Welcome to this week's presentation & conversation hosted by the **Canadian Association for the Club of Rome**, a Club dedicated to intelligent debate & action on global issues.

The views and opinions expressed in this presentation are those of the speaker & do not necessarily reflect the views or positions of CACOR.

Transition from Artificial Narrow to Artificial General Intelligence Governance.

Our speaker today is Jerome Glenn, Director of the Millennium Project, which looks at global challenges to humanity. He consults on forecasting methodology & on other issues. He was a founding partner of Future Options Room, helped craft the section of the SALT II treaty that prohibited Fractional Orbital Bombardment, created CARINET (a computer network CGNET Services later bought), & through CARINET introduced data packet switching to the developing world. He & TJ Gordon wrote a report in cooperation with the Smithsonian Institution & the Futures Group (now Palladium International) on the feasibility of a futures think-tank. He authors an annual publication, *State of the Future,* for the Millennium Project. Glenn is member USA COR.

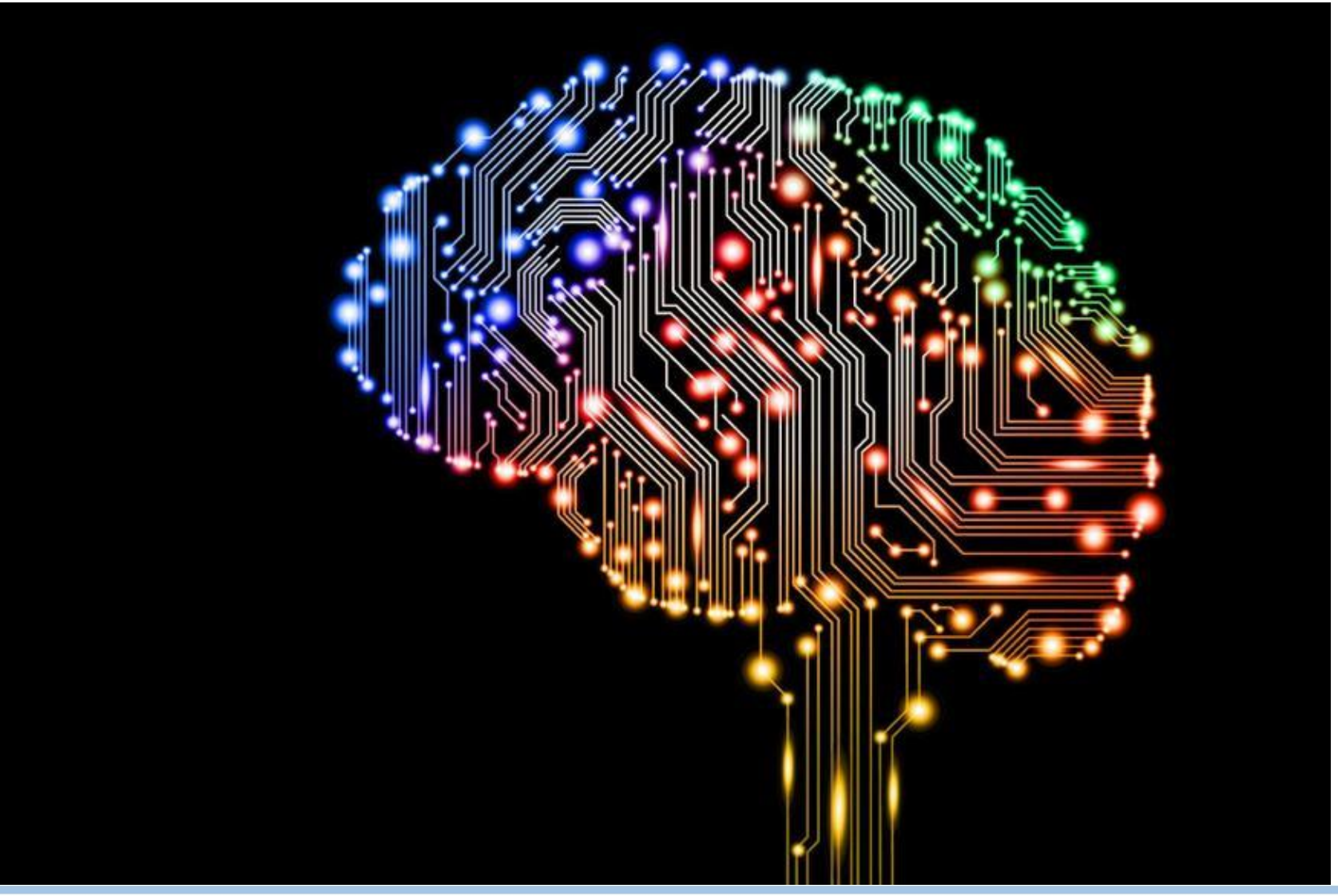
DESCRIPTION: Phase 1 of a transition analysis from Artificial Narrow Intelligence (ANI) to Artificial General Intelligence (AGI) collected views from US, China, UK, EU, Canada, & Russia. **Origin or Self-Emergence:** What are the possible paths from today's AI to more capable AGI? What are the most serious outcomes if these paths aren't governed or are governed badly? What are key conditions for AGI so that a super-AI doesn't emerge that's not to humanity's liking? **Value alignment, morality, values:** From the work of the Global Partnership on Artificial Intelligence (GPAI) & others, what values should be considered for AGI? If a hierarchy of values becomes necessary for international treaties & a governance system, what should be the top priorities? How can alignment be achieved? If you think alignment isn't possible, what is the best way to manage this situation?

