AUTORITEIT
PERSOONSGEGEVENS

Global AI Governance from a Supervisory Perspective

Discussion paper in response to the interim report of the UN Secretary-General's AI Advisory Body[1]

April 2024 | Dutch Data Protection Authority (Autoriteit Persoonsgegevens) | Department for the Coordination of (Supervision on) AI and Algorithms (DCA) |
Contact: dca@autoriteitpersoonsgegevens.nl

Executive Summary

- In response to the interim report of the UN SG's AI Advisory Body, this brief discussion paper outlines how national and regional AI supervisors should be involved in the global AI governance framework.
- The contribution provided by this paper is based upon the national experience in the Netherlands, where since 2023 the Autoriteit Persoonsgegevens has a role as National Coordinating AI Supervisor (NCAIS). In the Netherlands, the NCAIS focusses on the risks of AI on fundamental rights and public values through bi-annual risk reporting, supervisory cooperation and the development of guidance and supervisory policy perspectives.
- Based upon the experience as NCAIS, we advocate the establishment of a global AI institution that facilitates risk monitoring, standard setting, knowledge exchange and peer review on the international level.
- National and regional AI supervisors should play a central role in such a global AI governance structure, due to the direct involvement in AI risk identification and mitigation. Where feasible, they should cooperate in supervisory colleges for systemic AI models and systems. In addition, AI supervisors should also contribute to a global AI incident register to facilitate knowledge exchange and a common understanding of current and upcoming AI risks.

[1] This discussion paper aims to provide insight into the analyses conducted for this policy issues. The views expressed in this paper are preliminary.

1. Introduction

This discussion paper contributes to current discussions on the international governance of AI. The paper provides a reflection on the United Nations Artificial Intelligence Advisory Body's ("the advisory body") Interim Report: *Governing AI for Humanity*. It discusses the proposals to govern AI on a global level from a supervisory perspective, advocating for the establishment of clear (coordinating) supervisory functions as part of the global AI governance.

In the Netherlands, the Dutch Data Protection Authority (Autoriteit Persoonsgegevens, "AP") has the role of national coordinating supervisor on the use of algorithms and AI ("National Coordinating AI supervisor", NCAIS). This function is carried out by the Department for the Coordination of the Supervision on AI and Algorithms (DCA). The aim of the Dutch NCAIS is to protect fundamental rights and public values in light of the development and deployment of AI. As one of the world's first supervisors with a mandate to identify and address overarching societal and fundamental risks related to AI-systems, we believes global governance institutions for AI necessitates strong involvement of coordinating AI supervisors and regulators.

The structure of this paper is as follows. First, the paper will discuss the supervisory perspective on governing AI in a broad sense, based on the experience and activities of the AP in its capacity as NCAIS so far. It then discusses the chapter International Governance of AI of the interim report and, building further on this chapter, provides for proposed institutional framework to govern AI globally.

2. Global AI governance from a supervisory perspective

Algorithms and AI have become integral to our society, presenting significant opportunities for innovation but also potential risks to both individuals and societal structures. The global challenges and opportunities presented by AI call for global governance. The Advisory Body's Interim Report analyses and advances recommendations for an international governance of AI.

The interim report advocates for a governance framework as a key enabler to ensure AI is deployed for the common good, which incentivized participation from the private sector, academia and civil society. We share the view that AI governance should ensure access to opportunities created by AI on one end while taking action to prevent and mitigate potential harms on the other. This duality mirrors the current practices executed by the AP through the NCAIS-function in The Netherlands. The NCAIS performs overarching supervisory oversight of public values and fundamental rights in light of the rapid development and deployment of algorithms and AI. This supervisory risk monitoring is executed through a macro- or meta-based approach, which implies that risks are identified which arise on the collective or societal level (for example: based upon analysis of the usage and availability of such systems, what are the risks associated with adaptive learning in education?) and which subsequently need to be addressed through policy and legislative measures and supervisory requirements on the level of individual AI systems. To this end, the NCAIS-function focusses on:

(i) early AI risk identification and periodic overarching AI risk reporting on a national level (see the AI & Algorithmic Risk Report Netherlands, that is published bi-annually),
(ii) strengthening collaboration on AI risks between all relevant supervisors and stakeholders (industry, academia and civil society), and
(iii) providing guidance on managing the risks of algorithms and AI.

