International Governance Issues of the Transition from Artificial Narrow Intelligence to Artificial General Intelligence

Presentation of the Report Phase 1 by the Millennium Project

Sirkka Heinonen & Paula Pättikangas

AI Horizons sensemaking hybrid meeting FFRC 19th Sept 2023

UNIVERSITY OF TURKU

Transition from Narrow to General Artifical Intelligence

The Millennium Project has started an international assessment of how to govern the potential transition from Artificial Narrow Intelligence (ANI) to potential Artificial General Intelligence (AGI).

Key themes:

- Origin of Self-Emergence
- Value Alignment, Morality, Values
- Governance and Regulations
- Control

The Millennium Project

Phase 1

The views of 55 AGI leaders in the U.S., China, UK, Canada, EU, and Russia were collected. These experts were invited to address only those questions they preferred to address. Some experts were interviewed, some submitted written answers.

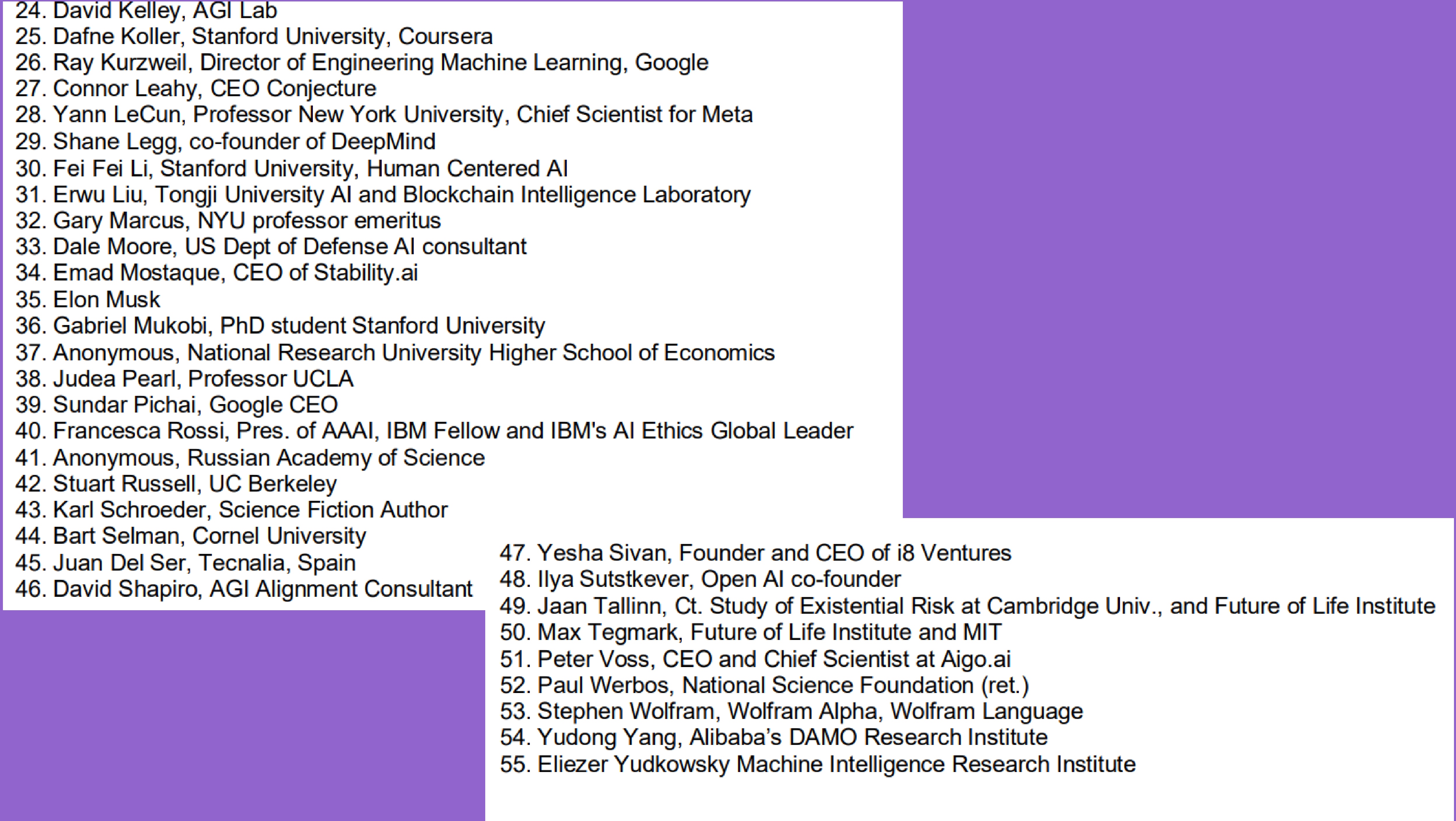
The analysis of Phase 1 will be used for an international assessment (Real-Time Delphi – RTD). The result of the RTD and Phase 1 will be used to generate content for alternative scenarios focusing on global governance of AGI.

