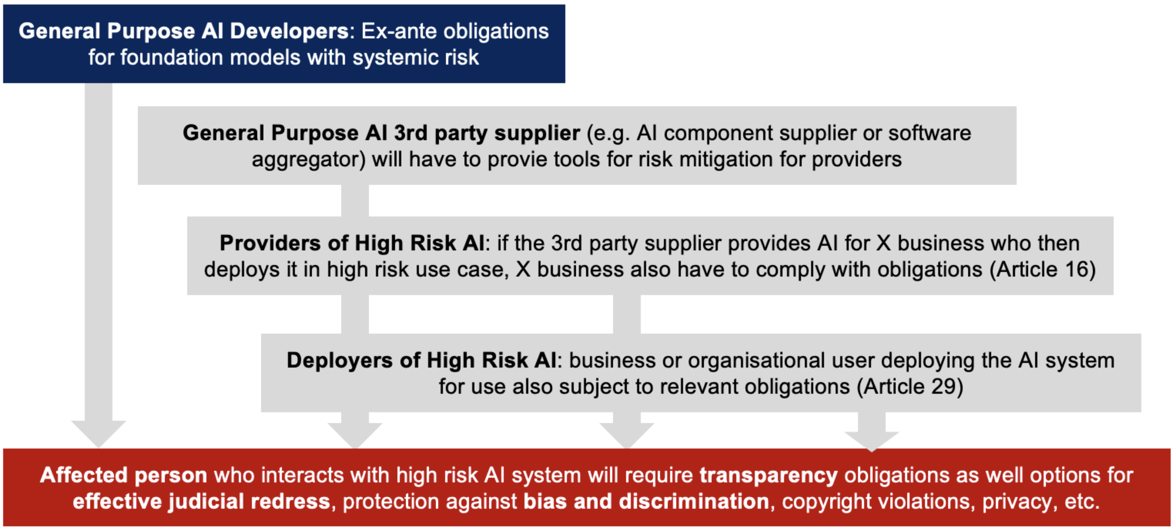
The recent US Executive Order creates a new definition of a “dual-use foundation models" as “an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters (…)”. This definition also includes a number of factors like for e.g. substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; helping enable “offensive cyber operations” or “deception or obfuscation” as criteria to classify “high risk” foundation models and is evidence of a rudimentary “risk based approach” being applied.

There is a need to build on the idea of “AI adequacy” as the next step to expanding on the risk based approach, given the limitations of the initial EU AI Act proposal. The conformity assessment mechanism (which is an ex-ante regulatory tool, grounded in EU safety regulation) only applies to high-risk AI systems, leaving lower-risk systems largely unregulated. This means that potentially harmful AI systems may still be deployed without undergoing the necessary risk assessments and conformity checks. One common example is AI systems that can interact with humans, like chatbots (especially when these chatbots can affect the emotions of end-users), or even AI systems producing “deep fakes” (highly realistic images, audios or videos that are artificially generated). These applications are considered “limited risk” by Article 52 and so no ex ante conformity assessment is needed for them. (Malgieri & Pasquale, 2023).

The EU AI Act trilogue discussions go further with a leaked compromise text containing ex-ante “justifications” for “general purpose AI models with systemic risk”. Therefore, such a proposal would create two levels of ex-ante risk-based obligations, one at the level of the foundation model provider, and another for the deployer in those defined high-risk use cases. The “systemic risk” here emanates from the “frontier capabilities” of certain general purpose AI models which is determined by a set threshold of amount of compute used for training (measured in floating point operations per second or FLOPs) with the European Commission empowered to regularly update this threshold. This idea of “systemic risks” mirrors EU efforts in the realm of content regulation (e.g., the designation of “very large online platforms”) and takes into account the main characteristic which contributes to systemic risk – in the case of social media platforms, this is number of users and spread of harmful content, while in the case of foundation models this can be adequately explained by the complexity, sophistication and vastness of the training the model has received. Incidentally, the inspiration to take into account amount of compute used for training in FLOPs can be found in the US Executive Order, which recommends the same threshold as the leaked EU compomise text (10^{26} FLOPs). Note also that the US Executive Order makes a further distinction between single and distributed computing systems, and creates a lower threshold of 10^{20} FLOPs for computing clusters that are physically co-located in a single data center.

