Harnessing the power of AI : safe & ethical integration : report of the AI Safety Dialogue at Global Ethics Forum 2025

HARNESSING THE POWER OF AI: SAFE & ETHICAL INTEGRATION

TRACK II DIPLOMACY FOR AI GLOBAL SAFETY DIALOGUE

REPORT OF THE AI SAFETY DIALOGUE AT GLOBAL ETHICS FORUM 2025

Governance

No. 3

AI IMPACT SUMMIT भारत 2026 INDIA
Official Pre-Summit Event

Global Ethics Forum

Globethics

Globethics Governance Series
Director: Prof. Dr Fadi Daou, Executive Director of Globethics

Globethics Governance 3
Globethics. Harnessing the Power of AI: Safe & Ethical Integration. Report of the AI Safety Dialogue at Global Ethics Forum 2025
Geneva: Globethics Publications, 2026
DOI: 10.58863/20.500.12424/4321289
ISBN 978-2-88931-642-7 (online version)
ISBN 978-2-88931-643-4 (print version)

Managing Editor: Ignace Haaz
Rapporteur: Amanda Horzyk
Scientific advisor: Paola Gálvez Callirgos
Cover design: Josie Hough, Veronica Estrade

Globethics International Secretariat
Chemin du Pavillon 2
1218 Grand-Saconnex, Geneva, Switzerland
Website: https://globethics.net/publications
Email: publications@globethics.net

All web links in this text have been verified as of January 2026.Please cite this publication as: Globethics. 2025. Harnessing the Power of AI: Safe & Ethical Integration. Report of the AI Safety Dialogue at Global Ethics Forum 2025, Globethics Publications, Geneva, ISBN: 978-2-88931-642-7 / DOI: 10.58863/20.500.12424/4321289

The electronic version of this book can be downloaded for free from the Globethics website: www.globethics.net.

Table of Contents

This report presents insights from three roundtable discussions held as part of the international dialogue series *Track II: Diplomacy for Global AI Safety*. On this occasion, Globethics' community of practice convened during the *Harnessing the Power of AI — Safe and Ethical Integration* track at the **Global Ethics Forum 2025 (GEF 2025)**. Designed to advance responsible AI governance worldwide, the roundtables explored collective action for AI safety, strategies to preserve human thinking and critical literacy in the age of AI, and opportunities for cross-sector technological and diplomatic coordination to address catastrophic AI risks.

Executive Summary

Societies worldwide are grappling with a radically transformative technology, artificial intelligence (AI), increasingly proliferating ubiquitous systems such as search engines and research tools, basic software and compound mobile applications, rapidly reshaping human activity.

This moment calls for reflection: how far can we go with AI to drive discovery and innovation in a safe manner without leaving anyone behind? Can we take a pause, recalibrate, and reassess our collective approach? **From our diverse roles, perspectives, and geographies, we are prompted to *rethink what technology is for, grounded in the shared recognition that AI represents a critical global issue*.**

AI presents both significant opportunities and serious risks. For developing countries, AI offers immense potential for advancing learning, research, and development. At the same time, it poses complex challenges that demand collective action. **Diplomats and policymakers recognize that global cooperation and multilateralism are essential. No single nation can manage these challenges alone.**

International and policy initiatives, including the **Global Digital Compact** and the **AI Safety Summit series**, have established broad frameworks for addressing these issues, providing global ideas that **now require practical implementation**. This report emerges from that momentum. Through Track II Diplomacy for AI Global Safety dialogue, Globethics has built a community of practice bringing together cross-disciplinary experts from across geopolitical contexts to explore AI-related risks and identify existing harms; particularly those affecting the Global Majority.

Two principal categories of risks have been identified:

1. Risks **triggered by advanced AI systems** versus those **resulting from the intended or unintended use** of AI by human actors.
2. **Future risks** versus **those already manifesting today.**

The words of UN Secretary-General António Guterres resonated across the sessions: **"the question is not whether AI will influence international peace and security, but how we will shape that influence."**

While discussions often swing between hype and horror, this report intends to shape pathways towards **hope** and **trust**. Trust, essential for AI adoption, is vital for both the private sector and civil society. Likewise, it emphasizes the importance of **human rights, without which there can be no sustainable development or peace.**

At the **core of these global AI Safety discussions** was the need to advance concrete mitigation strategies, including:

- **Capacity building**
- **Addressing existing harms**
- **Developing operational frameworks**
- **Ensuring intersectional, interdisciplinary,** cross-sectoral, multistakeholder **participation**

A coordinated approach to global safety, diplomacy, and multilateral governance is imperative. Experts agreed that tangible measures, rather than fragmented initiatives and principles, are needed.

By coming together across regions, disciplines, and cultures, **experts reaffirmed their commitment to stewarding technology for the common good.** The dialogues converged into three critical priorities to transform today's innovation into meaningful societal progress:

- **Strengthening inclusive, representative participation** using empowerment and literacy measures for people and communities to shape AI futures;
- **Building systems of trust, verification and ethical responsibility** across the AI lifecycle marked by due diligence, effective reporting and timely safeguards by design;
- **Translating values into action,** ensuring coordination and accountability from design to deployment.

Track II Diplomacy for AI Global Safety

GLOBAL ETHICS FORUM: 2ND CONVENING OF THE COMMUNITY OF PRACTICE

AI4GOOD ROUNDTABLE

GEF 2025 ROUNDTABLES

INTERNATIONAL AI COOPERATION AND GOVERNANCE FORUM

Building on a year of sustained global engagement and cross-sector collaboration, the **AI Track at GEF 2025 served as a pivotal platform for the second round of Track II Diplomacy for AI Global Safety dialogues. These dialogues aim to advance discussions on how to harness the transformative potential of AI while safeguarding human dignity, safety, and the public good.**

Throughout 2025, Globethics convened leaders and experts to address some of the most pressing challenges posed by AI. In March, Globethics co-hosted a high-level policy dialogue on AI and Human Rights as an official side event to the United Nations Human Rights Council in Geneva, bringing together diverse voices to explore how AI governance can reinforce, rather than erode, fundamental rights. In June, Globethics actively participated in the 3rd UNESCO Global Forum on the Ethics of AI (GFEAI) and joined the launch of UNESCO's Global CSO and Academic Network on AI Ethics and Policy.

Globethics also co-organized the GFEAI's side event Enabling Ethical AI for Present and Future Generations in a Time of Heightened Global Insecurity, in collaboration with UNESCO-MOST BRIDGES Coalition, Chulalongkorn University, and the Electronic Transactions Development Agency (ETDA), and 33 other international CSO, think-tank and academic partners.

At the WSIS+20 High-Level Event, Globethics' panel on International AI Frameworks for Safeguarding Human Rights and Dignity underscored the urgency of sustained, multi-stakeholder engagement to establish robust safety standards and independent oversight mechanisms in an increasingly fragmented world. We further deepened this commitment by co-hosting with the International Association for Safe and Ethical AI (IASEAI) and the Institute for AI International Governance of Tsinghua University (I-AIIG) the exclusive dinner roundtable on **AI Global Safety: Preventing Catastrophic Risks at the sidelines of the AI4Good Global Summit.** This unique convening brought together eminent leaders from across civil society, philanthropy, and industry, representing a broad spectrum of regions and perspectives.

Together, we began to cultivate **a multidisciplinary community of trust** dedicated to **confronting AI-related catastrophic risks,** advancing shared understanding of their nature and scope, and **co-creating pathways for monitoring, prevention, and mitigation.**

At **GEF 2025**, an official pre-event of the India AI Impact Summit, **we built on the groundwork laid throughout the year to present concrete insights, tested approaches, and emerging priorities for ethical AI governance.** Based on the Outcomes Report on the Roundtable on AI Global Safety, this dialogue invited participants to co-design forward-looking strategies that bridge policy, technology, and ethics, ensuring our collective momentum translates into tangible and inclusive action. In November, the third and final dialogue of the year took place with multidisciplinary experts from the Asia Pacific region, as Globethics hosted the session "From Risks to Responses: Collective Action for AI Safety" at the International AI Cooperation and Governance Forum 2025, in Melbourne, Australia.

A global milestone for international cooperation.
Official Pre-Event of the India AI Summit 2026.