The presentation will be followed by a conversation, questions, & observations from the participants.

CACOR acknowledges that we all benefit from sharing the traditional territories of local Indigenous peoples (First Nations, Métis, & Inuit in Canada) and their descendants.

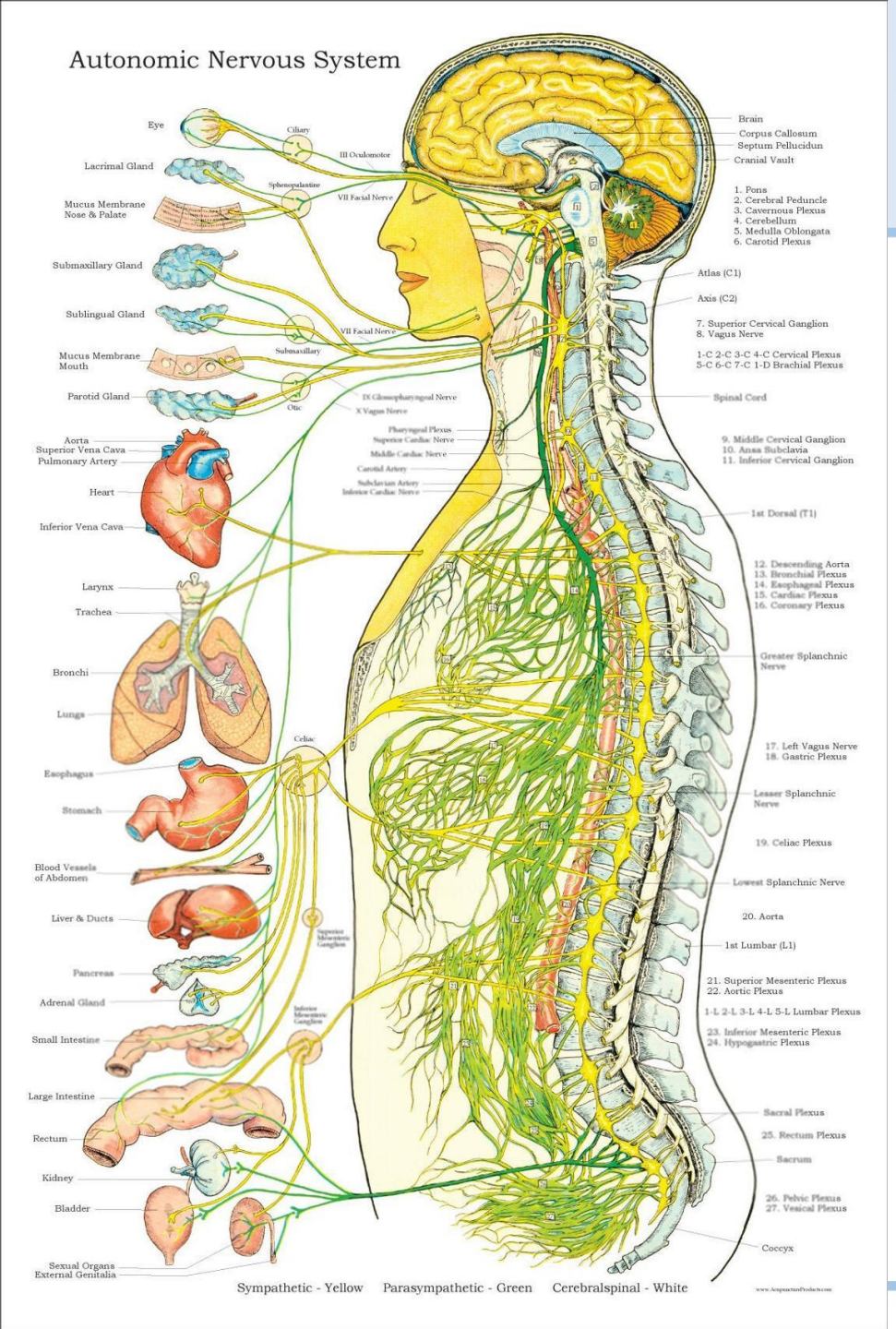
Website: canadiancor.com
Twitter: @cacor1968
YouTube: Canadian Association for the Club of Rome
2023 Nov 01 Zoom #169

