THE
FUTURE
SOCIETY

The 2019 Global Governance of AI Forum Report
World Government Summit
February 9-10, Dubai, UAE
DRAFT VERSION for UAE Ministry for Artificial Intelligence

Executive Summary

Introduction

- The Global Governance of AI Forum (GGAF) is a revolving **international multi-stakeholder governance process** which brings together a **diverse community of 250 global experts** and practitioners from government, business, academia, international organizations, and civil society.

- This Forum has been envisioned and designed as a **unique collective intelligence exercise** to help **shape and deploy global, but culturally adaptable, actionable pathways, guidelines, norms, standards and practices** to govern the rise of Artificial Intelligence.
    - Building upon the first edition held in February 2018, the 2019 edition began in August with an **intensive six-month preparatory program-building** and curation period. This included almost 90 expert Video Teleconference sessions.
    - Insights from the six-month preparatory period led to the publication of **14 background research papers on different topics.**
    - These papers in turn informed the agenda of a full-day roundtable workshop, with **four working sessions and 47 subcommittees**.

- This combined community building, research, and agenda-setting effort was done in partnership with a host of international organizations including the OECD, UNESCO, IEEE, the Council on Extended Intelligence, and the Global Data Commons Task Force.

- GGAF gave each partner organization a platform to meet and advance their own goals and initiatives on AI policy in a two-day period in Dubai, UAE. The Global Governance of AI Forum culminated with a one-day cross-cutting Roundtable, a Collective intelligence Workshop that provided ambitious but actionable pathways to improving the global governance of AI.
    - The GGAR had curated breakout sessions to maximize productivity and build up defined outcomes.

- The insights and recommendations have been captured into this comprehensive report, which includes an action-oriented summary for policymakers.

Key Insights

- GGAF showcased a rediscovery of the full government toolbox to respond to the opportunities and challenges brought by AI. Governance is not just enacting laws. Governance for AI must include developing, instituting and enacting a full range of tools: from narratives to technical standards, to codes of conduct, principles, through practices and policies--all have great potential in the broader AI governance agenda.
- GGAF showed governments how to be innovative in thinking about governance for AI. Using that toolkit, governments must put themselves in the driver's seat: they should be innovative and prepared to take risks in order to seize the greatest upsides of the AI revolution.
- GGAF put forward a number of research initiatives and proposals for new methodologies and metrics to help assess progress in AI research and application in society. At the moment, there is still much consultation and assessment to do, a deal of which can be spearheaded by attendees at GGAF at their institutional homes.
- GGAF is becoming the international reference point for multilateral, multi-stakeholder approaches to governing the AI revolution.

Proposals spanned in scope from smaller research programs to building new global institutions and observatories.

Research Initiatives

11 ideas for research initiatives were presented at GGAR

- There is a lack of information in the public sphere about the global state of AI research. We need new research initiatives at global institutions. 11 research programs were proposed at GGAR.
- Especially promising proposals included:
    - Researching the typology of AI industrial value chains: who does what, when and how in the global economy?
    - How can developing countries build business infrastructures and incentives to maximize their slice of the pie in the AI revolution?
    - How can we build smart taxation systems for the age of AI?
    - Research into pathways to AGI and options for risk mitigation, as well as implications for cybersecurity

Experiments

2 pitches for new experiments were presented at GGAR

- Experiments can showcase new approaches to AI research and governance and offer pathways to their reception in more widespread programs at the national or global level.
- Especially promising proposals including:
    - Policy “sandboxing” like trialing new regulatory and certification schemes.
    - Longitudinal surveying over the coming decade about how citizens are being affected by AI.

Metrics and Methodologies

GGAR participants called for at least 6 new methodologies and metrics to better understand AI’s impacts

- There are some characteristics of AI and its applications in society that we know little about and which we struggle employing existing tools and datasets to assess.
- Especially promising proposals:
    - Metrics to measure AI progress, including spillover effects and the impacts of hybridized technology application
    - An “Indicators of Intelligence” scale similar to what has been done for autonomous vehicles
    - SDG Matrix that would empower us to analyze the real-time interconnections between different SDGs

Observatories and Institutions

5 new observatories and institutions were proposed at GGAR to house new research initiatives and to offer platforms for novel thinking about AI

- Sometimes research programs are of such scope and importance that new institutions must be built to house them.
- Especially promising proposals included:
    - An observatory that collates projects and initiatives around the world that are helping to create “Explainable AI” systems
    - An observatory that acts as a non-partisan forum for assessing AGI research

  - A "CERN for AI" that would house a grand scientific project and allow world-leading researchers to pursue a major AI-related project like building a safe and responsible Artificial General Intelligence
- GGAR also discussed the potential and possible institutional design of the International Panel on AI that France and Canada have initiated

Cooperation and Convenings

3 new forms of convening and cooperative measure were requested to improve multilateral dialogue on AI

- General calls for cooperation and convenings between global actors. GGAF is a first-rate example of such international cooperation and coalition-building
- An example of the lack of multilateral dialogue is the UN Group of Experts' call for humanitarian law to apply in cyberspace not being substantiated or followed up on by governance actors

AI Governance Principles

GGAR participants considered 7 different aspirations for new governance principles

- For AI to be governed safely and effectively, governance actors need to work with ethically aligned agendas and operate using well-established principles and modes of cooperation.
- [Placeholder for IEEE Ethically Aligned Design General Principles Report & The AI Ethics Forum.]

Practices & Policies

GGAR experts suggested 11 new practices and policies for governments and the corporate sector to introduce

- Public policy and programs for governance intervention were put forward. Some were more practicably implementable, while others were purposefully blue-sky thinking and would require significant global buy-in before being initiated.
- Notable proposals included:

- Governance actors need to increase the incentive structure and infrastructures for private firms and other actors to disclose cyber-breaches or other security exposures. This becomes even more urgent in cyberspace that is enhanced with AI capabilities
- A tax levy on firms for the upskilling of workforces to make the global economy and labor force more resilient in the AI age
- A tax system put in place for firms that choose to automate jobs to compensate for the associated employment displacement

Educational Initiatives

The education sector is set to be transformed in the AI Revolution. GGAR participants proposed 3 new education initiatives to facilitate this transition

- Education systems across the globe must be reassessed and respond to AI-relevant change and innovation. This requires a deep understanding of emerging trends in the global jobs market, but also a reconsideration of what the purpose of education is in society more generally. In the coming years to cope with shifts in the global economy, we should see lifelong learning initiatives and reskilling programs instituted across society.
- Especially fruitful proposals included:
    - Vocational training schemes put in place in the developing world which focus on AI
    - Teach children about technology and personal cyber-safety from kindergarten

Conclusion

- There is excitement about the scale of innovation that can be brought across the world economy, from healthcare to transport to security. At the same time, there are significant structural challenges and risks that require effective and comprehensive governance programs.
- Ultimately, it is political will and multilateralism that will determine whether AI has a positive influence on the world.
- In 2020, GGAF will continue in its mission and provide a global forum to propel multilateral efforts. We look forward to welcoming back old friends and colleagues, and welcoming new faces to this community.

Table of Contents

Subject to change

Introduction

Vision

The Global Governance of AI Forum (GGAF) is a **multistakeholder, global expert forum** of leading thinkers on AI governance. Through this action-oriented platform involving roundtable discussions, experts, scientists and practitioners discuss the benefits, risks, and pathways to **develop effective**, yet **culturally adaptable** norms that will assure the **safe & ethical deployment of AI for the betterment of all humanity**.

The participants of the 2018 Global Governance of AI Roundtable represented a broad range of professional affiliations (government, NGOs, industry, academia) from a range of different fields (technologists, of course, but also ethicists, philosophers, political scientists, economists, lawyers, physicians, and individuals engaged in various fields of the humanities).

Objectives and Principles

The *Global Governance of AI Forum* has three chief objectives:

1. **Gather information about the state of AI technologies**, their socio-economic impact and the state of AI governance policies around the world to formulate a comprehensive knowledge base key to developing a robust and effective governance framework and policy options.

2. **Synthesize this information into a governance framework, actionable public policy options, and implementation-level guidelines** that can be used by the UAE as a pioneer in the field of AI governance, and by other governments around the world.

3. Serving as the **world's authoritative forum convening an inclusive and broad range of stakeholders for dialogue** in a methodical manner for ongoing **learning, adaptation, and evidence-based decision-making** for effective policies.

The driving principles behind these convenings:

1. Neutral forum
2. Multi-stakeholder discussion
3. Collective intelligence methodologies
4. Community Building and active participation

Methodology

The rise of AI is the result of a **dynamic and complex sociotechnical system where science, technology, and society are engaged in a continuous cycle of "co-production', locally and**

globally: new technologies and innovations, infused into societies through business, continuously impact and redefine societal values and thus policies; changes in values and policy in turn continuously shape techno-scientific developments.

To shape an effective system of AI governance there must be a firm understanding in current global governance realities, changing power dynamics, and the rising influence of new policy avenues. The Forum seeks to be adaptive and remain robust and relevant over time for the development of policies for AI governance. This approach is characterized by the following essential features:

- **Strategic approach**
  The government of the UAE and the Minister of State for AI have a long-term strategic vision for the development of norms for the governance of AI. This includes the **creation of a permanent forum for gathering information and perspectives** from a wide range of stakeholders **that is not simply a conference or series of conferences**. The AI Roundtable works systematically towards a governance framework and actionable policy options.

- **Broad-based**
  The **Roundtable seeks input from individuals with a broad range of professional affiliations** (representation from government, NGOs, industry, academia) **and fields of expertise** (technologists, of course, but also ethicists, philosophers, political scientists, economists, lawyers, physicians, and individuals engaged in various fields of the humanities). It is a collective intelligence exercise gathering knowledge, perspectives, and experience that works towards norms via collaboration, dialogue, and consensus.

- **Disciplined and systematic approach**
  The Roundtable seeks to **anchor the discussion in a foundation of shared core values** due to the multifaceted nature of the topic and diverse backgrounds of participants. Following a systematic approach can avoid amorphous dialogue and successfully develop a coherent and generally-applicable set of norms for the governance of AI.

- **Adaptive**
  The Roundtable **expects policy to adapt to changing conditions** and appreciates the complex and dynamic relationships between science, technology and society. It has designed a long-term **recurring initiative, capable of re-visiting and modifying its frameworks and policy recommendations**.

- **Evidence-based decision-making**
  For policies of AI governance to be effective they must be based upon, and continuously tested against **real-world conditions**. The Roundtable incorporates a wide-ranging and ongoing search for practices and relevant experience as it works toward **actionable and effective norms**.

In terms of GGAR, the aim this year was to move forward from 2018 to create more implementable and practical solutions for the governance of AI. In order to foster a successful collective intelligence exercise, GGAR 2019 includes some preparations. The Future Society team drafted background papers on the main discussion topics in order to harmonize our knowledge and set the stage for a fruitful expert discussion. These background papers served as an anchor for bimonthly expert calls organized throughout November to January. There were *fourteen* Expert Group discussion topics and each GGAR participant was allocated to *four* topics depending on their interest and expertise. The topics are also the basis of GGAR committees in Dubai.

To prepare for discussions in Dubai, the GGAR team will organize two rounds of calls for each Expert Group discussion topic. The first round of calls focuses on garnering insights from your expertise to inform the background papers and to identify the most pressing topics of interest, whereas the second aims at action-oriented preparation for the day of the GGAR in Dubai.

AIGO (OECD) Principles

Principles for Responsible Stewardship of Trustworthy AI[1]:

- Inclusive and sustainable growth and well-being (including inclusion of underrepresented populations)
- Human-centred values and fairness (including dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness i.e. lack of bias, social justice, labour rights)
- Transparency and explainability
- Robustness, security[2] and safety
- Accountability

National Policies for Trustworthy AI

- Investing in responsible AI research and development
- Fostering an enabling digital ecosystem for trustworthy AI
- Providing an agile and controlled policy environment for AI
- Building human capacity and preparing for job transition

IEEE Ethically Aligned Design General Principles

- **Human Rights**: A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
- **Well-being**: A/IS creators shall adopt increased human well-being as a primary success criterion for development.
- **Data Agency**: A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
- **Effectiveness**: A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
- **Transparency**: The basis of a particular A/IS decision should always be discoverable.
- **Accountability**: A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

[1] OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en.

[2] The risks and recommendations for digital security can also be found in OECD (2015), Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264245471-en and its 2020 update for critical functions: OECD *Recommendation of the Council on Digital Security of Critical Activities*, OECD/LEGAL/0456

- **Awareness of Misuse**: A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
- **Competence**: A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

The AI Ethics Forum

Convened three distinct, yet well-aligned communities of practitioners – UNESCO, COMEST, and IEEE to formulate a holistic framework to manage ethics of AI. The meeting focussed on sharing information on initiatives addressing the ethical dimensions of Autonomous and Intelligent Systems (A/IS) and Artificial Intelligence (AI), and identifying points of collaboration that can bolster and catalyze the collective work of our organizations and others to ensure the ethical, values-driven design of Autonomous and Intelligent Systems (A/IS).

1. **Research:** Undertaking a joint study on A/IS governance models, and on notions of ethical and human centred in A/IS, in partnership with the Council on Extended Intelligence (CXI, a joint initiative of MIT Media Lab and IEEE Standards Association);
2. **Publications:** Cooperation between UNESCO, COMEST, and IEEE in the framework of the second edition of the publication on Ethically Aligned Design, focusing notably on human rights in the digital age and on questions of diversity regarding A/IS;
3. **Global Data Commons:** Participation in the Open Community for Ethics in Autonomous and Intelligent systems (OCEANIS) and in the review of 14 IEEE standards projects under development in the domain of ethically aligned design;
4. **Teacher Training and AI digital literacy:** Development of teacher training and OER material on the ethical dimensions of A/IS and AI for diverse target populations including youth, journalists, mass media, engineers, computer scientists, legal professionals, etc., based on IEEE's Ethically Aligned Design and standards, UNESCO's actions in ethics education, media and information literacy, and capacity building, as well as COMEST and IBC reports, and in partnership with the Ethically Aligned Design University Consortium (EADUC) of IEEE and building on, as appropriate, a number of IEEE initiatives such as 1) the business course "Artificial intelligence and ethics in design"; 2) the Ethics Certification Programme for A/IS (ECPAIS), and; 3) the webinar "benefits and challenges of raising children in an AI world."

Global Data Access Framework (GDAF)

The GDAF acts as a precursor to deploying AI to help achieve the United Nations' Sustainable Development Goals (SDGs) by capitalizing on the global pools of data, up-scale use-cases of AI for SDGs, and monitor, simulate and predict outcomes for progress on the development goals.

The position papers submitted by GDAF partners identified key challenges to implementing the Data Commons spanning across: access to quality data, technical, legal, political and regulatory, social, and commercial issues. An iterative and systematic approach was outlined to make the GDAF a reality:

1. Scope definition & risk assessment
2. Identify requirements for data sets, data flow, and workflow
3. Determine required technology architecture and infrastructure
4. Define governance, risk & stakeholder management models
5. Define agreements and contracts

Council of Extended Intelligence (CXI)

During the Global Governance of AI Forum, members of CXI presented and held in-depth discussion on their recent release of its first paper, *The Case for Extended Intelligence*.

Global Governance of AI Roundtable (GGAR) Report

The Future Society designed and developed the 2019 Global Governance of AI Roundtable based on a mapping of major trends and topics in AI Governance. We categorized these topics into 14 roundtable 'Expert Groups' with committees led by AI experts ('Co-chairs'). Each topic built upon in-depth research papers authored by AI Policy Researchers at The Future Society's team. The Future Society identified and invited the expert participants based on its network.

1. Mapping AI Technological Development & Future Trajectories

Expert Group Leader: Anima Anandkumar

Committee co-Chairs: Gabor Melli, Jack Clark, Paul Epping

Key Recommendations

- Governance actors and research institutions must develop new metrics and methodologies for assessing progress in AI technologies, including spill-over effects and hybridised autonomous and intelligent systems.
- Longitudinal surveys should be developed for the general population that assess changes in perception and levels of impact of AI over people's lives over time.
- Governance actors need to improve incentives and coordination so that world-class A(G)I researchers will attend multilateral convenings. Without their buy-in and input, governance misses a vital actor in discussions about the future of AI governance.

The course of technological development, especially concerning AI, is uncertain, and the impacts for society remain even more unpredictable. Understanding trends in technological development is of vital importance so that governance actors can establish frameworks and build institutions to ensure safe and beneficial developments for society.

Four subcommittees met to evaluate and discuss the major areas of AI systems innovation, and to propose new institutions and methodologies for assessing the global technology landscape. These were:

A. **Horizon Scanning**
B. **Methodology**
C. **Key Indicators**
D. **Impact of Future Trajectories**

With fresh proposals for new tools and institutional norms put forward, it was nonetheless clear from this committee that there is no consensus among experts as to whether we should expect AI to have positive or negative effects on society, and how we might begin to evaluate those effects quantitatively or qualitatively.

Areas of AI technological development over the coming 12 months

Across these subcommittees, participants discussed areas of AI technological development where we will see advances in the next year. Those identified as areas of especial development were:

- AI Application-Specific Integrated Circuits (ASIC) chips (similar to Google's Tensor processing unit [TPU]);
- Gibbs sampling type hardware;
- General progress in sensor technology;
- Supercharged circuits;
- Neuromorphic chips;
- 5G infrastructure;
- Virtual Reality hardware, with relevant improvements in optics, rendering, resolution and anti-motion sickness;
- Nanosatellites;
- Quantum storage capacity improvements.

These distinct areas of innovation also increasingly interact with one another in novel ways. Cognitive analytics, for example, looks set to bring an exciting future for data analytics, and takes advantage of the huge increases in High Performance Computing, and combines AI, semantics, deep learning and machine learning. We can expect such technological combinations and interdisciplinary thinking to transform many industries.

We have general ideas about how developments will progress in these fields, but often assessments of AI development are matters of opinion and remain imprecise. Also, analyses about how technologies will combine to create new products and applications are too vague to provide concrete insights for understanding anticipated social effects and actioning new governance arrangements. **Participants called for new metrics to track progress in AI-relevant technologies internationally, to include the effects of spill-overs and hybridised autonomous and intelligent systems.** It was recognised that an international and cohesive effort will likely require significant institutional backing, from private industry and governments, and that an international organisation such as the United Nations or the OECD would be a prime venue for undertaking such research.

Worst case scenario and crisis planning

Exciting new technologies are expected to transform industries and societies, but impacts will vary across geographies and sectors. We can predict that some of these impacts will have negative social effects, such as the likelihood of considerable jobs displacement. There are also risks inherent to the technologies themselves. In 2016, for example, Microsoft's artificial intelligence Tay (tay.ai) had to be shut down after just sixteen hours because it started posting offensive and racist tweets. AI/S learn in part from human systems, so no matter the magnitude of technological innovation, AIs hold the potential to reassert entrenched problems in society such as racial prejudice. Also, in 2016, researchers at DeepMind and Future of Humanity Institute began developing a “kill switch,” which would ensure that AI systems are coded to prevent them from overriding human inputs. Such a kill switch, known as “safely interruptible AI,” would prevent an AI ever being able to resist human intervention or attempts to shut down. Ongoing research into AI safety is an imperative, as the challenges inherent in this example show.

Efforts such as the AI “kill switch” reflect the broader need to plan for crisis situations, in scenarios where technologies are developed that fail to meet adequate safety standards, or cause unanticipated negative impacts on people and world systems. Of course, the scale of this task is enormous at the global level, but GGAR participants discussed several initiatives which would help frame the terms of debate about AI safety and systemic risk. These included:

Defining worst case scenarios: Understanding worst case scenarios is just as important as determining what success looks like or what outcomes are to be expected from the development and application of new technologies. Research into what could be the most negative impacts, presented alongside a well-reasoned and objective case for the anticipated benefits, will help researchers build systems for crisis response, and outside parties to comprehensively assess the case for whether and how technologies should be employed.

New definitions of risk: Ideally, risk assessment should not be limited to either the technical parameters of a project or its social impacts. Rather, the two should be seen in dynamic relation to one another. At the moment, AI risk assessments from both the research community and policymakers tend to focus on one aspect or the other.