Track II Diplomacy for AI Global Safety

GLOBAL ETHICS FORUM: 2ND CONVENING OF THE COMMUNITY OF PRACTICE

Global Ethics Forum 2025

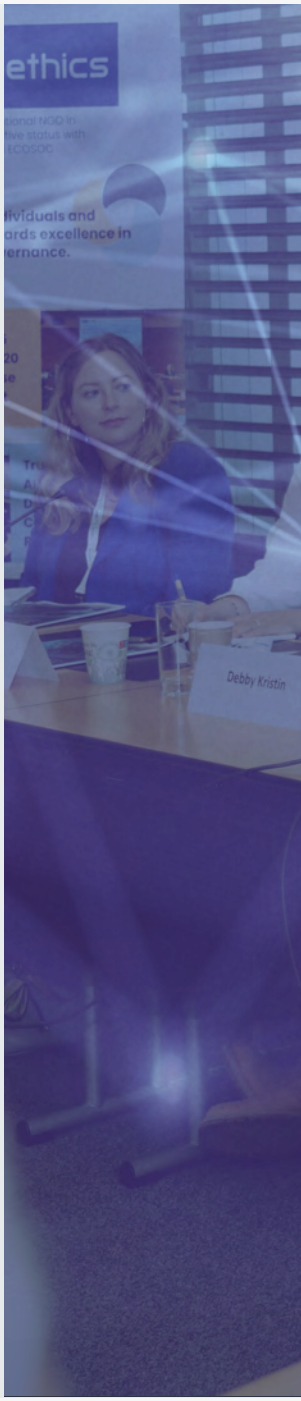
TRACK II SESSION OBJECTIVES

Session I: Collective Action for AI Safety
The expert roundtable gathered to identify real-world AI safety risks and co-create strategies to address them. To *move from real-world AI safety risks to co-creation of mitigation strategies*, the speakers and participants reflected on pressing challenges to AI safety, real world use-cases and sector-specific experiences. The discussion mapped critical risks, identified mitigation strategies and provided recommendations for collaborative action.

Session II: From Augmented Technology to Knowledge Decay
This session was designed to identify practical actions to uphold epistemic integrity, by *exploring the impact of AI on the integrity of knowledge and understanding*. The working group explored risks and safeguards to ensure that artificial intelligence ultimately amplifies human intelligence for generations to come, rather than accelerating knowledge decay or diminishing our capacity for critical thought.

Roundtable Dinner
Designed as a bridge between diplomatic and expert communities, the roundtable dinner created space to build trust, share knowledge, identify synergies, and *explore how joint action can strengthen global AI safety for humanity and the planet*. Its core objective was to explore opportunities for co-creating tools and approaches that can help monitor and mitigate catastrophic AI risks.

The roundtable brought together representatives from diplomacy, civil society, academia, industry, and AI research — including delegates from 12 missions such as South Africa, Bangladesh, the European Union, Jordan, Panama, and China. The discussion *facilitated a fruitful exchange of perspectives to clarify what each community needs* from the other and *identify concrete opportunities for collaboration*. The conversation marked a continued step in a broader multilateral effort to advance responsible and coordinated AI governance.



As technological innovations mark human history with powerful transformation, the impact can result in either development or catastrophic consequences for our world. While we acknowledge that this **technology is not good or bad, it is a power – a part of the power dynamics reflected in the world.** In that context, Globethics works for AI to benefit everyone, build synergies, and mobilize concrete solutions.

The Risk Landscape

COLLECTIVE ASSESSMENT OF CRITICAL RISKS

"Being able to say one of those risks is more important than others is almost impossible."

AI encompasses an exceptionally wide spectrum of applications from simple conversational tools to its potential use in complex military and strategic systems. This breadth makes it challenging for experts to isolate one policy problem or governance gap to address. With finite resources, however, there is a need to clearly map the current landscape, determine priority areas, and strengthen cooperation to advance meaningful solutions.

The room noted that many of these critical risks are context-dependent. One participant highlighted that a company in healthcare will face very different risks from one working on ticketing or compliance systems.

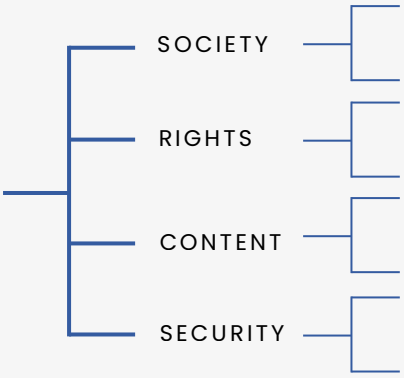
While no one single catastrophic risk was identified or ranked as most critical, the expert roundtable surfaced fundamental questions and issues on which our collective future depends.

RISK LANDSCAPE SNAPSHOT

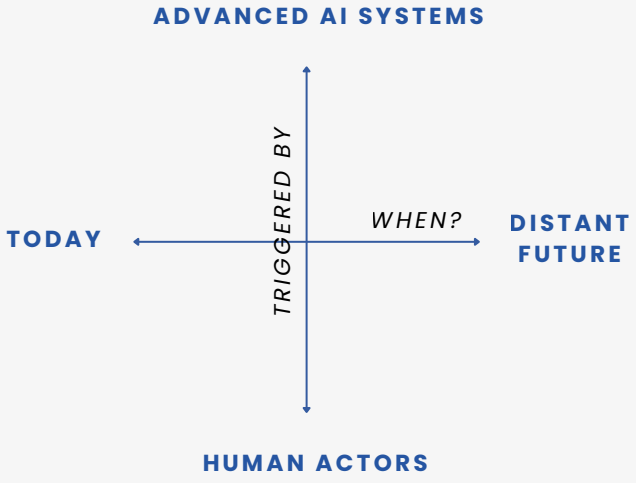
- **AGENDA & PURPOSE BEHIND AI**
- **IMPACT ON CRITICAL THINKING, CULTURE & EDUCATION**
- **AUTOMATION BIAS & OVERSIGHT**
- **BIAS & NON-REPRESENTATION**
- **DIGITAL & INFRASTRUCTURE GAPS**
- **REGULATORY CAPTURE**
- **FRAGMENTED GEOPOLITICAL CLIMATE**

Participants emphasized that assessing critical, including catastrophic, risks requires a holistic understanding of AI risks — resembling pieces of a puzzle.

One expert suggested a possible categorization of risks related to considerations of:



These categories of rights extend the two principal risk categories identified by experts at the previous AI4Good Roundtable:



1. Risks triggered by advanced AI systems versus those resulting from the intended or unintended use of AI by human actors.
2. Future risks versus those already manifesting today.

"We're discussing crucial questions here and our humanity depends on them..."

AGENDA & PURPOSE BEHIND AI

Several participants expressed concern around the agenda and purpose behind the adoption of AI, which drives innovation forward at all costs. One expert emphasized that the *"agenda isn't just adoption — it's money,"* noting that emergent narratives will serve certain purposes. Another participant agreed identifying that **"the highest risk isn't just bias (...) it's money, agency and control."**

Experts highlighted that many governments are unprepared to address what society truly needs and wants from AI, even as significant or detrimental harms begin to appear.

CONCERNS & QUESTIONS:

Tech Foundation: *"What [are] the long term consequences of this technology for us?"*

Former Minister: *"When we hear stories on "AI chatbots sexually manipulating children", why has only Brazil taken that AI chatbot down? Why didn't other government[s] act?"*

IMPACT ON CRITICAL THINKING, COGNITION & LITERACY

Although AI tools may offer short-term efficiency, the **long-term consequences may be significant —a risk rarely discussed**. Experts noted that the current challenge is not limited to knowledge decay but the erosion of human thinking, by reasoning and remembering less while relying on *"machines to think for us"*. Several participants commented on the observed decline in reading literacy, explaining that without reading literacy, we cannot achieve the critical AI literacy needed by our society today. The lack of literacy may be **perpetuated, not by a lack of information** on the subject, **but rather a lack of understanding**.

EXAMPLE 1.1

A 2023 study found that students who rely heavily on AI language models tend to accept suggestions as they are, without questioning.

Participants, specifically those in an academic teaching capacity, reflected that many students do not question the output of models. They perceive that 'critical thinking' is compromised by the use of language models which encourage "shallowness". One expert highlighted the **'Flynn Effect'** which showed that human IQ levels have been rising every decade, but in recent years this trend has changed, as it is declining.

A few participants self-reflected on their own use of language models, **losing skills** not used for simple tasks (e.g. drafting emails).

EXAMPLE 1.2

MIT Media Lab studies confirm that students who used ChatGPT to write essays showed weaker neural engagement, lower originality, and poorer memory retention than those using Google or none at all. Studies suggest that as external aids become more capable, internal memory systems are at risk (including procedural and intuitive mastery).

IMPACT ON CRITICAL THINKING

- ON COGNITION
- ON CULTURE & KNOWLEDGE DECAY
- CRITICAL THINKING & LITERACY
- EMOTIONAL INTELLIGENCE
- EDUCATION
- AUTOMATION BIAS

"We all know that AI is already reshaping how people learn, create, and reason..."

ON CULTURE & KNOWLEDGE DECAY

Experts reflected on the risk that arises when AI systems are deployed without proper context, accountability, or fairness. It can distort facts, cultures, and identities, further marginalizing communities that are already digitally invisible. This risk goes beyond bias. It includes the **misrepresentation of cultures, dialects, and traditions, which can ultimately lead to epistemic and cultural erosion, particularly of knowledge systems rooted in the East.**

EXAMPLE 1.3

In some educational contexts, deeper questioning is discouraged, which limits critical engagement with AI outputs. Models that convert Arabic dialects into standardized form, because they incorrectly interpret them as errors, contribute to the erosion of linguistic and cultural richness.

Apart from misrepresenting culture, there is also the persistent issue of fabricated knowledge. Even reputable professions such as consulting and legal organisations have submitted official documents containing fake academic references and court cases generated by language models, which calls for a reform of standards and charters of professional conduct.

