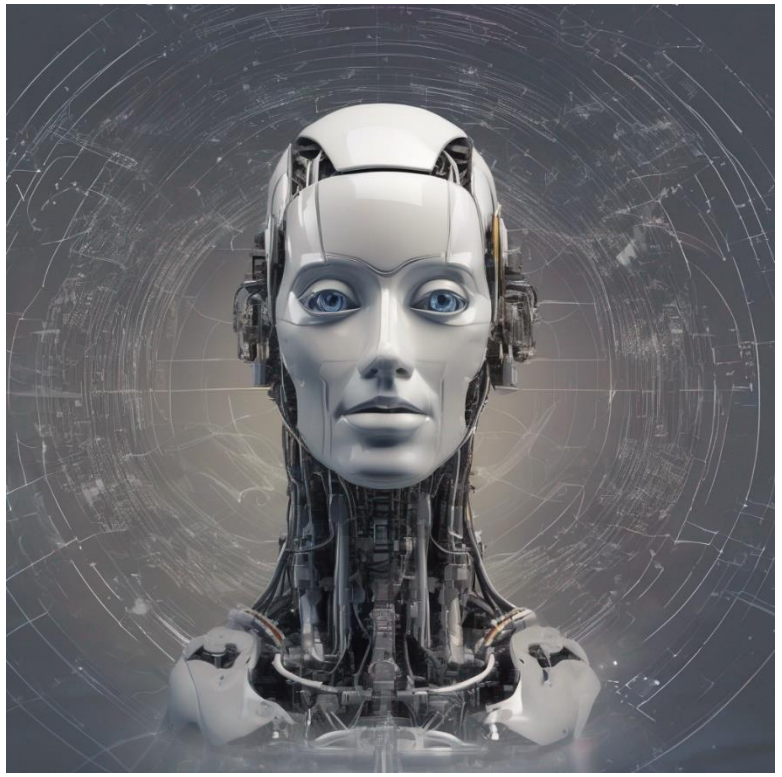
Governing the transition

To Tomorrow's Artificial General Intelligence



Jerome C. Glenn
The Millennium Project

UN Security Council session on AI Security. SG Guterres supports idea of a UN AI Agency and establishes a working group to create a framework for UN AI governance