The establishment of common risk-assessment standards by governments hinges on their adept incorporation of the "risk-based approach" articulated in Principle 5 of the G7 Code. Moreover, taking inspiration from the underlying framework, rationale, and architecture of the EU AI Act is important to drive regulatory convergence. An illustrative instance of can be found in the United States before the recent Executive Order wherein the National Institute for Standards and Technology (NIST) formulated the AI Risk Management Framework (version 2, released in August 2023) with the risk-based approach in mind. Notably, the recent U.S. AI Executive Order has taken this further and tasked NIST with the adaptation of this framework to accommodate generative AI, achieved through the creation of a corresponding "secure software development companion resource."[34]

The European Parliament is pushing to include an ex-ante “licensure” approach for foundation models that *could* be used in high risk contexts. As such it is clear that foundation models are moving towards increasing complexity and require explainability for decision making, especially when deployed in high-risk use cases. However, this is not always possible: while for traditionally data-based decision-making, it might be easier to give adequate explanations, in more complex AI-based decisions, it might be hard to reach such a level of explainability (Malgieri & Pasquale, 2023). This type of an ex-ante approach that focuses on foundation model developers is also endorsed in the US Executive Order, with the note that foundation models will be categorised as “dual use” even if they are provided to end users with technical safeguards that attempt to prevent unsafe use. This is an implicit recognition that safety measures at the deployer level might be insufficient to prevent the risks that could have been more easily addressed at the developer level further up the value chain[35]. The intent behind the European Parliament position is given in the chart below:



General Purpose AI Developers: Ex-ante obligations for foundation models with systemic risk

General Purpose AI 3rd party supplier (e.g. AI component supplier or software aggregator) will have to provie tools for risk mitigation for providers

Providers of High Risk AI: if the 3rd party supplier provides AI for X business who then deploys it in high risk use case, X business also have to comply with obligations (Article 16)

Deployers of High Risk AI: business or organisational user deploying the AI system for use also subject to relevant obligations (Article 29)

Affected person who interacts with high risk AI system will require **transparency** obligations as well options for **effective judicial redress**, protection against **bias and discrimination**, copyright violations, privacy, etc.

Given the powerful nature of some foundation models, a type of ex-ante licensure regime makes sense to require baseline requirements that address concerns like the creation and spread of harmful content, transparency, explainability for the user and to enhance security. However, a technical threshold that is set in stone in law naturally risks being rendered obsolete due to rapid technological evolution, and will therefore require close monitoring and regular updation.

[34] See Section 4.1 of the US Executive Order which tasks the Director of the NIST to within 270 days, develop “companion resources” to the AI Risk Management Framework, NIST AI 100-1, as well as Secure Software Development Framework for generative AI and foundation models.

[35] As part of their negotiating position, the European Parliament advocated an approach where responsibilities are shared by all actors in the generative AI value chain. See this OECD AI blog “A law for foundation models: the EU AI Act can improve regulation for fairer competition”.

Towards AI Adequacy as a Regulatory Tool for Convergence

To balance out space for innovation for a nascent technology like generative AI, and to avoid placing an innovation penalty on foundation model developers, while at the same time ensuring that all the risks outlined above are sufficiently addressed is a difficult exercise for policymakers all over the globe. This is why it will be important to engage with industry and create a common understanding of "AI Adequacy" for foundation model providers, as well as articulate clear obligations, enshrined as codes of conduct and embedded through international technical and governance standards.

Some civil society organisations have warned against the exclusion of these systems from the EU AI Act, as this means that the burden of making these systems compliant with the regulation falls entirely on the users of the AI systems instead of developers[36]. In addition, scholars have recently pointed out that the EU AI Act (European Commission proposal) is an "example of ex-ante justification of AI systems" where the AI providers need to "justify", through some technical documentations, that their system is adequate according to specific principles (transparency, accountability, human oversight, accuracy, security)"[37].

Adding an ex-ante layer of obligations for foundation or "general purpose AI models with systemic risk" could sufficiently enhance this notion of "adequacy" while ensuring industry and civil society participation in developing the codes of conduct could simultaneously preserve innovation. One example of this is by preserving open-source releases and thereby prioritising access of generative AI to startups. Malgieri & Pasquale's analysis is based on the European Commission proposal, and borrowing their approach of using the EU General Data Protection Regulation's own protections as inspiration to bolster AI regulation, we also propose the idea of establishing "AI Adequacy" across borders through standardised tools (similar to data protection adequacy tools like SCCs) that could also be useful to enhance regulatory certainty.