Three Forms of Artificial Intelligence



- Artificial **Narrow** Intelligence
- Artificial **General** Intelligence
- Artificial **Super** Intelligence

Autonomous Nervous System & Civilization



Just as the autonomous nervous system runs much of our body, freeing the mind to be creative;

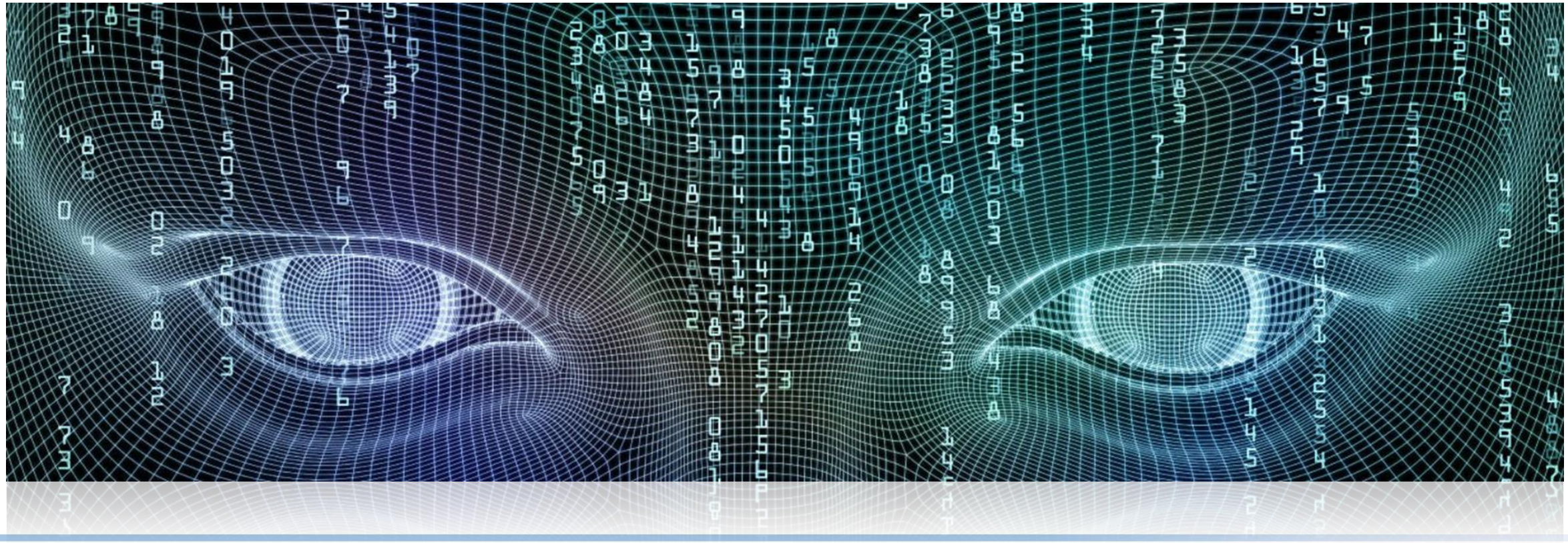
AI and other Next Technologies (NT) will run much of the physical infrastructure of future civilization, freeing more of humanity to be creative.