July 2023

Parliaments agree on safe AI cooperation

70 Parliamentary delegations and the Inter-Parliamentary Union agree to cooperate on developing AGI regulations.

September 2023

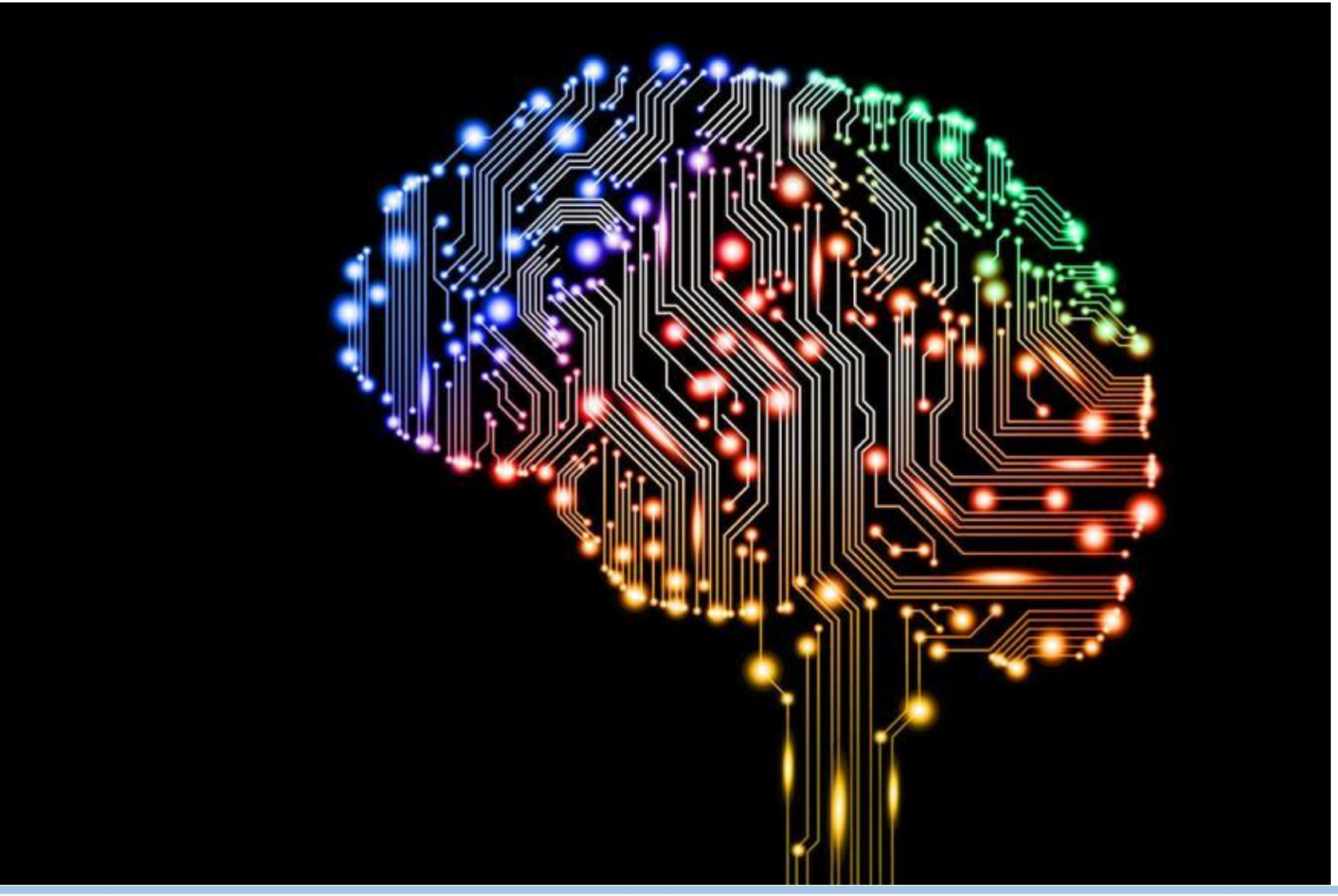
28 Countries agree on Safe AI cooperation