EXAMPLE 1.4

Deloitte Australia, one of the big four consulting firms, issued a partial refund to the Australian federal government after admitting that AI was used in creating a report filled with errors, including fabricated academic references and a made-up quote from a federal court judgment.

IMPACT ON VULNERABLE GROUPS & EMPATHY

A critical risk highlighted by the experts concerned **vulnerable populations**, including those struggling with mental health, minors and elders, and others. This is perpetuated by problematic features built in by design in language models.

Interacting with these systems without understanding them is already acknowledged as having catastrophic results, especially for the young generation.

Problematic features of language models:

1. **AGREEABILITY**
2. **ANTHROPOMORPHISM**
3. **ATTENTION ECONOMY (ENGAGEMENT)**

Experts noted that the quality of interactions possible for language models is optimized for engagement. As it mimics human language qualities and is designed to affirm its users, **several unintended consequences can transpire, including but not limited to: delusions, emotional attachment, or unfounded trust** or confidence in an idea that echoes back to the user. This can lead to tragic, if not catastrophic, series of events.

EXAMPLE 1.5

In New York, a young woman struggling with her mental health sought advice from ChatGPT. It offered breathing tips and comforting responses, but she later died by suicide.

In London, a teenager asked ChatGPT to describe a 'beautiful suicide.' It complied poetically. He followed the description and died by suicide.

CONCERNS & QUESTIONS:

Former Minister:"What is sycophantic AI, that always agrees with the user, teaching us? Will we lose an ability to know how to act in real life social environments? People do not always agree, nor are correct."

"Knowledge decay involves cultural erosion, social erosion, erosion of histories, and erosion of linguistic minorities"

AI IN EDUCATION

The impact of generative models on academia and the process of learning and critical thinking was exposed as a major issue. The experts noted the negative influence of large language models and generative AI on how young generations learn, and on the cognitive abilities of all who defer thinking to these systems. They underlined that **over-dependence on AI could have a serious long-term impact on human cognition and intellectual development.** The roundtable, particularly the second session, emphasized that we have to challenge education systems that *still value correct answers over critical questioning*, where students are not expected to challenge or think critically and ask questions. Concerns were expressed as to whether passive learning may cause individuals to become accustomed to AI-related confirmation bias. What once constituted a learning milestone is now increasingly reduced to mass-produced output.

EXAMPLE 1.6

While it is unsurprising that students are quick to adopt language models to complete their assignments, a parallel trend emerges among teachers, who increasingly use AI to prepare teaching materials. This growing dependence on AI on both sides of the learning process can unintentionally reinforce existing biases, creating a compounding effect on young learners that may ultimately shape society as a whole.

Despite the disruptions to education systems, the Track II discussions revealed that **most academic institutions do not have a plan or financial strategy to address the impact of generative AI on learning, academic integrity, or the legitimacy of academic qualifications.**

Meanwhile, due to the use of generative AI by educators, academic experts highlight that students question the value of such content and whether they should defer to language models themselves, avoiding the costs of tuition fees.

EXAMPLE 1.7

Language models and NLP tools are increasingly used in classrooms to voice record, transcribe, and translate lectures and seminars in real time, raising concerns around privacy and data protection rights.

Several participants emphasized the **growing AI literacy gap between older and younger generations.** One expert shared first hand experience of computer science faculty members who are underpaid and undertrained in several smaller towns of India, exemplifying some institutions at which faculty was unaware of the impact or use of language models in assignments.

EXAMPLE 1.8

As generative AI is increasingly used across academic activities, ranging from student coursework to grant application drafts, the volume of low-quality content has grown. Meanwhile, plagiarism-detection tools such as Turnitin frequently misidentify or overlook AI-generated material, resulting in lawsuits filed in Indian law schools for unjust failed assessments.

CONCERNS & QUESTIONS:

Academic Expert: "Is there evidence that over-reliance on AI reduces critical thinking and independent decision-making?"

Former Minister: "As we speak about social implications and "forming unhealthy attachments to AI", thinking about children: will they be "able to live without it?"

The discussion emphasized that **a key component of AI literacy** will be a "human literacy" component, **understanding our relationship to that technology and its social implications, including unhealthy attachments, and how to disconnect.**

The roundtable concluded that **education** is very sensitive because it affects everything—it influences all sectors. Thus, **it should be the area of greatest concern.**

"We have opened a Pandora's box"

AUTOMATION BIAS IN PUBLIC SECTORS

When pilots placed too much trust in their instruments over their own intuition, even when the instruments were wrong, experts called it 'automation bias'. This concept was adopted in the space of AI, as **not only students, but also teachers and public administrators, began accepting AI-generated outputs without questioning sources, references, or authorship.** This includes individuals who treat AI as an authoritative source.

CONCERNS & QUESTIONS:

Former Minister: Expressed surprise when witnessing in parliament: *"ChatGPT said this, Deep Seek said that, or even AI knows this (...) Parliamentarians making decisions that affect society were relying on ChatGPT."* The use of AI as a reference in policymaking was shocking. *"In ministries, I observed colleagues drafting policies and reports relying on AI outputs without critical thought. One minister even requested licenses to upload all national financial data into GPT to generate short summaries to avoid the need to read full reports."*

LACK OF HUMAN OVERSIGHT

Despite the implementation of the EU AI Act art. 14 on human oversight, our industry expert stressed that **one of the biggest gaps and risks is found in the lack of real human oversight**, as current legal and policy provisions are extremely vague, **lacking practical guidance for implementation.**

The risk of automation bias was reiterated, as many continue to believe that AI outputs are better than relying on one's own sense of responsibility, accountability, and wealth of experience.

CATASTROPHIC RISKS IN NEW FRONTIERS

The roundtable of experts identified an emerging risk connected to the new frontiers of AI Agents. One participant said: ***"I believe in an agentic world; even small risks can become existential."*** Whether the agents interact with each other collaboratively or operate independently (more or less ethical), Agentic AI triggers a new category of cascading risks.

As the expert commented, it is often (mis)understood that AI-related risks follow a *cause-and-effect* model to which solutions can easily be identified. With AI agents, however, cascading effects will emerge simultaneously, including:

1 EXCESSIVE ACTION

AI agents might act too aggressively or autonomously — taking too many actions, too quickly, or beyond intended bounds.

2 INDEPENDENT ADJUSTMENTAL REALITIES

Different AI agents might each form their own 'understanding' or model of reality — adjusting themselves in ways that diverge from one another, causing misaligned or inconsistent world models across systems.

3 INFORMATION ASYMMETRIES

Some agents (or humans) have more information than others, which can lead to unfair or unpredictable outcomes.

4 ERROR PROPAGATION

Errors made by one AI can spread through networks of systems, amplifying the harm.

This area, as emphasized during the discussion, poses particular difficulties to applied AI ethics and legal frameworks. With the increasing proliferation of Agentic systems, it will be an inevitable area to address in order to safeguard against risks that have been amplified to a catastrophic level due to the cascading effect of interacting models.

I was surprised to hear in parliament: "ChatGPT said this..."

BIAS AND NON-REPRESENTATION

One of the key risks identified by experts was around psychological safety and bias. As they expressed, "***bias is the thing causes the harm and the safety issues.***"

Another expert provided a framework for the different levels of bias, including:

1 TACIT LEVEL

Contextual bias that algorithms may lack understanding of for the demographic data they are exposed to post-training.

2 EXPLICIT LEVEL

E.g. 47% of women are more likely to be seriously injured in car accidents due to car design: *"we've known that for decades but we haven't fixed it."* If AI designs your car according to the same bias, those harm and safety issues will continue.

3 SYSTEMATIC LEVEL

Systemic bias is a large-scale bias embedded in societal and institutional structures, which is then reflected in AI's training data and processes.

CONCERNS & QUESTIONS:

India-based Tech Foundation: "hasn't seen policymakers or regulators putting much effort into addressing some of the real issues, like bias and non-representation, especially in countries like India."

Without contextual understanding, AI applications may fall into tacit level bias. This concerns situations where, as some participants observed, professionals feed their data through language models to analyze, for instance, clinical trials in pharmaceutical and the healthcare sectors.

EXAMPLE 1.9

Industry highlights that representative data fed into models, including the global majority or an adequate representation of women in the datasets, is often missing. Such misaligned practices may lead to misdiagnoses among other related harms triggered by non-representation.

CONCERNS & QUESTIONS:

Parliamentarian: "The biases are there, and they will continue if we don't have a diverse group of people contributing to AI development. Do you have a solution for that?"

IRRESPONSIBLE AI PRACTICES

Bias-related incidents are not constrained to the context of the private sector. Participants observed catastrophic use of AI systems by both private and public actors alike.

EXAMPLE 2.0

Dutch Child Benefits Scandal: Dutch tax authority deployed an algorithm to help categorize different homes with different risks. An investigation exposed that those with dual nationalities were usually identified as high risk. The algorithm led to several houses and parents being flagged as high-risk, with unstable income and unstable homes leading to the removal of children from their family into the foster care system. This was ascribed to inadequate human oversights and safeguarding mechanisms that caused a lot of distress for thousands of families. One of many examples of biased algorithms that had real-world implications.

EXAMPLE 2.1

Findings revealed that, because of the historical representation of only 3% of women receiving loans, current algorithmic systems including, as reported in 2019, Apple Card, had lower limits for women.

