Operationalising AI governance: A review of industry best practices

AI Assurance Club
A Global Digital Foundation Initiative

September 2023

Executive summary

This briefing paper provided advance context for the 4th annual International AI Governance Roundtable that took place on 19th September 2023 in Barcelona and Shanghai, hosted by Global Digital Foundation. This year the focus was on the operationalisation of AI governance practices, and considered how we can take an engineering approach. As in previous years, the event took place under Chatham House rules.

Our analysis indicates a high degree of **consensus about the overall direction of AI governance** at a high level. Several leading companies have adopted or developed **general frameworks** for managing the design, development and use of AI, and are voluntarily implementing a range of governance and assurance techniques. The market for AI assurance tools, such as compliance audits and model evaluation techniques, is continuing to mature. Some sectors, such as healthcare, show signs of more comprehensive and effective structures and practices. However, **significant evidence gaps** remain. With many **policy and standardisation processes still under development**, there remains **significant challenges** for the development of compliant assessment frameworks, and there are limited case examples on which we can draw to evaluate how well practices demonstrate **effective risk management** and **regulatory compliance**.

There is also little evidence of methodologies to support r**esponsibility sharing** across the AI value chain. Similarly, it is **hard to find evidence for quantitative approaches** (i.e. measurement methods) that might deliver **evaluation indicators**. The conformance standards supporting the AI Act are yet to emerge, and it may well be that those standards will provide the necessary impetus for quantitative **process assurance techniques** to be developed.

Nevertheless, this presents an opportunity to build the evidence base and shape the substance of such instruments.

Scope and methodology

The purpose of this briefing is to provide an overview of emerging industry best practices in AI governance and assurance. These practices include many different approaches and measures. Some of these, such as performance testing, are based mostly on objective and quantifiable criteria, while others, such as impact assessments, involve a greater degree of subjective expert judgement. A taxonomy of assurance techniques is provided in Annex 1.

We analysed policies and practices of fifteen leading technology and engineering companies (see Annex 2), two public databases of AI governance and assurance tools, and relevant academic and policy literature. It should be noted that all information relied on is in the public domain. While these publicly available sources allow us to generate insight about what is likely to be expected of different AI market actors, they will not reflect all practices currently undertaken.

To gain a more comprehensive account of current practice, future research would need to employ an alternative methodology based on primary research such as key informant interviews with relevant actors across industry. This could be segmented into sectors to offer a comparative perspective.

The emerging regulatory and standardisation landscape

Increasingly, governments are opting for so-called “hard” approaches to AI regulation. In multiple jurisdictions, enforceable rules governing the development, supply, and use of AI technologies have either been enacted or proposed. In many cases, existing laws are also applicable. These include pervasive legal regimes such as data protection law, as well as sectoral regimes in highly regulated domains like healthcare. There is significant cross-jurisdictional variation in many of these approaches. One common feature, however, is the central role that complementary assurance tools will play in giving effect to new rules.

In parallel, important developments in standardisation are taking place. Designated standards will form the backbone of EU AI-specific regulation once enacted. In addition to the compliance implications this presents, new standards such as ISO/IEC 42001, once published, will likely alter expectations placed on market actors. It is therefore timely to assess what different actors are doing to prepare themselves, and their customers, for emerging regulations and standards.

Organisational and risk management frameworks

Since the beginning of 2023, two significant standards offering strategic frameworks for organisations to manage risks associated with AI have been published.

In January, the United States National Institute of Standards and Technology (NIST) launched the first version of the Artificial Intelligence Risk Management Framework

(AI RMF 1.0). This is a voluntary framework that has been developed in collaboration with the private and public sectors. It allows organisations to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. This framework has been widely endorsed by key US-based industry actors, including: Microsoft, IBM, Google, Amazon Web Services, Partnership on AI, the Information Technology Industry Council, and the Alliance for Automotive Innovation. Beyond these endorsements and commitments, there is very little published literature setting out what organisations have done so far to incorporate AI RMF 1.0 into their own risk management practices.

The AI RMF 1.0 has already been used to assess efforts by companies to make their AI systems safer. The Federation of American Scientists, for example, used the framework to assess how well efforts by OpenAI to test and improve GPT-4's safety before release conform to current best practice. They found some alignment in the process to map, measure, and manage risks, and in specific measures used such as *red teaming.* However, they found that, whilst NIST's resources provide a helpful overview of considerations and best practices to be taken into account when managing AI risks, "they are not currently designed to provide concrete standards or metrics by which one can assess whether the practices taken by a given lab are "adequate.""

