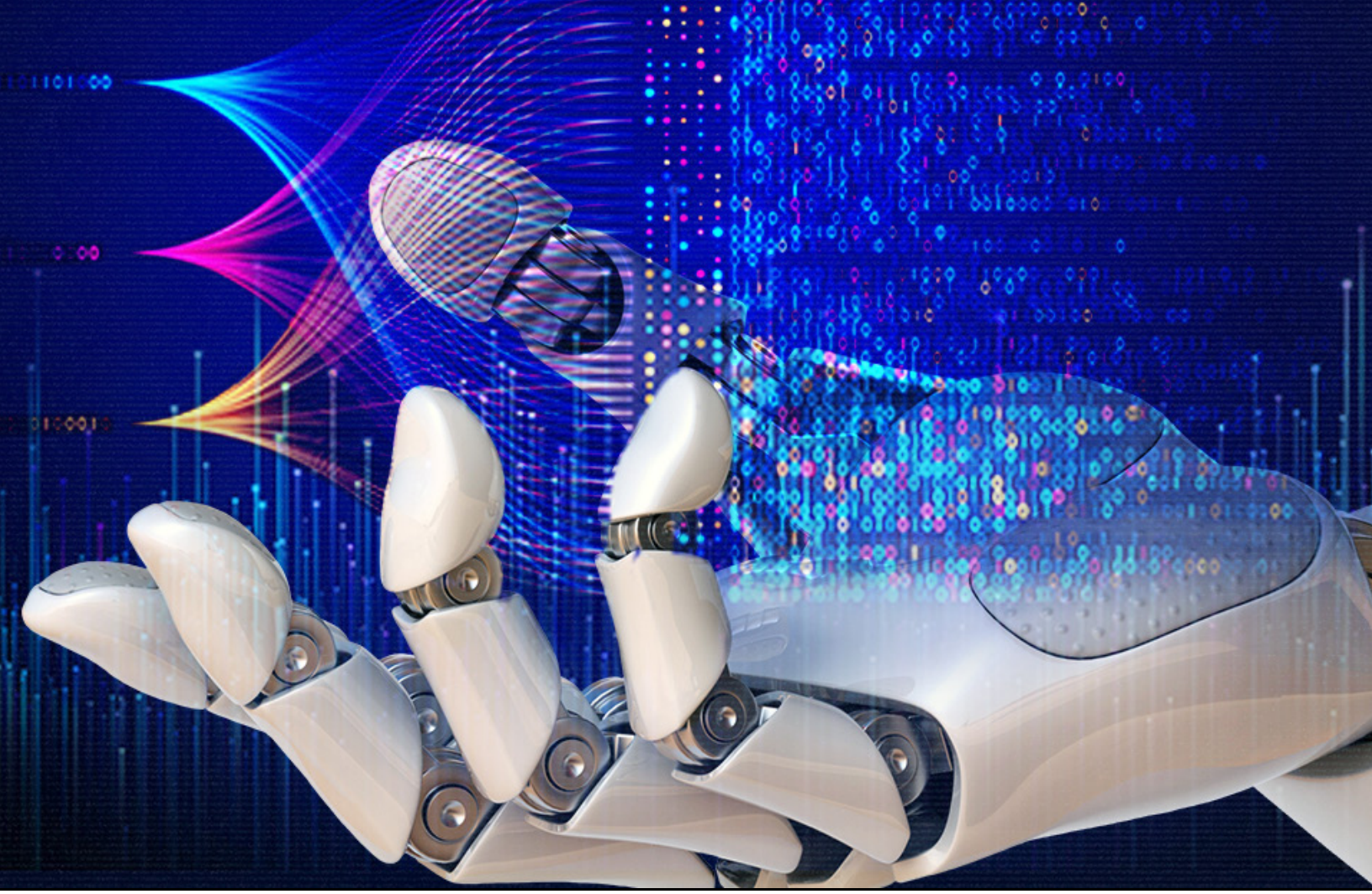


# Trustworthy AI Needs Trustworthy Data

## Why we need to talk about data if we want to unlock AI's full potential

**PREPARED BY EURASIA GROUP & LSEG**

10 FEBRUARY 2025

LSEG
Make more possible

eurasia group politics first

## Introduction

Artificial intelligence has shot to the top of the global agenda, as the international order is mired in what Eurasia Group founder Ian Bremmer calls a "geopolitical recession." Global efforts to agree on principles and guardrails for AI are taking place against a backdrop of intense geopolitical competition, disruption, and a deficit of international leadership.