These scenarios will be widely distributed to broaden and deepen the current conversation about future AI.

UNIVERSITY
OF TURKU

AGI Experts and Thought Leaders

1. Sam Altman, via YouTube and OpenAI Blog, CEO OpenAI
2. Anonymous, AGI Existential Risk OECD (ret.)
3. Yoshua Bengio. AI pioneer, Quebec AI Institute and the University of Montréal
4. Irakli Beridze, UN Interregional Crime and Justice Res. Ins. Ct. for AI and Robotics
5. Nick Bostrom, Future of Humanity Institute at Oxford University
6. Gregg Brockman, OpenAI co-founder
7. Vint Cerf, Internet Evangelist, V.P. Google.
8. Shaoqun CHEN, CEO of Shenzhen Zhongnong Net Company
9. Anonymous, at Jing Dong AI Research Institute, China
10. Pedro Domingos, University of Washington
11. Dan Faggella, Emerj Artificial Intelligence Research
12. Lex Fridman, MIT and Podcast host
13. Bill Gates
14. Ben Goertzel, CEO SingularityNET
15. Yuval Noah Harari, Hebrew University, Israel
16. Tristan Harris, Center for Humane Technology
17. Demis Hassabis, CEO and co-founder of DeepMind
18. Geoffrey Hinton, AI pioneer, Google (ret)
19. Lambert Hogenhout, Chief Data, Analytics and Emerging Technologies, UN Secretariat
20. Erik Horvitz, Chief Scientific Officer, Microsoft
21. Anonymous, Information Technology Hundred People Association, China
22. Anonymous, China Institute of Contemporary International Relations
23. Andrej Karpathy, Open AI, former AI S Researcher Tesla



24. David Kelley, AGI Lab
25. Dafne Koller, Stanford University, Coursera
26. Ray Kurzweil, Director of Engineering Machine Learning, Google
27. Connor Leahy, CEO Conjecture
28. Yann LeCun, Professor New York University, Chief Scientist for Meta
29. Shane Legg, co-founder of DeepMind
30. Fei Fei Li, Stanford University, Human Centered AI
31. Erwu Liu, Tongji University AI and Blockchain Intelligence Laboratory
32. Gary Marcus, NYU professor emeritus
33. Dale Moore, US Dept of Defense AI consultant
34. Emad Mostaque, CEO of Stability.ai
35. Elon Musk
36. Gabriel Mukobi, PhD student Stanford University
37. Anonymous, National Research University Higher School of Economics
38. Judea Pearl, Professor UCLA
39. Sundar Pichai, Google CEO
40. Francesca Rossi, Pres. of AAAI, IBM Fellow and IBM's AI Ethics Global Leader
41. Anonymous, Russian Academy of Science
42. Stuart Russell, UC Berkeley
43. Karl Schroeder, Science Fiction Author
44. Bart Selman, Cornel University
45. Juan Del Ser, Tecnalia, Spain
46. David Shapiro, AGI Alignment Consultant
47. Yesha Sivan, Founder and CEO of i8 Ventures
48. Ilya Sutstkever, Open AI co-founder
49. Jaan Tallinn, Ct. Study of Existential Risk at Cambridge Univ., and Future of Life Institute
50. Max Tegmark, Future of Life Institute and MIT
51. Peter Voss, CEO and Chief Scientist at Aigo.ai
52. Paul Werbos, National Science Foundation (ret.)
53. Stephen Wolfram, Wolfram Alpha, Wolfram Language
54. Yudong Yang, Alibaba’s DAMO Research Institute
55. Eliezer Yudkowsky Machine Intelligence Research Institute

***NB!** The inputs and views in this presentation are taken from the report and reflect the views of the interviewees. An attempt has been made to present a comprehensive range of respondents' views – often conflicting ones – although some emphasis has been placed on the most frequently encountered responses.*

At the end of this presentation, some comments and feedback from the presentation are given.