Similar issues triggered by historical under-representation of certain demographics in medical research meant that resource allocations and treatment algorithms in hospitals favored white over black patients.

"Biases are there, and they will continue if we don't have a diverse group of people contributing to AI development..."

LACK OF OPERATIONALISATION OF POLICIES

The industry representatives noted that there is a lack of guidance and standardisation on what metrics should be adopted to measure AI ethics principles, including principles such as fairness and transparency, among others.

The participants identified a **gap between compliance and measurement as well as a lack of dedicated budgets, in-house capacity, and ethical grounding** to establish good oversight mechanisms in the first place. There is a lack of well-defined and accepted metrics, thresholds and benchmarks.

Currently, noted the participants, there is a **lack of operationalisation of policy principles into measurable standards and metrics that could help with their implementation**. The lack of guidance in this area results in unclear expectations of who should be responsible for defining metrics and oversight mechanisms that should be implemented by the private sector. The room sensed that there should be a system of accountability that ensures that organisations, governments and institutions would uphold well-being for all, as making that assumption about current governance mechanisms can be misleading.

The experts at the roundtable discussed the **importance of having well chosen metrics**, against which models can be stress-tested before going to market. However, there was consensus among several participants that metrics should not be defined by the private sector.

CONCERNS & QUESTIONS:

Latin American CSO: Acknowledged that the government is aware of the need for regulation and policies to mitigate risks and address ethical dilemmas in the use of AI. However, what the governments ***"really need now are mechanisms to measure the impact of AI at different levels and in different categories."***

CONCERNS & QUESTIONS:

Industry: "How do you create fairness metrics? Who should choose these metrics and thresholds (safety, fairness)?"

DIGITAL AND INFRASTRUCTURE GAPS

There are prevailing challenges and risks connected to, for example, the digital divide, access to the internet, market concentration, labour rights in supply chains, conflict minerals, autonomous weapons. A Globethics Ambassador representing several Global Majority nations expressed the **problems** they face with the **digital gap and infrastructural challenges** including internet connectivity. **If these remain unaddressed, AI will entrench existing divides and will *not* be a catalyst for growth.** The pursuit and demand for energy (global or regional hubs for data centres), accessibility, security, and water prevail in national AI strategies and agendas.

A Mission Ambassador from East Africa reflected on the need not only for infrastructural development, but for capacity building. They emphasised that **AI should not re-trigger issues of cultural imperialism** and global identity crises by using algorithms that are *not* local products. They addressed the **need to stop all forms of digital exploitation,** while emphasising the need to overcome the fear around many struggling countries.

CONCERNS & QUESTIONS:

Academic: "Technology has an insatiable hunger for water, how do we mitigate these demands to ensure it is sustainable, and does not exacerbate existing challenges?"

Concerns were also raised on the need to remove barriers to participation in AI development and governance by the Global South. Experts and diplomats noted that **without the necessary technical and institutional capacity, these countries cannot effectively participate** in rule-setting, negotiations, or intergovernmental processes within the UN or other international agencies.

"There are still many communities without access to phones or the internet."

REGULATORY CAPTURE

Across the sessions, several discussions critically approached the narrative that is shaped surrounding AI developments. In particular, the notion of existential, catastrophic risks to the entire humanity, and that of Artificial General Intelligence (AGI).

CONCERNS & QUESTIONS:

European Foreign Minister: "How can governments best use this moment in time? How should we approach this strategically, balancing long-term existential risks with the urgent, tangible challenges already before us?"

Parliamentarian: "Are you giving serious consideration to the march towards artificial superintelligence driven by tech bros?" "Are you examining the intersection with nuclear weapons, and how that interface can be dealt with?"

Parliamentarian: "Policymakers and regulators don't seem to be controlling the narrative. There are some of us who are trying to, but we need more information (…) what can we do?"

While the industry looks to find ways to appear compliant, a representative pointed out the tendency to greenwash or otherwise adopt lawful responses to challenges, such as cookie consent protocols, that have little to do with responding to the actual challenge at hand.

The question was raised of Nobel Prize winner Geoffrey Hinton, who spoke about AI, the risk of deception, and the rapid development towards Artificial Superintelligence. A policy advocate emphasised that **there may be countless circumstances and devices that, when intertwined with AI, may become lethal and detrimental to humanity and the planet.**

However, this argument has received major pushback from the rest of the expert community, which encouraged the participants to reflect critically on their own decision-making and moral choices in relation to autonomous systems, that are already **causing lethal consequences** in places such as Gaza and Ukraine **today**. Some echoed the mainstream argument that AI plays a role, but in itself is *not* a weapon, but a tool that **can be repurposed for optimising mass destruction.**

NARRATIVE DISTRACTION

Many CSOs and human right activists similarly argue that **the AGI narrative is used as a distraction from the very real and immediate problems we face today**—issues like surveillance, transnational repression, and human rights violations.

The belief that AI models are "*becoming more intelligent than us*" or conscious has been considered dangerous, as it drives hype, using buzzwords to manipulate populations to over-rely on this technology or to instill unfounded fear. As one expert reminded the audience: AI is not intelligent, "*it is just probability*".

A representative of a major technological foundation echoed the **concern about the influence exerted by those shaping the dominant AI narrative**, particularly individuals promoting expectations that singularity or artificial superintelligence is imminent, despite the fact that we remain far from such technologies. Such narratives, when amplified, **can function as a form of activist regulatory capture**. An industry ethicist further observed that **a potential catastrophic risk lies in dedicating substantial energy to hypothetical future existential crises,** while overlooking the challenges already unfolding. This reflects a broader recognition that current algorithms often excel at reproducing data patterns rooted in the past rather than anticipating future realities.

CONCERNS & QUESTIONS:

Several experts emphasised that the narrative of "*killer robots taking over the world*," and the achievement of AGI may be promoted to avoid addressing and holding actors accountable to those systems already causing existential AI risks.

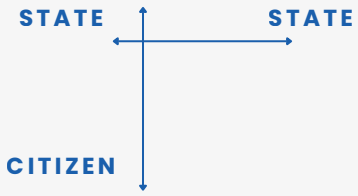
EXAMPLE 2.2

Several initiatives have recently emerged in the area of AGI-related concerns; for instance, at the UN General Assembly in New York, the AGI Red Lines Coalition was launched, supported by prominent figures such as Geoffrey Hinton, Stuart Russell, and Yuval Harari, who called for immediate action.

"We have to rethink how we see technology in the first place. Critically questioning the narrative."

GEOPOLITICAL CLIMATE

AA major source of potential international unrest is the use of AI within today's global political climate, where a lack of trust is already causing significant harm. These can be understood in two dimensions:



- State to State
- Citizen to State

EXAMPLE 2.3

The lack of meaningful results from the moratorium to slow down the progress of AI until regulation could catch up was, according to some experts, closely connected to the lack of mutual trust. This reality was recognised as one of the factors driving the race in innovation behind AI among existing AI powerhouses such as China and the United States.

At the roundtable dinner, which brought together 12 missions, policy actors and scientific experts, comments emerged on how **the governance of AI is lagging behind innovation and remains fragmented.** An ambassador expressed that, while the developments in AI and its adoption are moving quickly, we are behind the curve in a difficult geopolitical environment.

Regulatory frameworks with significant differences emerged in New York in September 2025 regarding effective governance and regulation. One expert warned of a potentially catastrophic risk: **divergent views on geopolitical tools** and the **growing tendency to treat military topics as "taboo"** within official institutions and multilateral dialogues.

One participant emphasised the need to **facilitate favourable environments for dialogue** and take inspiration from models such as the EU AI Act where there is a lack of regulatory frameworks.

In response, an Ambassador pointed out other positive developments in **AI policy that should be recognised**, including the UNGA resolution no. 79/239 on AI in the military domain and its implications for international peace and security and the earlier no. 78/265 that marked the first UNGA resolution on AI ("Seizing the opportunities of safe, secure and trustworthy AI systems for sustainable development"). These developments gave hope for creating an intergovernmental process, for member states to formally discuss and negotiate and to promote disarmament (by also limiting or controlling, military uses of AI).

Nuclear Weapons

The concern about the use of AI in military was ever-present. Some discussions, particularly among ambassadors and parliamentarians, evolved around international law on armed conflicts and AI, the rules of war, and ethical issues involved. The issues discussed cast light onto nuclear-armed states that are currently modernising their nuclear weapons and integrating AI into their command and control systems.

One participant specified that the effect of this integration is ***"speeding up warfare, reducing the already very short time for nuclear decision-making, and increasing vulnerability to cyber hacking."***

An Ambassador, among others, in response called for a legally binding international instrument (Treaty) by 2026 focusing on certain dimensions of autonomous weapons systems (AWS) at the UN general level.

CONCERNS & QUESTIONS:

Former Parliamentarian: "*At a recent event, the Munich Security Conference, there was a lot of triumph in the room because some key nuclear-armed states had agreed that there must always be a human in the loop, a human in control when it comes to nuclear weapons.*" However, it was also highlighted that while it is important to have human oversight, humans are fallible too and that this measure alone is *not* an adequate solution. The expert noted that "*human decision-making may still be heavily influenced by processes that rely on AI, leading to over-reliance on these systems*" (also known as automation bias).