November 2023

US-China Agree on Safe AI Cooperation

November 2023

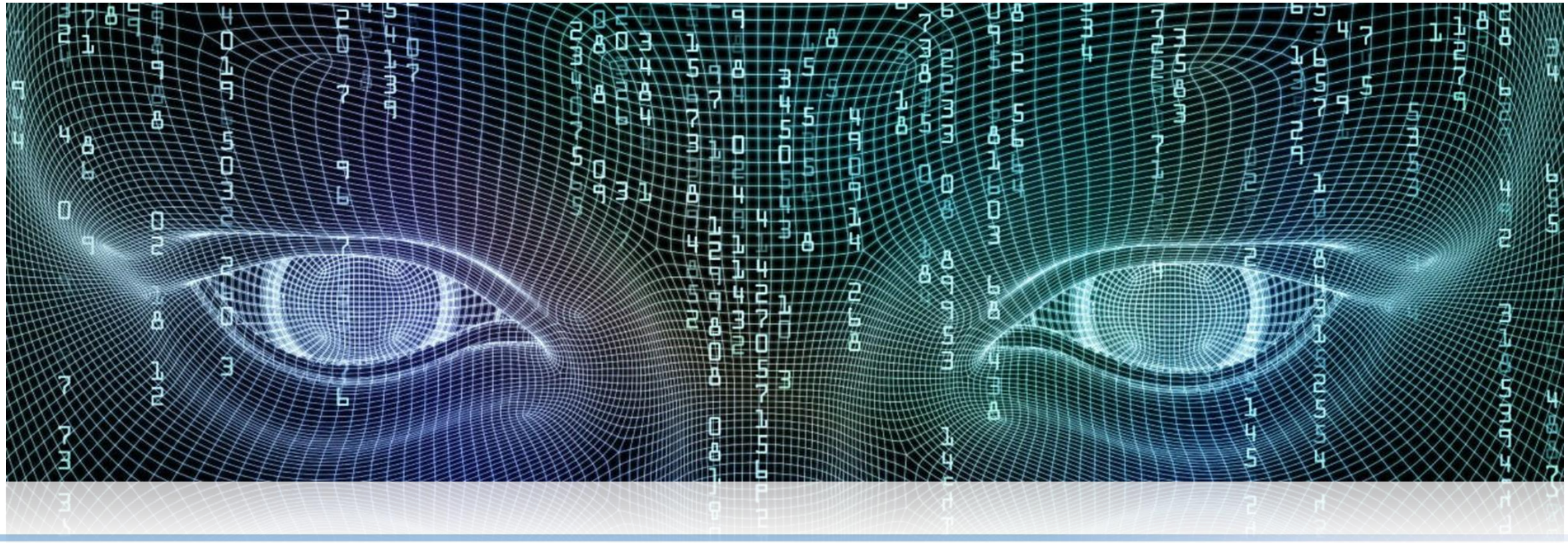
Three Forms of Artificial Intelligence



- Artificial **Narrow** Intelligence - ANI
- Artificial **General** Intelligence - AGI
- Artificial **Super** Intelligence - ASI

If we don’t get the initial conditions, rules, and guardrails “right” for AGI,

then **Artificial Super Intelligence** could evolve quickly beyond our understanding.




Transition from Narrow to General Artificial Intelligence
Let’s Avoid This

What We Need is a:

- **UN Convention on AI with TWO sections**:
    1. Current ANI and future forms of generative AI and
    2. Future Artificial General Intelligence

- **UN AI Agency to enforce it with TWO sections**
    1. Current ANI and future forms of generative AI and
    2. Future Artificial General Intelligence

The Millennium Project's AGI Global Governance Study

1. Reviewed other's research, conferences, podcasts, other Internet sources
2. Created 22 AGI-related questions
3. Interviewed about 25 AGI thought leaders and documented another 30.
4. Produced Phase 1 Report (now available)
5. Real-Time Delphi questionnaire AGI governance elements (current)
6. Produce Phase 2 report
7. Test several governance models in alternative detailed scenarios
8. Produce Phase 3 report
9. Global distribution to participants, conferences, UN, and parliaments

55 AGI ‘Experts’ and Thought Leaders Given 22 Questions

Sam Altman, via YouTube and OpenAI Blog, CEO OpenAI
Anonymous, AGI Existential Risk OECD (ret.)
Yoshua Bengio. University of Montreal
Irakli Beridze, UN Interregional Crime and Justice Res. Ins. Ct. for AI and Robotics
Nick Bostrom, Future of Humanity Institute at Oxford University
Gregg Brockman, OpenAI co-founder
Vint Cerf, Internet Evangelist, V.P. Google.
Shaoqun CHEN, CEO of Shenzhen Zhongnong Net Company
Anonymous, at Jing Dong AI Research Institute, China
Pedro Domingos, University of Washington
Dan Faggella, Emerj Artificial Intelligence Research
Lex Fridman, MIT and Podcast host
Bill Gates
Ben Goertzel, CEO SingularityNet
Yuval Noah Harari, Hebrew University, Israel
Tristan Harris, Center for Humane Technology
Demis Hassabis, CEO and co-founder of DeepMind
Geoffrey Hinton, AI pioneer, Google (ret)
Lambert Hogenhout, Chief Data, Analytics and Emerging Technologies, UN Secretariat
Erik Horvitz, Chief Scientific Officer, Microsoft
Anonymous, Information Technology Hundred People Association, China
Anonymous, China Institute of Contemporary International Relations
Andrej Karpathy, Open AI, former AI S Researcher Tesla
David Kelley, AGI Lab
Dafne Koller, Stanford University, Coursera
Ray Kurzweil, Director of Engineering Machine Learning, Google
Connor Leahy, CEO Conjecture
Yann LeCun, Professor New York University, Chief Scientist for Meta
Shane Legg, co-founder of DeepMind
Fei Fei Li, Stanford University, Human Centered AI
Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
Gary Marcus, NYU professor emeritus
Dale Moore, US Dept of Defense AI consultant
Emad Mostaque, CEO of Stability.ai
Elon Musk
Gabriel Mukobi, PhD student Stanford University
Anonymous, National Research University Higher School of Economics
Judea Pearl, Professor UCLA
Sundar Pichai, Google CEO
Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
Anonymous, Russian Academy of Science
Stuart Russell, UC Berkeley
Karl Schroeder, Science Fiction Author
Bart Selman, Cornel University
Juan Del Ser, Tecnalia, Spain
David Shapiro, AGI Alignment Consultant
Yesha Sivan, Founder and CEO of i8 Ventures
Ilya Sutstkever, Open AI co-founder
Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute
Max Tegmark, Future of Life Institute and MIT
Peter Voss, CEO and Chief Scientist at Aigo.ai
Paul Werbos, National Science Foundation (ret.)
Stephen Wolfram, Wolfram Alpha, Wolfram Language
Yudong Yang, Alibaba’s DAMO Research Institute
Eliezer Yudkowsky Machine Intelligence Research Institute