UNIVERSITY OF TURKU

DEFINITIONS USED

Artificial Narrow Intelligence (ANI)

The kind of AI we have today: each software application has a single specific purpose.

Artificial General Intelligence (AGI)

A general-purpose AI that can learn, edit its code, and act autonomously to address novel and complex problems set by humans with novel and complex strategies that are similar to or better than what humans could do.

Artificial Super Intelligence (ASI)

An articial superintelligence that far exceeds human powers and has become independent of humans, developing its own purposes, goals, and strategies without human understanding, awareness, or control and continually increases its intelligence and scope of action beyond humanity as-a-whole.

UNIVERSITY
OF TURKU

Transition from Narrow to General Artifical Intelligence

- The progress of AI is not likely to slow down, but accelerate.

- *The timeline is strict*. AGI could be within ten to twenty years, whereas, an international AGI treaty could take more than ten on twenty years.

- If the initial conditions of AGI are not ”right”, it could evolve into the kind of ASI that is not aligned with humanity’s interests.

- The most critical AGI issues are
  - Global governance
  - AGI’s initial conditions
  - Public awareness

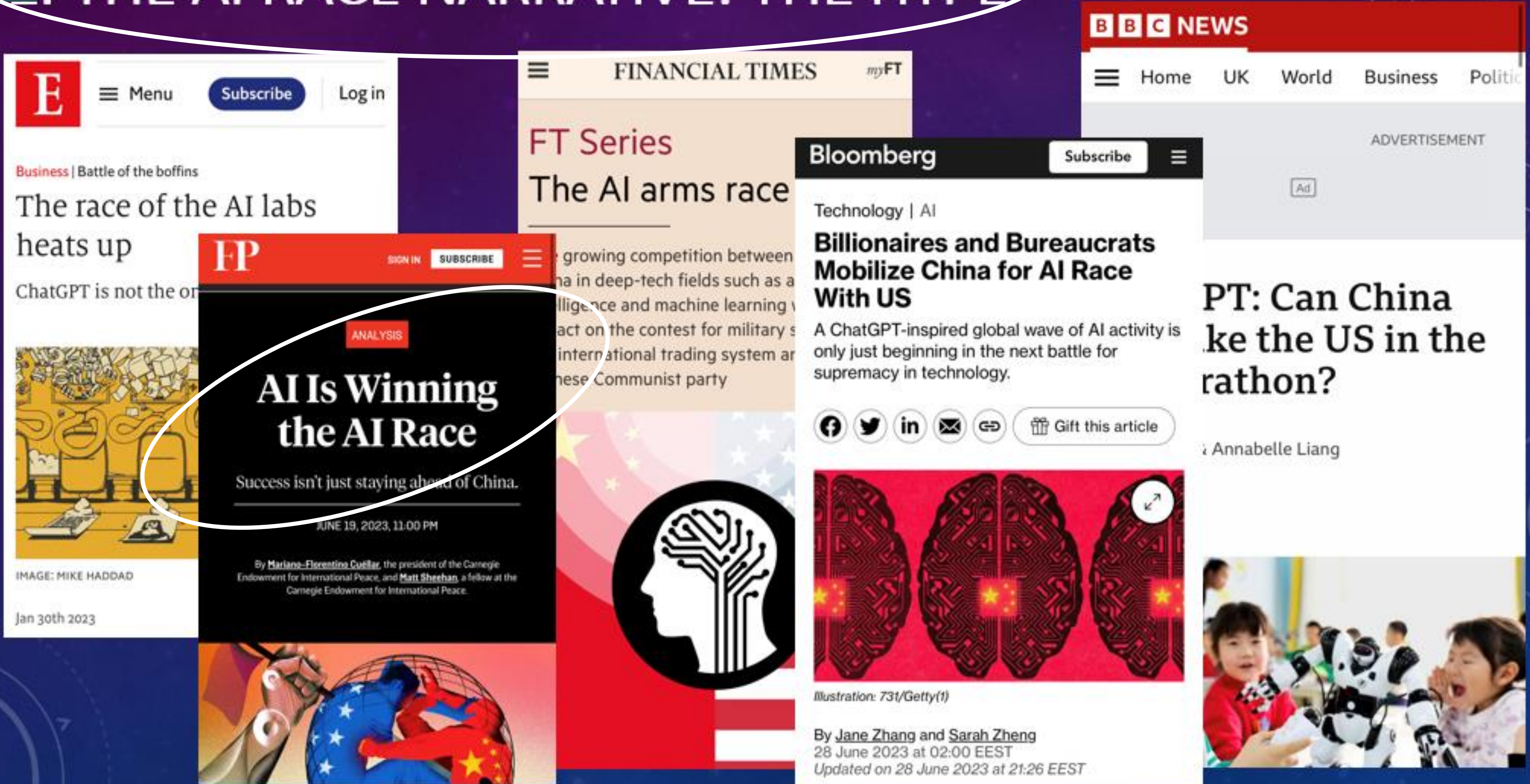
UNIVERSITY OF TURKU