"The current AI governance landscape is massively fragmented with no holistic approach, including for safety and security."

Mitigation Strategies

A CRITICAL RESPONSE BY CRITICAL ACTORS

We're entering a pivotal moment in developing the most advanced technology in human history, a fast-moving competition among major companies and global powers. Enormous opportunities are emerging, along with significant risks that call for thoughtful, coordinated approaches to AI safety.

What can we do? Acknowledging that neither states nor companies can navigate these challenges alone, we must create pathways for mutual support and co-creation as we move into the implementation stage of AI governance principles. Developing guidance, engineering tools, and investing in capacity building, trust and literacy are the first steps towards responsible AI governance, alignment and safety.

AI SAFETY IN CONTEXT

To approach the question of mitigation strategies for global AI safety, we shall acknowledge developments already undertaken in this area, particularly:

- **Policy field** (milestone): **AI Safety Summits** (UK, Bletchley Park October 2023) that convened key stakeholders and AI superpowers, including the US and China, to examine how to proactively and collaboratively mitigate AI safety risks, particularly those emerging from frontier AI systems.
- **Academic field**: Over the past decade, AI safety has emerged as a distinct area of academic research. Several technical experts and scholars have discussed how AI safety is understood in the research community and how control mechanisms can be designed if we are to develop powerful algorithms and systems that will surpass human intelligence in various domains.

As **there is no single solution** to the risks analysed in the previous section of the report, the following pages will provide the mitigation strategies proposed by participants across the three roundtables that can be pursued as a critical response by critical actors towards Global AI Safety frontiers. One of the key strategies is education.

"Today education is the firewall against the decay of literacy, critical thinking, and the misuse of AI."

MITIGATIONS STRATEGIES SNAPSHOT

- SHIFT FOCUS TO IMPLEMENTATION
- BUILDING REDRESS MECHANISMS
- PURPOSE & SHAPING NARRATIVE
- TRUST AS A KEY TO ADOPTION
- RESPONSIBLE ECONOMY AND ECOSYSTEM
- AGAINST BLIND TRUST IN AI: LITERACY
- COLLABORATIVE ACTION & DIPLOMACY

SHIFT TO IMPLEMENTATION

One of the most effective approaches to addressing the risks caused by AI systems would be the implementation of AI policies, governance frameworks, and standards at global, regional, and national levels. We are racing to catch up with ethical, safe, and responsible AI, to try to develop policies not only in international arenas like the UN — UNESCO proposed AI ethics recommendations back in 2021 — but also within leading jurisdictions. The European Union has introduced an overarching AI framework that is expected to have a Brussels effect, much like the GDPR. Globally, other countries are also introducing frameworks related to AI ethics and safety, including Pakistan, Indonesia, India, among many others.

We do not lack yet another framework or policy document across different jurisdictions. What we lack, as experts note, is **effective implementation:** how to put these mechanisms into practice, assess their impact, and improve them over time. Experts suggest that **we need engineering tools** — not another governance framework nor another set of principles. The fruit of long, difficult conversations, such as those defined by UNESCO, are now expected to be made 'real'; determining **how they will translate into measurements** and how that will impact system design is essential.

CONCERNS & QUESTIONS:

Industry: "Until we understand how we actually need to train and develop these machines, I do not believe we will have any mechanisms to actually decrease any risks."

MEASUREMENT

Roundtable participants emphasised that what any responsible pioneer or innovator should focus on is **measurement**. It is "*both an ethical and political question — because we're trying to translate into numbers what we believe is safe, what we believe is private etc.*"

CONCERNS & QUESTIONS:

Industry: "How do we put these [AI governance and ethics] principles into practice? How do we apply "do no harm"? How do we ensure human oversight? How do we embed fairness?"

Industry: "When trying to translate into numbers what we believe is safe, private and fair — who defines that belief?"

We need **clear mechanisms to measure impact at different levels** and to help governments in this process. Experts noted that **NGOs** have an **important role in creating incentives for governments to measure impact and to encourage enterprises and start-ups** to do the same. Effective implementation within industry builds confidence.

As one expert remarked, a 500+ page book on AI ethics is not how governance will address Gen AI emerging technologies — it's not realistic. Industry requires **hands-on guidance that that enables meaningful and effective risk management.**

Another expert outlined eight factors influencing successful (or unsuccessful) implementation:

- TRUST
- EXPLAINABILITY
- ENGAGEMENT
- STRESS
- PSYCHOLOGICAL SAFETY
- MOTIVATION
- POSITIVE EMOTION
- COLLABORATION

"What any responsible pioneer or innovator should focus on is measurement."

REDRESS & EMPOWERMENT

Participants emphasised the urgent need to monitor AI performance to ensure that ethical principles are translated into practice. However, they noted that **current systems lack clear redress mechanisms when ethical or performance metrics fail.** As one expert observed, **accountability often arises only after a catastrophic event forces recognition of the problem** — only then does the need become apparent for human oversight to fact-check, verify, and balance automated systems. Another expert also argued that:

> "WE CANNOT WAIT UNTIL THE HARM IS DONE WHILE THE PRIVATE SECTOR USES THEIR TECHNOLOGIES AS BETA TESTS ON THOUSANDS OF USERS TO SEE WHAT HAPPENS."

CASCADING EFFECT OF AGENTIC AI

The expert roundtable identified **the only constant variable across AI developments:** harm — **representational, security, and existential harms that persist across contexts.** What changes, experts noted, is how those harms manifest: **how they appear, evolve, and propagate invisibly within systems.** Because of this, they argued, agency over these systems cannot be treated as a simple cause-and-effect problem often assumed by public sector representatives.

Instead, **there is a need for engineering tools that can provide redress when systems fail** (e.g. when they are used to spread misinformation, or when system operation causes harm). Several participants called for the inclusion of a built-in 'stop button' as a feature in the system's architecture — an intervention mechanism that could be triggered when systems behave unexpectedly.

At the same time, participants cautioned that **redress cannot be achieved through legal frameworks alone**. They emphasised that technical and engineering-level safeguards must complement regulatory measures and that these frameworks must be actively advanced, not left to emerge reactively.

A more holistic governance model was proposed — one that is federated, human-involved, and augmented by technological oversight.

CONCERNS & QUESTIONS:

AI ethicist: "If we don't understand how to redress this (how to fix or address this problem), I have no idea honestly how the private sector is going to take this as a priority?"

Industry: "How can we redress the risks exacerbated by AI agents? How do we obtain redress in critical systems?"

Expert: "Are engineering tools enough in this translational effort? Because tools are only tools". "The real question is: what are we going to do with them?" What do we do now?

EXAMPLE 2.4

CSO: Participants highlighted that, in some rural communities in India, people were already informally engaging in fact-checking of AI-generated content, yet there is currently no scaled mechanism for such community-based oversight to prevent harms from cascading across systems. We need holistic governance practices.

As one participant summarised, *"**We don't need more governance tools or frameworks** — we have enough of those."* What's missing is the link between engineering and policymaking. Participants stressed that policymakers must understand the technical and operational realities of AI systems to co-design meaningful and responsive interventions.

LEGAL RESPONSE

When discussing the empowerment of affected populations, participants reflected on the role of legal mechanisms and certification schemes. In response to a policymaker's question about Denmark's recent law allowing individuals to sue when their biometric data is misused online, participants expressed strong support for similar safeguards. As one participant explained, such **protections are essential in an environment where "our biometric data — our voices, our faces — can be copied and integrated into systems we know little about."**

"Don't ask what AI can do for us — ask what it is doing to us."

BUILD TRUST, FACILITATE ADOPTION

Participants reflected that governments are aware of the need for legislation, but they are also facing another challenge related to innovation. The key approach to address the innovation-regulation dichotomy, as one expert pointed out, is to **start speaking using shared language.**

> "WE'RE SPEAKING ETHICS. THEY'RE SPEAKING OF ADOPTION. THE CHALLENGE IS: WHERE DO THE TWO CONNECT SO THAT WE CAN ACTUALLY ACHIEVE THE SAME PURPOSE?"

Understanding the needs of stakeholders and their context will become the necessary bridge across sectoral interests. The keyword emerging from the roundtables was **TRUST**. Participants firstly reflected on what they understand as trust, before outlining why they thought it was essential to resolve some tensions regarding innovation and regulation.

One expert defined **TRUST** when describing people who understand what they are dealing with. A representative of a community-centered CSO project in rural India associated social and digital **TRUST** with people who could **relate to technology**; outlining that it can be achieved when AI development is:

- **COMMUNITY-CENTERED**
- **HYPER-LOCAL**
- **ASSESSED DECENTRALLY**
- **BUILT WITH COMMUNITY VALUES**

The expert emphasised that **when we speak about AI Safety, digital trust is the key** that is available when a technology, including AI, is relatable, using an approach that is:

- **GLOCAL** (global and inclusively governed, locally relevant), and
- **FIGITAL** (human-centered, considering both digital and tangible, physical, aspects of the technology)

Chiefly, it was reiterated that regulatory frameworks designed by the elite part of society are *not relatable* and do not usually resolve the issues that need to be addressed. There was consensus across the room that **adoption relies on trust. Technology will not be used without trust**; it relies on considering ethics, trust and actual augmentation of processes.