In February 2023, the ISO/IEC 23894 — Risk management standard was published. It offers strategic and sector-neutral guidance for managing risks connected to the development and use of AI, and guidance on how organisations can integrate risk management into their AI-driven activities and functions. It is possible that this standard will form the basis of one of the harmonised standards to be developed by European Standards Organisations (ESOs) for the implementation of the proposed EU AI Act. The European Commission's draft standardisation request includes a requirement for a standard or deliverable on *Risk management system for AI systems.*

ISO/IEC FDIS 42001 is a horizontal standard under development. It provides specifications for integrating an AI management system within an organisation's existing structures. A management system is defined as "interrelated or interacting elements in an organisation to establish policies and objectives, as well as processes to achieve the objectives." Annex B sets out implementation of AI-specific control measures. For example, the standard requires documentation of the categories data held by the organisation to be used for machine learning, as well as the process used for labelling and training. This standard is complemented by the published technical report ISO/IEC TR 24028.

Many commentators are anticipating that the publication of 42001 will represent a landmark in AI governance. In a recent interview, the Director of Australia's National AI Centre, Stela Solar, summarised the significance of this standard:

"It's in essence going to be identifying which organisations are more mature with AI governance. By default, organisations who embrace those standards will be demonstrating that they're more mature in their AI practice and governance, those who are not may be seen as higher risk [...] Whether organisations are adhering to

global standards and best practices is going to start to determine who you partner with and who you choose in your supply chain."

Microsoft has its own internal Responsible AI Standard. This is structured around various sets of goals such as Accountability, and Privacy and Security and is the framework used to implement Microsoft's AI principles. Within each set of goals there are specific requirements, and tools and practices recommended for meeting those requirements. While comparator companies, such as Google, have their own established processes for AI development, Microsoft appears to lead in terms of the clarity of the standard it requires AI practitioners to work to across the AI life cycle.

We found that many companies have adopted sets of AI Principles and Codes of Ethics. BMW, for example, developed a set of Principles in 2020. There is almost no information about how these principles have been operationalised, or criteria available with which to evaluate implementation. Recent press releases on the use of AI technologies make little or no mention of the principles, or trustworthiness considerations.

Some companies also highlight their certification to non-AI quality assurance standards. For example, CARIAD, described as the "software powerhouse of Volkswagen Group" does not mention any AI-specific standards, but does state that it is certified to ISO 9001 – Quality management systems requirements.

Dimensions of trustworthy AI

Terminology for AI governance and assurance varies. While some bodies like the OECD and European Union have generally used the language of *trustworthiness*, many companies talk about their activities under the banner of responsible AI. This generally includes all AI governance practices and wider social impact activities (e.g., equality, diversity and inclusion initiatives).

We have divided governance and assurance practices into the following categories: Explainability and Transparency, Fairness, Reliability and Safety, Privacy and Data Governance, and Security. For each of these dimensions, we reviewed existing practices of fifteen companies, along with policy reports and recommendations, and relevant tools and platforms available in the burgeoning market for AI assurance. These practices can be broadly classified as either technical or process-based.

Explainability and Transparency

The provision of clear and meaningful explanations of an AI system's outcomes is widely considered to be a **crucial component of trustworthy AI**. Explainability and interpretability are closely connected to both transparency, which is about providing information and disclosure about an AI system to the appropriate stakeholders, and **traceability**, which refers to the ability for humans to follow elements of an AI system before, during and after its deployment.

The European Union's proposed AI Act contains provisions on both explainability and transparency. Article 13 is particularly relevant. It mandates that high-risk systems be designed and developed to meet appropriate levels of transparency that

enable users to interpret the system's output. Exactly what is considered appropriate, and the steps that should be taken to reach it, will be determined in standards.

Relevant standards on explainability and transparency are still in the early stages of development. At a workshop held earlier this year, participants were still at the stage of discussing appropriate definitions of key terms such as interpretability. In the public sector, there has been some progress with the iteration of the Algorithmic Transparency Recording Standard which helps public sector organisations in the UK provide clear information about the algorithmic tools they use, and why they're using them. While representing good practice for public sector organisations, this standard has narrow applicability.

Many different approaches to mitigating risks to transparency and explainability have been developed or recommended in research findings. Technical approaches can be model-specific or model-agnostic, while process-related approaches tend to be focused on documentation.