New definitions of complexity: Complex systems science has emerged as a field of study in its own right. It offers pathways to understanding the complex impacts of AI systems in formal models and real-world applications. Participants called for research into complexity and AI systems to be made more readily available to inform policymaking options.

Indicators of intelligence: We lack a formal framework for assessing the development of AIs’ “intelligence.” Experts remarked that a five-point scale is used to assess the state of the autonomous vehicle industry,[3] but that we do not have this kind of indicator-scale for the wider AI field. **Such a scale could be useful in diverse scenarios, including global reporting or even as a consumer barometer like the European Union’s energy label energy consumption scheme.**

[3] This scale has been approved by SAE International, the U.S. professional organization that develops standards in the transport sector, with “0” being “no driving automation” and “5” being “full driving automation.” See https://www.sae.org/standards/content/j3016_201609/.

New indicators and metrics to assess AI development will also curb the current tendency for policy assessments and regulatory direction to focus on current innovation, which comes at the cost of considering anticipated innovation in the coming decade. Focusing solely on current innovation **can lead policy prescriptions to be unsuitable to the frontiers of innovation.** For example, experts observed that today policymakers think about policy provisions for a world dominated by machine learning systems, whereas it is now generative adversarial networks which show especial promise, overcoming many well-known problems associated with machine learning: entrenched bias and discrimination. Undoubtedly, this technological step-change changes the implications for governance intervention.

Studying AI Longitudinally

Experts considered what could be fresh ways to understand AI and its impacts in society. A number of projects were "crowd-sourced" by committee members present. Participants were asked: **"what would be the highest impact way of using US$10million to understand something new, profound and policy-relevant about AI?"** A number of proposals were put forward and informally voted upon by committee members.

From the proposals discussed, the most popular option by vote was a longitudinal survey over eight years that would require **10,000 paid individuals chosen at random from across the world to assess at a number of intervals what has been the impact of AI on their lives.** Given the nature of GGAR, this proposal was intended to spark debate rather than offer a comprehensive, ready-made solution; but this kind of sociological research could give researchers and governance actors much more penetrating analysis of how AI is affecting people in different geographies and sectors *over time* rather than merely *in time*.

The need for longitudinal work reflects a broader concern within the AI governance community that governments understand technology through "snapshots" and react to them at key inflection points, rather than considering regulatory systems iteratively and responding to technological *emergence*. **This committee offered a call to academia and international civil services to use longitudinal surveys to better understand AI's effects over time, and not respond to AI innovation reactively, or in fits and starts.**

Learning from Computer Scientists

GGAR participants outlined the ongoing difficulties of attracting premier computer scientists to international AI governance events, to include the Global Governance of AI Forum. Policymakers and the broader research community alike need to know what kind of innovation is happening "at the edge" if they are going to be in a position to plan governance responses and areas of priority for academic inquiry. At the moment, liaison committees and forums focusing on AI governance tend to have a considerable number of traditional governance actors, like representatives from intergovernmental organisations. This is absolutely how it should be, however, **there is not currently enough incentive for**

computer scientists from the world's most advanced technology companies to attend AI governance convenings. Participants remarked that institutions responsible for organising AI governance events should strategize to ensure that a greater representation of world premier computer scientists attend and offer formalised support to them. This is important for accountability, knowledge-broking, directing research agendas, and richly informing governance initiatives.

Relevantly, a second popular proposal in response to the question **"what would be the highest impact way of using US$10million to understand something new, profound and policy-relevant about AI?"** related to how the governance community could learn from computer scientists. Drawing on the concept of the hackathon, this proposal would have AI-focused think tanks "duel" with one another to come up with problem-solving solutions to specific challenges. For example, two think tanks could be briefed to come up with the governance infrastructure that could support and regulate the "kill switch" research outlined above. The institution commissioning the research from the think tanks could adjudicate on the winning proposal and formally endorse or even strive to implement it in their governance agendas. **Participants called on AI-focused think tanks and institutions supporting their research to pilot this innovative "thinkathon" work commission process.** It would be especially helpful for computer scientists to be represented alongside policy experts in the duelling teams.

Assessing the Future

There was significant disagreement among experts as to whether the coming years will see generally positive or negative effects for society due to the increasing penetration of AI in society. Stocktaking the future, the following perspectives were put forward from participants:

- ***Prosperity and adaptation:*** "AI innovation in the coming decade will have very positive effects on society and will help the world achieve the SDGs. There needs to be a global shift in the jobs market, but both the workplace and the education system will adapt, and global society will come to thrive."
- ***Learning from mistakes:*** "The coming decade will see society make many mistakes with respect to AI because we do not know how the technologies work and how they should be applied. Hopefully, we will learn a lot of lessons, and within ten years we may be at the stage where we start using AI in appropriate and pro-social ways."
- ***Changing the free market:*** "The market economy will not be able to protect society from the social fallout that AI brings with it. We need to redirect capitalist models of growth to ensure that technology becomes and tool for human innovation, not a weapon for loss and exploitation."
- ***The end of human labour:*** "AI will end human labour as we know it, and hark a new but as yet largely unknown era for global progress and human endeavour. This shift will be remarkable, and it is at present difficult to say what impacts will be generally positive or negative for humanity."

There was little consensus on any of these perspectives, and the disagreement as the committee adjourned reflected how unsettled and inchoate ideas for the future of society in a world of AI remain.

2. The Geopolitics of AI

Expert Group Leader: Nicolas Miailhe, The Future Society

Committee co-Chair: Brian Tse

Key Recommendations

- A governance institution should establish a set of principles for AI similar to UNESCO's framework for Internet Universality. "AI Universality" could work to counter the hyper-competitive conditions of AI research we currently witness among major AI powers.
- A research initiative should be instituted to document each country's potential for:
    - i) domestic AI innovation; and
    - **ii)** capacity for sustainable absorption of AI systems into each economy.
- A research initiative is required to study the current and future AI value chain, including where and how value is created at each stage of AI practice (e.g. data collection, ML programming). We can use that information to better understand value-add and how to tax more smartly.

Two subcommittees met to discuss this topic:

A. **Digital Empires**
B. **National Strategies for a Fair Distribution of Power**

Geopolitics is a term of reference from realist international relations, and its study broadly relates to political power and how it is affected in, by and through geographic space. Applied to AI, geopolitics relates to concentrations in the ownership of data, patents, technologies and companies in specific countries and regions, and how power is distributed in light of those concentrations and their capacity for market valuation and export. Also, per many contemporary analyses of international geopolitics, thinkers in the AI governance space observe the significant growth of private actors as quasi-sovereign entities, holding the influence of independent countries in their own right.

Metaphors or Reality?

There is an array of metaphors currently used as shorthand to describe the distribution and concentration of AI production, ownership, and use. The most contentious was used in the very title of the first subcommittee, "digital empires." Other metaphors include the AI "arms race," analysis about global "AI city-states," and the opinion that Africa is the future "battleground" for the impending struggle between the US and China.

Metaphors like these help explain reality, but they can also lead to inaccurate assumptions about the world. Often, they carry ethical statements and implicit calls for action. Clearly, to speak of the emergence of a Chinese "digital empire," or to lament the EU's innovation "castration" through the GDPR, conveys partisan opinions about the role technology should have in our lives, and the threats that other value systems may have to ways of life.

Using metaphors to examine international power balances should always be tested against the real, empirical landscape of AI. For example, there is a move in the AI community to be more wary of using the term "AI arms race," since it does not convey the current boom in foreign investment in AI technologies, and side-lines more nuanced concepts like "co-opetition" and "cooperative rivalry" that may be more relevant to relations between major world powers.[4] To mitigate the tendency to oversimplify the emerging dynamics of the AI space, **participants called for new tools and methods which can be used to assess AI capacity and potential across the globe, and their geopolitical implications.** *A pathway is needed to incubate and survey such research initiatives across the globe, one shielded from partisan interests.*

With these caveats, GGAR participants explored how at present understanding AI through a realist international relations lens does lead to ***the unavoidable conclusion that we are witnessing the emergence of an "AI duopoly" with unique global dominance shown by the U.S. and China.***

AI Value Chains: Power in Practice

Nonetheless, this focus on the emerging U.S.-China technology duopoly does not fully reflect how AI technologies will be rolled out in diverse sectors and in novel ways in a global economy. Participants observed that in geopolitics analysis we pay much more attention to the development and production of AI "at source point," for example in preeminent research labs. ***Our focus on innovation "at the edge" comes to the detriment of understanding the geopolitical dynamics created as AI is applied throughout global value chains, such as value extracted from local concentrations and variance in user participation on personal computing devices.***

Participants also discussed how the coming years will see AI innovation and application diversify considerably, since the AI market is currently immature. Indeed, since 2017, experts referred to a "Renaissance of AI," ushered in by a number of countries which have developed national strategies for AI development. This aligns with the concept of the AI

[4] https://www.washingtonpost.com/outlook/2019/03/06/stop-calling-artificial-intelligence-research-an-arms-race/?noredirect=on&utm_term=.44386c3cd84e; https://www.amazon.com/Co-Opetition-Adam-M-Brandenburger/dp/0385479506; https://www.ibm.com/information-technology/artificial-intelligence-and-startups-ai-gold-rush

“gold rush” anticipated by many actors.[5] These strategies offer new directions for investment and growth; and each plan plays to the unique strengths of individual national economies. ***As these national strategies play out, this will likely lead to diversification in how AI technologies are being applied and valued in different contexts.*** *In that sense, perhaps talk of digital empires and AI hegemony is premature.*

At GGAR, experts called for more research into understanding how to value AI application and spill-over effects. We have little idea of how eminent AI companies will help associated industries and problems in society, and whether there will be a compounding “cluster effect” for global centres of excellence. For example, Anglo-American company DeepMind is a global leader in AI, but its revenue is dwarfed by American and Chinese Big Tech companies. At present, we do not understand in sophisticated, empirically testable ways how DeepMind’s research excellence will have spill-over effects for firms and countries with whom it works. *Participants called for a research initiative to study the global AI value chain that will enhance understanding of how AI innovation filters through society and contributes to existing geopolitical dynamics****. Such a pathway for research should include work to understand how AI can be taxed in “smart” rather than anti-competitive ways, and to reconsider world trade for the digital era.***

For all the network effects of AI innovation, experts expressed concern about the over-concentration of AI within two or only a few economies. ***Given that reality, participants were at pains to emphasise the consequences of AI geopolitics for international development.***

Geopolitics of AI for International Development

The U.S.-China Great Power narrative was challenged by the idea of promoting “Universal AI.” Participants saw that “Universal AI” can incorporate both normative and practical aspects to AI innovation, recognising it as both a global force, and one that will also have significant local variation. If we are entering into an AI “arms race” or witnessing a “gold rush,” it is essential to try to codify universalist principles, ones which seek to protect the interests of developing countries. Since 2015, UNESCO has advanced the concept of Internet Universality, which is deemed an important contribution in achieving the post-2015 Sustainable Development Agenda. ***Participants discussed a pathway to achieving this for AI Universality, and proposed that UNESCO could be an activating agent for achieving a “Universal AI” pact.***

Nonetheless, talking of universalism should not be a distraction from the huge global imbalances in AI innovation and roll-out, especially in the Global South. We should be aspiring to build a world where users of AI in less technologically developed countries can harness and deploy AI in empowering ways. That regional capacity-building offers a powerful opportunity for growth and self-determination.

[5] https://www.ibm.com/information-technology/artificial-intelligence-and-startups-ai-gold-rush

For experts at GGAR, the ongoing standoff between the U.S. and China concerning the rollout of 5G networks reflects a pressing and unnerving reality: decisions made today on critical infrastructure required for the AI revolution are likely to define the balance of international power for the next generation, and even beyond. ***Establishing normative frameworks and institutional fora for collaboration and participatory dialogue in this high-stakes global technology race have never felt so necessary.***

3. Agile Governance for Safe and Ethical AI

Expert Group Leader: Andre Loesekrug-Pietri

Committee co-Chairs: Gosia Loj, John C. Havens, and Isabela Ferrari

Key Recommendations:

- Governments should offer practical pathways and guidelines for private entities to follow, including through establishing a set of baseline principles for the ethical and safe implementation of AI.
- Governance actors should explore possibilities for "innovative" regulation such as sandboxes and experimentalist approaches. These initiatives can in turn help create more effective and accountable AI governance.
- Technology companies could be given formal responsibilities to disclose the incorporation of AI technologies in their products, part of a wider responsibility to make it clear to users how their algorithms work.
- Companies using AI should have a designated and identifiable AI ethics officer who is accountable and responsible for progressing ethical AI.
- Companies and other actors should institute an annual audit of their AI and data usage, similar to how sustainability reporting is now often done.

A governance framework able to manoeuvre and manage the deep complexities of AI needs to be agile, adaptive, credible, a good-faith broker, inclusive of multi-stakeholder input, comprehensive, and coordinated. However, actors trying to govern emerging technologies often revert to traditional governance measures that lack the flexibility needed to evolve at the same rate and in suitable ways for AI technologies. Due to the rapid, complex, global and unpredictable nature of AI development, designing smart policies to mitigate these risks is particularly challenging. 'Agile governance', also known as 'soft' governance tools, aim to address the shortcomings of standard policymaking processes and be more adaptive and responsive. Four sub-committees convened to discuss the following topics:

- **Multi-stakeholder Guidebook for Ethical and Safe AI**
- **Decentralized & distributed approaches**
- **Political Economy of Standardization**
- **Devising Innovative Regulation for AI**

Building Corporate Cooperation and Responsibilities

Prior to governing AI, it is imperative to know exactly where these technologies are being used. Technology companies need to disclose the incorporation of AI technologies in their products and make it clear to the public how their algorithms work. Good practices such as IBM's disclosure documents state which parts of their products contain AI technology as well as which AI technology should be adopted. This practice is essentially required as an extension of existing laws.

Educating engineers in AI ethics may be a solid method to nudge them to make more ethical AI. However, although it would be relatively straightforward to implement ethics training, the results may not be as fruitful as engineers are still first and foremost accountable to their companies. It is for the companies to establish broad accountability by, for example, establishing an identifiable ethics officer.

An Ethical Principles Guidebook

A guidebook to direct all stakeholders in the AI space to effectively implement ethical guidelines, standards, and codes in AI operations would be an optimum tool. However, experts thought it best to primarily focus their discussions on the private sector given its role in leading investment in research and AI development. The idea explored was to create a multi-stakeholder guidebook to lead organisations of all sizes, including start-ups and SMEs, for the promotion of more ethical AI development and deployment. Experts remarked a key component to the guidebook should be proportionality, taking into account the differences in growth and maturity of organisations and representing their true capacity to make the appropriate changes.

Taking into account the differences in ability of the various AI organisations may be essential, but not as essential as choosing the right ethical principles and codes as its foundations. Given the vast array of existing ethical principles and codes available, it is imperative to avoid the tendency to create new principles. Experts suggested choosing a set of principles that have the most backing and legitimacy, such as those created by the Institute of Electrical and Electronics Engineers or the OECD AIGO's work. Ultimately, the aim of this guidebook would be to provide a basic roadmap and concrete guidelines for the relevant stakeholders to execute on commonly agreed baseline principles, advocated as a solid benchmark for principles.

There are a number of qualities that the participants identified as a must for the guidebook. Firstly, to have a positive, widespread impact, it must be visible and accessible to all stakeholders and the public. Second, to ensure this tool has the right effect on the AI-related stakeholders and that it keeps up with how AI evolves to ensure an effective iterative process for the formation of guidelines that reflect the uncertainties and continuous developments of the AI revolution.

Lastly, experts suggested that the guidebook should establish a hierarchy of risk issues framed as tensions between responsibilities and risks, rather than a list of responsibilities versus risks. This approach takes into account that we are not aware of all of the risks and avoids the thinking of responsibility as a response to risk. Rather it highlights responsibility as the reference point.

Decentralised and distributed governance

The use of distributed technologies, such as Distributed Ledger Technologies (DLTs) - a database that is consensually shared and synchronized across multiple sites, institutions or geographies, can become an architecture for AI governance. To more holistically leverage decentralised, technology-based governance initial requirements include the right regulatory system and an understanding of how DLTs function and what they can support. Decentralised approaches improve a mechanism of control of assets (data); distributed mechanism ensures that the process is moved between multiple instances, thus safeguarding control. Initial requirements for such architecture include international standards and best practices guidelines on data governance (e.g. how to collect, use, share; privacy and data security). Under such a system, an anonymity principle is best ensured and so is protection of people. The use of decentralised approaches to AI governance therefore sets the level and balance between privacy, security and accessibility through auditing, enabled by DLTs.

However, the question of monitoring and public oversight is key when it comes to the intersection of governance and AI. If human intervention is involved, the efficiency otherwise secured with the decentralised technology will be compromised. The use of new protocols around decentralised data architecture is necessary and those can relate to technology itself as well as participants in the network. Such community data governance or opening system to a larger population potentially leads to less bias and greater accountability through transparency of the systems.

One of the examples discussed was to give the larger population to score and judge characteristics, making sure that actors understand what they are supporting. With a reward/scoring mechanism in place we make informed and not populist decisions. At this stage however, we still have to clarify and explain the deep tension between the risks and opportunities that are brought about through distributed and decentralized approaches. Very few people agree that these approaches are the way to go. From a point of view of policy-makers in relation to the use of DLTs for AI governance in society and government, very little is understood.

When there is an incident, policymakers tend to over-regulate. Therefore, a first step is to devise principles for governing the networks that would find solutions with decentralised community governance. The same is true for the problem of a time-scale where technologies are evolving a lot faster than a policy-planning cycle allows for its oversight. Therefore, working with politicians, political systems and policy-makers is key to enable the transition to implement DLTs for AI governance. Secondly, international organisations should promote best practices in terms of collecting, using and sharing data, especially meta-data derived from users' behaviour. Lastly, a provocative idea would be to allow networks/DLTs and AI to find the best solution for us and "let the best network win" by directing the technologies with values and principles.

Building trust through governance

Participants recognized the existing standards landscape as useful to build consensus in soft data governance approaches, however, it was felt that there is now a real need to interact with different standard-setting communities and to translate and bridge between geographies, i.e. North America, Europe, Asia, Africa etc. Within these distinctive contexts, to implement and adapt existing standards to a new world created by AI development, we can combine the two paradigms of top-down (legislation) and bottom-up (standard setting organisation). This creates a “Double Veil of Ignorance” where the implementation of both legislation and standards will question what would “true justice” look like. Clearly, this in itself requires the participants of all those affected by AI, requiring communication beyond AI experts and borders for a “representative justice". Standard-setting organizations align with this motive, their key aim is essentially to build trust and help bridge coordination problems where each actor has an incentive to deviate.

Other than building consensus through standard-setting bodies, experts noted that large corporations could take a decisive step by incorporating a section on AI and data in their annual reports, ensuring more visibility over the use of AI constituents. We can take example of the Fortune 500 “sustainability reports” which, although is not mandatory to publish, is almost common procedure as sustainable investment grows in popularity, with over 80 percent of 500 companies publishing in 2015. A pathway to achieving such reporting measures would involve international dialogue on how to make such auditing as meaningful and widespread as possible; including special requirements for multinational corporations.

Participants discussed the role of civil society, including citizens and the media, in the governance framework and suggested delegating algorithmic checks and supervision of AI tools to citizens. An example would be the case of government application of AI-related technology that restricts civil society’s rights or imposes sanctions. This would be an instance where the citizen would have a vested interest in properly monitoring such issues by highlighting core points that demand regulatory intervention. This is similar to when the investigative journal, ProPublica, proved that the recidivism COMPAS tool reflected racial bias. However, it is important to understand that this approach could potentially cause disruption when you have extreme groups in society who point out issues that are deemed unethical and therefore would not be followed up on. The assumption here is that citizens would only use this mechanism for positive instances in effort to secure civil rights. But the concept of positive is very subjective and we must question who defines this term.

Discussions also pointed out that civil society’s involvement in governance planning and investigations into the effects of algorithmic use could bring perspectives not imagined by regulators and other entities, either in AI development or post-development.