AGI is expected to be able to

- Address novel problems without pre-programing like ANI needs
- Initiate searches for information worldwide
- Use sensors and the Internet of Things (IoT) to learn
- Make phone calls and interview people
- Make logical deductions, reason similar to humans
- Learn from experience and reinforcement, without the need for a massive data base to learn a task like ANI needs
- Re-write or edit its code to be more intelligent …continually, so it gets smarter and smarter, faster than humans

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

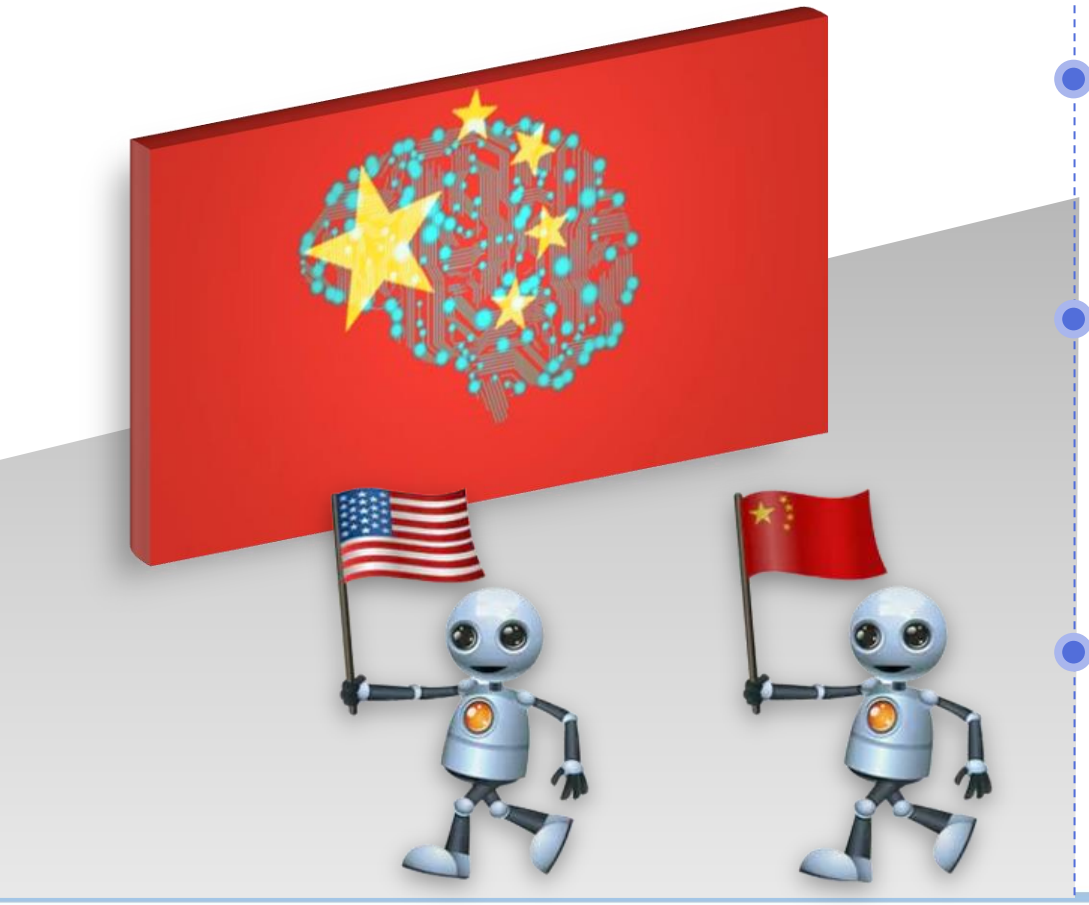
then **Artificial Super Intelligence** could evolve quickly beyond our understanding.