Popular process-based approaches include documentation tools like Model Cards (Google), Datasheets for Datasets (Microsoft), Transparency Notes (Microsoft Azure/ OpenAI), and System Cards (Meta). In 2022, Google's People + AI Research (PAIR) team launched the Data Cards Playbook. It aims to help teams create structured transparency artefacts for datasets. It comprises four modules designed with participatory activities to define "long-term transparency" for datasets in their contexts.

Developing effective and widely applicable technical approaches for AI transparency and explainability is recognised as a **particularly challenging task**. Leading scholars on interpretable AI have argued that while there are increasing numbers of model-agnostic interpretation techniques for models such as partial dependence plots (PDP), permutation feature importance (PFI) and Shapley values, that provide insightful model interpretations, if these are applied incorrectly, incorrect conclusions can easily be reached (Molnar et al. 2021). It is also now widely accepted by AI researchers that complex trade-offs exist between explainability and other properties such as accuracy and privacy.

Fairness

The NIST AI RMF 1.0 divides biases into "systemic, computational and statistical, and human-cognitive." Biases can affect legally protected characteristics such as ethnicity, gender, or age. It is widely considered best practice to apply measures promoting fairness at all stages of the AI life cycle, and to analyse performance to inform any decision making about model retraining. However, **fairness is a moral concept** and as such tied closely to societal values. **Assessing fairness in a cross-cultural environment is a significant challenge** without quantitative tool support.

From a regulatory perspective, provisions related to fairness tend to take the form of impact assessment and reporting requirements. In 2023, a New York City law (Local Law 144 of 2021) requiring all employers to conduct third-party bias audits of any

algorithm involved in hiring decisions came into effect. There is a growing market in providers of third-party audit services. Many of the items featured in both the OECD Catalogue of Tools & Metrics for Trustworthy AI, and the UK CDEI portfolio of AI assurance techniques are third-party platforms and services for auditing AI systems for bias. BABL.AI which conducts third-party audits for automated employment decision tools is a typical example. It serves as an independent, third-party auditor to certify that an auditee has performed sufficient testing to meet the minimum requirements of the NYC law.

Beyond legal requirements, AI providers have developed different tools for their own and their customers' use. Amazon SageMaker Clarify is a feature for Amazon SageMaker that includes both explainability and fairness functionalities. It is aimed at machine learning developers and aims to support bias detection across the entire life cycle, including during data preparation, model evaluation, and post-deployment monitoring. The researchers consider this tool to be a “scalable, cloud-based bias and explainability service designed to address the needs of customers from multiple industries.” It has been used by customers including Bundesliga.

Reliability and Safety

AI system reliability is critical for ensuring systems function appropriately whether they are being used as intended or misused, and for ensuring they do not pose unacceptable safety risks. In many AI use cases, existing consumer and industrial safety regulations will already apply.

We found surprisingly few examples of innovative or leading practice in AI governance in safety critical domains, particularly in engineering environments. We reviewed the available policy and governance documents of companies such as Siemens, ABB, Bosch and Thales, and found **very limited information about AI related safety issues**. In the automotive industry, despite widespread interest and adoption of machine learning based software, there are very few detailed examples of AI assurance in practice. In our small sample - BMW, Volkswagen, and Tesla, we did not find examples of cutting edge AI assurance research and practice.

Whilst the use of formal (mathematical) methods has been applied to the problem of software safety in traditional symbolic AI systems, **probabilistic approaches to machine learning** combined with the **‘black-box’ nature of neural network based systems** makes this a currently challenging research topic. Much of the state of the art in assurance for autonomous systems appears to be coming from academia and government.

A recent report by the UK CDEI set out its proposals for a responsible and trustworthy regulatory and assurance framework for autonomous vehicles. It cites the work of a project at the University of York in partnership with Lloyds Register Foundation. The Assuring Autonomy International Programme (AAIP) aims to help those in industry to follow processes that prove their autonomous systems are safe, and to support regulators in setting consistent safety standards worldwide. Their work is translated into practical guidance providing methods and processes to give confidence in the safety of autonomous systems. This includes “the first

methodology that defines a detailed process for creating a safety case for autonomous systems."