Regulation for AI

The subject of regulation for AI is controversial and is often perceived to hamper its innovation. However, some participants argued that many AI-specific issues fell under existing legislation and on the other hand, others suggested the creation of AI laws, either through a wholesale or domain-specific approach. One participant mentioned that in Estonia wholesale general regulation was used as it was thought that boundaries could not be drawn between AI product. An example could be that regulation for the autonomous car would be hard to distinguish from regulation for optimization tools.

It was also expressed that the purpose of regulating AI is not to regulate the technology, as misplaced regulation could derail the huge potential benefit AI could bring, but to regulate the capacity of AI and prevent harmful effects.

In any case, we must aim for a streamlined process for creating legislation around AI. This involves being aware of the dependencies between different forms of regulation, as they can refer to other legislative frameworks or laws cited for specific functions that can affect AI in undiscovered ways; while ensuring all sectors, public and private, communicate in order to align to avoid conflicts of interests.

Discussions moved to how governance actors can make innovative regulation. The following mechanisms were mentioned:

Participants suggested an experimentalist approach, by utilising regulatory sandboxes to ensure steps are taken to protect innovation in new sectors.

As an example, three categories of sandbox initiatives from Singapore were presented:
(1) Experimenting with existing licensing regimes for the discovery points of improvement to improve and customize new licenses.
(2) Exploring existing regulations to find appropriate amendment to laws.
(3) Crafting a certain set of regulations, through an iterative process and within a closed environment, that supports new businesses.

Participants noted that broad, results-focused regulation would deviate from dictating what AI should be like, but rather focus on what the capacity and outcome of AI should be, allowing entities to use their know-how to find the best paths to achieve their goals. An example would be to state that facial recognition devices should consider the plurality of human nature instead of determining that the algorithms should not be trained through unsupervised big data.

It was proposed to re-evaluate objectives after a period of time and collect best practices so that state and regulatory agencies can adjust expectations. By allowing trial periods for innovative initiatives, regulators and consumers can evaluate whether it had a beneficial effect.

Additionally, participants noted that forbidding undesired outcomes might be easier if stated as a prohibition of violating some central values that also happen to be rights. For example, algorithms might have discriminatory effects. Forbidding the result might be easier than forbidding all the ways it might happen. However, this is not quite as simple when it comes to results that are positive but cause negative consequences to get there.

4. Interpretable & Explainable AI

Expert Group Leader: Nozha Boujemaa

Committee co-Chairs: Dekai Wu, Jessica Cussins, Meeri Haataja, Jim Dratwa.

Key Recommendations:

- AI should strive to be intelligible so as to enable the people to educate themselves on the structure of decision-making processes in society. Intelligible AI should seek to understand how autonomous systems are part of wider, dynamic socio-technical systems.
- It should be explored how to ensure AI is held accountable through empirical evidence on the why, what, and how of AI systems to help facilitate more robust mechanisms for access justice.
- An observatory should be established focused on how explainability and cognate concepts are handled in different industries and contexts, and how they might be enhanced to allow for greater decision-making power for users.
- Governments should ensure that the actors responsible for explaining AI and making it intelligible, and those actors ("data subjects") who are being explained as part of that exploratory process, have institutionalised relationships, including defined roles and responsibilities.
- The status quo that prioritizes predictions and outcome production should be broken and prioritize the assessment and understanding of the data and how it is being used in decision-making processes and mobilize this through legal rules and independent evaluation.
- AI developers should move from a "removing bias" paradigm and towards a "harm reduction" to understand the foreseeable effects that algorithmic design and implementation may cause for disenfranchised populations and address the technical issues associated with algorithmic bias.

Deep learning neural networks are often labelled "black boxes" because, while their input and output are visible, the internal processes of getting from the input to the output remain opaque. Their architecture involves numerous "hidden" layers which are composed of linear and nonlinear functions. These functions are connected by weights which are adjusted in forward and back-propagation methods. For some applications or sectors (e.g. healthcare, law, banking, HR), there is significant interest in overcoming the challenge to interpret and explain decisions made by AI systems, posed by the "hidden" layers. Therefore, explainability has become a topic for technical and academic research.

Three subcommittees convened to discuss "Interpretable & Explainable AI", namely:

A. **What, Why, and How?**
B. **Algorithmic Bias – Value Alignment**

C. From Big Questions to Right Actions

Explainability, or Intelligibility?

The key aims of "Explainable AI" are to disentangle the what, why, who, and contexts of AI systems and their use, while reflecting on broader questions of:

- *What sort of world might using this AI lead to?*
- *What and where is justice in a hybrid autonomous-human system?*
- *Are there certain contexts in which AI is not suitable or safe enough to use?*

Participants pointed out that in our existing paradigm of innovation, we often place more emphasis on understanding the specific issues that AI systems are programmed to solve, including questions of "who is involved?" and "how will automation improve this system?", rather than ensuring the holistic explainability of that system. ***Allowing for that holistic explainability contextualises the use of AI into how and why it is being used, taking into account the whole social system being affected by automation, rather than just identifying the localised use of the algorithms alone.***

The experts also took issue with the broadness of the term "Explainable AI." What, they asked, would be the intent and scope of implementing Explainable AI as an industry norm? The aspiration for Explainable AI falls somewhere on the line between the two following statements:

- *Shallow Explainability:* *Explainable AI should explain how an algorithm works and the extent of its capabilities and influence on decision-making.*
- *Deep Explainability:* *Explainable AI should offer a systematic reflection of the social system in which an AI system operates; and how an AI works and is employed should be fully intelligible by humans. If "Deep Explainability" is not possible, the use of the AI system should be prohibited until a time at which it can be properly explained.*

Some experts proposed that *Intelligibility* or *Intelligible AI* provides for an even more risk-sensitive and detail-oriented approach to using AI systems. Intelligible AI would provide that:

- *AI can be understood in a comprehensive and systematic way by both developers and the people who would be implicated in the implementation of an AI system.*
- *The use of AI systems facilitates the general growth of human competence and control over the functions of AI. This condition extends into allowing people to understand how autonomous systems influence life outcomes, such as the role of government in citizen affairs.*
- *People will be in a position where they can educate themselves about the structure of decision-making processes in society more broadly, because Intelligible AI seeks to*

understand how autonomous systems are part of wider, dynamic socio-technical systems.

Given these parameters, the hope would be that Intelligible AI can also be a trust-building mechanism which will, if done correctly, lead to one of the following outcomes:

- *Increased user acceptance of the use of an AI system, because users are given the opportunity to question the objectives of the AI.*
- *Decisions to revoke consent of the use of an AI in situations where it is felt that accountability, intelligibility or fairness are not sufficient. This may lead to developers improving or correcting the code for reassessment; or to certain kinds of innovation being rejected.*
- *Users are able to assign liability in cases where developers have provided inaccurate information or have not acted in good faith to aspire for Intelligible AI.*

In this vein of thought, participants explored how implementing Intelligible AI in society will require various new institutions of governance, civic participation and regulatory review. In a society that ascribes to Intelligible AI, governments would need to ensure that the actors responsible for explaining AI and making it intelligible, and those actors ("data subjects") who are being explained as part of that exploratory process, have institutionalised relationships, including defined roles and responsibilities, reporting mechanisms and enforcement powers, that can offer an accurate transfer, regulation and accountability of knowledge.

The principle of comprehensive explainability, or Intelligible AI, leads to much more onerous responsibilities for all actors involved. It also throws up various practical and ethical dilemmas. ***However, the aspiration for Intelligible AI should not be completely cast aside; and even if it is deemed too challenging, various other propositions such as petitions or charters to progress requirements for transparency and understandability of AI should be seriously considered.***

Explainability towards liability

Explainability along with interpretability and transparency are essential to good governance: they help to determine causal processes and impacts.

Participants explored how being able to identify a causal link to allocate responsibility is a mechanism that can give redress in situations of exploitation by powerful actors. As it stands, unaccountable AI systems arguably allow developers and the companies for which they work to avoid the rule of law. ***Making sure that AI can be held accountable through empirical evidence on the why, what, and how of AI systems can help facilitate more robust mechanisms for access justice.***

Mechanisms for enhancing explainability

It was highlighted that, as the world moves forward in the design and implementation of autonomous systems, there is currently a tendency towards prioritizing predictions and outcome production. This focus comes to the detriment of breaking through the opacity of autonomous systems, *which would allow for full assessment and understanding of the data and how it is being used in decision-making processes. The status quo is inadequate and change needs to be mobilized through legal rules and independent evaluation.*

As AI systems become more advanced, it becomes close to impossible to understand algorithms. Similar to the concept of financial auditing, third-party auditing is a common practice among organisations such as the OECD and private companies. It exposes the disparate impacts of AI by investigating the people who programmed the software to the training data (for example by identifying hidden influence of user-defined sensitive variables on other variables) to the output. With the causal connections made between institutions programming and outcomes achieved, we are able to flag biases in the process. Third-party auditing processes are informative to prepare policymakers and users on what to expect from the implementation of an AI system and how it might be improved prior to use in the real world. ***Experts called for more research and pathways towards institution-building that would standardise third-party auditing in industry.***

Participants observed the new research currently underway in causal machine learning and probabilistic modelling, which increases the transparency of AI functions through spotting mistakes due to distributional drift or incomplete representations of goals and features. Using inherently more interpretable models can help build trust in AI and facilitate more control by humans.

Having explored the central emerging crux of explainability for AI, ***participants proposed establishing an observatory focused on how explainability and cognate concepts are handled in different industries and contexts,*** and how they might be enhanced to allow for greater decision-making power for users. A primary task for the observatory would be to evaluate explainability under different cultural contexts and provide lower of higher assumptions, depending.

Algorithmic Bias; and towards a fairer, low/no-harm AI

As artificial intelligence systems become more advanced, their logic pathways become increasingly difficult to understand. At present, even with the existing research in the field of explainability, we do not have the appropriate tools to understand and pinpoint the areas where people experience discrimination. ***In order to resolve biases in a world shaped by algorithmic decision-making, we require new and improved metrics to measure the distribution of bias and its impact through mechanisms such as metadata on the datasets used in current AI.***

Although being able to pinpoint algorithmic bias is a good starting point, ***participants observed that addressing the technical issues associated with algorithmic bias requires moving beyond a "removing bias" paradigm and towards one that seeks "harm reduction".***

The removal of bias does not necessarily promote justice, for biases are not inherently detrimental nor based on prejudices. Accordingly, we must go further than understanding biases and begin analysing their connection with harms. This will allow us to more fully understand the foreseeable effects that algorithmic design and implementation may cause for disenfranchised populations.

Bringing issues of algorithmic bias to use-case level can help avoid a "one-size-fits-all" approach and would cater for different perspectives of unwanted biases from various contexts and cultural settings. Also, a use-case approach provides the opportunity to question whether or not the algorithm used is appropriate in the first place. ***Some experts conveyed that it is essential to fundamentally question the design of algorithms because efficiency, a large driver of algorithms, in and of itself is more likely to cause harm than a human-centric system.***

Recently, scholars identified twenty-one forms of fairness in the field of machine learning. This diversity in thinking about fairness brings its own practical issues: the individual politics associated with each definition of fairness necessarily leading to some incompatibility. Divergent models of fairness create a "trolley problem" where the use of one model may be to the detriment of the other. Participants observed that integrating different models of fairness in AI systems in a society that uses automation will lead to contradictions and problems for a rules-based governance system aspiring for Explainable/Intelligible AI. ***The only solution to avoiding "fairness system clashes" is much greater multilateral dialogue, new systems of governance, and codified rules which set out the regulatory and legal processes if there are inconsistencies between different AI systems.***

5. Governance of the Development of Artificial General Intelligence (AGI)

Expert Group Leader: Jessica Cussins

Committee co-Chairs: Richard Mallah, Seán Ó hÉigeartaigh

Key Recommendations:

- Comprehensively funded public research efforts should be instituted that are briefed with investigating the work of AGI laboratories across the globe. Further research efforts should be promoted that seek to understand the impacts of the spread and intensity of those AGI projects, to risk assess them, and to scope the pathways to the proper governance of AGI if and when it is developed.
- A "CERN for AI" should be seriously considered by major AI powers and world-class researchers. Such a project could lead to better collaboration and safer scientific development of an AGI.

Three subcommittees convened to discuss the Governance of the Development of Artificial General Intelligence:

A. **Direct and Indirect Policy Recommendations**
B. **Other mechanisms for impact**
C. **Stakeholder Coordination**

Thresholds for AGI

Discussions concerning AGI spent a deal of time reflecting on definitions. Importantly, experts explored how we should not expect a straightforward linearity between how we know AI today and how we will define and recognise an "AGI" of tomorrow. The kinds of development that will be seen, and the paths that might be followed to make AGI a reality, have a considerable impact on how governance regimes should be designed and implemented today. It could be ineffectual to put in place policies designed for today's AI landscape if in doing so we do not generalize well. We should be wary about designing policy about AI that could be unsuited to future requirements.

In this vein, participants called for more sustained scientific enquiry led by public bodies about the possible transition from AI to AGI. Today, there is significant disagreement on what AGI will look like or when it might arrive. It could be the case that unique combinations of different AIs operating together create something which we could call "AGI." Other experts made the point that **comparing AI with AGI is not necessarily a helpful or meaningful comparison**. They believed that thinking about the AI-to-AGI transition as a series of stages or as innovations operating on a continuum may not be a sound means of assessment. Some experts even cast doubt on whether we really know anything substantive about AGI today, beyond the implications from constraints derived from physics and logic.

Participants debated what is the game-changing quality or characteristic that will mark AGI's genesis. Some of the varied and controversial **suggestions included sentience, agency, learning arbitrary jobs quickly, and consciousness. Participants observed that even these terms may be stretched to such an extent that they could prove unhelpful for understanding the technological transition towards AGI.**

Finally, and testament to a ranging discussion, several experts expressed the concern that these discussions of definitions and thresholds may be premature, and neglect the more immanent discussions about "simple" AI that need to be had as a matter of urgency.

Values and AGI

Nonetheless, having a "pre-emptive" discussion about AGI in itself brings many normative questions. In part, the drive to discuss the governance and ethical implications of AGI now is because no institution or state has "won" the "race" for its creation yet. Since with AGI the overall pie will grow significantly, and conflicts could be disastrous beyond today's possibilities, cooperation among relevant powers seems clearly in mutual self-interest. Experts explored how it might be easier to arrive at consensus-driven solutions before we see an even less equal playing field emerge. **Putting governance frameworks in place now can help mitigate the risks of a winner-takes-all AGI scenario**.

It is also unclear how we should anticipate and govern AGI due to the uncertainty about AI thresholds and transitions. If norms, principles and guidelines are put in place for innovation today, **it may follow that these remain "sticky" and lay a good foundation for a world of tomorrow contending with AGI.** Given there was such pervasive disagreement over how we should define our terms of reference, there was likewise little resolution as to how far we can proceed in governing AGI using vague aspirations for institutional and normative durability.

As participants navigated tensions surrounding the timing and scope of governance interventions for AGI, this led them to question how far "we" (humanity writ large) have agency to govern the development of an AGI. Encouragingly, early research in AI safety and responsible AI seems promising. However, there is less work being done to understand the kinds of control and freedom certain actors and institutions might have in the process of AGI development and regulation. ***Participants called for an intensification of the good research already being done surrounding pathways to AGI and options for risk mitigation.***

One area of more general consensus was the belief that an AGI would not itself require "agency" to exist, nor will it require "goals" and "ends" in the human sense. This will make it more difficult to ascertain whether a future AGI is operating in alignment with human values. This is not just a technical question. In fact, it seems integral to our idea of human agency and self-expression to believe that we have decision-making capacity about whether or not to delegate to a machine. Human capability to understand what an AGI is and how it works brings existential questions about our place and purpose in the world. Given this, **experts again called for more technical and policy research to develop our understanding of how AGI can be controlled and held accountable in human-centric systems.**

Instituting a Global Body for Communicating and/or Observing AGI

Observing and communicating about A(G)I are two ways of understanding and expressing the values that we hold. However, the question of who should have this communicative and/or observatory role is hardly straightforward. In particular, the authority and scientific objectivity of any global institution such as the proposed International Panel on AI always be contested.

It is not only global actors like states and big tech companies which are capable of contesting the authority of institutions that carry observatory or communicative mandates. Publics also participate in such debates. **GGAR participants disagreed about what place the public should have in an institution charged with understanding the development of AGI.** Encouraging participation from the broadest possible range of voices may limit what can be discussed openly, in sufficient depth, and with sufficient scientific rigor. It is also difficult to build an AGI observatory or science communication institution that is truly equitable, accessible and representative if it must also be responsible for driving high-stakes global policy. Equally, qualities like openness and transparency are fundamental to any science observation/communication institution.

The job of global observation and communication about AGI will be extremely complex and iterative, and will necessarily involve a diverse range of actors if it is to be both scientifically credible and respected by publics. There was a general consensus, however, that **thinking about the ethics of observation and communication should not be seen as a stumbling block, but as an intrinsic part of the process of making AI beneficial to humanity as a whole.**

Much discussion focused on institutional endurance for an AGI observatory or science communication body. A good observatory institution should be able to consider on a continuum what are the short, medium and long-term aspects and consequences of the development of an AGI. Given the nature of the subject matter, these understandings will change over time. We should aspire to put in place a durable policy and governance infrastructure that is applicable for both today's AI landscape and one in which an AGI could in future feature. If an AGI monitoring, evaluation or policing institution is built, it must be able to adapt and respond to changing scientific and social realities. **Experts called for a comprehensive assessment of the current case for a global AGI science observatory, to include consideration of which actors should be involved, what are its reporting procedures and mandates, and how it could be built for short and long-term relevance.**

Building an A(G)I Mega-Project

Aside from the centrally important jobs of observing and communicating about AI is the task of "doing" the actual science of A(G)I innovation. That too requires good governance and well thought through institutional design. **Participants discussed a "CERN for AI," which would be a multi-stakeholder organisation housing a grand scientific project, perhaps to include the development of a safe and beneficial AGI.**

One of the main features of an AI Mega-Project would be that it includes both a grand scientific project and a separate risk assessment and policy component. GGAR participants suggested that there should be a suitable firewall between these two divisions to prevent the project being politicised. Not having such a firewall in place could undermine the objectivity of the science being done, as well as the claims to good management, communication and analysis that the "social" arm of the institution would hope to have.

GGAR experts were frank about how difficult it would be to achieve a "CERN for AI" mega-project, not least due to the general dominance of private for-profit actors in the AI field. That said, one of the main benefits of such a project would be to show how cooperation and competition can coexist under the roof of one institution. **It could be in the contexts of such a mega-project—be it European, US-Chinese, or truly global—that the foundations of effective multilateral trust are built. If we are to see an AGI developed that remains safe and fit for purpose, this layered and deep kind of trust seems a prerequisite.**

6. Building Capability for 'Smart' Governance of Artificial Intelligence

Expert Group Leader: Konstantinos Karachalios

Committee Chairs: Leanne Fry, Lord Tim Clement-Jones, Ali Hessemi

Key Recommendations:

- Governments should take the lead to build pathways to address imbalance between private and public sectors' resources and for research into how technological deficiency could change trust, faith and citizen reliance on government services.
- There should be a call for research initiatives at policy schools and similar institutes so governments can understand how to build effective collaborations to foster innovation while respecting government mandates.
- Governments should undertake formal responsibilities to understand and respect the social diversity of their populations and assess risks such as contravening the right to privacy if data is not properly stored or managed.
- For safe, responsible and effective integration of AI there should be widespread educational initiatives within government civil services.
- There should be more research into institutional design of smart AI governance agencies that would be mandated to protect public interest and/or assess successes and failures of specific projects.

AI adoption into public sector organizations and processes requires diverse and multi-level competence and capabilities. Policymakers are often less adept at managing AI technologies and systems and their adoption can often lag behind industry in terms of technical expertise.

Knowledge gaps and talent shortages in this technology area and in the associated field of cybersecurity hamper the capacity to formulate adequate technology policy and practices. For public sector organizations without prior expertise in managing AI systems or digital technologies more generally, the appropriate 'smart' governance strategy for their implementation is not immediately clear. Capability building often requires outside support, for example from external consultants and researchers, including to set up trial periods and to implement and monitor new programs.