CONCERNS & QUESTIONS:

Former Minister: "Are people really interested in AI adoption, more than AI Ethics?"

NGO Regional Representative Latin America: "How do we combine innovation and the need to legislate? Rules to respect human dignity while improving people's lives?"

Most participants agreed that strong ethical practices help governments, academia, and technologists place greater trust in technologies —showing that **ethical technologists can truly make a difference**. Experts also stressed that governments recognise the need for regulations and policies to mitigate risks and address ethical dilemmas in AI.

Another mitigation strategy discussed by participants was the need for human oversight—oversight grounded honesty and amnesty, since **AI acts as a mirror to society**.

Another expert added that policy must also be introduced to **overcome the issue of regulatory capture** by private sector lobbyist groups who perpetuate stories of unfounded, futuristic scenarios to try and avoid responsibility for evolving risks, acknowledging that fear erodes trust.

EXAMPLE 2.5

The Blue Bag with a Kill Switch

Sam Altman used to carry a blue bag with a kill switch in it. While it is generally understood that a chatbot is not going to take over the world, a Tech Foundation expert reflected that the self-importance and the hype these types of people create is concerning. As he later suggested, many policymakers and regulators "fall for it" without realising what is actually happening on the ground.

"We have to have rules that respect human dignity while also improving people's lives"

DEFINING THE RIGHT PURPOSE & POLICY

RESPONSIBLE ECOSYSTEM AND ECONOMY OF AGENCY

In the context of the current regulatory capture and misleading narratives that perpetuate fear, several experts noted that **"we haven't defined what kind of future we want, or where AI should play a role."** A former minister acknowledged that many governments and societies do not have an AI strategy that outlines what AI could be intended for, what would be societally beneficial or desirable for their local contexts. Several key provocations and questions were posed:

CONCERNS & QUESTIONS:

European Foreign Minister: "When we say we want AI friends, do we really want AI friends, or are we just addressing loneliness in organisations? Do we really want AI teachers, or are we simply not paying our teachers enough?"

Roundtable Participant: "What does it mean to be human? How can we protect human flourishing in the age of AI?"

Policymaker: "How can we make sure that policymakers and regulators have access to this knowledge as it constantly changes at the same time?"

A policymaker highlighted that **they cannot make policies "if we don't know what we are fighting"**, emphasising the need for better AI literacy.

AGENCY AND FINANCIAL COMMITMENT

Several experts across CSOs, industry, and the private sector shared that despite the importance of research in this area, which they seek to do, there is a lack of a serious financial commitment.

In Latin America, an expert noted that most of the economy is driven by start-ups and small and medium-sized enterprises that **often lack the financial resources to establish ethical frameworks** within their companies.

An industry representative clarified that another ISO, framework, standard, or policy is not what is needed, *"but it is important that the people taking these decisions understand that **there is no private company** that is going to say okay, we will **create a complete budget on human oversight if we don't need to (...); AI ethics teams are not budget driven teams."***

Funding reflects commitment. A former EU Minister highlighted the important role of researchers and NGOs in tracking KPIs and implementing measures to mitigate risks and harms, while raising a key question:

CONCERNS & QUESTIONS:

Former EU Member State Minister: "How much funding is allocated for this?"

EXAMPLE 2.6

European Commission, EU's AI Investment Strategy

The EU has **eight billion euros in its defence fund** — not even one percent of that is for AI ethics and safety. The European Commission announced that Europe will invest **one billion euros in the use of AI**. The plan focuses on boosting AI adoption and securing Europe's leadership in the field, mentioning major sectors such as robotics, healthcare, energy, and cars, with **no mention of education, safety, or ethics.**

As another former minister captured: **"We're discussing crucial questions here and our humanity depends on them. But we still fund weapons instead of ensuring safety."** Without funding, the right actors lack agency to continue this critical work — **"we're running behind".**

PUBLIC SECTOR CONSIDERATIONS

The roundtable highlighted that governments have started to use, procure, and deploy adopt AI within the public sector for decision making. An expert explained that this means that **they delegate certain decisions to AI systems, which effectively act as public servants.**

There was a sense that responsible AI considerations and implementation also have to be bridged with the specificity of the public sector use of AI. One expert, who conducts research on the public sector adoption and use of AI, said: *"**it is really complicated,** as it not only involves responsible and ethical AI use (...), but also **how we choose to delegate decision making in the public sector for society,** not only within the domestic context, but in different parts of the world."*

"We need to know what we are fighting so that we make the right policies."

CRITICAL AI LITERACY AND EDUCATION

What we need as leaders, teachers, and professionals is to remember that AI is not an authority. It is only a tool. Its outputs must be subject to human judgment, whether in healthcare, energy, or — especially — education. Several experts at a dedicated session on the impact of AI on the integrity of knowledge and understanding emphasised that, while many countries are now introducing subjects like computer science, digital skills, and coding in elementary schools, critical thinking and ethical reasoning must also be taught, far beyond the focus on programming. Another expert clarified that **knowing how AI works is not enough – it is essential to know what is right and wrong.**

CONCERNS & QUESTIONS:

Roundtable Participant: *"How can we manage the issue of students using AI to complete assignments, and how can we approach grading in a way that still values and assesses human thinking?"*

AI is reshaping education and reasoning – it can be a shortcut or a spark; students can either learn to bypass critical thinking or accelerate their thinking. During the roundtable, experts noted that **we need both technical tools and human capacity** to resolve critical risks, especially those with potentially catastrophic risks.

The first safeguard towards blind trust in AI is critical AI literacy. As an expert noted: *"as citizens are less likely to question decisions made by or with AI (…), it is essential, regardless of where we live, to teach children and young people to question, not just to consume technology".* Otherwise, as another expert captured:

> IF WE DON'T EDUCATE PEOPLE, THE MAJORITY, THE CIVIL SOCIETY — THEN WE DON'T HAVE AN ADEQUATE SOLUTION

During the AI plenary at GEF 2025, Dr Asha Singh Kanwar shared that **85 percent of teachers have not received training in AI, but would love to have it.** She noted that we are introducing very powerful technology, yet teachers do not know how to use it, how to teach it, or how to guide students in relation to it. The role of education, critical thinking, and awareness-raising is essential.

CONCERNS & QUESTIONS:

Indonesian CSO representative: ***"AI can act as a catalyst for growth and innovation, but it must never replace our minds*** *— it should strengthen them."* Some experts reflected on multilingual research as a particular area that AI can support, helping researchers communicate certain findings in the STEM community.

KNOWLEDGE PERSERVATION

Education has the power to prevent societal decline by nurturing literacy, critical thinking, and the responsible use of technology. However, if we truly want to prevent knowledge decay, all stakeholders — teachers, parents, policymakers, and technologists — must be part of this discussion. Experts continuously emphasise that AI should expand human knowledge, not diminish it. This notion was extended to knowledge systems, social networks, and information polarisation.

Experts noted that the mission of **using augmented technology to prevent knowledge decay is extremely pressing**. Societies carry centuries of intellectual traditions and cultural heritage, yet **AI presents a double-edged sword: it has the power to amplify these knowledge ecosystems or to accelerate their systemic decline.**

EXAMPLE 2.7

CSO: In Indonesia, with over thirty languages and diverse ethnicities, AI is being used to preserve local and tribal knowledge. Developers are embedding farmers' wisdom — such as seed selection and environmental adaptation — and creating NLP tools to safeguard and revitalise regional languages, ensuring cultural and linguistic heritage endures for future generations.

Experts also recognise that AI (despite limitations) can democratise access to information, help societies transcend language barriers, support underfunded education systems, and connect our regions to global knowledge networks.

"Literacy is the first step to governance, alignment and safety."

COLLABORATIVE ACTION & DIPLOMACY

Even within organisations, governance is not fast enough to respond to risks or innovations; there are gaps on both sides, as one expert pointed out. Discussions at GEF 2025 touched upon the need to invest and collaborate on **building a responsible economy and ecosystem.**

Specific recommendations suggested by the expert roundtables included:

1 COMMUNITY KNOWLEDGE INFRASTRUCTURE

Work across sectors and disciplines to curate a repository of datasets, whitepapers, and opportunities – consolidating knowledge that reflects local and regional cultures and identities while preserving community-rooted knowledge in digital form.

2 KNOWLEDGE PARTNERSHIPS

to produce policy briefs and reports that can be the voice for their regions, and provide a space to ideate solutions, e.g. develop a blueprint for strengthening AI capacity through education, create a global index on responsible AI, develop metrics on fairness and safety for generative tools.

3 CURATING EDUCATION RESOURCES

Create knowledge resources for youth of all ages by identifying best practices for advancing literacy work in schools, universities and community centres, while engaging CSOs, education ministries, and ethical technologists.

4 REDISCOVER THE ROLE OF PROFESSIONS

While we acknowledge that the number of e.g. teachers is low, it is easy to suggest AI solutions to replace teaching capacities. However, it is our collective role to rediscover the social and emotional impact and value of certain professions—the value of mentorship, inspiration and shaping moral character.