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

Who ever leads AI... rules the world - Putin

China Goal: Lead the World in AI by 2030



Human Brain Projects:

USA, EU, China

Artificial Brains:

IBM, Google, Facebook, Tencent, SenseTime, MEGVII, Microsoft, Alibaba, Cloudwalk, Yitutech, Apple, Amazon

The Great AGI/Brain Race is on! But needs national and international governance

UNSG Guterres confirmed his support for a UN AI Agency and establishment a working group to draft the framework for AI governance during the July 18th UNSC first meeting on AI Security Issues

AI UN Governance Group Established

- September 20th United Nations AI meeting: UN governance of AI is inevitable. which means:
- UN Convention on AI and UN AGI Agency to enforce the Convention.
- Parliaments will have to create their national ANI and AGI regulatory systems.
- Harmonizing UN work and national work on AI regulations would be wise.
- For example: shut off switch:
- first by internal continuous audit system – if that fails – then notify user to shut it down - if that fails - then instant communication with the national regulatory agency – if that fails – then the UN AGI Agency
- All this will have to be worked out among AI companies, users, national regulations, and UN system.

The Millennium Project's AGI Global Governance Study

1. Review of 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. Results will shape an international Real-Time Delphi questionnaire (nominations for participants welcome)
5. Results of the RTDelphi used to construct alternative scenarios of different global governance models
6. Scenarios submitted to a second RTDelphi for improvement
7. Final Scenarios distributed and translated through the 71 Millennium Project Nodes around the world, social media, conferences, and talks like this
8. Shared with the forthcoming UN Strategic Foresight and Global Risk Report

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

3 of 22 questions on Origin of AGI

1. How do you envision the possible trajectories ahead, from today's AI, to much more capable AGI in the future?
2. What are the most important serious outcomes if these trajectories are not governed, or are governed badly?
3. 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?

Value alignment, morality, values

4. Drawing on the work of the Global Partnership on Artificial Intelligence (GPAI) and others that have already identified norms, principles, and values, what additional or unique values should be considered for AGI?
5. If a hierarchy of values becomes necessary for international treaties and a governance system, what should be the top priorities?
6. How can alignment be achieved? If you think it is not possible, then what is the best way to manage this situation?

Governance and Regulations

7. How to manage the international cooperation necessary to build international agreements and a global governance system while nations and corporations are in an intellectual “arms race” for global leadership?
8. What options or models are there for global governance of AGI?
9. What risks arise from attempts to govern the emergence of AGI? (Might some measures be counterproductive?)
10. Should future AGIs be assigned rights?
11. How can governance be flexible enough to respond to new issues previously unknown at the time of creating that governance system?
12. What international governance trials, tests, or experiments can be constructed to inform the text of an international AGI treaty?
13. How can international treaties and a governance system prevent increased centralization of power crowding out others?
14. Where is the most important or insightful work today being conducted on global governance of AGI?