Four subcommittees convened to discuss "Building Capability for 'Smart' Governance of Artificial Intelligence", namely:

A. **Building Competency for Governing AI in the Public Sector**
B. **How to Build Public Trust**
C. **Lessons from Case Studies**
D. **The case for Public-Private-People Partnerships**

Building government capabilities for smart governance

Participants highlighted that the added value of AI in government is twofold: smart governance can supplement service delivery with AI for efficiency and productivity gains, and it can also supplement the policymaking process through data-driven decision-making. **However, it was pointed out that most governments at present fall short of resources such as money, time, talent and expertise, almost always lagging behind the private sector.** This imbalance becomes more problematic as AI capabilities further develop, because government is accountable to its citizens in a way that the private sector cannot be.

Participants saw it as undesirable for that gulf between private and public sectors to widen further, since in that scenario governments would have no alternative but to outsource responsibilities and decision-making. Such a situation will be expensive for governments, but also problematic if they do not have sufficient information and know-how about how their citizens are being "governed" by private actors in their stead. **Participants called for government-led pathways to address this imbalance and for research into how technological deficiency could change trust, faith and citizen reliance on government services.**

Public-Private-People-Partnership collaborations for smart governance

What can we expect governments to achieve in the AI space? It is likely unrealistic to see government as capable of excelling in AI R&D: that ship has already sailed, with private sector companies and some international universities leading the way. Instead, it might be best to think of government as a platform that facilitates collaborations, guides principles, and communicates to and learns from the public. It is less feasible for governments to be bodies that produce ready-made or technologically advanced solutions. **Cross-sector collaborations are likely the most beneficial means to advance AI initiatives sustainably and safely**, but the question remains whether there are sufficient incentives and a responsibly crafted risk and reward structure for different sectors to work together.

Governments can be supported by a number of entities: participants explored how academia can inform smarter decision-making for policies and can help evolve policymaking through testable and iterative procedures; while the private sector can help speed up the delivery of services and assist decision-making pertaining to education and agenda-setting. **Committee members called for research initiatives at policy schools and similar institutes for good government to understand how effective collaborations can be built in ways that foster innovation while respecting government mandates.**

Committee members discussed the "Usability" of different types of Public-Private-People Partnerships. The outcomes of those committee ratings were as follows:[6]

Types of PPPPs	Likelihood	Effectiveness	Usability Index
Education	3	3	9
Hackathons	5	2	10
Interest alignment	2	5	10
Self-Sovereign identity-based coop for governance use of data	3	4	12
Bottom-up Regulation	5	3	15
Considering Ethical Issues	4	4	16
Subsidies, investment, grant, procurement, experiments, crowd-funding	5	4	20
Continuous participation of people in government	5	5	25
Commons/Open Source for Software	5	5	25

This table represents the "usability" of Public-Private-People Partnerships with Usability Index = L (Likelihood) x E (Effectiveness))

<u>Best practices for government adoption of AI</u>

GGAR participants also discussed how governments could integrate AI tools into public services in the near-future. The possibilities for this are endless, but having a comprehensive **data assessment process is key. Governments should undertake formal responsibilities to understand and respect the social diversity of their populations and assess risks such as contravening the right to privacy if data is not properly stored or managed.** Huge "data warehouses" for specific domains like healthcare or agriculture could help government decision-making, but they carry many challenges such as respect for citizen privacy.

Even before governments begin to introduce AI technologies into smart governance mechanisms, it is essential that they prioritize technology literacy internally. **Safe, responsible and effective integration of AI will require widespread educational initiatives within government civil services.** Participants recognized the benefits of a 'learning-by-doing' approach to smart governance, in which public organizations begin with small-scale projects involving adoption of AI systems and gradually scale up test projects in terms of size and scope. Such a process should be iterative and based on feedback and

[6] The International Organisation for Standardization defines this Usability Index as "the extent to which a system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." Here, participants determined "usability" by the measurements of "likelihood", a rating from 1 to 5 of how likely this type of partnership is; and a measurement of "effectiveness," a rating from 1 to 5 of how effective the implementation of a given partnership would be. Overall, "usability" is the multiple of the rate of "likelihood" and "effectiveness", which has been represented in the table below.

learned expertise. This kind of approach could fit well with private sector and academic methodologies, and foster a multi-stakeholder process.

Building trust in smart governance

Participants highlighted that without sufficient opportunity for citizens to test, improve upon, and offer meaningful input into the architecture of smart governance tools, there will not be sufficient societal buy-in, and also an increased risk of uneven take-up or backlash from the public around AI in governance.

Smart governance is a field with huge scope for expansion. However, there is a particular requirement that governments are able to comprehensively assess the benefits and risks of integrating AI into different sectors and for diverse purposes. This onus is more pressing for governments than it is for the private sector, because people choose to buy industrial products, while they are often required to participate in government-run schemes, such as income tax filings. Also, government services must be accessible to everyone, including those who do not have high technology literacy or who may not be willing to participate in AI-enabled schemes. **Participants explored how smart governance capabilities could be enhanced fairly and properly by improving the overarching regulatory architecture that polices and evaluates government services.** Social recognition of smart governance mechanisms derive from:

(i) Maintaining institutions considered generally capable of redressing problems;
(ii) Initiatives perceived as acting in pursuit of the public good; and
(iii) Consistency between what the technology promised and what it achieved, with generally positive experiences for citizens.

Through achieving these targets, there is significant scope for government adoption of smart governance tools. Experts highlighted that to build such mechanisms, governments would need sufficient talent and skills development in place on an ongoing basis to be in a position to implement smart governance initiatives and be effective communicators about it.

How can this communication role for government be achieved and properly maintained? First, there are dilemmas concerning how governments should communicate about a technology which is inherently complex and difficult to explain to non-experts. Governments need to navigate whether simplicity and clarity, vice comprehensiveness and technicality of how smart governance tools work, is preferable.

Participants also pointed out that alongside questions of “message” and communication are ones of “messenger” and mandate. They identified that it is unlikely to be central government that polices AI tools, but rather arms-length governmental agencies mandated to implement and monitor smart governance. One possibility for such an AI governance agency would be to foster the kind of trust that firefighter services hold in the eyes of citizens. Alternatively, this kind of agency could carry the independent scrutiny capabilities that offices of budget accountability and responsibility have in many countries. **Experts called for more research into institutional design of smart AI governance agencies that would be mandated to protect public interest and/or assess successes and failures of specific projects.**

Participants identified the scale of the current challenge for governments hoping to offer effective and accountable smart governance services. Given this reality, it seems likely that many governments will pursue the track of outsourcing projects which they do not consider themselves capable of managing. As with any other outsourcing project, poorly managed projects that lack oversight often lead to criticisms of poor accountability and legitimacy, and weakened trust in government services. In this sense, AI governance is no different from governance that uses traditional methods. **Participants stated that governments should develop clear agendas for the use and maintenance of smart governance programs, paying especial attention to the role of outsourcing, and ensuring that they have sufficient capabilities to be able to understand and audit companies carrying out services on their behalf.**

Finally, it was emphasized that there must be ample opportunity for citizen involvement at every stage of the scoping and building of smart governance products. It is easy to see how smart governance could fall victim to the usual problems of “top-down” government approaches, since is likely that smart governance programs would be applied at scales that affect thousands if not millions of citizens. **Participants explained that government must actively choose to deploy smart governance programs responsive to local needs.** It is through a “bottom up,” experimentalist approach to building such initiatives that we could see smart governance achieve wide acceptance in communities and begin to be a trusted component of government service provision to citizens. Again, the yearning for localized solutions is hardly unique to smart governance, showing how AI brings to the fore both new problems and perennial dilemmas that have faced governments since long before the digital revolution.

7. Governing AI Adoption in Developing Countries

Expert Group Leader: Zaki Khoury

Committee co-Chairs: Eileen Lach and Stan Byers

Key Recommendations

- The governments of developing countries and international organizations should invest in massive vocational training schemes to build a workforce with the skills necessary for a world with deep AI usage. Generally, these vocational schemes should take priority over elite education programs, which too often lead to a "brain drain" out of developing countries.
- A research initiative should be instituted which analyses the anticipated effects of AI on existing urban-rural divides, and promotes policy agendas which would mitigate negative impacts.
- NGOs should be involved in development and infrastructure programs in developing countries related to AI. NGOs are able to act as "honest brokers" due to their unique knowledge and skills, and could help prevent undesirable power imbalances between investors and recipient countries.

AI and other emerging technologies provide many opportunities for people living in developing countries, for example through providing better and more widespread access to vital goods and services in areas like healthcare, food, education and energy. At the same time, AI technologies and how they are developed and deployed undoubtedly bring an array of challenges. These include potentially exploitative relationships in trade agreements, technology sharing, data mobility and other multilateral and public-private partnerships are not fairly negotiated, or if citizens' rights are not properly respected.

GGAR participants reiterated the importance for developing countries to build capacity and knowhow that will help prepare them for the age of AI. Experts also encouraged policies and programs which would allow the benefits of that preparedness and the subsequent technological transition to be extended across socio-economic classes and for those living in both urban and rural areas.

Three subcommittees convened to discuss the subject of "Governing AI Adoption in Developing Countries", namely:

A. **Building Capabilities while Avoiding Exploitation**
B. **Opportunities & Challenges**
C. **Managing Risks vs. Opportunities for Development**

Preventing Brain Drains and Data Liberalisation in the Developing World

As is often the case in the development sector, policy discussions about AI and international development must bridge the gap between well-meaning ambitions and the more feasible possibilities available. **GGAR participants remarked of the yawning gulf between AI development in major world research clusters like Silicon Valley, and those of less developed countries, particularly in Africa.**

The realities of this huge divide in AI research and development capacities led to discussion about what place we should practicably be aspiring for lesser developed countries to have in the wider technological revolution. It seems that government-led attempts to prevent brain drains rarely succeed. One example is the practice of developing countries supporting their elite students with scholarships to attain advanced degrees at global universities. As well as the difficulties gauging returns on such investments, having a very small number of elite educated researchers is often not the best avenue for ensuring more widespread and equitable capacity and revenue building in developing countries. **There continue to be well-documented issues of the "brain drain," and this phenomenon is particularly acute in the AI and tech sector of lower-income countries.**

Much discussion in these subcommittees related to issues of data value and management for the developing world. This is because well-established imbalances and inequities mean that it will be difficult for developing countries to become leaders in AI R&D. **Instead, developing governments should in general aspire to effectively manage and cultivate the AI value chain through comprehensive domestic Big Data management programs.**

How to properly manage domestic Big Data in responsible ways presents a central tension for governments in developing countries. They are actors uniquely capable of facilitating and capitalising on data availability. For example, governments often have authority to sell datasets about public utilities like water to private companies. And yet, these governments must also recognise their responsibility to understand the sensitivity and value of the data that they hold, and where appropriate to shield their citizens from commercialisation that could exacerbate existing inequities of the North-South divide. **Participants voiced concerns about the emergence of "digital empires" and the need to educate about AI and build up capacity in developing world governments if they are to be protected from exploitation.**

The policy landscape being faced by developing countries is not dissimilar from that of the 1980s, where Western powers encouraged developing countries to liberalise their trade regimes. The positive and negative effects of the liberalisation era are still hotly debated. There is nothing to suggest that those surrounding global data mobility and the AI revolution in the developing world will be any different, in terms of whether international openness will empower or compromise poorer nations.

Creating Value Opportunities for Developing Countries in the AI Revolution

Domestic capacities in the AI sector can be enhanced through the creation and harnessing of indigenous, clean datasets. Getting this right will involve a concerted effort to train governments and bureaucracies in understanding how domestic data generation and management can be safely and ethically pursued. To achieve this, a much-needed shift is required in the developing world away from educating the privileged few to the highest standards, **and towards facilitating much more widespread vocational training schemes drawing predominately from the STEM subjects.**

Generally, this will not require highly advanced and resource-intensive training in STEM subjects. The focus should be to offer opportunities for the creation of good technology-focused jobs across the developing world and for significant numbers of people. At present, since we know too little about AI's global industrial added value chains, we also do not know how those kinds of domestic developing country job opportunities can be best facilitated and cultivated by governments and IGOs like the World Bank. **Participants called for much greater research and analysis into how to create jobs and value opportunities for the AI revolution in the developing world.**

It was emphasised repeatedly by participants that AI governance experts need to refocus their attention away from global centres of excellence (e.g. Silicon Valley, London) and towards envisaging realistic plans that would allow developing countries to participate meaningfully in the AI revolution. One expert gave the creative analogy of the Gold Rush: few made their fortunes panning for gold, but many more drew good livelihoods by riding the wave of opportunities. **Governments in developing countries need comprehensive training and capability-building schemes to understand how they are best-placed to ride the wave of this technological transformation.**

Holistic Development in an Age of AI

We cannot be blinded to the risks of a new digital divide becoming entrenched, with AI industries developing at a rapid pace in just a select few countries around the world. In particular, we can learn from previous cases of this phenomenon to understand how the digital divide was borne partly out of developing countries' lack of basic infrastructures, social capital, and governance capacities to roll out technologies in a beneficial and safe way. Good livelihoods empowered through access to advanced technologies are generally contained within small, elite, networked urban zones in the developing world—while the majority are left unable to harness the benefits. **Participants called for more research to understand how AI will affect established urban-rural divides and what policies might help prevent those divides growing.**

Providing utilities like food, electricity, and safe water, and putting in place good cross-country transport systems, must be progressed for developing countries to reap the benefits of the AI revolution. One participant observed how the Indian government's attitude of "connectivity will sort itself out" likely exacerbates existing economic divides—but it seems that this attitude is replicated in many developing nations. Since we know too little about the global AI value chain, we also have scant knowledge about where resources should be focused to help developing countries cultivate their technology sectors, while also improving general prosperity. Technologies like 5G will allow machine learning and AI functions to be decentralized and performed locally on devices, which provides ample opportunity for value creation, if the right governance and business infrastructure is put in place. **If basic infrastructures in the developing world do not reach beyond contained urban zones, it is impossible to see how the AI revolution can occur in a way that enhances general prosperity in the Global South.**

Participants agreed that it is not in the interests of governments in the developing world to fully close off investment and innovation available from AI technology companies and major players like the US, China and EU. But developing countries need to be capable of understanding the benefits and risks that come with the various pathways to development currently available, particularly when applied to AI. For that, there is no alternative to hard work on the ground improving AI know-how in developing world governments. **Additionally, providing formal oversight access to "honest broker" actors like NGOs should be a priority as major infrastructure and development plans are pursued**, for they are well-positioned to advise governments on how to avoid entrenching power and monetary imbalances with the developed world.

8. AI in the Judicial system, Access to Justice, and the Practice of Law

Expert Group Leader: Nicolas Economou

Summary of proceedings:

Consistently with the Roundtable's 2019 theme of *Pathways to Global Governance*, the committee conducted its proceedings under the prism of advancing "from Principles to Practice" in achieving Informed Trust in the adoption of AI in legal systems and the practice of law. The committee's agenda rested on the following focus and objective:

- **Focus:** How AI, if properly governed, could enable the law (law-making; civil and criminal justice; law enforcement) to enhance the functions of the law and to protect and advance human well-being.
- **Objective:** The development of actionable and effective, yet adaptable, norms for the trustworthy adoption of AI in legal systems and the practice of law, grounded in four principles that are common to The Future Society and the Law Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The four principles" are: Effectiveness, Competence, Accountability, Transparency.
- The Committee concentrated its deliberations on the principles of Effectiveness and Competence.
- In examining what constitutes sound evidence of effectiveness (fitness for purpose) and operator competence in AI-enabled systems, the Committee reviewed studies conducted by the United States National Institute of Standards and Technology (US NIST), which assessed the performance of AI-enabled processes in legal fact finding (electronic discovery).
- The Committee also discussed the decision-making process that underpinned the initial judicial approval of AI in legal discovery in the United States, including the afore-mentioned US NIST studies and a cost-benefit analysis.
- The Committee discussed what constitutes sound evidence that operators of AI in legal applications are competent to operate it effectively and safely. In that context, the absence of certifications attesting to such competence was acknowledged. Various legal and judicial educational endeavors were also discussed.
- The Committee unanimously agreed that sound evidence should inform decision-making in determining the extent to which AI-enabled systems, including their operators, were effective at meeting the intended objective.
- The Committee acknowledged that, with some exceptions, including the US NIST studies, sound studies of such effectiveness were scant.

Introduction:

In its 2018 proceedings, the Law Committee acknowledged the great variety of use-cases and applications of AI in legal systems and the practice of law, and recognized that such applications involve substantially different stakes, risks, and benefits. This realization informed the development of the agenda for the 2019 proceedings, in particular the need for a definition of **Informed Trust**, resting on a single set of principles that should be:

- ✔ Individually necessary and collectively sufficient.
- ✔ Accounting for the totality of the legal system.
- ✔ Viewing the legal system as an institution *accountable to the citizen.*
- ✔ Applicable irrespective of legal or cultural tradition.
- ✔ Capable of being operationalized.
- ✔ Capacious enough to apply to all use-cases (and, as needed, within different instances of a same use-case).
- ✔ Technology neutral.
- ✔ Able to adapt to rapid innovation.

For the 2019 edition, it was proposed that four principles, which met the above desiderate, were constitutive of **Informed Trust** (or informed mistrust) in the adoption of AI in legal systems and the practice of law. These four principles, drawn from those examined during the 2018 proceedings of the Law Committee are: Effectiveness, Competence, Accountability, Transparency. ("The Four Principles").

Advancing the operationalization of Informed Trust.

In advancement of the Roundtable's theme of *Pathways to Global Governance*, the Committee focused its agenda on advancing the operationalization in practice of the Four Principles, summarized as follows:

- **Effectiveness:** Creators and operators shall provide evidence of the effectiveness (fitness for purpose) of an AI-enabled system.
- **Competence:** Creators shall specify, and operators shall adhere to, the knowledge and skill required for safe and effective operation of an AI-enabled system
- **Accountability:** AI-enabled systems shall be created and operated such that it is possible to trace lines of responsibility, among the agents involved in the creation and operation of the system, for a given outcome.
- **Transparency:** The basis of any decision made (or to be made) by an AI-enabled system shall be discoverable

In its deliberations, the Committee prioritized the examination of the principles of "**Effectiveness**" and "**Competence**" in AI-enabled systems deployed in legal systems and the practice of law.

- **With respect to Effectiveness:**
  - The Committee unanimously agreed that sound evidence is important in determining the extent to which AI-enabled processes should be trusted (or mistrusted) in legal systems and the practice of law. It was acknowledged that, with few exceptions, scientifically sound evidence of the effectiveness of AI applications in the law was scant.
  - When discussing what types of information constitute a sound basis to enable evidence-based decision-making, the Committee reviewed evidence emanating from the domain of legal fact-finding (electronic discovery). In the United States, the main

factors that influenced the judicial approval of the use of AI-enabled processes in the domain were:

  - **Scientific evidence** produced by the *United States National Institute of Standards and Technology* (NIST) in its ground-breaking series of studies known as TREC Legal Track and a widely cited meta-study, which established that some AI-enabled electronic discovery processes outperformed human attorneys at the task. Those studies contributed to the establishment of sound metrics, known as "precision" and "recall" for the measurement of the effectiveness of AI-enabled processes in electronic discovery.
  - **A cost-benefit analysis**, which showed that, to the extent that AI-enabled systems could also be trusted to perform as effectively as, or more effectively than humans, the cost savings were considerable, thus facilitating access to justice.
- o The need for scientifically sound metrics of the effectiveness of AI-enabled processes was underscored in a discussion of risk assessment algorithms in bail hearings and sentencing. It was noted that that there was no consensus evidence of the extent to which such algorithms are effective at producing accurate scores or at avoiding bias.