In the EU, general purpose AI models used in high-risk use cases could require third-party conformity assessment, and it will be important to develop common standards on transparency measures, risk assessment, watermarking of generated content, etc. Therefore ensuring compliance with principles of fairness, lawfulness, transparency, etc. will be a key focus in terms of next steps for both European regulators and European standards organisations. In addition, industry should follow suit by working on developing and implementing appropriate codes of conduct based on these mutually agreed upon standards.

The work done as part of the G7 Hiroshima process is therefore very timely, as it seeks to establish exactly the kind of legislative framework that could include as a next step, concrete work on common

[36] : For e.g. position paper from the Future of Life institute titled "General Purpose AI and the AI Act" (2022).

[37] : Malgieri, G., Pasquale, F. (2023), "Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology", URL: https://www.sciencedirect.com/science/article/pii/S0267364923001097

taxonomies and standards to ensure compliance with the principles articulated, including on implementing a risk based approach for developers of “advanced AI systems”.

Conclusion: Path forward for the G7 Hiroshima AI process

It is imperative that governments now focus on developing common standards as guidance for industry that properly defines the risk-based approach in Principle 5, as well as to implement measures like enhancing transparency through reporting in Principle 3, cybersecurity controls articulated in Principle 6 and provenance and authenticity (watermarking, etc.) described in Principle 7.

Ensuring the implementation of these measures through the development of common standards will help achieve regulatory and technological convergence, increase legal certainty, and reduce compliance costs for industry, especially smaller players. Governments should work towards developing these "AI adequacy principles" in a similar vein to the concept of data protection adequacy in the EU General Data Protection Regulation (GDPR) as well as build the tools for compliance by providing guidance on common processes and standards.

Finally, to help with moving ahead from the 11 principles agreed upon in Hiroshima, we also propose the following policy actions for key countries:

G7 Hiroshima AI Principles for companies		Policy actions/recommendations
1	Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.	Ensure that risk-based approach is fundamentally applied in key jurisdictions including in the US, EU and other countries (Japan, India, etc.). This includes increasing policy coordination at the multilateral level.
2	Patterns of misuse, after deployment including placement on the market.	Devise risk assessment standards through international standardisation bodies and strong bilateral/regional coordination (e.g., US-EU Trade and Technology Council).
3	Publicly report advanced AI systems capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.	Link reporting to existing reporting mechanisms for cybersecurity and data protection. Encourage and standaridse practices including publication of a “model card” (similar to a product specs or data sheets) for substantial model releases.

4	Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.	Transparency reporting duties should be devised. Governments should facilitate industry-led standardisation for incident reporting and coordinated vulnerability disclosure practices.
5	Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems.	Focus on tasking standards bodies to devise harmonised standards for risk assessment, cybersecurity and privacy protection for foundation model providers, especially in high risk use cases.
6	Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.	Governments should provide further guidance on internal cybersecurity practices (e.g. red-teaming) while companies should regularly share results of any third party security audits, as well as information about cybersecurity safeguards.
7	Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content	Work with industry and civil society to develop common standards through codes of conduct and through tasking standards bodies for watermarking and other techniques for identification. Require deployers of GPAI systems to enhance transparency, especially if tools are available widely for use.
8	Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.	Work with researchers to make data available for research on the risks outlined in this paper and invest in tools for effective mitigation. Invest to build capacity in civil society and academic institutions.
9	Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education.	Invest in skills development to build the talent pool and enhance labour mobility. Build awareness of the use of generative AI in different sectors and invest in key use cases to solve challenging problems.
10	Advance the development of and, where appropriate, adoption of international technical standards	The EU and US should task their respective stnadards bodies (NIST, JRC as well as bodies like CEN/CENELEC and ETSI) to develop bespoke technical standards for general purpose AI providers for demonstrating compliance, with third-party conformity

		assessment reserved to deployments in high-risk use cases.
11	Implement appropriate data input measures and protections for personal data and intellectual property	Link these measures to developing a notion of “AI Adequacy” that will allow foundation model providers to place their products on the market after rigorous self-asessment for data input quality, and other ex-ante mitigations through a code of practice.