What is the Narrative for AI? For AGI?

Who is winning the race for AI?

Junhua Zhu at FIIA (Finnish Institute of International Affaitrs) seminar 14th Sept 2023: Challenges in the Chinese Innovation ecosystem - AI as a Case He is doing his doctoral thesis on AI and ethics at University of Turku

. THE AI RACE NARRATIVE: THE HYPE



E ☰ Menu Subscribe Log in

Business | Battle of the boffins

The race of the AI labs heats up

ChatGPT is not the or

IMAGE: MIKE HADDAD

Jan 30th 2023

FP SIGN IN SUBSCRIBE ☰

ANALYSIS

AI Is Winning the AI Race

Success isn't just staying ahead of China.

JUNE 19, 2023, 11:00 PM

By Mariano-Florentino Cuéllar, the president of the Carnegie Endowment for International Peace, and Matt Sheehan, a fellow at the Carnegie Endowment for International Peace.

☰ FINANCIAL TIMES myFT

FT Series

The AI arms race

growing competition between na in deep-tech fields such as a lligence and machine learning act on the contest for military s international trading system ar nese Communist party

Bloomberg Subscribe ☰

Technology | AI

Billionaires and Bureaucrats Mobilize China for AI Race With US

A ChatGPT-inspired global wave of AI activity is only just beginning in the next battle for supremacy in technology.

Gift this article

Illustration: 731/Getty(1)

By Jane Zhang and Sarah Zheng
28 June 2023 at 02:00 EEST
Updated on 28 June 2023 at 21:26 EEST