Even in such an environment, global leaders' efforts to create governance frameworks for AI have been impressive. From the UN Global Digital Compact to the Paris AI Action Summit, governments have recognized the mutual imperative to ensure that AI's potential is harnessed safely and responsibly. Now, the challenge is to build coherence and consensus around the expanding web of AI initiatives and principles that have already been crafted by a wide range of organizations, including the OECD, UNESCO, the Council of Europe, the G7, and more. **Despite the flurry of global, regional, and national policymaking activity around AI in recent years, longstanding global governance challenges involving the use and handling of data—which is critical to ensuring trustworthy AI—remain unresolved**.

**It is often noted that AI runs on data. But data, and specifically data governance, are often treated as separate matters from AI in policy conversations**. As a result, multilateral discussions on the free flow of trusted data and global data governance have not fully benefited from the momentum surrounding AI. If anything, the trend risks moving in the opposite direction, with governments considering the creation of policy barriers and implementing data localization and data sovereignty measures that would challenge the movement of data. That would impede the ability of governments, regulators, industry, and society to harness the benefits of AI technologies. This trend must be reversed.

**To ensure AI can reach its full potential, we need to connect the dots between AI, data inputs, and data policy**. That means refocusing on ensuring the data used to power AI systems is accurate, trustworthy, and able to flow freely and securely—wherever and whenever needed. It means breathing new life into global initiatives designed to achieve these objectives, despite growing geopolitical tensions. And it means empowering the access, use, and reuse of data, with robust controls in place, but without unnecessarily complex regulatory requirements.

This report outlines why data matters for AI and the main issues at stake. It highlights some innovative solutions that businesses are deploying, and it suggests considerations for both government and industry to foster innovation and a pro-growth environment.

## Why does data quality matter for AI?

*"The true value lies in data and metadata, the oxygen fueling AI's potential."*
**Marc Benioff**

Data is the raw ingredient of AI. Together with advances in algorithms and hardware, today's AI has been powered by massive datasets that make possible modern techniques such as machine learning. The availability of vast quantities of data has been one of the principal drivers of AI's astonishing growth in recent years. By 2026, it is estimated that more than 221 zettabytes of data will exist—roughly enough to store 110 trillion movies.[1]

AI training datasets typically contain nearly all public data available on the internet, in addition to other sources, such as industry and government datasets that are made available for researchers and entrepreneurs. The indexed web contains about 500 trillion words of unique text, which is projected to increase 50% by 2030.[2] Though the sheer volume of data continues to expand rapidly, it will still be outpaced by AI developers' appetite for even more data as AI models grow in size.

As this growth occurs, developers are increasingly turning to other kinds of data, such as multimodal datasets (consisting of video and images) as well as synthetic (AI-generated) data.[3] At current growth rates, data availability constraints will place a ceiling on AI's skyrocketing trajectory before 2030. Whereas chip availability and energy and grid constraints present more immediate bottlenecks, data (specifically high-quality data) are essential for sustaining AI's long-term upside.[4]

Outside the world of frontier model development, however, the availability of *high-quality* data is an even more important consideration. Sheer quantity is less useful for specialized use cases and small language models (a smaller, less complex version of large language models, or LLMs). This is because an AI system is only as good as the data that goes into it—good AI needs both the AI model *and* the AI data to be of sound quality and governed effectively. Ironically, generative AI is also contributing to an overall decline in the quality of data available on the web.

Inaccurate, incomplete, or biased data can skew the inner workings of an AI system, leading to hallucinations, model drift, and poor-quality outputs. These can in turn lead to errors and financial losses. For example, a major US real estate online marketplace overpaid for houses based on flawed algorithmic predictions of future prices, costing it more than $500 million and prompting it to lay off 25% of its workforce. This company over-relied on historical data, failing to take into account local variations between markets, while its nationwide scale exacerbated the impact of the flawed data. Even minor inaccuracies in valuation data, when multiplied across thousands of transactions, resulted in big financial losses.

These risks are heightened in sensitive use cases, such as healthcare, transportation, and financial services. In these environments, the accuracy of AI systems is paramount, but this is not always the case for many off-the-shelf models, which come with limitations. And not all models are created equally or designed to solve every problem, leading some users to take a multimodel approach.