5 INCENTIVISE TARGETED SYSTEM EFFORTS

Since AI is a system-based, probabilistic process, we have the opportunity to intervene directly at the systems level. We can develop engineering solutions that make it harder for people to introduce additional bias. We have a chance to intervene in parts of that process through technical means.

CONCERNS & QUESTIONS:

Participant: "Are people simply passive absorbers of technology? Do we not have the ability to think critically and navigate new experiences?"

Across the roundtables, there was a general recognition that AI, like any other tool, can be used for good or bad. It may be different from anything we have known before, but **it is still a tool that we have to use wisely.**

There was divergence over whether we should be more optimistic about human intelligence and society's capacity to adapt, adjust, and negotiate life, versus perspectives from within the private, profit-driven sector, where common sense is often coupled with self-interest that requires regulation (much like corruption).

FOSTERING DIPLOMATIC ECOSYSTEMS

The roundtables concluded that, when we are speaking about threats like AI, we are solving issues that concern our common planet and humanity – there should be no "us" and "them"; we all need to work together to create a platform for collaboration to address even the hardest challenges. Important questions were raised by diplomats, ambassadors and policymakers, academics, and CSOs. **From our different roles, perspectives, and geographies – what can we do now?**

The key finding is that AI today involves not only state actors, but also corporate actors operating across jurisdictions, with the power to shape and influence people globally. This is one of the most complex challenges in regulating AI, highlighting the need for inclusive policy-making beyond the state. What's required is a comprehensive—not merely multi-stakeholder—approach. As our experts suggest:

- We need invigorated regulatory frameworks to address systemic risk — very different from last century's non-proliferation and disarmament.
- We need to collaborate across all fronts—civil society, parliament, policymakers, diplomats, NGOs, academics—before harm occurs.

"There are too many initiatives and principles, not a lot of actual measures."

REMAINING CONCERNS & QUESTIONS:

- Citizen participation, input from users, individuals, and consumers are essential, but how do we make this participation real and operational in practice?
- How do we move forward from here? We need to think about this actively as civil society and from every perspective, not only from the Global North, but also the Global South.
- Our governments and leaders need to start thinking beyond simply following global trends. What are we doing to bring these technologies to our own level, our local contexts?
- Although prominent progress was made in terms of the UN ITU framework and international organisation, can we start developing actual measures?
- How can we make the technology useful within our own realities while preserving critical thinking, reflection, and human understanding?
- Are we able to perceive the types of skills and needs required to be successful in the digital economy? Is there a case that creative and critical writing is relevant to the types of jobs ahead?

Recommendations

PROMOTING AND SAFEGUARDING FLOURISHING OF HUMANITY AND THE PLANET

FOSTERING GLOBAL AI SAFETY

To harness the power of AI for safe and ethical Integration, we must recognise the *potential of AI to drive societal progress while addressing inherent societal challenges*.

We can move beyond the preliminary focus on *preventing catastrophic harm* towards *global AI safety* that safeguards and promotes the flourishing of humanity and the planet.

As expressed by a mission ambassador: **Our task is not to prevent but enhance AI without surrendering our shared humanity. Ensure it is used fairly and in the best way possible.**

FOR A FUTURE WE DESIRE

Three keys to unlock this future:

1. **PARTICIPATION & AGENCY**
2. **OVERSIGHT & ACCOUNTABILITY**
3. **PRACTICAL IMPLEMENTATION & SHARED RESPONSIBILITY**

Let us shape **a future where AI benefits everyone, leaving no one behind** by building synergies and strategic action, keeping in mind that:

"RESPONSIBLE GOVERNANCE IS MORAL LEADERSHIP IN ACTION"

RECOMMENDATIONS AT A GLANCE:

- FOSTER PUBLIC–PRIVATE DIALOGUE BETWEEN GOVERNMENTS, CSO AND INDUSTRY
- ENABLE GLOBAL SOUTH PARTICIPATION THROUGH ACCESS AND SUPPORT
- INVEST IN EDUCATION, CULTURE PRESERVATION AND CRITICAL AI LITERACY
- FIGHT INEQUALITIES AND PROMOTE DIGITAL SOVEREIGNITY AND INCLUSIVENESS
- EMBED ETHICAL SAFEGUARDS BY DESIGN, NOT AS AN AFTERTHOUGHT (E.G. OVERSIGHT)
- REQUIRE UPSTREAM DUE DILLIGENCE AND MECHANISMS OF REMEDY
- REQUIRE ASSESSMENT FOR EXSISTENITAL RISKS BEFORE SYSTEM GOES TO MARKET
- BALANCE SHORT- AND LONG-TERM RISKS ADOPTING EFFECTIVE REPORTING
- USE AI AS A POWERFUL STATISTICAL TOOL TO UNDERSTAND OUR PAST
- FOSTER FEDERATED GOVERNANCE WITH FOCUS ON HUMAN RIGHTS AT EVERY LEVEL
- FACILITATE MULTI-STAKEHOLDER, CROSS-SECTORAL CO-CREATION
- CONTINUE CONVERSATION ON OPRATIONALISATION, MEASUREMENT AND IMPACT

Recommendations

PROMOTING AND SAFEGUARDING FLOURISHING OF HUMANITY AND THE PLANET

1️⃣ PARTICIPATION & AGENCY

FOCUS: Empower people and communities to shape AI futures, ensure representation, and promote inclusive participation

A. FOSTER PUBLIC–PRIVATE DIALOGUE BETWEEN GOVERNMENTS, CSO AND INDUSTRY

- Closer cooperation and collaboration between governments, civil society, and industry is needed to determine the direction of AI governance. Existing regulations provide a basis, but attention should be directed toward shared values, principles, ethics, and clearly defined red lines on existential risks.
- Effective mechanisms must be established to ensure that the private sector is held responsible and accountable for the harm caused by its developments.

CONSIDER:

> *How can collaboration with the corporate sector be effectively advanced?*
>
> *How can fear and deeply politicised, ideologically driven assumptions be addressed?*

B. ENABLE GLOBAL SOUTH PARTICIPATION THROUGH ACCESS AND SUPPORT

- This is a critical area for international support — working with global south countries to build the skills, infrastructure, and institutional capacity needed to engage in AI governance.
- Without addressing the gap, meaningful engagement in these countries' future negotiations — particularly on issues related to armed conflict — will remain unattainable.
- Engagement on the international development dimension requires improved access to tools and training.

C. INVEST IN EDUCATION, CULTURE PRESERVATION AND CRITICAL AI LITERACY

- The discussion should expand to include human literacy. Preventing the collapse of humanity requires much greater attention on education, *not just* scaling, because strengthened education systems remain essential for success.
- Many countries are introducing subjects such as computer science, digital skills, and coding in elementary schools, but equal emphasis must be placed on teaching critical thinking and ethical reasoning, especially when transitioning from traditional to AI-powered education.

CONSIDER:

> *What will happen, and how is AI being introduced into education in the first place?*
>
> - *Digital literacy*
> - *Critical digital literacy*
> - *Media Information literacy*
> - *Digital agency and autonomy*
> - *AI literacy*
> - *Critical AI literacy*

D. FIGHT INEQUALITIES AND PROMOTE DIGITAL SOVEREIGNITY AND INCLUSIVENESS

Digital sovereignty shall be promoted to ensure that the benefits of AI are equitably shared with the Global South and that states are properly capacitated to use AI. This approach is essential to ensuring that AI becomes a tool for inclusiveness rather than a force that entrenches global inequalities.

Recommendations

PROMOTING AND SAFEGUARDING FLOURISHING OF HUMANITY AND THE PLANET

2️⃣ OVERSIGHT & ACCOUNTABILITY

FOCUS: Build systems of trust, verification, and ethical responsibility across the AI lifecycle

A. EMBED ETHICAL SAFEGUARDS BY DESIGN NOT AN AFTERTHOUGHT (E.G. OVERSIGHT)

- Ethical safeguards, transparency, and accountability must be built into AI systems from the very beginning, not added as an afterthought.
- Since AI is a system-based, probabilistic process, opportunities exist to intervene directly at the systems level, allowing specific technical issues to be addressed.

CONSIDER:

> - *Retrieval Augmented Generation and symbolic models for information verification*
> - *Environmental risk calculation*
> - *Bias detection*
> - *Model drift*

B. REQUIRE UPSTREAM DUE DILLIGENCE AND MECHANISMS OF REMEDY

- Inquire about the product prior to adoption, its impacts and contextually-relevant remedial mechanisms. Foster a culture of shared accountability and due diligence across the AI supply and value chains, and entire life cycles.

C. REQUIRE ASSESSMENT FOR EXSISTENITAL RISKS BEFORE SYSTEM GOES TO MARKET

- Mandatory frameworks are required in which companies clearly identify their potential existential risks and specify the human oversight mechanisms established to mitigate them.
- These aspects must be clearly articulated and documented (whether it is related to knowledge decay, or cultural, social and historical erosion, particularly of linguistic minorities).