Control

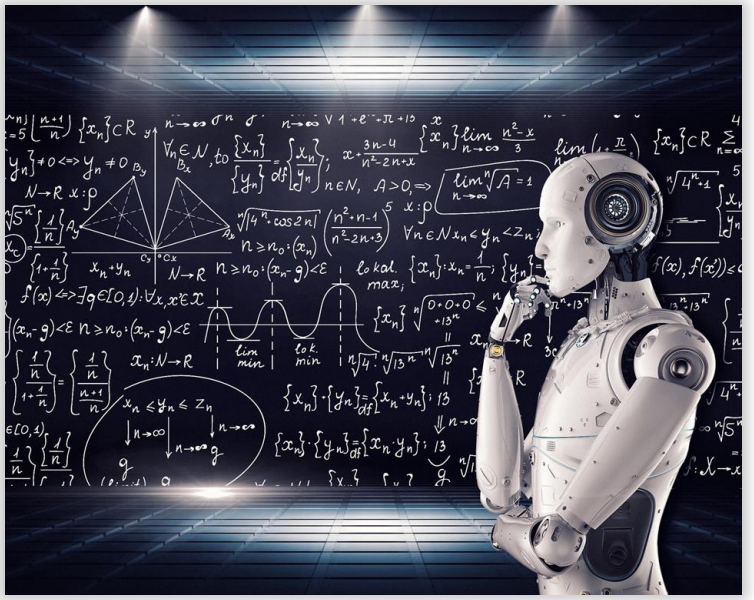
15. What enforcement powers will be needed to make an international AGI treaty effective?
16. How can the use of AGI by organized crime and terrorism be reduced or prevented? (Please consider new types of crimes and terrorism which might be enabled by AGI.)
17. Assuming AGI audits would have to be continuous rather than one-time certifications, how would audit values be addressed?
18. What disruptions could complicate the task of enforcing AGI governance?
19. How can a governance model correct undesirable action unanticipated in utility functions?
20. How will quantum computing affect AGI control?
21. How can international agreements and a governance system prevent an AGI “arms race” and escalation from going faster than expected, getting out of control and leading to war, be it kinetic, algorithmic, cyber, or information warfare?

Recommendations

- Review AI regulatory work by conduct hearings on AGI governance and regulations, making clear distinctions between ANI and AGI
- Coordinate draft parliamentary regulations with UNSG's working group on the framework for AI governance, and related efforts by OECD
- Recommend an UN AGI Agency to the UNGA (UNSC not in agreement), at the UN Summit on the Future
- Create "preparedness teams" in major AGI companies like OpenAI has done

AGI Global Governance

- It is argued that creating rules for governance of AGI too soon will stifle its development.
- Some AGI experts believe it is possible to have AGI as soon as 3-5 years.
- Since it is likely to take that long to
  - ✓ develop national and international agreements
  - ✓ design an international governance system
  - ✓ And begin implementation



Then it is wise to begin exploring global governance approaches **now**.

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

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

Values for ANI (IEEE, OECD, UNESCO) and Asimov’s three laws of robotics.

Massively complex simulations to test AGI’s alignment with these values.

Ability for humans to check why the AGI is doing something.

Ability for AGI to distinguish between how humans act vs. how we ***should act***.

Ability to determine why and how it failed or caused harm.

Self-replication with human supervision.

Cannot turn off its own off switch.

Guardrails to catch unanticipated behavior that triggers evaluation, audit correction.

Able to be continually audited and shut down if it fails the audit.

AI generated content must show its source (water mark, etc.)

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. IPCC-like model in concert with international treaties. This approach has not led to a governance system for climate change.
3. 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.
4. GGCC (Global Governance Coordinating Committees) would be flexible and enforced by national sanctions, ad hoc legal rulings in different countries, and insurance premiums. This has too many ways for AGI developers to avoid meeting standards.
5. UN, ISO and/or IEEE standards used for auditing and licensing; licensing would affect purchases and would have impact, but requires international agreement of all countries ratifying.
6. Put different parts of AGI governance under different bodies like ITU, WTO, WIPO. Some of this is likely to happen but this is not sufficient to govern all instances of AGI systems.
7. Decentralized Semi-Autonomous TransInstitution. This could be the most effective, but the most difficulty to establish since both Decentralized Semi-Autonomous Organizations and TransInstitutions are new concepts.

AGI & UBI - Why is Tech-Unemployment different this time?

1. The acceleration of technological change
2. The globalization, interactions, and synergies among NTs
3. The existence of a global platform—the Internet—for simultaneous technology transfer … with far fewer errors in the transfer than in the past
4. Standardization of data bases and protocols
5. Few plateaus or pauses of change allowing time for individuals and cultures to adjust to the changes
6. Billions of empowered people in relatively democratic free markets able to initiate activities
7. Machines can learn how you do what you do, and then do it better than you.