AI safety considerations have recently taken on increased importance with the rapid diffusion of powerful generative AI models. A report by Schett et al. (2023) outlines various measures being considered by leading experts from frontier AI labs, academia, and civil society. The authors found a great deal of consensus, with almost all of the 50 proposed practices supported by a majority of respondents. The authors claim this finding will help to identify best practice for so-called AGI (artificial general intelligence) safety and governance. However, the statements used in the survey are vague and ambiguous. For example, "AGI labs should take extensive measures to identify, analyze, and evaluate risks from powerful models before deploying them." Further research, with more stakeholders and a more rigorous methodology, would be needed to establish what best practice in risk management would look like for the development of large generative models.

Privacy and Data Governance

Given the reliance on huge volumes of data, **machine learning presents serious privacy concerns** relating to the risks of data leakage, and failure to comply with data protection law. A variety of privacy-enhancing techniques have been developed to mitigate risks to personal or sensitive data across the AI system lifecycle. Unfortunately, however, this is another area where trade-offs exist. Practices for preserving privacy may come at the expense of explainability, robustness, and fairness.

Xu et al. (2021) offer a summary of the state of the art in Privacy Preserving Machine Learning (PPML) solutions. One of the most widely cited solutions is federated learning. The term was first coined by Google researchers in 2016, and describes an approach for developing models by distributing training data across devices. According to IBM, federated learning is "becoming the standard for meeting a raft of new regulations for handling and storing private data."

Health is one area in which PPML solutions are relatively mature. Given the risks of de-anonymisation and the highly sensitive nature of patient data, researchers have developed various technical tools. For example, Kaissis et al. 2021 develop PriMIA (Privacy-preserving Medical Image Analysis) - a free, open-source software framework for differentially private, securely aggregated federated learning and encrypted inference on medical imaging data. Even in this relatively mature domain, however, gaps remain in research and in practice.

Security

AI is recognised as presenting a double-edged sword for cybersecurity. While AI techniques can be used to support and automate cybersecurity operations and controls, the **application of AI can also open many new avenues for attack methods**. As a result, cybersecurity features prominently in AI legal, policy and standardisation instruments, including the proposed AI Act. It requires that high-risk AI systems have appropriate levels of robustness, accuracy and cybersecurity which

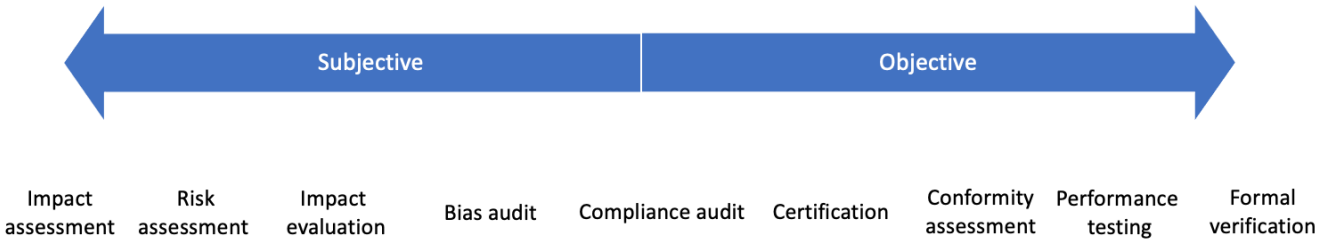
must be maintained throughout the entire lifecycle. The precise technical solutions to be employed will, however, depend on the specific circumstances.

Industry wide **standards and practices for securing AI are still evolving**. In April 2023, the European Union Agency for Cybersecurity (ENISA) published an assessment of standards for the cybersecurity of AI. It featured recommendations to support the implementation of upcoming EU policies on AI, including development of technical guidance on how existing standards related to the cybersecurity of software should be applied to AI.

In July 2023, ETSI launched a series of reports developed by its Securing AI group (ISG SAI). One of these reports sets out a security framework for AI computing platforms to protect valuable assets.

Annexes

Annex 1. The spectrum of AI assurance techniques



Adapted from the Centre for Data Ethics and Innovation (CDEI) AI Assurance Guide (BETA).

At one end, impact assessments are designed to account for a greater degree of indeterminacy of potential future harms. They require professional expertise and subjective judgement to account for these factors, and they enable standardised processes for *qualitatively* assessing potential impacts. At the other end of this spectrum, formal verification is used for assessing trustworthiness for subject matters which can be measured more objectively and with a high degree of certainty.

Annex 2. Companies selected for analysis

ABB
Amazon Web Services
Anthropic
BAE Systems
BMW
Bosch
Google
Google DeepMind
IBM
Meta
Microsoft
Siemens
Tesla
Thales Group
Volkswagen Group