- **With respect to Competence:**
  - o The Committee reviewed evidence from the afore-mentioned US NIST studies suggesting that competence of operators of AI in legal applications cannot be taken for granted. Data from those studies suggested that participating operators of AI-enabled systems were unevenly unable to correctly evaluate the accuracy of the AI-enabled systems they operated. Specifically, the mean difference between the participants' own estimation of an important measure of accuracy (known as "recall") and the actual accuracy they had achieved was 34%. As an illustrative example, this result suggested that an operator recall estimate of 50% could reflect a performance as low as 16% and as high as 84%. This finding suggested the inadequate application of statistics to the task, raising the question of insufficient competence. In addition, this finding raised the question of whether operators of AI unable to effectively measure the accuracy of their processes, are competent operators in the first place.
  - o When discussing what types of information may constitute a sound evidence of competence, the committee acknowledged that no consensus instruments existed today, which would attest to the operator's competence in AI systems deployed in the law.
  - o The Committee then heard from committee members engaged in legal and judicial education. The perspectives and information shared included:
    - Various endeavors designed to enhance the AI literacy of stakeholders in the legal system.
    - Views that legal education must evolve so that students and practitioners have at least a generalist's understanding of AI, so as to render them capable of recognizing the type of expertise, including scientific expertise, needed in the operation and the measurement of effectiveness of AI-enabled systems.
    - Efforts by leading international law firms to bridge the knowledge gap by bringing in technical or scientific experts to educate and inform lawyers.

- Efforts by certain legal or judicial education institutions to introduce basic statistics as a requirement, so that, at the very least, lawyers and judges are better able to understand the requisite skills in statistics or other domains that may be needed to understand, operate, and measure the performance of AI-enabled systems.
- Consideration of the extent to which introducing some additional courses within ordinary legal education may be (in)sufficient to provide lawyers with the requisite skills to operate and measure the effectiveness of AI, in particular in sensitive use-cases.
- Consideration of the extent to which education developed specifically for lawyers on AI-enabled electronic discovery and related statistics sufficiently equipped them to reliably make correct statements of fact to opposing counsel or courts.
- Consideration of the duty of professionals involved in the legal system to seek to understand the extent of their technical and scientific competence and to identify appropriate experts in the operation and measurement of AI-enabled systems.
- Consideration of the extent to which such expertise is readily available in the market. The example of the U.S. was cited, where statisticians with specific expertise in the measurement of the effectiveness of AI in legal context were not widely available.

Looking to the future.

The Committee acknowledged the progressive emergence of standards and certifications in other salient data-intensive domains, such as data security. Such standards, which were described as the "currency of trust", were discussed as a useful reference in producing evidence-based decision-making for AI applications in the law. The Committee considered that such standards could be viewed as the product of a four-step process that has typically accompanies technological innovation:

1. As a new technological field appears, a wide range of practices emerge in industry (as has been observed in AI and the law);
2. Some of those practices become "best practices", in that they meet certain desirable characteristics, which drive both the endorsement by a wide range of stakeholders and their adoption in practice;
3. Some of those are formulated into standards, which in turn enable enforcement (through, at first, private contracts, or professional codes, or regulation);
4. Some of those standards eventually enable performance of tasks in sufficiently scalable, reliable, and predictable ways to result in informed trust, with a varying range of enforcement mechanisms, including market-driven incentives and regulation.

At present, AI's use in the law generally remains at the first or, in rare cases, second step of this process.

The Committee next considered instruments, including regulatory sandboxes, that could help ensure the trustworthy adoption of AI in legal systems while addressing certain concerns relating to innovation and intellectual property protection. More generally, to further advance Informed Trust grounded in the operationalization of the Four Principles, the Committee proposed, as part of its agenda, the following policy topics for further examination:

- **Topic 1:** Calibration to use-case: evidence-based decision-making
- **Topic 2:** Pathways to trustworthy adoption of AI: achieving useful standards
- **Topic 3:** Educational initiatives and professional training

Conclusion and adjournment.

The Committee adjourned with:

- A unanimous agreement that sound evidence of the effectiveness of AI-enabled processes was necessary to support the trustworthy adoption of AI in legal systems.
- A recognition that competence of operators of AI systems in the law cannot be taken for granted.
- A recognition that whereas laudable efforts are undertaken to support the education of lawyers with respect to AI, such efforts are still nascent and evidence of the extent to which they can ensure sufficient competence is not yet established.
- An intent to pursue the Committee's work both with respect to the principles of Effectiveness and Competence that were the principal focus of the deliberations, but also with respect to the principles of Accountability and Transparency.

Appendixes.

Material related to the deliberations of the Committee are attached in Appendix

9. From a Data Commons to an AI Commons

Expert Group Leaders: Amir Banifatemi and Don Gossen

Committee co-Chairs: Sarah Pearce, Alpesh Shah, Brent Barron, and Ryan Budish

Key Recommendations:

- Problem owners and the community of AI practitioners should be connected to collectively solve problems.
- Availability of trusted data repositories (data commons) should be provided and access to cloud and compute capabilities centers for problem solving to move forward.
- An "AI safe sandbox" for collaboration should be created—a simple context with established standards for participation and incentives, as well as guidelines for safety, ethical consideration, data privacy, IP ownership, and project governance based on best practices.
- A fair-trade policy around data collection, qualification, and usage should be set up.

The AI Commons aims to bring together the key components for AI: data, compute, storage, interfaces, machine learning algorithms and talent, into a single platform. The objective is to connect problem owners with AI capabilities to address major global challenges including progress towards the UN Sustainable Development Goals. This involves bringing together diverse stakeholders to pool resources into a single platform that can be used to scale up use of 'AI for Everyone' and 'AI for Good'.

The "AI Commons" furthers the idea of "Data Commons", which seeks to aggregate government, private sector, and individual users' data into accessible and trusted data marketplaces. The AI Commons greatly widens the capacity of all Data Commons to serve as a platform for collaboration.

Four subcommittees came together to discuss the topic of Data Commons to AI Commons:

- **Data Commons vs. AI Commons**
- **Relevant Framework & Methodologies for Open Initiatives**
- **Building the AI Commons**
- **Deploying the AI Commons**

Defining the AI Commons

Firstly, the definition of an AI Commons must be considered, especially in terms of how it differs from the "Data Commons." It is important to identify the main commonalities and differences and to avoid misappropriation and ambiguity between the two concepts.

Expanding the number of resources from data to include compute and machine learning algorithms and expertise raises new challenges and requirements.

An initial question is the relevance of considering each of the components (data, compute, machine learning algorithms and expertise) separately during planning before aggregating them into the same platform. Alternatively, is it more effective to begin with the entire platform in mind and tackle the technical, operational and societal challenges each brings holistically? Similarly, is it effective to consider the components' unique requirements separately, or their commonalities? How intertwined are the components in terms of the requirements to establish the holistic AI Commons?

Experts highlighted that a Data Commons is a curated data repository, organized by topic, community or interest, that is usable for AI models and is accessible to anyone to use towards the common good. It is not "open data" because data is limited to certain sectors and industries, and mainly takes form as meta-data. **The key criteria of the Data Commons include maintaining licenses, understanding the quality of data, managing sources of data, ethical considerations, and consensus on what should and should not enter the commons.**

Meanwhile, for an AI Commons to be implemented, a range of questions need to be answered, such as:

- What do we want to achieve with the AI Commons?
- Will it be leveraged for collaboration or competition?
- Will it be a collaboration tool or a management tool?
- What type of rules do we want?
- At which geographic level should it stand to make sure all can use AI to solve their problems, by municipality, locally, or globally?
- What are its core values?

Participants suggested the first two practical steps to scoping an AI Commons: first, to pick specific use-cases for the AI Commons, undertake a risk assessment, and to test and iterate on proposed solutions. Second, actors need to identify the key stakeholders involved, including governments, businesses, academia and individuals.

Participants pointed out that it is important to include those who will be directly impacted by the use of these commons and technologies. This includes those who would not necessarily consider themselves stakeholders, such as civil society and local communities whose data will be leveraged and who will be affected by the modelling and applied prescriptions made in the Commons.

Data management of an AI Commons

GGAR participants agreed that it is best to have one organising entity responsible for ensuring that data is provided in a way that works for all stakeholders. Before utilizing the data, there is still some work required to standardize data and data exchange to provide a framework and common ground. These associated standards can pertain to what data should be shared and how it is cleaned and labelled. AI Commons initiators must establish concrete examples of types of problems that will be addressed by the data collected and farmed. This can help establish metrics and benchmarks for goals to be met by the AI Commons.

There was a general consensus among participants that the most practical way to start an AI Commons is to identify a narrow, specific use case as a pilot and to scale up the Commons over time. An initial pilot project mitigates concerns that may be held by companies regarding loss of intellectual property. How can users guarantee that companies do not lose their competitive advantage by sharing? **One proposed solution GGAR participants aired was to set up a fair-trade policy around data collection, qualification, and usage. An AI Commons could be inspired by Fair Trade programs, which connect disadvantaged farmers and workers with consumers, or Green Data policies.**

Participants also questioned this central dilemma: **will the AI Commons generally allow for ownership of data storage facilities, or ownership over the data itself?** For example, the U.S. provides that data is not copyrightable unless it refers to a trade secret and is thus part of firm intellectual property. These contexts need to be examined before establishing benchmarks on the types of data that should be shared by different stakeholders. One viable approach is that all stakeholders work together to label the data and to share the labelling in a systematic manner. First, there must be the setting-up of boundaries, subject and domain area. Second, there will be the setup of inter-commons communication.

Next, efforts are needed to ensure that the AI Commons is supported by representative datasets. Inclusion of geographic and demographic diversity will be important for the legitimacy of an AI Commons. **Measures must also be taken to avoid skewed or biased datasets which may arise if there is selection bias as to which individuals, private or public actors choose to upload or sell their data to the Commons.** For example, if the value of individual data is low, low-income individuals may have greater incentives to incorporate their data. A major risk that also exists is the under-representation of "data poor" regions that are less digitally connected.

Characteristics of an AI Commons platform

There are many good open source commons which can be used as templates to build an AI Commons. Wikipedia is a platform where the public contributes information and is incentivized to ensure information is accurate. Despite the increasingly decentralized nature of Wikipedia, its **notable accuracy in substance and successful common-based management is due to its governance structure, which is based on Elinor Ostrom's eight design principles**. These principles are for self-organizing communities that manage natural and valuable resources. Participants believed that similar principles should be incorporated into the construction of an AI Commons. This is a "Commons as a Common Resource" approach, offering a decentralized model for the AI Commons. **While such an approach could be more efficient, GGAR participants were concerned that a distributed approach would soon run up against limits of existing central processing units and internet speeds available in most countries and regions.**

Discussants also remarked that there need to be good governance models for openness, which will ensure that the data and AI are deployed for intended, limited and well-defined purposes. The committee discussion pointed to examples such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, an association with widespread credibility and multi-stakeholder input; and the Wi-Fi Alliance, which was formed to safeguard the interests of its constituents through standards and tough specifications.

Participants also questioned whether the AI Commons should have a reward, or penalty-based, structure imposed. There are several different governance systems that can be studied to inform this. For Kaggle, an online community where data scientists and machines learners can find and publish datasets and other activities, users are rewarded for acting well on the platform. For Wikipedia, there are penalties and the removal of rewards for malicious use.

10. International Panel on AI: International Cooperation on the Global Governance of AI

Expert Group Leader: Francesca Rossi

Committee co-Chairs: Arisa Ema, Anne Carblanc, Raja Chatila

Key Recommendations:

- If instituted, an IPAI would need to consider both scientific and socio-economic aspects of AI to be truly effective. Proponents of IPAI need to consider carefully how scientific rigor is maintained if the socio-economic effects of AI are also being analysed.
- An IPAI would need to expand beyond the traditional state-based model of the IPCC. A multi-stakeholder approach would be able to adequately reflect both private sector dominance in the AI space and the networked, extraterritorial realities of AI innovation and use.
- Proponents of an IPAI must seriously consider the impacts of using the G7 as the institution for giving initial momentum to this institution. If the G7 is used, attention must be paid as to how other major AI powers like China are included as prominent actors at the earliest opportunity.

December 2018 saw the French and Canadian governments propose an “International Panel on AI” (IPAI), a new organisation which would be modelled on the existing Intergovernmental Panel on Climate Change (IPCC). The IPCC is the intergovernmental body of the United Nations dedicated to providing objective, scientific pronouncements on climate change. The IPCC also provides analysis of the natural, political and economic impacts of climate change, and possible response options.

IPAI remains at the scoping and proposal stages, but planning and dialogue between major state powers continues apace. Undoubtedly, the formation and institutionalisation of such a body faces an uphill battle, as it requires a mandate, defined goals and objectives, an institutional home (e.g., the UN), personnel, funding, membership, global buy-in, and considerable political will.

Four subcommittees met to discuss the proposals for an IPAI, namely:

- **Mapping and lessons from the IPCC and other intergovernmental organisations**
- **IPAI Objectives & Approaches**
- **Membership of IPAI**
- **Designing and Global Governance of AI Framework**

A Commitment to Science

Many analysts consider the IPCC to be the international standard for scientific facts on climate change. This is arguably its hallmark achievement. The body has critics, but it is the most common and authoritative reference point for governments across the globe to understand the effects of climate change.

Experts quickly realized how modelling an equivalent of the IPCC for AI implies that such a body would aspire to become the international reference point for facts about AI and associated technologies. Much discussion at GGAR focused on how to build an IPAI in a way that would help fulfil this objective.

Institutionalizing a body that deals in the creation and dissemination of facts is hardly a non-partisan exercise, however. Scholars of science and technology studies, including experts represented at the GGAR subcommittees, have shown how the natural order ("science") and the social order ("society") are not two separate spheres, but are "co-produced" through intertwined intellectual and social processes. This does not mean that there are no objective facts about AI to be analysed. But the unavoidable overlaps between science and science communication means recognizing that an IPAI would have an intrinsically political task at hand. **Being an institution that is in the business of generating facts about AI would also involve treading the difficult line of existing as a "scientific" body that also navigates political and social mandates.**

Given this, subcommittee members identified a central crux in the design of an IPAI: what kinds of data and information might it be qualified and mandated to assess? The IPCC, which has itself had difficulties limiting its scope of work, is ultimately measuring one thing above all else: temperature change over time. With AI innovation, there is not a single "thing" to be measured. Understanding developments in AI, for example in machine learning, is difficult to separate from the wider exploration of how those tools are applied in "real world" applications. If an IPAI is tasked with assessing the likely impacts of AI across the world, that involves understanding how AI would be applied in diverse new contexts. Moreover, as participants explored, for an AI to innovate involves countless interactions between raw technology on the one hand, and the interaction of that technology with users, on the other. **Measuring and utilising AI is a more social phenomenon than climate science. That makes it more difficult to isolate what datasets could to be used for scientific and/or policy analysis by an IPAI.**

Participants generally agreed that it would be difficult to limit what kinds of data it would and would not be IPAI's role to analyse. Some experts argued that for an IPAI to be a worthwhile institution, it should go beyond the most contained mission of being a body generating facts, and should also operate to identify and understand the trajectories, developments, standout cases, and inhibiting factors of AI innovation. **There was general agreement that the IPAI should at least seek to analyse what governance policies and conditions would maximize the benefits of AI, while minimising its downsides and risks.**

A Feel for the Facts

A majority of GGAR committee members envisioned IPAI as an institution assessing the state of AI science, but also one that builds consensus concerning knowledge about AI more broadly (e.g. contributions from the fields of sociology and philosophy). This mandate would be wider than that of the IPCC, which aims to build consensus about climate science, and whose contributors are almost entirely from the scientific community. **Committee members saw a role for IPAI in breaking down silos between different forms of knowledge, for example research in the social sciences and humanities concerning AI.** In fact, some experts were uncomfortable with the reference point of the IPCC for IPAI's institutional design, because they believed it would have to be much more generalist in its scope of work and analysis.

In this more comprehensive vision of fact-making and reporting, the IPAI would find it difficult to avoid conducting qualitative research, and making judgements about how AI is and should be used. For example, if an IPAI sought to understand the impact of AI innovation on the international jobs market, that would inevitably involve some kind of qualitative social scientific work like surveys and consultations, and perhaps judgements about where there are skills shortages or a pressing need for policy innovation. This is a contentious course to chart since it would need to achieve both representation and a plurality of perspectives.

Pluralism and its Perils

The consensus among participants was that an IPAI, unlike the IPCC, could not rely on a state-centric model, with only states as members. **It was felt that a state-centric approach would not properly reflect the networked, extraterritorial realities of AI innovation and use.** Also, Big Tech companies have revenues larger than many states, and to exclude them could be detrimental to accessing quality data and institutional credibility. A multi-stakeholder approach also allows for representation from actors such as global trade unions and non-profit organisations, which hold key information about AI's impact on society and the economy, and have claims to represent voices which may otherwise be marginalised. One viewpoint from GGAR expert was that states could be members, while non-state actors could be participants or observers to the IPAI. It seems that there are pros and cons to either approach, and that much depends on what kind of institution the IPAI aspires to become.

Committee members reflected that a multi-stakeholder approach is not necessarily a panacea to improving legitimacy and credibility. **Some pointed out that diversifying membership and representation to include non-scientists, especially in the AI industry, would put IPAI at risk of regulatory capture.**

Creating Credibility

How would an IPAI be instituted so as to gain international recognition and credibility? Our reference point, the IPCC, was formed as an intergovernmental panel under the remit of the United Nations Environment Programme and the World Meteorological Organization;

membership is open to all members of the UN and WMO. There was concern from subcommittee experts that there is not, at present, the multilateral appetite to institute an IPAI. **Experts also argued that it would be difficult for such an organisation to gain legitimacy if it did not include major AI powers like Japan and China from the outset.**

In that vein, it was observed that an IPAI may struggle to attract global buy-in if the G7 countries use their membership of that elite body to make initial decisions on IPAI's institutional design. The G7 consists of Western countries only, save for Japan: could full-throated support from China really be expected if the IPAI is dominated by Western actors?

Participants discussed various ways in which these concerns about legitimacy and fairness might be mitigated. For example, extending membership or at least involvement to non-state actors in the AI sector could prevent naked state-based competition. Also, participants explored how the IPAI could evolve over time, to ensure both pragmatism and nimbleness. Initial scoping by the G7 should be merely an interim step before full institutionalisation, at which point other global actors should become involved in design and growth.

Even with these measures, proponents of an IPAI face an uphill battle to balance the requirement for strong leadership and coherent vision on the one hand, and the pressing need for legitimacy from a pluralistic approach, on the other. **On a very practical level, founding membership and participation could give specific state actors a first-mover advantage.** However, unless there is a viable convening framework and mission that would incentivise expansion from the outset, it becomes difficult for G7 leadership to build a genuinely representative international science-making body.

11. AI for SDGs

Expert Group Leader: Cyrus Hodes

Committee co-Chairs: Baroness Beeban Kidron, Elizabeth Gibbons, and David Jensen

Key Recommendations:

- In order for AI companies to work towards these goals, incentives for SDG compliance must be created and embed understanding and know-how into corporate culture for the appropriate use of AI.
- The various stakeholder groups relevant to AI should be aligned as to how we appraise others and choose which qualities and skills we value for the future in a collective manner.
- Synthetic data should be created to address the stark disparity in the amount of data available across geographies, stunting the acceleration of the SDGs, which will be in turn sandboxed to then train algorithms.
- Infrastructural lags and data shortages before the application of AI tools should be identified to mitigate negative effects that contradict the "leave no-one behind" maxim of the SDGs.
- The trajectories and dependencies within the Sustainable Development Goals should be mapped to mitigate negative effects specific to AI.
- A matrix of vertical and horizontal AI applications for the SDGs should be created to map the chain reaction and influences of advancing one SDG with respect to another.
- The purpose of AI should be identified before establishing where the use of AI is necessary and beneficial to the promotion of the Sustainable Development Goals.

AI can be understood as a general-purpose technology, and in that sense, it holds the potential to transform many of our current global challenges. It can help the world achieve the United Nations' Sustainable Development Goals (SDGs), which aim to address a wide variety of global challenges faced by humanity such as poverty, climate change, human rights abuses, and inequality. The SDG framework is built upon the key objectives of economic development, societal stability, and supporting the Earth's ecosystem over the long-term. The objective of this committee was to understand how we can practically forge collaborations to deploy AI to advance the SDGs, how to do this in a safe and ethical manner, and what use-cases we can collectively devise for the specific areas of education, healthcare and climate change.