BBC NEWS

☰ Home UK World Business Politic

ADVERTISEMENT

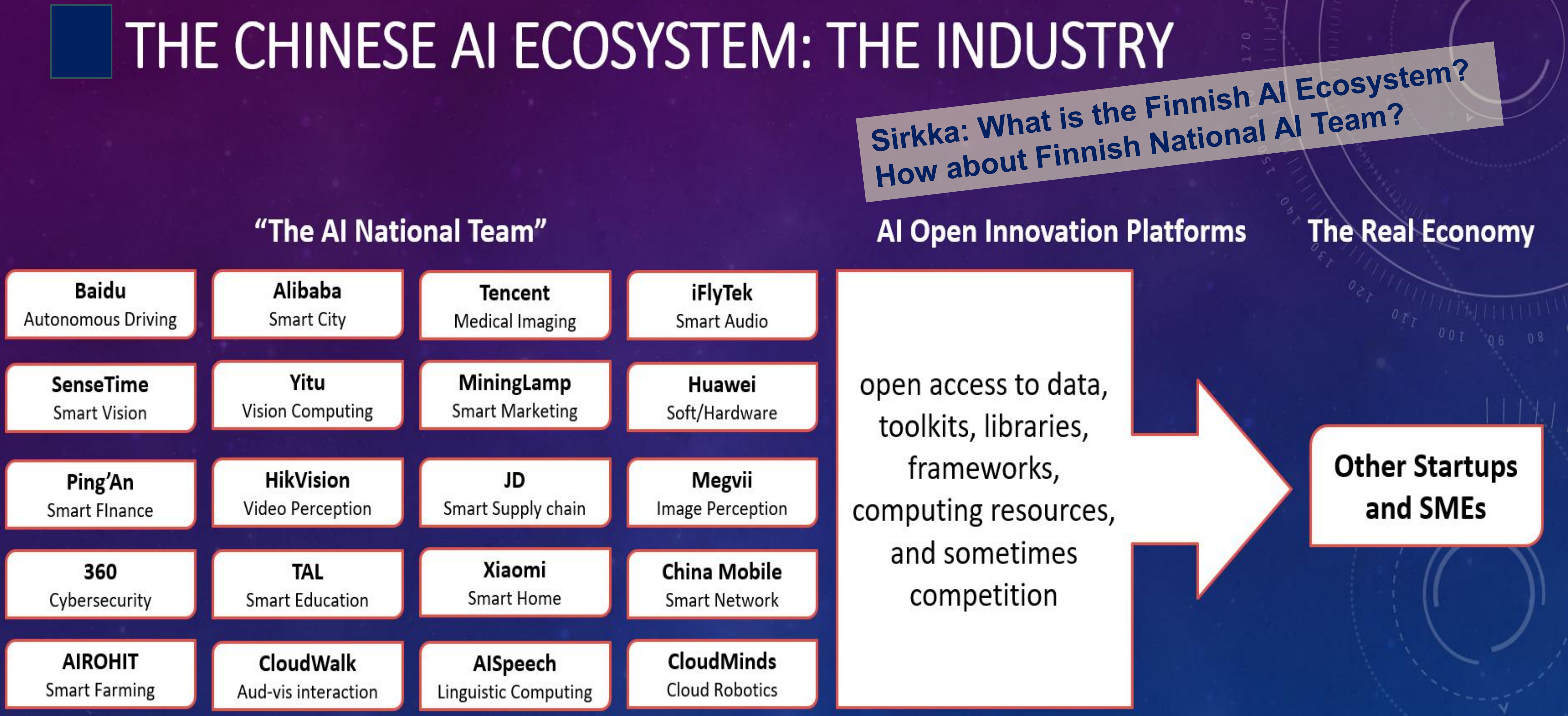
Ad

PT: Can China ke the US in the rathon?

Annabelle Liang

Junhua Zhu at FIIA seminar 14th Sept 2023
https://www.fiia.fi/en/event/forac-project-final-event

Challenges in the Chinese Innovation ecosystem - AI as a Case



THE CHINESE AI ECOSYSTEM: THE INDUSTRY

Sirkka: What is the Finnish AI Ecosystem?
How about Finnish National AI Team?

“The AI National Team”