Appendix: CERRE’s recommendations for the G7 Hiroshima AI process

Governments must offer clear guidelines for deploying generative AI in high-risk contexts, assigning specific responsibilities to key players in the AI industry. Shared responsibility is crucial for governments to keep pace with rapid technological advancements. Therefore, a flexible, multistakeholder governance model is essential to promote an effective "risk-based approach" for generative AI. These ideas underpin the policy recommendations submitted in October 2023 by CERRE as part of its response to the European Commission’s “stakeholder survey" on the draft International Guiding Principles for organisations developing advanced artificial intelligence (AI) systems”.

The G7 Hiroshima AI Process represents a vital step in the policy discourse on generative AI, to ensure its responsible development and deployment while reaping its transformative economic and social benefits. To achieve convergence on the implementing regulatory framework for generative AI, a global, inclusive, multi-stakeholder approach must continue to be pursued.

Considering the above, we propose the following four concrete policy recommendations as a path forward to build regulatory convergences on generative AI globally:

1. Governments should build on a “Risk-based Approach” when devising their own rules

G7+ countries should adopt a risk-based approach for all advanced AI systems in line with Action 5 of the code of conduct. This approach should place emphasis on high-risk use cases. This involves foundation model providers and other actors (deployers, distributors) through tiered obligations for risk assessment and mitigation depending on the level of their risks. More prescriptive rules should apply to situations at higher risks, such as financial advisory services or when deployed by public administration or judges, where citizens’ fundamental rights may be at risk. At the same time, it is imperative that governments devise the tools necessary to enable developers of “general purpose AI” or “dual-use foundation models” to conduct self-asssesments and comply withcodes of conduct.

2. Governments should focus on developing tiered, yet limited regulatory obligations for high-risk use cases.

To address the high-risk cases above, a tiered set of regulatory obligations should be established. This approach would include ex-ante fundamental rights impact assessments, stricter data input controls, advanced cybersecurity measures, and enhanced privacy, fairness, transparency, content moderation, and copyright measures. Governments should collaboratively determine, along with civil society experts and industry, criteria for identifying advanced AI companies, such as the compute power required to train foundation models (measured in FLOPs), number of active users, use in high-risk cases, and integration into existing designated "very large online platforms” or VLOPs. For all other cases, self-assessment should serve as the first line of defence, and international organisations can contribute to this by developing relevant governance and technical standards.

3. Governments should pick collaborative, multi-stakeholder governance models for AI.

The European Commission should take the lead in shaping global governance for generative AI by formulating standards for risk assessment, systemic platform designation, and other technical risk mitigation measures, such as watermarking while consulting with all relevant stakeholders, including representatives of highly impacted stakeholders. Systemic platforms should work closely with advanced AI companies to limit the production and dissemination of harmful content. G7+ governments should agree to set up a harmonised, ex-ante, multi-stakeholder, fundamental rights impact assessment process, one that is facilitated by regulators and industry. Governments should also focus on developing collaborative policy measures, including, next to regulatory collaboration, joint exploration of beneficial applications, creating regulatory and technical sandboxes, and investing in education for SMEs on compliance and AI innovation. In addition, policymakers and governments should invest in educational programs and workforce development initiatives to build a skilled AI workforce so that it remains relevant to a continuously evolving digital labour market. In general, it would be important to foster collaboration between academia and industry to bridge skill gaps.

4. Government should move towards developing “AI Adequacy” to build regulatory convergence globally.

As a next step to the articulation of the 11 principles or actions in the Hiroshima code of conduct, it will be important for G7 countries, as well as India, Brazil, etc. to agree on tools for implementation. In other words, countries will have to continue to work together to develop standards that will be comprehensive and sufficient to move towards legal convergence. This will require governments to specify “AI adequacy principles” for governance, which include measures for maintaining data quality, as well as to responsibly deploy AI products.

cerre Centre on Regulation in Europe

Avenue Louise 475 (box 10)
1050 Brussels, Belgium
+32 2 230 83 60
info@cerre.eu
www.cerre.eu
@CERRE_ThinkTank
Centre on Regulation in Europe (CERRE)
CERRE Think Tank