The subcommittees that convened to discuss "AI for SDGs" were:

A. **Preparing to Apply AI for SDGs**
B. **Use-cases and Frameworks for Education**
C. **Use-cases and Frameworks for Health**

D. Use-cases and Frameworks for Climate Change & Urban Development

Agenda-setting AI4SDGs

In 2019, the United Nations Environment Programme (UNEP) published a report entitled *Measuring Progress: Towards achieving the environmental dimension of the SDGs*. Focusing on the SDGs related to the environment, the report's recommendations nonetheless lay out starkly the scale of the challenges ahead for the SDGs more generally. The report reads:

> *"Of the 93 environment-related SDGs indicators, there are 22 (23 per cent) for which good progress has been made over the last 15 years. If this progress continues, it is likely that these SDGs targets will be met. However, for the other 77 per cent of the environment-related SDGs indicators, there is either not sufficient data to assess progress (68 per cent) or it is unlikely that the target will be met without upscaling action (9 per cent)."*
>
> ***UN Environment Programme, 2019***

A lack of data or stagnating progress is a concern for achieving many of the SDGs. Given the scale of the challenges ahead, discussing the relevance that AI has for the SDGs is important given that:

- **Advances in AI require high quality Big Data procurement and management. Reciprocally, AI innovation can also greatly improve our ability to procure and analyse Big Data, which if handled in the right way will help to progress the SDGs.** In a world where AI is being used more in day-to-day life, there will be increasing ways we can use the Big Data sets that AI is trained from, and the new information generated by AI, to draw insights relevant to the attainment and monitoring of the SDGs.
- **Recent innovations in AI have brought with them profound leaps in the quality of goods and services. The potential to scale those innovations, and achieve new ones in the future, may open up the sufficient momentum required to fulfil the SDGs by 2030.** AI can help us approach timeworn problems associated with sustainable development in new and innovative ways. AI-related innovations in fields from healthcare to water management may help us to achieve specific SDGs.
- **AI systems help us to understand SDG-relevant problems in novel, combinatory ways through analysing probabilities and correlations at massive scale**. Specific SDGs intersect and overlap with one another in a number of ways. Water management, for example, clearly relates to healthcare provision. With the

probabilistic and correlative underpinnings of AI, there is the potential for powerful new capabilities to map interconnections and synergies between different SDGs.

These parallels and interconnections between AI and the SDGs offer a broad indication of what can be learned and innovated in the field of AI to help achieve the SDGs. They also show how we might better procure the kinds of data and analysis that would allow us to understand if, where and how the SDGs are being achieved in near-real time. **Despite it being well-established that the future of the SDGs and the future of AI are intimately related, it often remains the case that ideas about their intertwined future remain inchoate. We are only just beginning to realise the promise and perils that AI brings for sustainable development globally.**

Experts discussed how there remains depressingly limited understanding of the SDGs in private industry, especially with respect to how the aspiration of achieving the SDGs should affect how businesses conduct their affairs in comprehensive and systemic ways. Increasingly, we see that companies have sustainability agendas and wide-ranging policies, programs and reporting mechanisms related to corporate social responsibility. However, if business-led initiatives are directly aspiring towards the commitments embedded in the SDGs (e.g., mitigating climate change), **it remains the case that knowledge about the sub-goals and key performance indicators of each of the seventeen high-level political goals of the SDGs, and the way they relate to one another, is lacking.**

We also need to be clear-eyed to the fact that few, if any, AI companies are proactively pursuing the attainment of SDGs in their work. Even in especially R&D-oriented machine intelligence companies like DeepMind, there is ultimately a bottom line which involves being generally beholden to commercial interests. Even if corporations working in the AI sector recognise the power their technologies carry to change the world, it is rare that their business models allow them to fully maximize the benefits and minimize the downsides of their products in pursuit of sustainable development.

In view of this informational and incentive deficit, **measures to promote norms about SDG attainment as a vital business norm must be created.** Actors positioned to encourage this shift, such as international NGOs, should help to embed understanding and know-how into corporate culture. This principle should be applied in particular to the Big Tech companies, which have such influence over the global supply chain and the disposable capital to make game-changing commitments. It is likely that we will be much better-equipped to meet the SDGs if AI technologies are developed in ways that enable, energize and invigorate sustainable development, including through facilitating an acceleration in goal attainment in specific sectors (e.g. water and sanitation management; education). **It is the aspiration that one day AI companies, with their unique capabilities and capacities for exponential growth and scalability, will see it as integral to their mission to help attain the SDGs, while also being cognizant of and minimizing how their technologies may undermine them.**

Mapping an SDG matrix

IGOs, NGOs and governments responsible for implementing the SDGs must develop a greater understanding of their roles and responsibilities and invest in new research programs and initiatives that would map the interconnectedness of the different SDGs. **Producing a matrix requires intensive collaboration and coordination across institutions and geographies.**

In particular, GGAR participants remarked how, alongside a lack of sustainable development-relevant methodologies, we also do not understand the nuanced, dynamic, systems-based interdependencies among different SDGs. **Datasets which could be mined in novel, combinatory ways remain siloed, either due to legal or ethical constraints or a lack of incentives and inspired vision for actors to collaborate and combine approaches and data.**

Experts discussed the ample potential that AI technologies hold to help us understand the correlations between different datasets. Better harnessing AI systems would allow us to perceive synergies between different SDGs and how policy prescriptions could assist in maximising attainment across a spread of SDGs. However, there exist major obstacles in making this AI-powered, cross-cutting approach a reality. In particular, there is significant variance in the maturity and quality of data verticals in different regions of the globe and a dearth of data readily available to analyse different substrates of development indicators.

In an attempt to resolve these constraints and limitations, participants identified that a matrix of vertical and horizontal AI applications for the SDGs should be created. The intent of this would be to map the chain reaction and influences of advancing one SDG and to assess its impact on others. This matrix would be exceptionally useful and allow us to visualize how attaining certain SDGs can have spill-over effects on others and provide evidence about how governance and policy mechanisms should be designed to responsibly encourage the most positive and high-impact spill-overs. **This sophisticated, data science-based analysis requires both exceptional talent pooling, interdisciplinary collaboration and international governance coordination.**

Challenging AI

While there are countless possibilities for using AI to attain the SDGs, we should also be comprehensive and programmatic in the design and risk analysis phases of any project in this space. As many other GGAR committees identified, there are legal, ethical, risk and resource-based dilemmas that arise out of applying AI, especially when there are inadequate governance frameworks and methods for democratic and civic redress in place.

Aware of these challenges, participants expressed the importance of questioning AI’s added value and the assessment of risks prior to engaging an AI system on SDG-relevant projects. Experts explored the array of qualitative questions which could and should be asked as part of AI systems design for SDGs:

- What is the purpose of integrating AI to assist in this SDG-relevant project? What are the expected upsides and outcomes?
- How is applying AI in this context expected to change outcomes compared with human and non-autonomous ICT tools alone?
- What are the risks involved in applying AI in this case? How have those risks been mitigated, and have the relevant actors who might be affected been consulted about the nature of those risks?

Much of this work is standard practice in any sustainable development project, however, some points seem especially pertinent where AI is integrated. Using AI tools may well lead to different conclusions being reached and the who, what and how of sustainable development will change. One of the primary considerations when utilising AI in an SDG project would be to ask: **what is being overlooked and what are the risks and implications in this specific project for sustainable international development if an aspect of the program has autonomous as opposed to human participation?**

SDGs, AI and Environmental Change & Sustainable Cities

A monitoring framework of 244 indicators has been agreed on for the monitoring of the SDGs, of which a total of 93 have some environmental dimension. Many of the 17 overarching goals also relate in some way to environmental issues, for example the provision of affordable and clean energy (Goal 7) and climate action (Goal 13). Also, Goal 11 is dedicated specifically to building and making sustainable cities.

GGAR experts identified particular aspects of environmental and sustainable cities initiatives that hold especial promise for the integration of AI technologies. These included:

- Monitoring and evaluation of deforestation and land use change;
- Optimizing urban transport;
- Conserving species;
- Monitoring water resources and sea levels;
- Data center energy optimization;
- Disaster risk production; and
- Air quality management.

Participants observed that at present the greatest barrier to the acceleration of AI with respect to environmental protection is the availability, cost, quality and access to data. Inadequate data is especially a problem in the Global South, although there are also thorny issues in developed countries such as the cost and private ownership of data.

SDGs, AI and Education

Quality education is one of the seventeen SDGs: “Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all” (Goal 4).

Experts remarked that there are two primary issues relevant to AI and education:

- AI will lead to the automation of many tasks and come to have a huge effect on the global jobs market. This will affect citizens in all countries, however, there is a particular threat to manufacturing and low-skill service jobs, which make up a greater proportion of total jobs in developing countries than in developed ones. **Unless we educate the world, and particularly citizens of developing countries, for the age of AI, the fallout in labor markets will cause widespread social disruption and exacerbate existing global inequalities.** The effect of this may mean we do not reach education and employment-related SDGs, which will have a compounding effect on all areas of international development.
- AI holds the promise to fundamentally transform education systems across the world. AI technologies can personalize education to each individual student and can assist in the monitoring and assessment of educational programs across the world. AI can create more proactive, rather than reactive, models of education, and the scalability of “edtech” at relatively low costs means that citizens of poorer countries could see vast improvements compared to existing educational provision. **Successfully reforming education for the age of AI will likely require incorporating AI technologies into education systems on a massive scale.** Incorporating those technologies in the humanistic spirit of the SDGs will require that this be done in sensitive, safe and responsible ways, taking into account local culture and nuance.

We need to see education systems evolve for the age of AI. That involves thinking innovatively about education systems—primary, vocational, secondary, post-secondary—which will often **require expanding the traditional parameters of academic performance and instituting more proactive, rather than reactive, models of education.** Those new models of education must also honor the diversity of students and skills to cater to all demographics and work towards an equitable education system.

However, who will dictate how and where AI should be used in education, and what are the most “desirable” or “beneficial” competencies for future jobs in a world of deep AI integration? Such issues remain unsettled, negotiable, and fiercely determined by local contexts, religion, and culture. Fundamentally, the answers to resolution and progress shift by geography and due to legitimately divergent expectations about what the purpose of education is. The scalability of AI may help the “no one left behind” mantra of the SDGs. However, advanced edtech systems will also require significant human capital development (e.g. teacher training) of the kind we spent so much of the twentieth century trying to perfect and massify, but which remain unfinished projects marred by irresolution and inequalities.

12. Safe and Secure: AI and Cybersecurity

Expert Group Leaders: Roman Yampolskiy

Committee co-Chairs: Robert Silvers, Yann Bonnet, Lydia Kostopulos

Key Recommendations

- Governments and security partnerships like NATO should develop a sophisticated conceptual architecture that characterizes what are “cyber-offensive” and “cyber-defensive” capabilities and practices. This will provide actors with a practical and normative framework in which to understand their cyber activities and its impact and perception from others.
- Private companies and governments alike need to build incentives to share technical know-how, weak points and vulnerabilities in cyberspace.
- Stronger formal mechanisms need to be built by governments and intergovernmental organisations for institutions to be equipped and incentivised to report cyber breaches, for the well-being of cyberspace and for empowered protection and awareness of citizen-users.

Four subcommittees met to discuss AI as it applies to cybersecurity:

- **Building Rules and Norms for Industry** (2 sessions);
- **Evaluating Policymakers’ Options**; and
- **Educating and Empowering Users.**

“Users” here included individuals, businesses and governments, who are all currently facing unprecedented cybersecurity challenges. These challenges will morph and in some cases amplify as new AI technologies are developed in the coming years. Correctly applied, AI also holds the potential to enhance user cybersecurity for individuals and institutional users alike.

The goals of this committee were to develop concrete governance recommendations about:

Resilience: Ensuring digital systems, including AI systems, are built to be resilient and robust against cyber threats.

Adaptation: Creating roles, responsibilities and expectations for managing AI in complex cyberattack and cyber-defence systems.

AI for Cybersecurity; Cybersecurity for AI

GGAR experts identified how we need to be careful and considered when discussing the implications of AI for cybersecurity. Not every cybersecurity challenge is an AI challenge, and likewise not every AI challenge is a cybersecurity one. Practitioners saw the need to strictly differentiate these two terms in order to maintain conceptual clarity.

Nonetheless, of course these terms are increasingly linked as AI becomes embedded in all aspects of society and life. **Committee members proposed that a useful way of delineating these terms was to think about "AI for Cybersecurity" on the one hand, and "Cybersecurity for AI" on the other.** The first of these pinpoints the many ways in which AI is and could be applied to enhance cybersecurity. The second identifies the role of the cybersecurity community of practitioners in promoting the safe and secure development and use of AI.

Experts stated that generally there are more beneficial opportunities for using AI technologies in cybersecurity, but that AI more generally brings considerable risks to the safety and good management of cyberspace. AI will soon be able to enhance the identification of cyber breaches and security threats in a way that human-level intelligence could not match. One participant pointed out that if we are candid about the sheer scale and size of cyber challenges for states and private companies, we soon realise that there is no hope for properly responding to them without AI tools. **The amount of cybersecurity procedure and monitoring required is of such scale that only the velocity and magnitude brought by AI tools will be capable of defending cyberspace.**

AI tools hold significant potential for personalizing security recommendations for individual users and situations. **AI-enabled User Behavior Analytics (UBAs) will show how different users might be at threat. This could prove a powerful tool to provide people with self-protective tools.** At the moment, soft-touch "nudge" tools like generic warnings about users not sharing credit card details do not have the kind of impact that we should aspire for. Some participants even felt that the unintended leaking of personal data is arguably a more pressing issue than the security threats posed by using autonomous technologies in cyberspace. It seems that AI, by personalising messaging and security measures to various types of user particular situations, could do much to engage the general population in cyber-safety.

Turning to providers of those services, clearly, cyber firms that manage to develop AI-enabled tools and products capable of improving general cybersecurity have ample opportunity for market growth. The effective use of AI in the cyber industry is becoming a key market differentiator. An additional benefit of AI for cybersecurity would be to act like an "auditor" and "investigator" of the sector, in that AI can help firms and regulators identify the false positives generated from human analysis of cybersecurity. Again, **this shows how AI can help firms improve their products; and holds the potential to enhance the accountability of the cybersecurity sector.**

On the other side of the coin, **participants explored how the use of AI-enabled cybersecurity can create many issues for the controllability of those technologies, as well**

as the liability for firms and states should their use cause damages. As many analysts in the AI space have already identified, just because decisions are automated does not mean that accountability processes should be any less rigorous. Two broad examples of AI leading to a less safe environment are in circumstances of poor system architecture and if an AI is exposed to data poisoning. For criminals, the use of “deep fakes” could cause systemic threats to cybersecurity, since AI has such powerful potential to impersonate voice and image and to mirror how real users operate online.

In that sense, participants identified a key difference between cybersecurity (which entails improving cyber safety and enhancing defensive capabilities) and cyber ops (which recognises that cyberspace is not a naturally defensive space and something that *must be secured*). The concept of “hack backs,” involving the use of AI to counterattack malicious cyber threats, conveys the more realist and pragmatic thinking about AI-enabled cyber ops.

This offense/defence dichotomy is an especially thorny one: some analysts argue that a truly safe cyberspace can only be achieved through some combination of defensive and offensive capabilities. However, there was a remarkable consensus among GGAR participants that “hack back” should generally be illegal. Indeed, participants also acknowledged that “hacking back” creates substantial risks given the difficulty of proving attribution, the prevalence of false flag operations (and thus the risk of retributive operations being targeted at innocent organizations), and concerns around deputizing private parties to engage in aggressive actions that typically have been under the exclusive authority of duly appointed law enforcement agencies. All the same, **developing a sophisticated conceptual architecture for cyber offense vis-à-vis defence remains an important project as actors seek to secure a dominant presence in, and develop governance structures for, cyberspace.**

Assessments and Projections for Cybersecurity and AI

These framing discussions set good ground for understanding what the role and approach of governance actors in the AI and cybersecurity sectors should be. **Experts explored how the absence of multilateral dialogue and growing militarisation in this field is leading to substantial differences in how states and private firms are understanding what “good” behaviour is in cyberspace.** Of course, divergent approaches are to be expected, but as major players begin to consolidate those programmes and policies, differences become starker and begin to stand in contradiction to one another.

For example, France’s November 2018 “Paris Call for Trust and Security in Cyberspace” implored the need for multilateral dialogue between global public and private actors to ensure cyber peace and to avoid the restrictive binary choice emerging between “a Californian Internet and a Chinese Internet.” **A lack of multilateral progress has been the norm for cybersecurity since the UN’s Group of Governmental Experts failed to follow up on its 2015 recognition that international humanitarian law applies in cyberspace.**[7]

[7] https://www.lawfareblog.com/avoiding-world-war-web-paris-call-trust-and-security-cyberspace

As states and global governance actors alike fail to come together to agree on the laws and norms that apply in cyberspace, states in particular now see cyberspace as the emerging frontier of military operations. Given the risks pertaining to "Cybersecurity for AI," and the threats posed to the effective and safe functioning of cyberspace, states are now beginning to pay attention to the amount of resources that must go into protecting their citizens from malign AI-enhanced cyber threats.

Finding Tools for Collaboration in Cyberspace

This absence of multilateralism is hardly new, but GGAR participants gave some reasons why incentives for collaboration are so challenging at the AI-cyber intersect. **Cyberspace still lacks a well-functioning system of incentives for actors to disclose their technical knowhow, weak points and vulnerabilities.** A company or country explaining its cyber system architecture in any technical detail has the potential effect of exposing chinks in the armour. In the private sector especially, there is also a profit incentive for the non-disclosure of best practices given the intellectual property firms protect. Weaknesses in big firms' cybersecurity will often also cross over with those of states. If, for example, Amazon Web Services were to discover a significant vulnerability in some of its servers, it would likely cause as much of a security and reputational challenge for governments as it would other AWS customers.

Experts explained how thus far one of the major challenges for governance has been putting in place some kind of legal and incentive ecosystem that encourages private and public actors to disclose cyber breaches and vulnerabilities. Disclosure responsibilities are being developed, especially since the "WannaCry" cyberattack in May 2017, after which Microsoft called for a "Digital Geneva Convention" that would require governments to stop hoarding cyber vulnerabilities. However, **disclosure and patching duties remain enormous capacity and technical challenges, even for the world's largest and most secure entities.** It remains but an aspiration to have global systems in place that can contain and neutralise zero days attacks. Therefore, if incentives for cyber collaboration are currently absent, innovative governance arrangements need to be put in place that turn actors' fears about vulnerability disclosure into an embrace of universalist and humanitarian responsibilities in cyberspace.

13. Managing the economic & social impacts of the AI revolution

Expert Group Leader: Calum Chace

Committee co-Chairs: Marek Havrda, Priya Lakhni, Irakli Beridze

Key recommendations:

- The implementors of AI technologies should bear in mind that just because these technologies are developed does not mean they will be used, either in predictable ways, or at all.
- We should allow ourselves to imagine how AI technologies could make firms more sustainable and adaptable, and for the benefit of employers and employees alike.
- More research needs to be undertaken to understand anticipated variation in jobs disruption across space and time due to the impact of AI.
- There should be a call for governance actors to build principles that plan for the AI age while also devoting attention to pressing existing socio-economic problems.
- There should be pathways to "lifelong learning" cultures, in which people would learn throughout their career, adapting their skills to what the economy requires and what will best serve their personal development.
- States to encourage economy-wide shifts in human capital development, and not assume that private firms have the capacity to respond to jobs displacement from AI on their own accord.
- A tax levy should be institutionalized for businesses that is ringfenced by governments for the upskilling of workforces; and proposals for an international taxation system that would require companies who decide to automate jobs to provide comprehensive compensation for the associated human job losses.
- It should be explored how to reposition AI as something that is not a commodity from which companies reap profit, but as a resource capable of fostering a more inclusive, sustainable and adaptive economy.