1 https://www.csis.org/analysis/operationalizing-data-free-flow-trust-dfft
2 https://epoch.ai/blog/can-ai-scaling-continue-through-2030
3 https://arxiv.org/html/2211.04325v2
4 Ibid.

To mitigate these risks, companies such as LSEG (London Stock Exchange Group) have adopted responsible AI principles and practices. These principles, which are regularly reviewed and updated, are consistent with global regulatory frameworks for AI, such as the EU AI Act, the OECD AI Principles and definition of AI, and the AI Risk Management Framework (RMF) from the US National Institute of Standards and Technology (NIST). These regulatory frameworks, among others, mean businesses must ensure that their data is not only accurate but also transparent and auditable. For companies in the financial services sector, pinpoint accuracy and trust in data are vital for doing business.

### LSEG's Responsible AI Principles

**Accurate and reliable:** Accuracy and reliability in AI involve the consistent performance or results of AI systems according to their intended functions and in varying conditions. The aim of this requirement is to ensure that AI systems perform correctly and dependably, delivering expected results to users**.**

**Accountable and auditable:** Accountability and auditability in AI reflect the extent to which information about an AI system and its outputs is available to individuals interacting with such a system. AI systems must have clear governance implemented and flag AI-generated outputs.

**Safe:** AI systems must be designed and tested to prevent harm to users and the environment, ensuring their operation does not pose undue risks. This principle involves identifying potential risks and actively working on their mitigation.

**Secure and resilient:** AI systems must be secured against unauthorized access and attacks, with robust measures to ensure their resilience and maintain the integrity of data and operations. AI systems must have protocols in place to avoid, respond to, or protect from attacks. AI systems must not reveal LSEG intellectual property to unauthorized users.

**Interpretable and explainable:** Interpretability and explainability in AI involve detailing how the underlying (AI) technology works and how the AI model reached a given output. This principle focuses on offering users information that will help them understand the functionality, purpose, and potential limitations of an AI system.

**Privacy-enhanced:** AI systems must prioritize the protection of personal data, ensuring that user privacy is upheld through robust data handling and anonymization techniques. This principle emphasises the protection of personal and sensitive information by AI systems, and compliance with existing privacy regulation and LSEG governance.

**Fair with bias managed:** Developers and users of AI should identify and mitigate biases in AI systems, which can otherwise lead to unfair outcomes. This principle focuses on the need to have fair AI systems that are in line with LSEG's values and culture.

## Why does quality data matter for the global economy?

In economic terms, AI is a general-purpose technology similar to electricity or the internet. Simply put, it is a tool that can be applied and adapted across all industrial sectors and, when used correctly, has the potential to unlock productivity, drive rapid economic growth, and solve real-world problems. Specialized datasets allow firms to fine-tune foundation models and adapt them for specific purposes within their business operations. For example, the agricultural machinery company John Deere uses specialized AI models trained on decades of geospatial and soil quality data to distinguish crops from weeds in real time. This enables precision spraying that reduces costs for farmers and consumers while also minimizing the industry's environmental impact.

Proprietary and industry-specific datasets allow traditional businesses to benefit from the AI revolution, withstand disruption, and unlock new business opportunities. In this way, non-tech firms, and businesses large and small, can capture some of the value and tangible benefits generated by AI, instead of the benefits accruing solely to Silicon Valley.

Similarly, companies with long histories of dealing with data have amassed quality datasets over decades of operations, which cannot be replicated with synthetically generated content. Combined with the power of AI, these datasets can open the door to new product offerings and benefits for customers. For example, LSEG has developed LSEG AI Insights in partnership with Microsoft, leveraging LSEG's data and analytics capabilities to help finance professionals summarize large volumes of information using natural language processing capabilities. As a result, if a user wants to assess the best-performing investment fund within a certain timeframe, they can do that in a matter of minutes without coding knowledge or English-language proficiency.

**LSEG AI Insights**

LSEG AI Insights is revolutionizing user workflows by leveraging natural language and generative AI to deliver seamless access to LSEG data and analytics via Workspace. By eliminating the need to navigate multiple applications and complex menus, it provides a unified, intuitive experience that significantly boosts productivity. Tailored to diverse professional needs, it supports tasks such as deal analysis by junior bankers, investment strategy recommendations by buy-side research analysts, and market updates prepared by wealth advisers. For instance, LSEG AI Insights enables wealth advisers to efficiently answer client queries on topics such as fund composition and performance, transforming how financial professionals interact with data and insights.