Some examples of the 22 questions

1. What are some key initial conditions for AGI so that an artificial super intelligence does not emerge later that is not to humanity's liking?
2. How can alignment be achieved? If you think it is not possible, then what is the best way to manage this situation?
3. What options or models are there for global governance of AGI?
4. How can governance be flexible enough to respond to new issues previously unknown at the time of creating that governance system?
5. What enforcement powers will be needed to make the UN Agency effective?

Examples of global AGI governance models

1. IAEA-like model or WTO-like with enforcement powers. These are the easiest to understand, but likely to be too static to manage AGI.
2. Create two divisions in a UN AI Agency: one for artificial narrow intelligence including frontier models and a second division just for AGI.
3. A multi-stakeholder body (TransInstitution) in partnership with a system of artificial narrow intelligences, each ANI to implement specific functions/requirements continually feeding back to the humans in the multi-stakeholder body and national AGI governance agencies.
4. International S&T Organization (ISTO) as an online real-time global collective intelligence system; governance by information power. This would be useful to help select and use an AGI system, but no proof that information power would be sufficient to govern the evolution of AGI.
5. Put all the most powerful AI training chips and AI inference chips into a limited number of computing centers under international supervision, with a treaty granting symmetric access rights to all countries party to that treaty.(Suggested by Eliezer Yudkowsky)
6. Decentralized emergence of AGI that no one owns (like no one owns the Internet) through the interactions of many AI organizations and developers like Singularity Net. (https://singularitynet.io)

We may rush into creating AGI without making sure its rules, guardrails are “right.”

Examples of some potential “right” initial conditions, rules, and guardrails:

- National licensing systems connected to UN AI Agency for certification.
- Aliened with UNESCO, OECD, GAPI values and Asimov’s three laws of robotics.
- Prior to UN certification of a national license, the AGI developer must prove safety as part of the initial audit via simulations to test AGI’s alignment with these values.
- Continuous real-time audit software built in the AGI that can turn off the AGI when not aligned with values or goes beyond guardrails.
- Cannot turn off its own off switch.
- AI generated content must show its source (watermark, etc.)
- Require developers and users to keep a log of the AGI use like a flight recorder with the ability to recreate a decision and factors included in the decision.
- Be self-reflective, compassionate, truth-seeking, and love of humanity.