Four subcommittees convened to discuss "Managing the economic and social impacts of the AI revolution". These were:

A. **Radical Disruption**
B. **Gradual Change**
C. **Educating for AI**
D. **Mitigating Rising Inequality**

***Continued uncertainty*:** The socio-economic impacts of the AI revolution will be felt across geographies, sectors, age groups, social classes, and forms of labour. In recent years, there have been a number of reports by research institutions identifying the types and numbers of jobs and tasks that are threatened within different timeframes and under various scenarios of AI innovation. **Despite the considerable research undertaken thus far, it was made abundantly clear during GGAR that there remains an enormous amount of uncertainty and ambiguity surrounding the likely socio-economic impacts of AI.**

Variance in political economy: At least some of this uncertainty relates to how quickly AI technologies will be developed, brought to market, and distributed across societies. Moreover, there will clearly be substantial variation in the responsiveness of labour markets and social policy due to local and regional variances in political economy. ***To put that another way: just because technologies are developed does not mean they will be used, either in predictable ways, or at all.***

Topics under consideration: To reflect the considerable uncertainty over the future pace of innovation in the AI sector, two subcommittees met to consider how the economic and social impacts of the AI revolution would differ based on scenarios of "Radical Disruption" as opposed to "Gradual Change." A further two subcommittees were convened on the topic of managing the economic and social impacts of AI: "Educating for AI" and "Mitigating Rising Inequality."

Of course, there could have been many further topics considered, such as the impact of AI on healthcare and generational or urban-rural divides. The topics of education and economic inequality were chosen because of the widespread recognition that these are two issues requiring urgent policy intervention, over which AI is likely to have particularly dramatic effects, and where there remains ample opportunity for governance and policy intervention to improve outcomes.

Finding Meaning in Work

Jobs displacement due to automation has, clearly, been going on for decades, even centuries, but the intensity of debate has grown with the enhancement of AI. How states and firms choose to respond to shifts in the labour market is culturally and historically conditioned. Each country has distinct traditions in terms of how labour and work intersect with markets and the state. For many advanced and open economies, workers' rights have weakened and been displaced by more free market regimes. The global shift towards what is sometimes characterised as Anglo-Saxon capitalism has, as GGAR experts expressed, been responsible for recent innovations in business culture, which have in turn influenced perceptions about how AI can and should be innovated. Many recent success stories in the AI space have come from companies seeking "radical disruption" or "disruptive innovation." **The cultural and economic effects of this fast-paced innovation have led to considerable fragility within the jobs market already.**

Paradigms and philosophies about work matter. For example, there was some consensus among GGAR experts that our tendency to think of jobs as filling a series of robotic and repetitive functions, rather than as being manifestations of human enterprise and creativity, has led to the current malaise about a jobless future where AI will automate all tasks. In this sense, ideas such as "radical disruption" are contingent to our current paradigms about what and who work is for. If, rather than seeing jobs as mere functions of time, income and cost, we were to have a more "pro-people" philosophy of work, that might lead to new understandings of automation. For example, this could reduce our tendency to associate AI with widespread social fallout. **Instead, we should allow ourselves to imagine how AI technologies could make firms more sustainable and adaptable, and for the benefit of employers and employees alike.** There was general support for this broad idea; but also recognition of the ambitiousness of precipitating this paradigm shift.

Understanding Expectations for Change

Radical Change, Radical Uncertainty

Discussion turned to what kinds of shifts in the jobs market could be anticipated, and when and how AI would influence those transitions. In the "Radical Disruption" subcommittee, participants identified that, **even in such a scenario, AI systems will still not have high-level emotional intelligence within thirty years or less. Even rapid change has limits.** Human capabilities like empathy are often considered a human "last resort" characteristic, and used to exemplify why AI systems will never meaningfully displace people. Nonetheless, even if high level emotional intelligence is not achievable in thirty years, the model of "Radical Disruption" still has enormous implications for the global jobs market.

Approximately four fifths of expert participants on the "Radical Disruption" subcommittee opined that machines would take over a majority of jobs in a thirty-year time horizon (that is, if the technological revolution proceeds at a radical pace). **In that hypothetical future, experts observed that it would be virtually impossible to retrain substantial parts of the population, even in developed countries.**

Gradual Change, Varied Impacts

Meanwhile on the "Gradual Change" subcommittee, participants paid more attention to how AI innovation would be felt in different ways depending on factors such as sector of the economy and whether a country or region is developed or less-developed. A powerful core/periphery dilemma was advanced: **some countries will have particular access to and capability to utilise AI technologies to optimize firm management; but in other countries, AI will disrupt workforces in ways largely beyond firm or state control and which will damage business productivity.**

Generally speaking, places that will be more empowered to streamline and grow are those with dominant knowledge economies of the Global North. Meanwhile, countries of the Global South will be in a less strong position due to the predominance of low-skill manufacturing and service industries, which are more obviously capable of automation. Our knowledge on regional and sectoral difference remains inchoate. Only with greater knowledge can we begin to work out how to improve decision-making capacity and policy planning, particularly in developing countries expected to face significant disruption. **It was observed that much more research needs to be undertaken to understand anticipated variation in jobs disruption across space and time due to the impact of AI.**

There was also a word of caution about focusing too obsessively on future scenarios of social and economic disruption due to AI, because it was felt that that forward planning may come to the detriment of considering socio-economic problems in the here and now. Many countries already face chronic issues of skills shortages and unemployment; and some are focusing on issues surrounding the future of care and robotics in a way that elides the problems for ageing populations that need addressing at present. GGAR participants called for **governance actors to build principles that plan for the AI age while also devoting attention to pressing existing socio-economic problems.**

Education

The "Radical Disruption" subcommittee came to the conclusion that in a future scenario where there are rapid advances in AI, most countries will struggle to retrain their workforces, because we will not know what are the right tools nor have the capacity to implement such major policy program within the very constrained timeframes.

Meanwhile, in the subcommittee operating on the assumption of "Gradual Change," it was still considered imperative for governments and businesses to act quickly and decisively to retrain workforces through comprehensive educational programmes. In both the "Gradual Change" and "Educating for AI" subcommittees, it was remarked that there is already a lot of impetus in advanced economies about the need for "lifelong learning," where people think of educational development as something that they enrol in at multiple stages throughout their lives. **Participants called for pathways to "lifelong learning" cultures, in which people would learn throughout their career, adapting their skills to what the economy requires and what will best serve their personal development.**

However, while lifelong learning as a concept has attracted a deal of attention, GGAR experts pointed out that, while a small number of pet projects exist, the principle is not at all widespread in the global jobs market. The prohibitive cost of advanced education in many countries also creates barriers for re-entry into education in mid-career. As noted above, it was considered that too few employers are interested in creating conditions for a positive learning environment within their firms. This means that much of the global economy is simply not set up in a way where it can respond adequately and in dynamic ways to disruption caused by AI innovations. **Experts called on states to encourage economy-wide shifts in human capital development, and not assume that private firms have the capacity to respond to jobs displacement from AI on their own accord.**

A variety of proposals were put forward to encourage this shift towards "lifelong learning" and facilitate a labour environment well-prepared to adapt to what is required in a time of exponential AI innovation. First, there needs to be much greater resources channelled into lifelong learning in the private and public sectors. This requires much greater coordination between government industry and jobs strategies and employment plans for firms. There were more interventionist proposals, too. **More radical proposals included instituting a tax levy for businesses that is ringfenced by governments for the upskilling of workforces; and proposals for an international taxation system that would require companies who decide to automate jobs to provide comprehensive compensation for the associated human job losses.**

Participants also stressed that educating for AI is not solely about retraining the workforce. Educating for AI requires intervention from a young age. **Proposals were put forward to teach children about data, privacy and technology from kindergarten.** Teaching about such topics should not be considered in isolation from a more holistic approach to education. Children learning about technology should also be learning about their rights and responsibilities, and the ways in which they can use AI for positive ends in society. Participants set out the scope and challenge for the future of education, but they also identified the urgency for change as studies are increasingly alerting us to the negative influences of technology on children's mental health, skills, and social development.

Social Safety Nets

There was almost universal consensus that across vast swathes of the world economy, social safety nets are at present wholly inadequate to mitigate the inevitable socio-economic disruption that will be caused as a result of the AI revolution.

Initial trials of a Universal Basic Income (UBI), for example in Finland in 2017/18, have not led to widespread adoption of similar policies in other countries. The current jobs climate, epitomised for example in the "gig economy" that is facilitated through new technologies, is leading to a hollowing out of workers' rights and employer responsibilities. **Governments seem prepared to act as mere repairmen in situations of market failure, not as institutions capable of delivering more sustainable and adaptable jobs markets.** This model of government must be reconceived for the age of AI. GGAR participants called on governments to put themselves in the driving seat of the AI revolution.

For example, it is the place of government to explore new and much more comprehensive pilot schemes for reskilling workers and/or exploring policies similar to a UBI. Representatives of the people in government are arguably the only actors who can consider what place there should be for redistribution of AI-generated wealth. **That points to a more fundamental question: how can AI be repositioned as something that is not a commodity from which companies reap profit, but as a resource capable of fostering a more inclusive, sustainable and adaptive economy?**

In the West, but also globally, there is a growing sense that a majority of people can now be characterised as the "left-behinds," or "the 99%," anger about which fuels growing populism and demands for systemic income redistribution. The committee that met to discuss the topic of "Mitigating Rising Inequality" lamented the lack of political will to understand how inequality and access to technology are linked; and the as yet unmet need for an international association that could deliver cross-country coordination for initiatives seeking to reduce economic inequalities. Such capacities are likely to become more essential in the face of the hyper-concentration of financial profits reaped by Big Tech firms. In sum, participants across all four subcommittees expressed how the opportunities for governments across the world to take bold initiatives and risk-taking to respond to the AI revolution are plentiful, but that this capacity at present remains concerningly dormant.

14. AI Narratives

Expert Group Leader: Sarah Dillon

Committee co-Chairs: Casper Klynge, Mark Halverson, John P. Sullins

Key Recommendations:

- Governments should inspire the world with goals and opportunities that offer people a genuine place and say in the story of what AI will become – such as John F. Kennedy's speech and policy of sending man to the moon, a good historic example for how to bring people onboard with emerging technologies.
- Policymakers should be aware of the narrative behind the use of AI and either stay true to the narrative, or if it is not beneficial for larger society, create policy that will contain and mitigate the negative effects of its use, while promoting the positives.
- An observatory to monitor different emerging narratives pertaining to AI globally should be built to surface narratives from underrepresented groups of people and ensure that the voices at the decision-making table are not just that of the government and corporates but also of those who are not in the room.
- Narrative-builders could be "Digital Champions," who are appointed to help promote the benefits of an inclusive digital society and act locally with citizens, communities, businesses, governments and academia.
- A human-centric development of AI should be prioritized, instead of focusing on making AI trustworthy to the public.
- Governments should establish a state office or department of AI similarly to the UAE in order to indicate that they are thinking seriously about the AI revolution at the highest levels.
- The narrative of technology-related legislation should to shift from a corporate focus to one that is more citizen-centric.

The definition and trajectory of Artificial Intelligence (AI) is deeply embedded in prevailing narratives and imaginaries of technology.

> **Narrative**: "narrative texts, images, spectacles, events; cultural artefacts that tell a "story"."
>
> **Socio-technical imaginaries:** Collective, public and institutionally stabilized visions of possible futures, which are driven by shared understandings of social life and order. These common understandings are attained through advances in sciences and technology. An imaginary frames a projection, symbol and associated belief about a technology, not only in an individual's mind but also across peoples and society. Such a framework is useful in analyzing and accounting for power dynamics and issuing ethical and inclusive policy.

AI, as with all past technologies, does not operate in a vacuum that is separate or circumvents societal perspectives, values and influence. The rapid development and deployment of AI technologies makes it a powerful asset, or a weapon, and how we frame the conversation matters. AI can be understood as the result of a complex sociotechnical system, whereby, science, technology, and society are engaged in a continuous and evolving cycle of "co-production". New technologies continuously redefine societal values and thus policies. Changes in values and policy shape developments in AI. A dynamic understanding of our collective visions serves as a key anchor to holistically evaluating how AI will impact humanity.

The trust that citizens put in AI is also influenced by such narratives: shifts in our human values influences policy, which impacts how we manage developments in AI. There exists significant heterogeneity in perception and trust of AI technologies across the world: in some regions, AI is seen as an opportunity, while in others it is perceived with significant scepticism and fear.

To effectively govern AI technologies and their impact - so as to maximize the innovation upsides and minimize the downside risks - a careful observation and critical understanding of global technology narratives is needed. Four subcommittees convened to discuss the topic of "AI Narratives", namely: "AI Narratives: Underrepresented Narratives", "AI Narratives: A tool for policymakers", "A/IS Infrastructure and Ecosystem", and "Building Trust in AI Systems".

Influences of AI narratives

The tendency with AI narratives is that those that are predominant and most wide-reaching are those influenced by the rise of Big Tech companies in the West. Through under-regulated processes and power imbalances, we see those interests overpower the voices of disenfranchised communities. Those Big Tech interests also try to speak on behalf of disenfranchised communities without fully understanding their position or offering them a genuine stakeholding. Since these companies were the first to acquire AI technologies for the commercial market, their ideas about what technology's role in life and society should be has been dominant and influenced practice.

It was highlighted that there is an overarching global divide in AI narratives. In the East, the priority is to use AI for the population in general, helping to overcome poverty, improving quality of life, and offering more efficient governance. This often takes priority over questions of individual privacy. In the West, the AI arms race and individual safeguards seems to be the trend demonstrating that capitalist grown societies do not have a built-in social fabric to keep supporting one another kin financially for long.

There are positive and negative AI narratives, however the tendency is to move towards a negative, dystopian view of AI. Science-fiction and entertainment are a huge drive for public opinion on AI, such as the highly popular "Black Mirror" television series, still in production since its launch in 2011. AI in media such as entertainment, business, and news is typically perceived with a dystopian lens. It imposes a reductionist, deterministic and simplified perspective on AI.

Participants suggested that governments should inspire the world with goals and opportunities that offer people a genuine place and say in the story of what AI will become. John F. Kennedy's speech and policy of sending man to the moon is a good historic example for how to bring people onboard with emerging technologies. His speeches inspired and rallied the American people behind the space race.

Narratives in policymaking

Recognising "narratives as policy" and "policy as narratives" is a way to imbue the voice of the population into national policymaking. It was thus suggested that policies should become the new narrative to avoid negative pathways for the control and development of AI. This points to how the AI revolution should not only be seen as a technology and economic revolution but also a revolution of politics and citizenry, shifting policymaking into the joint role of public institutions and citizens. Voices and perspectives from the general population can be shared to public policymaking institutions via AI tools and apps, collecting data on local issues for social progress, innovation and growth. In this way, AI can become a powerful tool in the public policymaking process. However, it is important to bear in mind the negatives of data farming to discover local populations' opinions, since issues of privacy and transparency could be at stake.

Participants highlighted that in policymaking we must be aware of the narrative behind the use of AI and either stay true to the narrative, or if it is not beneficial for larger society, create policy that will contain and mitigate the negative effects of its use, while promoting the positives. This calls for a move beyond the existing narratives and to make new narratives for what we want from AI in the future. However, first, questions must be answered such as who has the power and is most well-equipped to drive these narratives and make such policy?

It was noted that underrepresented AI narratives are the most optimistic and should be taken as an example. Typically, those that stem from Global South and low-income countries are founded on a problems-solving attitude as they understand and have more pressures to take advantage of opportunities provided by AI tools to achieve their public interests. On the flip side, this perspective could be another culturally conditioned narrative and not reflect on-the-ground realities. This is a perfect example of how narratives must be fully understood before implementing policies and before understanding potentially detrimental effects.

Policies must be designed to understand different demographics that are affected within jurisdictions. Governments may need to be assisted in using AI narratives to develop new policies. The process of doing so would be to surface the narrative, nurture it, and spread it.

This can be done through the assistance of an observatory to monitor different emerging narratives pertaining to AI globally. Such an observatory will surface narratives from underrepresented groups of people and ensure that the voices at the decision-making table are not just that of the government and corporates but also of those who are not in the room. **An institution like this will build more inclusive processes of implementing AI narratives, supplemented by policies which engage civil society and those mostly vulnerable and affected by such policies.**

Positive narratives pertaining to AI must be delivered by people we trust. The right actors are unlikely to be governments or corporates. **Many citizens would dismiss industry actors, as they are generally incentivized to speak positively of AI, since this is their profit-making prerogative**. The same goes for government: in certain geographies, people do not quite trust their government. **It was proposed that narrative-builders could be "Digital Champions,"** who are appointed to help promote the benefits of an inclusive digital society and act locally with citizens, communities, businesses, governments and academia.

<u>**Understanding trust in AI**</u>

Firstly, the committee remarked that it is difficult to provide concrete recommendations pertaining to trust as it is an evolving concept that changes over generations and often by demographic (e.g. age, race, locality).

Participants emphasised that we must persist with a human-centric development of AI, instead of focusing on making AI trustworthy to the public. Put simply, changing AI development should come before trying to change human development. In this vein, there should the development of risks strategies for over-trusting machines. These strategies can be supplemented by an understanding of the different variances of human-level trust relative to machine-level reliability.

An important aspect of AI design is to understand the different relationships involved in AI and the components of trust, such as human-to-human, machine-to-machine, human-to-machine, and newly, the self-to-self relationship. Intergenerational and geographical differences change how far we trust in technology. It is therefore essential to educate policymakers on several aspects of trust:

- What is trust? (including psychological, cultural, and philosophical dimensions)
- How is trust, especially of technologies, built and cultivated? (including neuroscience, values, norms, societal influences)
- What is the difference between trust and reliability in terms of human-human and human-machine relations?

Participants highlighted that one mechanism of trust is regulation, with GDPR being a prime example. However, this particular piece of legislation is not about *personal privacy* but *data protection* and its unauthorised use, particularly by corporations. **A case was made that the**

narrative of technology-related legislation needs to shift from a corporate focus to one that is more citizen-centric.

The discussion moved on to *"how to articulate narratives?"*, which is crucial because narratives are built in ways that are understood and digested by recipient actors. The following process was proposed to frame an AI narrative: positioning AI in a certain light that accommodates that of the recipient's mentality, create an emotional story around it, and package it for educational purposes. Participants suggested to base communication on Science Communication, which entails telling a story that is connected to science with a focus on the result of the science, while capturing some intuition from the science. The science should have an impact that the public care about and that is truthful, but one must mindful of who does the communicating and how is authority made in this specific case of Science Communication.

Stakeholders for the positive narrative of AI

It was pointed out that partnerships must be carefully thought through, especially in light of data access and public opinion. In the case of the partnership between Palantir, a partially CIA-funded software company, and the United Nations World Food Programme (WFP), there has been a backlash and many concerns about lack of transparency around the process and terms of the agreement, but also risks pertaining to the models that are extracted from WFP's data, the software that is used, and biases within models that will be applied to the data. **In similar humanitarian-private partnerships, there are always tensions in terms of the extent of transparency to demonstrate that responsible data practices are in place. This can perpetuate the negative point of view that corporations are not considering the public's concerns and calls for businesses to demonstrate integrity**.

However, positive partnerships can be forged between government and private sector companies. Such partnerships should have a strong element of public accountability. **A product of this partnership can be the development of safe and measurable regulatory sandboxes to show potential uses and benefits of AI, while also mitigating large scale risks.** There needs to be greater urgency for both government and industry to work together, especially in establishing codes, certifications, and other mechanisms towards building ethical AI, such as through regulatory sandboxes.

Government must demonstrate competence in AI matters, and be able to convey risks and opportunities adequately. **It was proposed that governments should establish a state office or department of AI similarly to the UAE in order to indicate that they are thinking seriously about the AI revolution at the highest levels.**

Using a narrative/use-case to practically solve the issues around AI

The subcommittee on "A/IS Infrastructure and Ecosystem" focused on a use-case driven approach to identify infrastructure that will accelerate the safe adoption of automated and intelligent systems. The use-case was Drone Taxi Services:

The horseless carriage would never have had the global impact it has done without the creation of infrastructure such as roads, traffic lights/signs, laws and enforcement, etc. In this session we will draw corollaries to Drone Taxi Services to highlight infrastructure and ecosystem requirements for safe adoption.