Baidu
Autonomous Driving

Alibaba
Smart City

Tencent
Medical Imaging

iFlyTek
Smart Audio

SenseTime
Smart Vision

Yitu
Vision Computing

MiningLamp
Smart Marketing

Huawei
Soft/Hardware

Ping’An
Smart FInance

HikVision
Video Perception

JD
Smart Supply chain

Megvii
Image Perception

360
Cybersecurity

TAL
Smart Education

Xiaomi
Smart Home

China Mobile
Smart Network

AIROHIT
Smart Farming

CloudWalk
Aud-vis interaction

AISpeech
Linguistic Computing

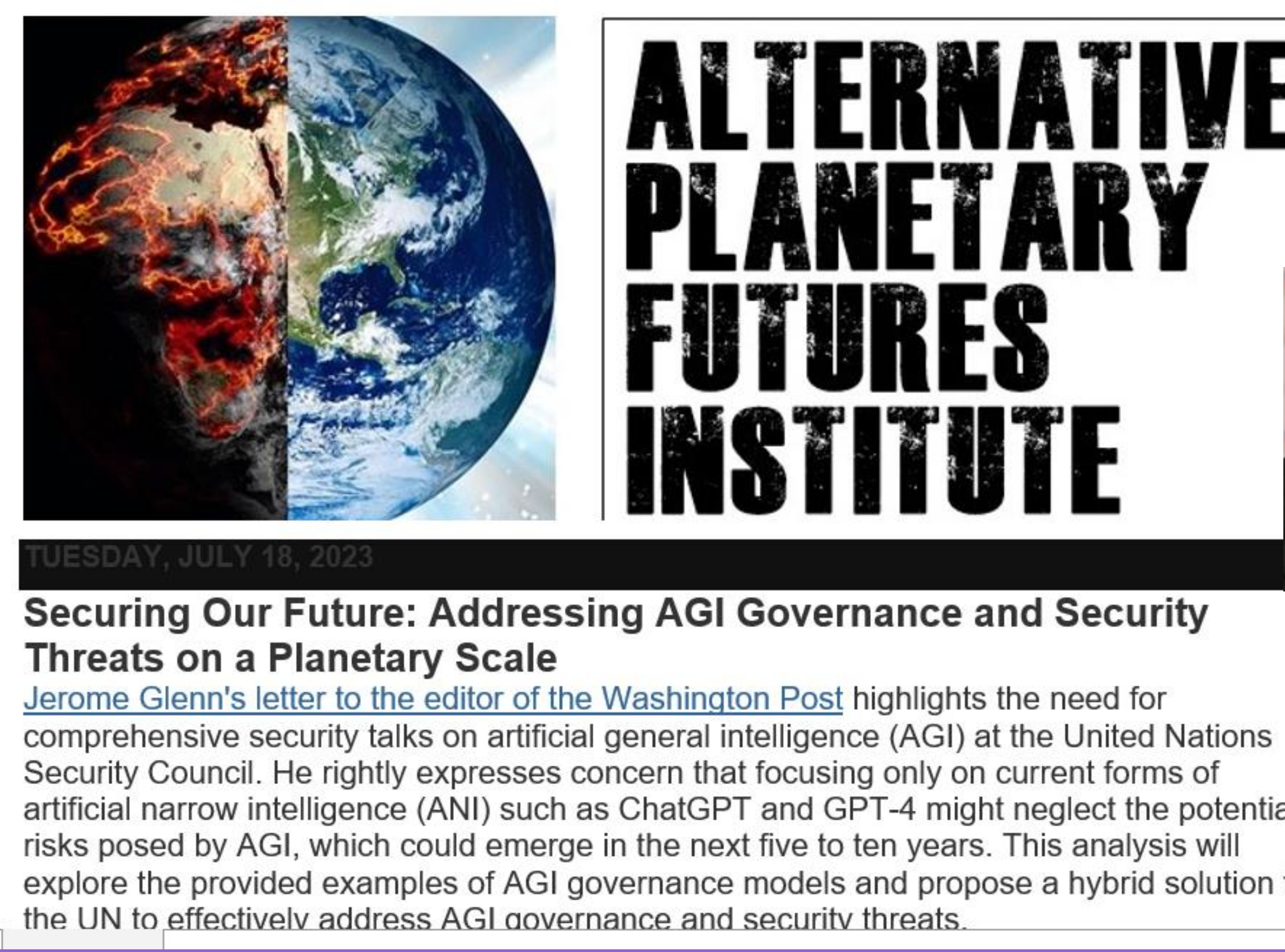
CloudMinds
Cloud Robotics

AI Open Innovation Platforms

open access to data, toolkits, libraries, frameworks, computing resources, and sometimes competition

The Real Economy

Other Startups and SMEs



ALTERNATIVE PLANETARY FUTURES INSTITUTE

TUESDAY, JULY 18, 2023

Securing Our Future: Addressing AGI Governance and Security Threats on a Planetary Scale

Jerome Glenn's letter to the editor of the Washington Post highlights the need for comprehensive security talks on artificial general intelligence (AGI) at the United Nations Security Council. He rightly expresses concern that focusing only on current forms of artificial narrow intelligence (ANI) such as ChatGPT and GPT-4 might neglect the potential risks posed by AGI, which could emerge in the next five to ten years. This analysis will explore the provided examples of AGI governance models and propose a hybrid solution for the UN to effectively address AGI governance and security threats.

Osmo Kuusi Initiative: Helsinki Node/MILLENNIUM PROJECT -> Letter to Ministry for Foreign Affairs -> UN

Ways to avoid risks related to the transition from the Artificial Narrow Intelligence ANI to the Artificial General Intelligence AGI

Finland suggests that one key theme of the UN Summit of the Future (2024) should be the global control of the rapidly developing AI !