Proprietary datasets will become more important as AI develops. Some analysts argue that as advanced frontier models converge in terms of capability and become widely accessible at low cost, they will eventually become commoditized and more fungible.[5] In such a scenario, unique high-quality data, not the models themselves, will become the key competitive differentiator for businesses developing and deploying AI.

This data will be the essential ingredient behind the shift from companies using off-the-shelf LLMs to integrating AI across all parts of the economy. It is this "AI diffusion" effect—the widespread AI deployment across and within firms—that will unlock genuine added-value and productivity growth over the long term, not just the training of ever bigger and more advanced foundation models.

In this new paradigm, greater access to data will become a strategic issue for businesses across all sectors and a key driver of productivity growth in the wider economy. This has geopolitical implications: Countries with more access to data will be more self-sufficient, benefit from increased innovation, and be better positioned to win in the global AI race.

## What are the obstacles facing data-driven businesses?

### Barriers to cross-border data flows

The growing recognition of data's strategic value, particularly when it comes to AI, has contributed to the emergence of barriers to data transfer and use. Paradoxically, these barriers undermine the value of the data itself: If data cannot be shared, it is of no use to anyone. Global data policy thus resembles a classic collective action problem, or tragedy of the commons: Without international coordination, increasing numbers of countries will adopt restrictive data policies, with a resulting net loss to society.

5 https://www.microsoft.com/en-us/worklab/llms-are-becoming-a-commodity-now-what

Collecting, processing, and transferring data is critical to companies of all kinds. Seventy-five percent of the value of free data flows is shared by traditional industries such as agriculture, logistics, and manufacturing.[6] According to Digital Europe, the manufacturing sector stands to lose the most in absolute value from restrictions on data flows.[7] For example, Airbus's latest A350 aircraft has 50,000 sensors on board that together collect 2.5 terabytes of data daily.[8] Cross-border data flows are likewise essential to the financial services sector: LSEG alone delivers about 300 billion data messages to customers across 190 markets every day, including 7.3 million price updates per second, while customer demand for LSEG's data has risen by about 40% per year since 2019.[9]

Not only are large firms affected: Small companies seeking to move human resources or research data from one country to another, or from a subsidiary to a parent company, also grapple with data flow restrictions. But in a geopolitically contested global environment, barriers to global data flows continue to proliferate, for reasons ranging from industrial policy to privacy, sovereignty, and security concerns.

China, for example, imposes data localization laws, which require personal, business, biomedical, and financial data to be stored inside the country's borders on the basis of national security.[10] India requires payment system data to be stored locally, as part of its efforts to build a sovereign tech stack. The EU controls data flows to countries outside the bloc based on their respect for personal data protections, and it could expand this approach as part of a broader technological sovereignty agenda via regulatory proposals and cybersecurity certification schemes. The US has imposed restrictions on data flowing to and from specified countries of concern, even as data privacy laws continue to vary state by state, creating a complicated patchwork of regulation for businesses.

At the same time, data management introduces new obstacles for businesses to consider. According to an OECD-WTO business questionnaire, data localization requirements could raise data management costs by 15% to 55%, depending on the types of measures put in place. Also, data processing must take place in expensive or potentially insecure locations, compounding data fragmentation. Therefore, data localization requirements, besides being a barrier to trade, could compromise the security of the data supply chain, leading to increased cyber incidents or data breaches.

With its high demand for data, AI exacerbates these issues. AI systems require large, diverse datasets to perform effectively across different contexts and use cases. Data localization rules can lessen the amount of data available for training. They can also narrow the cultural and geographic diversity of training data, thereby negatively affecting data quality, introducing biases, and reducing the performance of AI models.

For the financial services sector, data flow restrictions pose substantial challenges, affecting companies' ability to deploy AI at scale. Financial markets are inherently global and interconnected, making seamless data flows essential. When these are restricted, it creates blind spots in risk assessment and analysis. For instance, a trading algorithm trained only on US market data might miss crucial patterns that emerge from Asian or European markets, leading to suboptimal decision-making and more systemic risk.