Related to this, key priorities in the work program include
a) enhancing algorithmic transparency and explainability,
b) stimulating bias and fairness testing to minimize the risk of discrimination,
c) minimizing the risk of arbitrary outcomes in AI related processes and
d) preventing the misuse of AI for manipulation.

This focus contributes to the mitigation of negative impacts and the enhancement of the positive effects of the deployment of AI. The independent role of supervisory authorities in overseeing AI deployment is crucial for establishing a safe, responsible digital environment. This necessitates a robust, ongoing supervision framework ranging from the preliminary development stages through to post-deployment evaluations. On a global level, it is imperative to have a focus on these elements to contribute to safeguarding fundamental rights and public values in all compartments of the AI and algorithmic value chain, from data collection to training, application and reviewing AI systems. An additional focus should be on reaching consensus on use-cases where the application of AI-technology is undesirable and should be regulated or banned.

We therefore believes that a global AI framework – as discussed in the interim report of the advisory body – should include active participation and involvement of independent or coordinating AI supervisory authorities and regulators. Such coordinating national and regional regulators and supervisory authorities are best positioned to identify and assess current and future risks from the practical experience and expertise of supervisors.

3. Institutional functions and principles as a basis for global AI governance

There is a need for common guidelines and standards to govern AI on a global level. The cross-border application of AI systems entail that both opportunities and risks of AI manifest globally, affecting societies and international fundamental rights. Therefore, the interim report formulated recommended *principles* and *functions* that could serve as a foundation for AI governance on a global level. The principles form the underlying basis of global institutions for AI governence, and the functions reflect the tasks these institutions would need to perform.

Principles and institutional functions: Experience and recommendations from the Dutch NCAIS. The following principles have been identified in the interim report:

- AI should be governed inclusively, by and for the benefit of all citizens, including those in the Global South;
- AI must be governed in the public interest rather than private, commercial interests;
- AI governance should be built in step with data governance and promotion of data commons ;

- AI governance must be universal, networked and rooted in adaptive multi-stakeholder collaboration;
- AI governance should be anchored in International Law: the UN Charter, International Human Rights Law and other agreed commitments such as the Sustainable Development Goals.

Based upon the experience as NCAIS; these four additional key principles are needed for global governance institutions for AI:

1. **Focus on fundamental rights and public values** - Establishment of a common understanding about the need to protect fundamental rights and public values related to the development and use of AI systems. The reference to fundamental rights and public values is necessary to provide for a risk matrix that supports joint risks analysis and a common vocabulary / taxonomy on AI risks.

2. **Strong involvement of (coordinating) AI supervisors and regulators** - Active participation and involvement of (coordinating) independent national and regional AI supervisors and regulators. Such coordinating national and regional regulators and supervisors are best positioned to identify and assess current and future risks from the practical experience and expertise of supervisors. Examples of overarching period risk identification include the two "AI & Algorithmic Risks Report Netherlands" that the AP has published since our establishment in early 2023. We publish this report twice a year and the purpose of the report is to provide an overarching risk assessment and also to monitor the development of AI risks and AI incidents over time. Our first report is available here and our second report is available here.

3. **Contribute to colleges of AI supervisors -** For the most systemic AI models and AI systems which are deployed on a global scale, supervisors could work together through supervisory colleges. While respecting the independence of each national or regional supervisor, such supervisory colleges provide for information exchange, joint risk analysis and coherent supervisory action on specific AI systems from the perspective of ensuring safe and compliant AI systems.

4. **Provide for centralized global AI incident registers -** Global understanding of AI opportunities and risks will only be achieved when there are opportunities for sharing knowledge, best practices and a joint understanding of AI incidents globally. This could build upon the current work of the OECD.

We supports an international governance regime for AI that embodies the institutional functions as enlisted in the report, it is however necessary to strengthen the role of supervisory authorities. These must be taken into account and incorporated in the institutional functions. Monitoring AI can eliminate or mitigate risks early on, which can reduce the impact of such risks on individuals and society. Supervisory authorities can jointly contribute to establishing guidelines and norms related to responsible AI. To this end, the harmonization of management frameworks and practices should enable supervisors to effectively cooperate.