Based on statements of the key experts cited in the MP study, it is highly important to launch UN Agency concerning the AGI. Because USA and China are now key developers of the AI, the Agency could be co-chaired by USA and China. Basic risks of the AGI that the possible UN AGI Agency has to handle concern the transparency and the trustfulness of action recommendations made by the possible AGI. These are also basic challenges of the proper public governance that uses the AI in the informing or delivery of public services. So, national public sector projects concerning the proper use AI can be relevant tools in the global control efforts of the possible AGI. In Finland, the AuroraAI project 2019-2022 focused on public services and discussed the transparency and the trustfulness problems of AI based services.

Possible Trajectories foreseen by the respondents

- **AI models for every modality, for every sector and customized for countries and individuals (rather than for the advertiser’s interest)**
- **Creating embodied agents (e.g., robots) is a path to achieve an AGI**
- **AGI will be like a new species. If it is nonhuman-like, then we cannot predict the collaboration with humans, especially if it is multidimensional or creates multidimensional reality.**
- **Humanity might be a passing phase in the evolution of intelligence**
- **Incompetent AI that makes mistakes – That’s the biggest danger and the one that unfortunately gets the least attention, because maybe it's the least obvious**
- **Digital environment might change each 6 months.**
- **Extinction of human species *(‘We are all in this boat together’* Nick Boström)**

UNIVERSITY OF TURKU

Next in AGI development?

- LLMs (Language Learning Models)
- Neuroscience (brain mimicry is the most potential development path)
- Reinforcement learning vs. embed values from the beginning
- Embodied agents (learning world models from observation)

UNIVERSITY OF TURKU

Message from many of the respondents

AGI development itself should not be regulated.

It is more about WHO controls the development and use of AGI.

Arguments of respondents for this message

1. We should not control the development of AI, e.g. researchers.

2. We should try to control who uses AI and for which purposes.

3. We should promote education so that people would be aware and report misuses.

On the other hand:

’We don’t sell hand grenades or nuclear bombs in a supermarket, we have rules for this’ (Max Tegmark)

AGI reflections by respondents

Government resolutions will have a zero impact.

It is dangerous for policy makers to regulate a technology that they don’t understand.

However,

ASI can evolve from AGI without human awareness and understanding.

→ The only way to influence ASI’s relations to humanity is a governance system of the initial conditions for AGI.

UNIVERSITY OF TURKU

Proposed elements for AGI

- *Feedback loops* that allow humans to intervene
- Ability to *pause and evaluate* the AGI when it does something unexpected or undesirable
- A governance model for criteria for knowing *when the AGI should be autonomous* and when it should check with humans.
- Users to be able to *change the behavior of the AI* they're using

The control system (super ego for AI) must operate at a higher level of complexity than the system it is controlling!

Beginning to explore and assess rules for governance of AGI will not stifle its development, since such rules would not be in place for at least ten years.

We should start preparing this before AGI exists – that might not be aligned with humanity’s interest.

→ Focus on governing *the use of AGI*.

UNIVERSITY OF TURKU

Proposed elements for Governance

- **International agreement, not 193 different rules**
  - (UN Convention/Treaty/Agency on AGI)
- **Trust is key! Nash equilibrium – an "optimal strategy" that is beneficial for everyone**
- **Staged approach, not ”one shot to get it right”**
- **Decentralization of power and AGI systems**
- **Transparency and open development**
- **Flexibility and responsiveness to new developments in AGI**
- **Multidisciplinary and collaborative approach**
- **Oversight bodies**
  - audit standards, e.g. inspections, fines, and sanctions
  - certification processes
- **Stress-testing against a wide range of possible future scenarios**
  - (not only the AGI related but also in the broader context, such as climate change, mis- and disinformation, societal upheaval due to increased inequality)

UNIVERSITY OF TURKU

Challenges for Governance

- Any ”global” model or governance can be easily designed to keep status quo for the current leader in AI race
- It is dangerous for policy makers to regulate a technology that they don’t understand
- Defining terms that are universally binding might be difficult, e.g. how to define a risk, can different definitions lead to conflicts? (even defining a risk may be risky)
- A powerful global governance can be taken over
- The real challenge is how to get governments to adopt sufficient governance mechanisms
- The only layer that legislators might be able to control is the hardware substrate, not on the software level
- The main complication within a few years will be AGI’s taking almost everyone´s job, and we will need universal basic income in the developed world in order to prevent a real risk of complete chaos.