The impact extends beyond the financial sector. Global supply chains, which increasingly rely on AI for optimization, require frictionless data flows across multiple jurisdictions to function effectively. When data cannot be freely shared between participants, inefficiencies and transaction costs rise, weakening companies' global competitiveness.

6 https://iccwbo.org/global-insights/digital-economy/data-flows/
7 https://cdn.digitaleurope.org/uploads/2021/06/DIGITALEUROPE_Data-flows-and-the-Digital-Decade.pdf
8 Ibid.
9 https://www.lseg.com/en/insights/why-the-global-financial-system-needs-high-quality-data-it-can-trust
10 https://itif.org/publications/2017/05/01/cross-border-data-flows-where-are-barriers-and-what-do-they-cost/

Moreover, data localization requirements can create an uneven playing field in AI development. Major incumbent firms with the resources to maintain duplicate infrastructure in multiple jurisdictions gain an advantage over smaller competitors and innovative startups. For example, leading US cloud providers are able to construct local data centers around the world to serve international markets consistent with data localization rules, whereas smaller challengers are mostly confined to serving domestic markets.

### Transatlantic data transfers

According to the American Chamber of Commerce in the EU, cross-border data flows account for more than half of Europe's global data flows and for half of the US's total, and 90% of EU-based companies rely on transatlantic data flows. Despite this, transatlantic data flows have been a recurring source of uncertainty and compliance costs over the past ten years.

Two transatlantic agreements—Safe Harbor and Privacy Shield—designed to facilitate the free flow of data between the two jurisdictions while respecting EU privacy rules were struck down by the European Court of Justice (ECJ) following challenges by Austrian privacy campaigner Max Schrems, first in 2015 and then in 2020. These agreements were based on commitments from Washington to respect European data privacy rules, but which were deemed inadequate by the ECJ over concerns regarding US government access to EU citizens' personal data. Without a legal basis for free data flows, businesses were forced to use cumbersome alternative legal mechanisms to transfer data across the Atlantic, such as standard contractual clauses.

A third agreement—the Data Privacy Framework—was agreed in 2022 and has been in force since 2023. This has once again allowed businesses to freely transfer personal data between the EU and the US, including for processing by AI applications, without needing to resort to complex legal workarounds. However, it could be at risk from a "Schrems III" challenge, particularly in the event of a severe breakdown in transatlantic relations.

## Regulatory complexity

In addition to data flow restrictions, increased regulatory complexity and growing regulatory burdens are preventing greater AI adoption in many markets. This applies particularly in highly regulated sectors such as financial services, healthcare, and medical devices. These sectors, while already subject to extensive sector-specific regulatory frameworks, now face additional layers of AI regulation. According to the 2024 Stanford AI Index report, there were already 25 AI-related regulations in the US in 2023, up from just one in 2016.[11] In some cases, AI-specific regulations are duplicative: Many AI or data-related risks—such as those pertaining to privacy, data governance, cybersecurity, and operational resilience—are addressed in existing regulations.

Many businesses are already grappling with the task of implementing new or existing regulations across multiple jurisdictions. Now, they face the prospect of needing to deal with additional, and overlapping, AI rules adopted by governments globally. Compliance with these conflicting and sometimes duplicative regulations could come at the cost of investment in AI deployment and innovation. In a 2024 survey of chief information officers and chief technology officers conducted by PwC, 51% of respondents cited complying with new legislation and regulations generally as a significant challenge to their business priorities.[12]

11 https://aiindex.stanford.edu/report/
12 https://www.pwc.com/us/en/executive-leadership-hub/library/election-insights-2024-technology-leaders.html

### Fragmentation of global data rules

The fragmentation of global data privacy and intellectual property rules is one of the main drivers of barriers to cross-border data flows and regulatory complexity and fragmentation more broadly. For example, regulations such as the EU's General Data Protection Regulation and China's Personal Information Protection Law overlap in conflicting ways, taking different approaches to concepts such as data consent, processing, and transfer mechanisms. Whereas AI excels at processing unstructured data and extracting actionable insights, feeding raw data into an AI model runs the risk of license breaches and privacy violations, given the complexity of the regulatory environment.

Exactly how privacy rules should apply to AI has become a notable source of disagreement and uncertainty, including within the EU, where different data protection authorities have taken different interpretations of the law.