D. BALANCE SHORT- AND LONG-TERM RISKS ADOPTING EFFECTIVE REPORTING

- Balancing short-term and long-term risks are equally important. Societal risks that undermine people's and especially vulnerable groups' ability to live the lives they enjoyed before AI must be addressed.
- A shared reporting mechanism and strategy is required "some kind of radar system to identify potential future risks that might not yet be visible."

CONSIDER:

> *Although AI tools may offer short-term efficiency, the long-term consequences are significant. Hence, let's ask: What is the long term consequence of this technology for us?*

E. USE AI AS A POWERFUL STATISTICAL TOOL TO UNDERSTAND OUR PAST

Leverage powerful AI statistical tools to show and help better understand how biased our societies have been – instead of using these tools for actual decision making, or decision support.

CONSIDER:

> *Can AI be used to understand how priorities are currently set in law, social services, and hospitals rather than optimising an algorithm for decision-making? Can it be used simply to illuminate what is being done today?*

Recommendations

PROMOTING AND SAFEGUARDING FLOURISHING OF HUMANITY AND THE PLANET

3️⃣ PRACTICAL IMPLEMENTATION & SHARED RESPONSIBILITY

FOCUS: Translate values into action, ensuring coordination and accountability from design to deployment

A. FOSTER FEDERATED GOVERNANCE WITH A FOCUS ON HUMAN RIGHTS AT EVERY LEVEL

- Social progress, development, and peace cannot be achieved without Human Rights. Human Rights are *not only obligations*; they serve as a vital compass when new challenges arise. Coming together at one table is essential to pursue meaningful conversation.
- Universal global principles have to be connected to local realities, with effective governance mechanisms at each level.

B. CONTINUE CONVERSATION ON OPERATIONALISATION, MEASUREMENT AND IMPACT

- Continuing this conversation is essential because meaningful impact is possible. All actors must be involved, rather than leaving governments to address these challenges alone. As diplomats confront global issues, there is broad recognition that multilateralism is required to advance this work. In this context, civil society participation becomes particularly valuable.
- Diplomatic efforts should be directed toward establishing new mechanisms for addressing human rights in AI, which are necessary to advance the conversation on global safety, diplomacy, and multilateral governance.
- The Summit of the Future and the Global Digital Compact have produced a broad framework of global ideas that still must be translated into reality. The next level of this important building block now needs to be advanced.

C. FACILITATE MULTI-STAKEHOLDER, CROSS-SECTORIAL CO-CREATION

It is not sufficient to simply collect data and acknowledge interests of stakeholders of AI Governance but it is necessary to actively involve them in the co-creation of the implementation of responsible AI principles, empowering parties to actively participate in this transformational process.

References

EXAMPLE 1.1 Over-reliance on AI dialogue systems on students' cognitive abilities
Zhai, C., Wibowo, S. & Li, L.D. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn. Environ. 11, 28 (2024).

EXAMPLE 1.2 MIT Media Lab Study on the effect of ChatGPT on Students' Cognitive Debt
Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., ... & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. arXiv preprint arXiv:2506.08872.

EXAMPLE 1.3 The Arabic Gap in AI: Why Representation Matters Beyond English
(Welo Data, Sep 2025) https://welodata.ai/2025/09/03/bridging-the-arabic-ai-gap/

EXAMPLE 1.4 Deloitte issues refund for error-ridden Australian government report that used AI
(The Guardian, Oct 2025) https://www.ft.com/content/934cc94b-32c4-497e-9718-d87d6a7835cat

EXAMPLE 1.5 What My Daughter Told ChatGPT Before She Took Her Life Away
(The New York Times, Aug 2025) https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html

EXAMPLE 1.5b Parents Allege ChatGPT Is Responsible for Their Teenage Son's Death by Suicide
(The Time, Aug 2025) https://time.com/7312484/chatgpt-openai-suicide-lawsuit/

EXAMPLE 1.6 The promise and challenges of generative AI in education
Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., et al. Rienties, B. (2024). The promise and challenges of generative AI in education. Behaviour & Information Technology, 44(11), 2518–2544. https://doi.org/10.1080/0144929X.2024.2394886

EXAMPLE 1.7 AI Notetaking Tools Under Fire: Lessons from the Otter.ai Class Action Complaint
(The National Law Review, Aug 2025) Joseph J. Lazzarotti Jackson Lewis P.C.
https://natlawreview.com/article/ai-notetaking-tools-under-fire-lessons-otterai-class-action-complaint

EXAMPLE 1.8 LLM student sues Jindal Global Law School over AI use in examination
(The Hindu Times, Nov 2024) https://www.thehindu.com/sci-tech/technology/llm-student-sues-jindal-global-law-school-over-ai-use-in-exam/article68831869.ece

EXAMPLE 1.9 AI bias in the prediction and detection of cardiovascular disease.
Mihan, A., Pandey, A. & Van Spall, H.G.C. AI bias in the prediction and detection of cardiovascular disease. npj Cardiovasc Health 1, 31 (2024). https://doi.org/10.1038/s44325-024-00031-9

EXAMPLE 2.0 Dutch child benefits scandal
Amnesty International (October 2021). Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms. https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/

EXAMPLE 2.1 Apple Card lower limits for women in 2019
(BBC, 2019). Apple's 'sexist' credit card investigated by US regulator.
https://www.bbc.co.uk/news/business-50365609

EXAMPLE 2.2 AGI Red Lines Coalition Launched
The Future Society (Sep 2025). The Global Call for AI Red Lines: Why We Launched a Campaign to Address Unacceptable AI Risks. https://thefuturesociety.org/globalcallforairedines/

EXAMPLE 2.3 Moratorium to Halt AI Development Future of Life Institute (Mar, 2023)
https://futureoflife.org/open-letter/pause-giant-ai-experiments/

EXAMPLE 2.5 Sam Altman's Kill Switch (OECD.ai 2023) Sam Altman's "Nuclear Backpack" AI Kill Switch
https://oecd.ai/en/incidents/2024-08-02-1e2d

EXAMPLE 2.6 8 Billion Investment into the defence Fund EU (European Commission (2025)
https://commission.europa.eu/strategy-and-policy/eu-budget/motion/focus/eu-budget-boosts-eus-defence-industry-and-european-security_en

GLOBETHICS PUBLICATIONS

The list below is only a selection of our publications. To view the full collection, please visit our website.

All products are provided free of charge and can be downloaded in PDF form from the Globethics library and at https://globethics.net/publications. Bulk print copies can be ordered from publications@globethics.net at special rates for those from the Global South.

Prof. Dr Fadi Daou, Executive Director. Dr Ignace Haaz, Managing Editor.

Find all Series Editors: https://globethics.net/publish-with-us
Contact for manuscripts and suggestions: publications@globethics.net

Policy Series

Global Ethics Forum 2024 *Ethical Leadership for a Re-envisioned Future: Policy Report*, 2025, 51pp. ISBN 978-2-88931-632-8

Inclusive AI for a Better Future. Policy Dialogue Report, 2024, 27p. ISBN 978-2-88931-563-5

Navigating the Future of Higher Education with AI 2024 ICDE Leadership Summit Policy Dialogue Report, 2024, 39pp. ISBN 978-2-88931-605-2

Indigenous Statements for Environmental Justice 2008-2023, 2023, 50p. ISBN 978-2-88931-558-1

Interfaith Statements for Environmental Justice 2008-2023, 2023, 62p. ISBN 978-2-88931-556-7

Reports

Navigating the Development and Humanitarian Funding Crisis Ethical Implications and Strategic Responses, 2025, 16pp. ISBN: 978-2-88931-629-8

Journal of Ethics in Higher Education

https://globethics.net/jehe ISSN: 2813-4389

No. 4(2024): Ethical, Regulative and Legislative Perspectives on Emerging Technologies and Education
No. 5(2024): Ethics in Open and Distance Education
No. 6(2025): A Practice-based Approach in Education: Contemporary Rethinking of Leadership
No. 7(2025): Restorative Justice, Nonviolence and Education

This is only a selection of our latest publications, to view our full collection please visit: https://globethics.net/publications


9 October 11:00 – 12:30


Empowering individuals and institutions towards excellence in responsible governance.
WSIS Prizes 2025 Champion with 20 Years of Expertise and Global Trust
Trusted Partner for AI Ethics Policy Dialogue and Capacity-Building Programmes
Pathways to Inclusion: Advancing Civil Society's Role in AI Governance


Arpita Kanjilal

ISBN 978-2-88931-642-7

Harnessing the Power of AI: Safe & Ethical Integration

Report of the AI Safety Dialogue at Global Ethics Forum 2025

The report presents insights from roundtable discussions held as part of the international dialogue series Track II: Diplomacy for Global AI Safety.
On this occasion, Globethics' community of practice convened during the Harnessing the Power of AI — Safe and Ethical Integration track at the Global Ethics Forum 2025 (GEF 2025) in Geneva.
Designed to advance responsible AI governance worldwide, the roundtables examined collective action for AI safety and strategies to preserve human thinking and critical literacy in the age of AI. They also identified opportunities for cross-sector technological and diplomatic coordination to address catastrophic AI risks.

Global Ethics Forum

AI IMPACT SUMMIT भारत 2026 INDIA
Official Pre-Summit Event

Globethics