A few words about
The Millennium Project

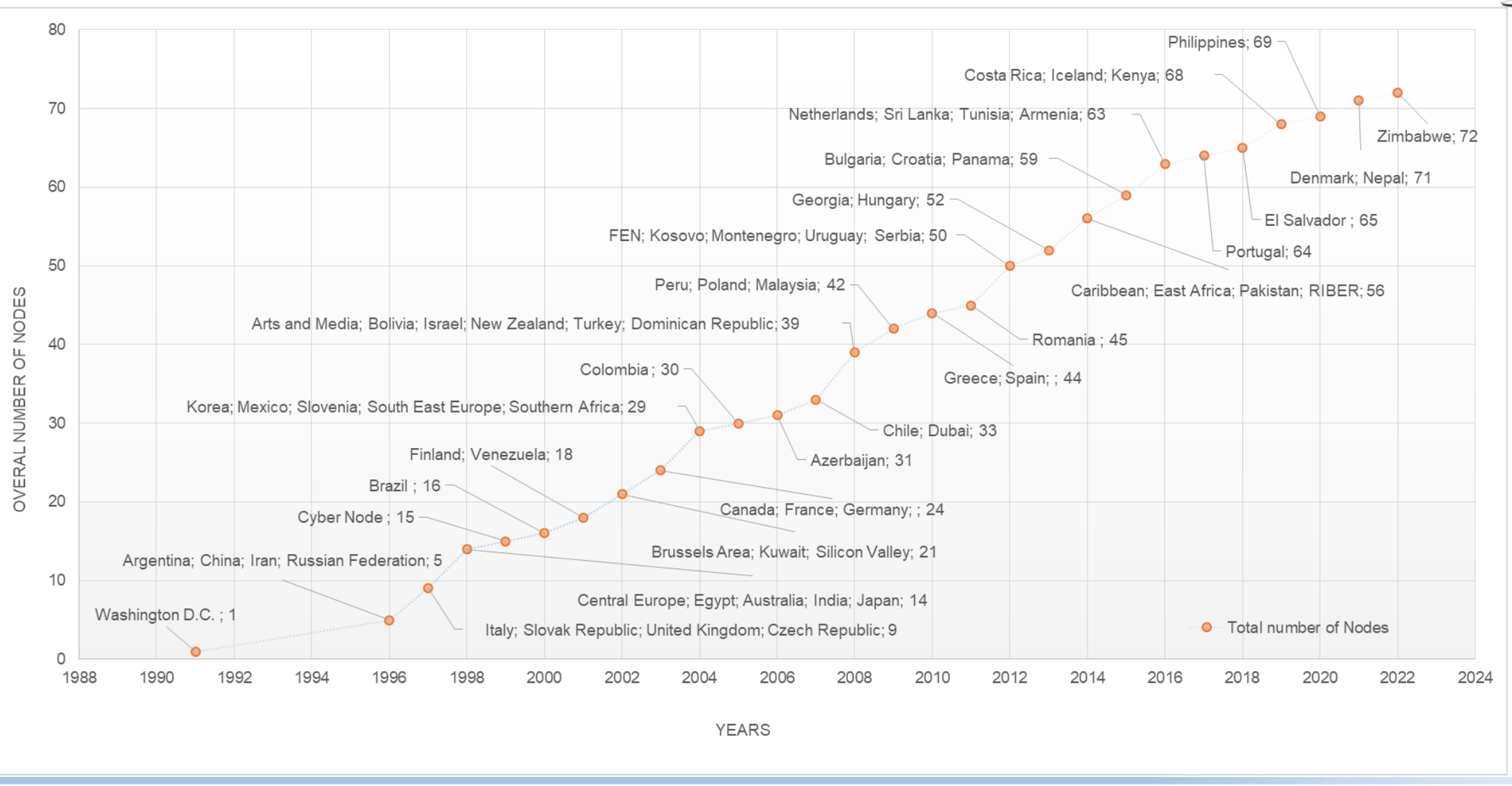
… Acts like a *TransInstitution*

Established in 1996

Three-Year Feasibility Study 1992-1995

One-year Pre-feasibility study 1991-1992

Evolution of The Millennium Project Nodes



For more information on The Millennium Project AGI study:

Phase 1 Report https://www.millennium-project.org/transition-from-artificial-narrow-to-artificial-general-intelligence-governance/

Phase 2: Real-Time Delphi deadline December 31st.

https://www.millennium-project.org/agi-global-governance-real-time-delphi-deadline-dec-31-2023/

European Commission AGI paper for EC's Horizon 2024-2027 planning

https://www.futures4europe.eu/blogs/artificial-general-intelligence-issues-and-opportunities

3-minute video https://www.youtube.com/watch?v=Xd6at9XCD3U

Paper for the UN Secretary-General's Envoy on Technology (available email Jerome.Glenn@Millennium-Project.org