In addition to longstanding problems around data privacy, the application of intellectual property rights in the context of AI is proving to be a new challenge for policymakers. While this issue is receiving attention in international policy circles, including at the G7 level and the Paris AI Action Summit, different political and legal traditions are leading to fragmentation in approaches among different states. For example, the US has a wide-ranging interpretation of "fair use" in copyright law, whereas European countries tend to be more favorable to the creative sectors.

In this context, policymakers are turning to transparency requirements, such as those in the EU AI Act, to better understand which datasets are used in AI training, and whether they are protected by intellectual property rights. Yet for companies deploying AI, particularly those relying on third-party services, it can be almost impossible to verify where or how a generative AI model has used data. Although protecting the rights of creators is highly important, this must be balanced with the objective of not further burdening AI deployers with unnecessary reporting requirements, since this would act as an additional drag on AI adoption.

## How can the international community address these issues?

These are hard problems—which is why they remain unsolved after many years. Questions around data touch on sensitive matters of national sovereignty, while diverging policies can reflect deep-rooted cultural differences. With rising geopolitical tensions challenging a rules-based international order defined by multilateral institutions, coordination among governments is trickier than ever. But without it, the full benefits of AI may never be realized.

The good news is that governments do not need to start from scratch to make progress on data governance. Digital trade chapters in modern free trade pacts, as well as self-standing digital trade agreements, contain provisions prohibiting unjustified barriers to data flows, including data localization requirements.

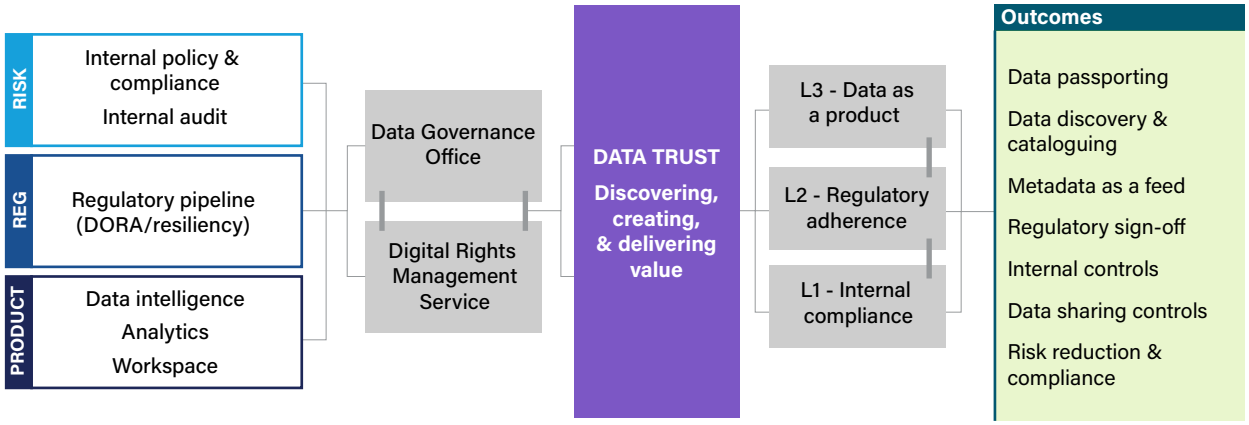
In addition to bilateral trade deals, multilateral initiatives such as **Data Free Flow with Trust (DFFT)** at the G7 also have a role to play. First proposed by Japan at the 2019 World Economic Forum, DFFT aims to establish common principles and standards for cross-border data flows while ensuring privacy, security, and intellectual property protection. In 2023, G7 leaders endorsed the "G7 Digital and Tech Ministers' Vision for Operationalising DFFT and its Priorities" and established an Institutional Arrangement for Partnership. The OECD is currently undertaking a multistakeholder process to promote DFFT and its policy objectives. Still, DFFT remains a work in progress. What is needed is to turn principles into practice.

**Under the Canadian G7 presidency in 2025, the group should commit to making progress on operationalizing DFFT, supporting the OECD's efforts to craft practical frameworks and standards**. In doing so, OECD member states can build on the achievement of the 2022 OECD Declaration on Government Access to Personal Data Held by Private Sector Entities. The declaration is proof that with sustained effort and a willingness to compromise, it is possible to reach a consensus on difficult and sensitive questions related to data. The G7 should now harness the momentum around AI to make progress on the other pillars of DFFT, including data localization, regulatory cooperation, and data-sharing.