Accordingly, multiple institutional functions are linked and need to be seen in conjunction to each other to create a framework for effective global AI governance. In the framework for international AI governance that could be considered, the institutional functions for the governance of AI operate in synergy to create a continuous cycle of (a) overarching risk and incident monitoring and assessment, (b) improvement of AI regulation and risk management and (c) peer review and evaluation.

The following institutional functions identified by the advisory body – some of them provided with recommended amendments - could support such a framework.

IF1: Horizontal scanning, building scientific consensus - The interim report advises to assess regularly the future directions and implications for AI; an expert-led process that continuously provides for scientific, evidence-based insights to inform policymakers about the future trajectory and implications of AI. This function would perform risk assessments and standards to measure impacts of AI.

- ***Not only future directions and implications of AI but also current overarching risks and incidents must be continuously assessed to form the foundation for ideally achieving consensus on risk mitigating measures****. Such an approach would be crucial for managing the development and use of AI systems globally. We would therefore suggest to formulate IF1 as I "IF1: Global horizon scanning of current and future AI-related risks and incidents to build supervisory consensus on risks and mitigating measures".*

IF3: mediating standards, safety and risk management frameworks - Horizontal scanning of current and future AI-related risks and incidents (IF1) would support consensus building towards minimum alignment of standards, safety and risk management frameworks.

IF6: Reporting and peer review - The interim report stresses the importance of the capability to monitor, report and respond to systemic vulnerabilities to global stability. Reference is made to the macro-prudential frameworks used in the financial sector to create resilience against risks to global stability. In this regard, we do agree with the interim report that a techno-prudential framework, akin to the macro-prudential framework used to increase resilience in central banking and financial sector supervision, is a promising avenue for the techno-prudential model (or "AI prudential model") that is needed.

- ***The execution of IF1, IF3, and IF6 within a global AI body can facilitate a continuous cycle.*** *Horizon scanning of current and future related-AI risks and incidents (IF1) would subsequently flow to consensus building towards minimum alignment of standards, safety and risk management frameworks (IF3). This subsequently provides a framework for reporting and peer review (IF6) of which the outcomes provide periodic input again for global horizon scanning (IF1). This closes the cycle.*

IF7: norm elaboration, compliance and accountability - In addition, the interim report finally discusses the need for legally binding norms and enforcement as well as the added value of non-binding norms to ensure compliance and accountability. In this regard, reporting and peer review to global governance institutions would help to ensure compliance and prevent accountability gaps. Such an institution equally needs to be held accountable itself, however. Governance efforts much therefore demonstrate trustworthiness, including transparency in objectives and processes.