A few words about
The Millennium Project

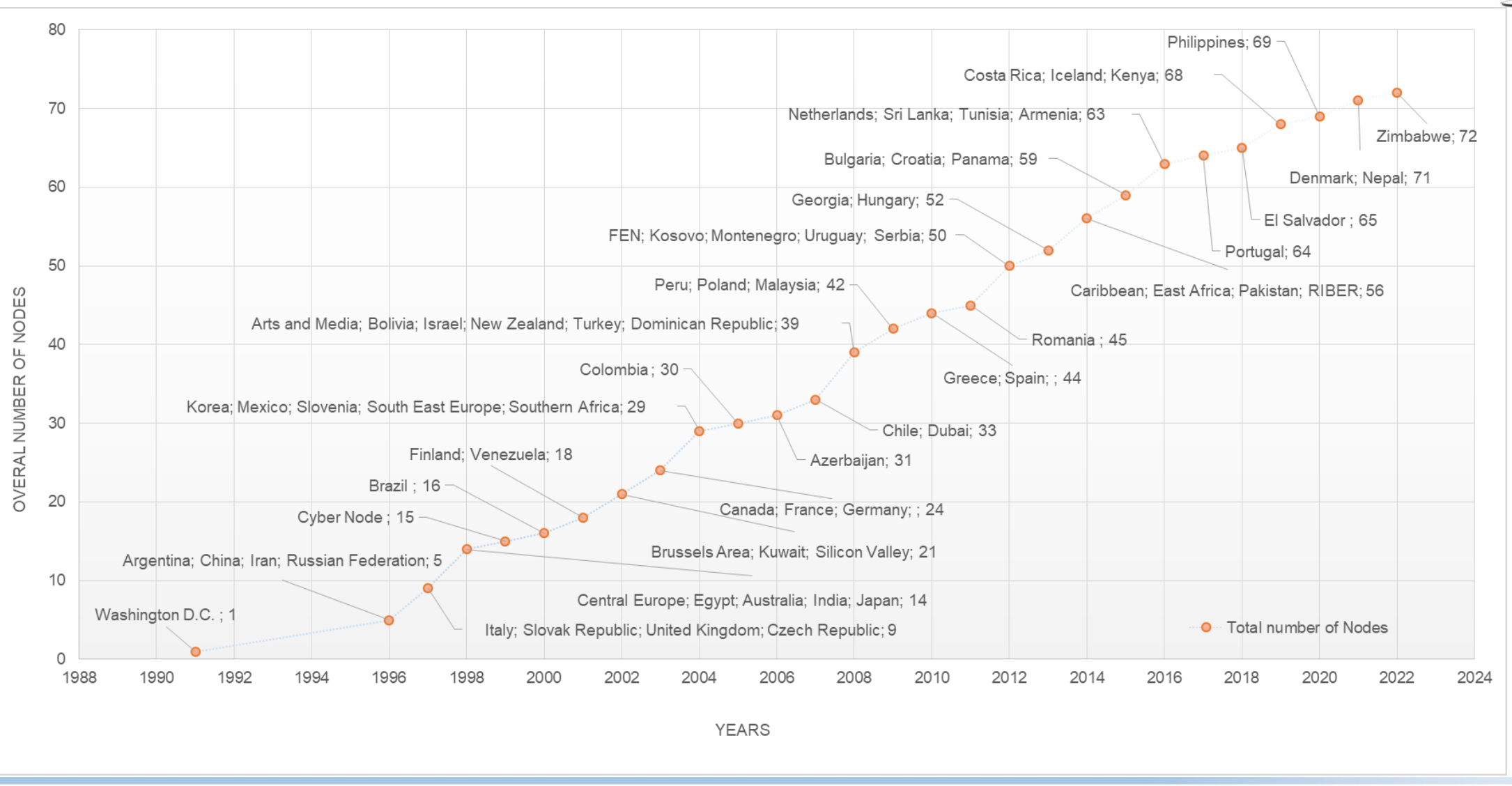
… Acts like a *TransInstitution*

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:

The Millennium Project AGI study:

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

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