When it comes to privacy, the Global (formerly APEC) **Cross-Border Privacy Rules (CBPR)** system has the potential to help bridge the gaps between different personal data protection regimes and is already operational. Voluntary frameworks such as the CBPR can help businesses, including those deploying AI, to navigate overlapping regulatory frameworks while maintaining privacy protections.

That said, while multilateral frameworks are an important part of the puzzle, they are not in themselves sufficient. Bilateral agreements—such as the EU-US Data Privacy Framework and the UK-US data bridge—will remain indispensable. Given this reality, it is important to maintain bilateral channels for dialogue and cooperation, such as the EU-US Trade and Technology Council, the future of which is currently uncertain.

Alongside political solutions, business has a pivotal role to play in building trust in data, as well as navigating regulatory fragmentation and establishing best practices across industries. LSEG's data passport concept, for example, helps clients understand the provenance and ownership of datasets, and whether this is subject to privacy or other rights and restrictions—giving them the confidence to trust the data.[13] Solutions such as the data passport concept can support AI deployers in satisfying regulatory requirements—for example, on data accuracy and transparency—while navigating a patchwork of global regulatory regimes.




Ultimately, policymakers and businesses will need to work together. Businesses can work to increase the quality of their data and build trust in data, by adopting best practices in line with existing regulations and international standards. At the same time, policymakers should aim to build a stable, interoperable, and flexible regulatory environment, prioritize internationally agreed-upon definitions relevant to AI and data, and devise fair yet light-touch methods of managing intellectual property and digital rights.

## Conclusion

AI has brought renewed urgency to questions of data governance, but too often these conversations occur in parallel with each other instead of in concert. As this report has shown, the full benefits of AI cannot be realized without ensuring the fuel that powers it—data—is trustworthy, accurate, and free-flowing.

The challenges to making progress on international data policy are significant, and solving these challenges could unlock new markets, new economic growth, and, importantly, new ways of keeping data secure. Geopolitical tensions and competing visions of digital sovereignty are leading to a proliferation of data localization requirements and fragmented regulatory frameworks, undermining AI's potential to drive innovation and productivity growth. In this environment, political and business leaders should work together to inject fresh political momentum into data governance discussions and make the global regulatory environment for data less burdensome and more interoperable.

13 https://plus.reuters.com/lseg-data-integrity-and-responsible-innovation-in-ai/p/1

How can this be done? Below are some considerations for how governments and the private sector can work side-by-side and help foster a pro-growth and pro-innovation regulatory ecosystem for the digital economy:

1) Data governance should be at the heart of international discussions on AI—including through the G7, the G20, and future AI Safety/Action summits.

2) Governments and industry should continue to support the OECD's work on facilitating cross-border data flows (including operationalizing DFFT) and building closer alignment on global data governance more broadly.

3) Governments, in consultation with the private sector, should pursue bilateral and multilateral agreements that enable cross-border data flows, drawing inspiration from other successful standalone digital trade pacts, such as the UK-Singapore Digital Economy Agreement, as well as broader agreements that include provisions on digital trade, such as the United States-Mexico-Canada Agreement or the potential UK-US digital trade agreement.

4) Where relevant, governments should examine domestic policy frameworks and consider making targeted amendments to increase interoperability with international partners and reduce barriers to the free flow of data, while also prioritizing and protecting data privacy and security.

5) Policymakers should prioritize access to secure, high-quality data as a cornerstone of economic resilience strategies, including by fostering data-sharing ecosystems that balance innovation with privacy, and establishing interoperable governance frameworks.

Making progress on these objectives will require political will and sustained focus. But there are reasons for optimism. The DFFT framework, the OECD's work on trusted government access to data, and various digital trade agreements provide a solid foundation to build upon, even in an era of heightened geopolitical competition. Industry is an enthusiastic partner in these efforts and is already working to develop standards, controls, and guidelines.

By combining progress on multilateral initiatives with business-led solutions, we can ensure that AI fulfills its promise as a transformative technology for the global economy, and that trustworthy AI is powered by ample supplies of trustworthy data.



New York | Washington DC | San Francisco | London | Tokyo | Singapore | São Paulo | Brasília

This report is based on the opinions of Eurasia Group analysts and various in-country specialists. Eurasia Group is a private research and consulting firm that maintains no affiliations with governments or political parties.

Report issued 10 February 2025 