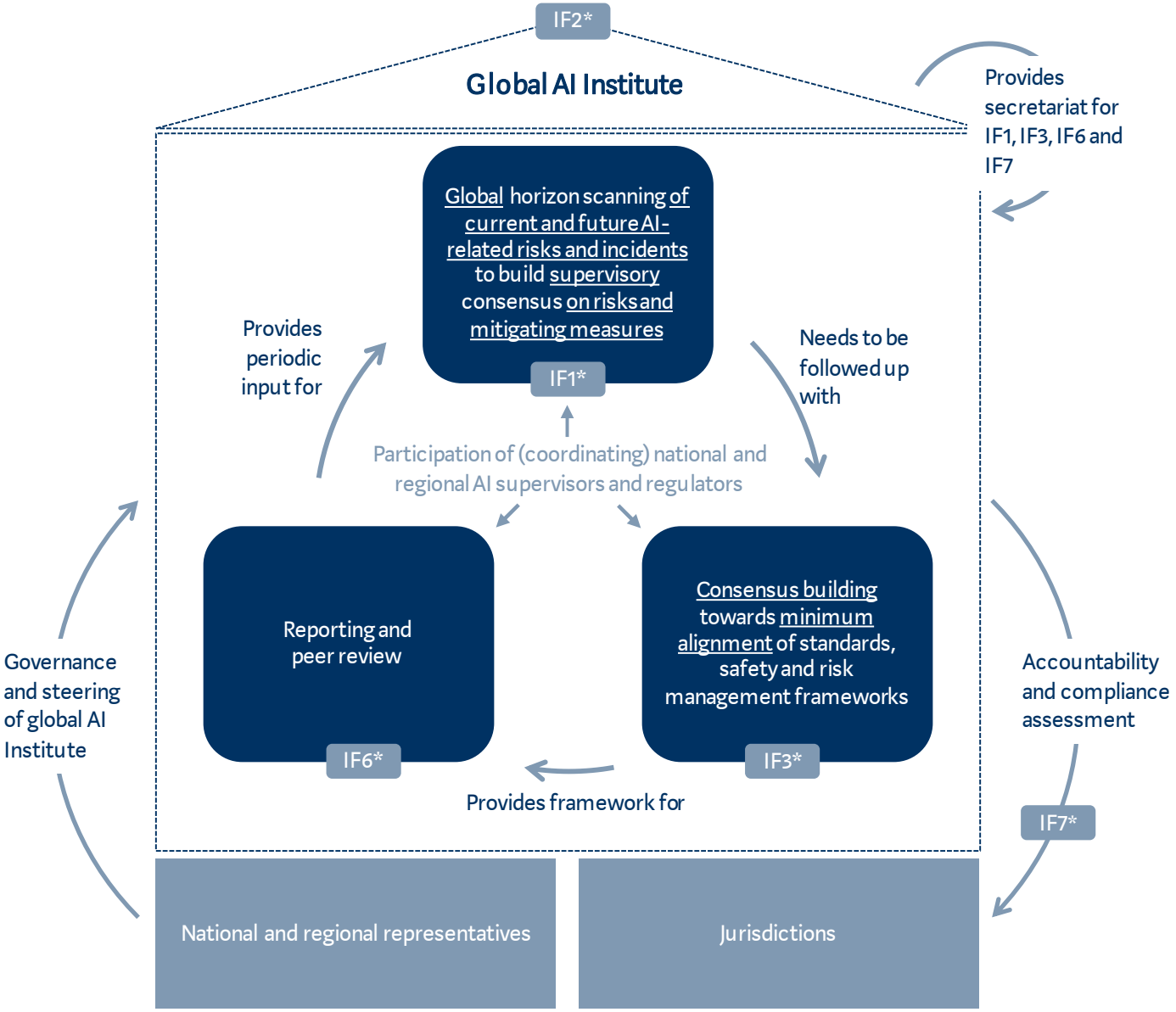
- ***We do agree that a global institute can support accountability and compliance assessment (IF7)*** *. This would relate to national and regional frameworks for regulation, supervision and infrastructure to support credible, trustworthy and responsible AI, which also allows for global inclusiveness. Efforts to report and perform peer reviews could enhance existing supervisory practices, and prevent accountability gaps.*

4. Structure of a global AI institution

The institutional functions can be institutionalized by the establishment of a Global AI Institute, which would realize IF2: a Global AI Governance Framework. Figure 1 provides a conceptual outline of how such an institution could function and be organized.

At the core of the global set up would be a Global AI Institute (IF2), which would serve as the hub for the cycle. The international institution supporting and hosting the Global AI Institute would provide the secretariat to support IF1, IF3, IF6 and IF7. The Institute would have national and/or regional representatives that govern and steer the global governance function of the Institute.

Figure 1 - Conceptual framework for a Global AI Institute



The Institute would be responsible for global horizon scanning of current and future AI-related risks and incidents (IF1). This forms the foundation for ideally achieving consensus on risk mitigating measures. Such an approach would be crucial for managing the development and use of AI systems globally. As explained in the previous paragraphs on the institutional functions, IF1 would lead to consensus building towards minimum alignment of standards, safety and risk management frameworks (IF3) which provides a framework for reporting and peer review (IF6) of which the outcomes contribute as periodic input for global horizon scanning (IF1).

In addition, the Institute can support accountability and compliance assessment (IF7) of national and regional frameworks for regulation, supervision and infrastructure to support credible, trustworthy and responsible AI, which also allows for global inclusiveness. It is pivotal that the roles of supervisors and regulators are emphasized and a central part of the institutional functions, in particular the global horizon scanning of risks and incidents and the follow-up through the development of standards, safety and risk management.

***