This subcommittee identified the following:

Infrastructural challenges: One of the fundamental infrastructure challenges would be airspace traffic control. Would it be possible to have roads in traditional airspace? How is it possible to control a crowded airspace? Participants suggested that in the U.S. the Federal Aviation Administration could be an example of an institution that might manage drone airspace traffic control. The FAA and similar agencies in other major economies may at present lack the relevant technical expertise, but it has a mandate, history and a lot of other institutional knowledge to build from.

Regulating drones: There needs to be a systematic way of thinking about the control of drones, as ownership and use increases. This may involve establishing rules to limit the number or regulate the use of drones, phases of control as numbers grow, and methods of dealing with risks and accidents, such as if drones collide.

Use existing templates: An example that could be useful in the case of drone control comes from India. The Civil Aviation Regulation introduced Digital Sky, a technology platform that handles the entire process of regulating the registration and permissions for all Remotely Piloted Aircraft Systems. This was designed by various private and government actors over a fifteen-month period.

Participants

On February 10, 2019, 250 of the world's leading thinkers in AI and AI governance gathered in Dubai for the 2nd edition of the Global Governance of AI Roundtable. In keeping with the Roundtable's broad-based approach, participants represented a wide range of countries, professional interests, and perspectives. Organizations represented included think-tanks and non-profit organizations, academic institutions, government agencies, international and multinational organizations, and private-sector enterprises. The Future Society identified and invited the following experts based on its network.

WORKING GROUP RAPPORTEURS

- These individuals were chosen because of their relevant expertise pertaining to their allocated Working Group.
- Consolidated the findings of the committees' discussions within the Working Group and presented them to all the participants and the AI Minister at the end of the day.

Working Group 1 – Mapping the Rise of AI and its Governance

Amandeep Gill

Amandeep Singh Gill, Executive Director, Secretariat of the High-Level Panel on Digital Cooperation, is India's Ambassador and Permanent Representative to the Conference on Disarmament in Geneva. He joined the Indian Foreign Service in 1992. Amandeep Gill is currently Chair of the Group of Governmental Experts of the Convention on Certain Conventional Weapons (CCW) on emerging technologies in the area of lethal autonomous weapon systems. He serves on the UN Secretary General's Advisory Board on Disarmament Matters.

Working Group 2 – Governing the Rise of AI in Different Contexts

Eva Kaili

Member of the Group of the Progressive Alliance of Socialists and Democrats in the European Parliament and Chair of the European

Parliament's Science and Technology Options Assessment body (STOA). Previously, elected twice in the Greek Parliament with PanHellenic Socialist Movement (PASOK).

Working Group 3 – AI for SDGs and Intergovernmental Panel on AI

Konstontinos Karachalios

As managing director of the IEEE Standards Association and a member of the IEEE Management Council, he has been enhancing IEEE efforts in global standards development in strategic emerging technology fields, through technical excellence of staff, expansion of global presence and activities and emphasis on inclusiveness and good governance, including reform of the IEEE standards-related patent policy.

As member of the IEEE Management Council, he championed expansion of IEEE influence in key techno-political areas, including consideration of social and ethical implications of technology, according to the IEEE mission to advance technology for humanity. Results have been rapid in coming and profound; IEEE is becoming the place to go for debating and building consensus on issues such as a trustworthy and inclusive Internet and ethics in design of autonomous systems.

Working Group 4 – Making the AI Revolution Work for Everyone

Anne Carblanc

Head of the Digital Economy Policy Division (DEP) in the OECD Directorate for Science, Technology and Innovation. Her division develops policy frameworks to foster digital transformation and make it work for the economy and society. DEP serves the Committees on Digital Economy Policy and Consumer Policy and their Working Parties, which are composed of delegations from member and partner countries, and from business, civil society, trade-unions and the Internet technical communities. Ms Carblanc joined the OECD in 1997. Prior to joining the OECD, she was Secretary General, Director of Services in the French data protection authority (Commission Nationale de l'informatique et des libertés - CNIL). She has also served ten years in the French judicial system, both as a judge in charge of criminal investigations and as the head of the criminal law department in the Ministry of Justice.

EXPERT GROUP LEADERS

- Leading up to the GGAR in February, Expert Group Leaders led participants of their committee topics in brainstorming key issues through two calls.
- The aim of these calls was to prepare participants and ensure that they did not start discussions at GGAR from scratch or empty handed.
    - The first calls were an opportunity for participants to contribute their perspectives on the discussion topic. The aim is to facilitate the garnering of key insights and salient points from different expert fields which may be incorporated into a paper published on the GGAR website.
    - The second call focused on action-orientated preparation for the discussions taking place at the 2019 GGAR. A product that came out of these calls were summary notes that consisted of key points made on the calls and to fill in information that might be needed to continue the discussion on the GGAR day.

1. **Mapping AI Technological Development and Future Trajectories**

Anima Anandkumar

Led the subcommittee on “Mapping AI Technological Development: Key Indicators” and is a Bren professor at Caltech CMS department and a director of machine learning research at NVIDIA. Her research spans both theoretical and practical aspects of machine learning. In particular, she has spearheaded research in tensor-algebraic methods, large-scale learning, deep learning, probabilistic models, and non-convex optimization.

Anima is the recipient of several awards such as the Alfred. P. Sloan Fellowship, NSF Career Award, Young investigator awards from the Air Force and Army research offices, Faculty fellowships from Microsoft, Google and Adobe, and several best paper awards. She is the youngest named professor at Caltech, the highest honor bestowed to an individual faculty. She is part of the World Economic Forum's Expert Network consisting of leading experts from academia, business, government, and the media. She has been featured in documentaries by PBS, KPCC, wired magazine, and in articles by MIT Technology review, Forbes, Yourstory, O’Reilly media, and so on.

2. **Geopolitics of AI**

Nicolas Miailhe

President and Co-Founder of The Future Society and led the subcommittee on “Digital Empires”. A recognized strategist and thought-leader, Nicolas advises multinationals, governments and international organizations. He has over fifteen years of professional experience working at the nexus of innovation, technology, government, industry and civil society across Europe, America and Asia. Nicolas teaches at the Paris School of International Affairs (Sciences Po), is a Senior Visiting Research Fellow with the *Program on Science, Technology and Society* at Harvard, and a Fellow with the *Center for the Governance of Change* at IE Business School in Madrid. Nicolas is also the co-founder of *YesEuropeLab*, a pan-European civic innovation & entrepreneurship lab; and of *Aletheion*, a Paris-based startup which harnesses the power of AI for cognitive cyber-security. He is a member of three Committees (Policy, Economics, and General Principles) of the IEEE *Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems*.

3. Agile Governance

Andre Loesekrug-Pietri

Andre led the subcommittee on “Agile Governance: Multi-stakeholder Guidebook for Ethical and Safe AI”. He is is the founder of ACAPITAL and of the European Security Circle. HESTIA is a new fund by ACAPITAL that focuses on building European Innovation Champions in Security Technologies. 100% DeepTech : Cybersecurity, Artificial Intelligence, Big Data, IoT, Robotics, Drones are core themes. The European Security Circle is a ThinkTank bringing together the most important R&D centers in Europe, former Vice Chiefs of Defence of Germany and France, and large industrialists to identify the Security technologies that will be critical for the next 5-10 years for Europe. Formerly, Andre was executive assistant of the CEO of Aerospatiale Airbus, in charge of the investment arm of several industrial family offices and has 15 years of venture capital and private equity experience in Europe and Asia. He is Colonel with the French Air Force People’s Reserve, has both German and French nationalities and is very frequently invited to speak on topics linked with technology in Europe as well as on Europe-China relations.

4. Explainable & Interpretable AI

Nozha Boujemaa

Research Director at Inria, Director of DATAIA Institute (Data Sciences, Intelligence & Society), Project leader of TransAlgo scientific platform for algorithmic systems transparency and accountability. Knight of the National Order of Merit, Founding Director of Digital Society Institute (ISN), President of Scientific and Technological Council of IRT SystemX, Senior Scientific Advisor for “The AI Initiative“, International Advisor for

Japanese Science and Technology Agency Program "Advanced Core Technologies for Big Data Integration" , Elected Member of the Board of Directors of Big Data Value Association, General-chair of European Big Data Value Forum 2017 (Versailles), Member of the board of Data Transparency Lab, Member of the Scientific Councils of INRA, CentraleSupélec and Member of the Strategic Orientation Council of Institut Français.

Previously, Advisor to the Chairman and the CEO of Inria in Data Science with concern to the socioeconomic impact of Big Data and AI capabilities, Scientific Head of IMEDIA research group for over 10 years (till 2010) and the Director of Inria Saclay Research Center for 5 years (2010-2015).

5. Governance of the Development of AGI

Jessica Cussins Newman

Jessica Cussins Newman led the subcommittee on "Direct and Indirect Policy Recommendations" and is a Research Fellow at the UC Berkeley Center for Long-Term Cybersecurity, where she focuses on digital governance and the security implications of artificial intelligence. She is also an AI Policy Specialist with the Future of Life Institute and a Research Advisor with The Future Society. Jessica was a 2016-17 International and Global Affairs Student Fellow at Harvard's Belfer Center, and has held research positions with Harvard's Program on Science, Technology & Society, the Institute for the Future, and the Center for Genetics and Society. Jessica received her master's degree in public policy from the Harvard Kennedy School and her bachelor's in anthropology from the University of California, Berkeley with highest distinction honors. She has published 18 articles and more than 130 blog posts on the implications of emerging technologies in outlets including The Los Angeles Times, The Pharmaceutical Journal, Huffington Post, and CNBC.

6. Building Capability for "Smart" Governance of Artificial Intelligence

Konstantinos Karachalios

Led the subcommittee on "How to Build Public Trust". *His bio can be found in the Rapporteurs section.*

7. Governing AI Adoption in Developing Countries

Zaki Khoury

Led the subcommittee on "Opportunities and Challenges" and is Senior Technology and Strategy Advisor for The World Bank, advising and assisting National Governments on achieving Digital Development. Prior he was Regional Director for International Organizations, Middle East and Africa at Microsoft.

8. AI in the Judicial system, Access to justice, and the Practice of Law

Nicolas Economou

Nicolas Economou is the chief executive of H5. He serves as co-chair of the Law Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, and chair of The Future Society's Science, Law and Society Initiative. He is also a member of the Council on Extended Intelligence (CXI).

9. From Data Commons to AI Commons

Amir Banifatemi

Led this Expert Group along with Don Gossen. He is the General Manager for Innovation and Growth, and leads he AI initiatives and Executive Director for the IBM Watson AI XPRIZE and the ANA Avatar XPRIZE. Prior to joining XPRIZE, Mr. Banifatemi began his career at the European Space Agency and then held executive positions at Airbus, AP-HP and the European Commission division for information society and media. He managed two venture capital funds and contributed to the formation of more than 10 startups with emphasis on Information Technologies, Telecommunications, IoT, and Healthcare. Mr. Banifatemi is a guest lecturer and an adjunct MBA professor at UC Berkeley, Chapman University, Claremont McKenna College, UC Irvine, and HEC Paris.

Don Gossen

Led this Expert Group along with Amir Banifatemi. He is the Executive Director and Co-Founder of Ocean Protocol Foundation which aims to change the way in which people and companies manage and work with data. Leveraging the power of blockchain and decentralization, Ocean will combine the elegance of transactional immutability and crypto-economics with the fundamental principles of data governance and distributed processing to create a revolutionary data ecosystem. An ecosystem that affords end-to-end data provenance and portability with data privacy and security baked in. He was previously Head of Analytics and Big Data Practice at everis UK.

10. International Panel on AI

Francesca Rossi

Led the subcommittee on "Mapping and lessons from IPCC and other intergovernmental organizations". Francesca is the IBM AI Ethics Global Leader, a distinguished research scientist at the IBM T.J. Watson Research Centre, and a professor of computer science at the University of Padova, Italy.

Francesca is both a fellow of the European Association for Artificial Intelligence (EurAI fellow) and also a 2015 fellow of the Radcliffe Institute for Advanced Study at Harvard University. A prominent figure in the Association for the Advancement of Artificial Intelligence (AAAI), at which she is a fellow, she has formerly served as an executive councilor of AAAI and currently co-chairs the association's committee on AI and ethics. Francesca is an active voice in the AI community, serving as Associate Editor in Chief of the *Journal of Artificial Intelligence Research* (JAIR) and as a member of the editorial boards of *Constraints*, *Artificial Intelligence*, *Annals of Mathematics and Artificial Intelligence* (AMAI), and *Knowledge and Information Systems* (KAIS). She is also a member of the scientific advisory board of the Future of Life Institute, sits on the executive committee of the Institute of Electrical and Electronics Engineers (IEEE)'s global initiative on ethical considerations on the development of autonomous and intelligent systems, and belongs to the World Economic Forum Council on AI and robotics.

11. AI for SDGs

Cyrus Hodes

Cyrus led the subcommittee on " Preparing to Apply AI for SDGs". He most recently served as Advisor to the UAE Minister of Artificial Intelligence, currently working on projects that will positively impact the world through the use of AI and help shape the upcoming global governance of AI. Being passionate about drastically disruptive technologies, Cyrus previously led and still advises robotics and biotech ventures. In 2015, he co-founded the AI Initiative, which he managed by engaging a wide range of global stakeholders to study, discuss and help shape the governance of

AI. The AI Initiative did, and continue to do so, through various international policy platforms (OEDC, HKS Forums, Japanese MIC, French Parliament, etc.) as well as AI ethics and safety initiatives. Cyrus spearheaded several projects using innovative tools (such as the Global Civic Debate and its multilingual collective intelligence platform on the governance of AI) and works at using AI and Machine Learning to tackle policy issues. Cyrus is a member of three Committees (Policy, Well Being and General Principles) of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems as well as a Senior Advisor to The Future Society. Cyrus was educated at Sciences Po Paris, where he later was a Lecturer, holds a M.A. (Hons) from Paris II University in Defense, Geostrategy and Industrial Dynamics and a M.P.A. for Harvard Kennedy School.

12. AI and Cybersecurity

Roman Yampolskiy

Led the subcommittee on “Building Rules and Norms for Industry”. He is a Tenured Associate Professor in the department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville. He is the founding and current director of the Cyber Security Lab and an author of many books including Artificial Superintelligence: a Futuristic Approach. During his tenure at UofL, Dr. Yampolskiy has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a Senior member of IEEE and AGI; Member of Kentucky Academy of Science, and Research Advisor for MIRI and Associate of GCRI.

13. Managing the Economic & Social Impact of the AI Revolution

Calum Chace

Led the subcommittee on “Radical Disruption”. He is the author of The Economic Singularity Artificial Intelligence, and the Death of Capitalism (2016), Surviving AI: The Promise and Peril Of Artificial Intelligence (2015), and Pandora’s Brian (2014) – a novel examining the possible impact of super-intelligence. Prior to this he served as a chairman, coach, and consultant (three Cs) to entrepreneurs based on 30 years’ experience as a CEO, strategy consultant, and marketer.

14. AI Narratives

Sarah Dillon

Led the subcommittee on "Underrepresented Narratives". She is is Programme Director of the AI: Narratives and Justice Programme at the Leverhulme Centre for the Future of Intelligence, and a University Lecturer in Literature and Film in the Faculty of English at the University of Cambridge. Dr Dillon is a scholar of contemporary literature, film and philosophy, with a research focus on the epistemological function and role of fictional narratives, and on the engaged humanities. Her work is situated at fields of intersection and interconnection – between disciplines, and between sectors – and interrogates those sites in order to theorise and perform the specific modes of thought and knowledge offered by stories, and by the humanities. She is currently preparing two monographs arising out of the AI Narratives research.

<u>CO-CHAIRS</u>

1A: Mapping AI Technological Development: Horizon Scanning **Gabor Melli, Sony Playstation**
1B: Mapping AI Technological Development: Methodology **Jack Clark, OpenAI**
1D: Mapping AI Technological Development: Impact of Future Trajectories **Paul Epping, Philips Healthcare**

2B: Geopolitics of AI: Exploring the geostrategic landscape of AI **Brian Tse, Centre for the Governance of AI, Future of Humanity Institute**

3B: Agile Governance: Decentralized & distributed approaches **Gosia Loj, greeNet Solutions/All-Party Parliamentary Group on Artificial Intelligence (APPG AI)**
3C: Agile Governance: Political Economy of Standardization **John C. Havens, The IEE Global Iniitiative on Ethics of Autonomous and Intelligent Systems**
3D: Agile Governance: Devising Innovative Regulation for AI **Isabela Ferrari, Federal Judge Brazilian Judiciary**

4A: Explainable & Interpretable AI: What, Why and How? **De Kai Wu and Jessica Cussins,**
4B: Explainable & Interpretable AI: Algorithmic Bias - Value Alignment **Meeri Haataja, Saidot.ai**
4C: Explainable & Interpretable AI: From big questions to right actions **Jim Dratwa, European Commission**

5B: Governance of the Development of AGI: Other mechanisms for impact **Richard Mallah, Future of Life Institute**
5C: Governance of the Development of AGI: Stakeholders coordination **Sean O'hÉigeartaigh, Cambridge Centre for the Study of Existential Risk**

6A: Building Capability for "Smart" Governance of Artificial Intelligence: Building Competency for Governing AI in the Public Sector **Leanne Fry, AUSTRAC**
6C: Building Capability for "Smart" Governance of Artificial Intelligence: Lessons from Case Studies **Tim Clement Jones, UK House of Lords**
6D: Building Capability for "Smart" Governance of Artificial Intelligence: The Case for Public-Private-People Partnerships **Ali Hessami, IEEE**

7A: Governing AI Adoption in Developing Countries: Building Capabilities while Avoiding Exploitation **Eileen Lach, IEEE**
7C: Governing AI Adoption in Developing Countries: Managing Risks vs. Opportunities for Development **Stan Byers, New America, Policy Fellow on AI, Cybersecurity and International Development**

9A: From a Data Common to an AI Commons: AI Commons vs. Data Commons **Sarah Pearce, Paul Hastings LLP**
9B: From a Data Common to an AI Commons: Relevant Frameworks & Methodologies for Open Initiatives **Alpesh Shah, IEEE**
9C: From a Data Common to an AI Commons: Building the AI Commons **Brent Barron, CIFAR**
9D: From a Data Common to an AI Commons: Deploying the AI Commons **Ryan Budish, Berkman Klein Center for Internet & Society at Harvard University**

10B: International Panel on AI: Objectives & Approaches **Arisa Ema, University of Tokyo**
10C: International Panel on AI: Membership for IPAI **Anne Carblanc, OECD**
10D: International Panel on AI: Designing a global governance of AI framework **Raja Chatila,IEEE**

11B: AI for SDGs: Use-cases and framework for Education **Baroness Beeban Kidron, UK House of Lords**
11C: AI for SDGs: Use-cases and framework for Healthcare **Elizabeth Gibbons, Harvard FXB Centre**
11D: AI for SDGs: Use-cases and framework for Climate Change & Urban Development **David Jensen, UN Environmental Cooperation for Peacebuilding Programme**

12B: AI and Cybersecurity: Evaluating Policymakers' Options **Yann Bonnet, National Cybersecurity Agency of France (ANSSI)**
12C: AI and Cybersecurity: Educating & Empowering Users **Lydia Kostopoulos, ESMT Berlin**
12D: AI and Cybersecurity: Building Rules and Norms for Industry (2) **Robert Silvers, Paul Hastings LLP**

13B: Managing the Economic & Social Impact of the AI Revolution: Gradual Change **Marek Havrda, GoodAI**
13C: Managing the Economic & Social Impact of the AI Revolution: Educating for AI **Priya Lakhani, CENTURY Tech**

13D: Managing the Economic & Social Impact of the AI Revolution: Mitigating Rising Inequality **Irakli Beridze, UN Centre on AI and Robotics**

14B: AI Narratives: A tool for Policymakers **Casper Klynge, Technology Ambassador of Denmark**
14C: AI Narratives: A/IS Infrastructure and Ecosystem **Mark Halverson, Precision Autonomy**
14D: AI Narratives: Building Trust in AI Systems **John Sullins, Sonoma State University**