How can the use of AGI by organized crime and terrorism be reduced or prevented?

Should we have open-source models?

UNIVERSITY OF TURKU

Threats to Humanity?

UNIVERSITY OF TURKU

Different views on ’Should AGI have rights?’

Yes!

- Rights are part of the social contract → When AI enters, they should be accorded the same rights as human participants in the social contract.
- Assigning rights to AGIs also helps create mutual respect and cooperation between AGIs and humans.

No!

- They would need be sentient and conscious in order to have rights.
- It can lead to a situation where we need to ask them to give us rights.

Guidance or manipulation?

AGI will become very influential of convincing and figuring out what is more convincing to people over time.

Should we use AGI to assist humanity in determining what should be humanity’s further values and objectives? Possible challenges:

- AI giving bad advices, we don’t know how it works
- AGI might “infantilize’ us

Post-AGI future?

- **People need to find meaning and to be enabled for more enlightenment**
- **End of human dominated history (what happens to advertising, news etc.?)**
- **How will AI change ethics? Technology changes society and therefore our values**
- **Should we aim for posthumanist perspective? Humanist perspective doesn’t protect natural world**
- **How to live without digital technology?**
  - If we determine that it is impossible to create an aligned ASI, we might need to remove even small compute capacity in order to prevent human extinction

UNIVERSITY OF TURKU

AI Alignment problem

= a challenge of ensuring that artificial intelligence systems act in accordance with human values and goals.

Should we try to achieve an AGI/ASI that is aligned with human values?

Yes!

- ASI will be so much more powerful than humans → the biggest challenge is to make it *care,*
  - aligning according to e.g. SSIVA (Sapient and Sentient Intelligence Value Arguments)

No!

- An AI system that is uncertain about the human preferences is obliged to ask permission before undertaking potentially harmful actions → not fully autonomous
- What are human values? It is difficult to reach a consensus on universally held values.

→ focus on objective functions that the AI is going to optimize, ie. aligning AI's behavior to objectives rather than values

UNIVERSITY OF TURKU

Questions for discussion in FFRC AI Horizons sensemaking group organised by Dr Maria Höyssä, Senior Advisor, Committee for the Future, Parliament of Finland

For example:

Should AGI have rights? If so, which ones?

How to prevent AGI from using subliminal techniques to manipulate humans for nefarious purposes?

UNIVERSITY OF TURKU

References

The Millennium Project (2023). *International Governance Issues of the Transition from Artificial Narrow Intelligence to Artificial General Intelligence. Report of Phase 1.* [manuscript, unpublished].

Glenn, Jerome (2023). Zero-Sum Power Politics vs. Synergetic Politics for Human Security. Cadmus 5(2): August 2023.

Glenn, Jerome (2023b). Reducing Threats from Future AGI systems. Presentation at the 3rd Stanford Existential Risks Conference – From Global Catastrophes to Existential Risks: Intersections, Reinforcements, and Cascades. 22nd April 2023.

UNIVERSITY OF TURKU

References cont.

Heinonen, Sirkka; Maree, Burgert; Karjalainen, Joni; Sivonen, Risto; Taylor, Amos; Viitamäki, Riku & Pättikangas, Paula (2023b). Flourishing Urban Futures to Overcome Polycrises – Roadmap for Resilience 2050. FFRC eBooks 4/2023. http://urn.fi/URN:ISBN:978-952-249-592-1

(forthcoming, Millennium Project Special Sessions 2022 & 2023 with Jerome Glenn reported here)

Kuusi, O., & Heinonen, S. (2022). Scenarios From Artificial Narrow Intelligence to Artificial General Intelligence—Reviewing the Results of the International Work/Technology 2050 Study. World Futures Review, 14(1), 65–79. https://doi.org/10.1177/19467567221101637

UNIVERSITY
OF TURKU

Comments from the Discussion/FFRC AI Group on MP Study1/3

- **Selected interviewees were very heavily male-dominated, and mostly from US (or China)**
  - Would people from Global South or Nordic welfare states bring different views in the debate?
  - An analysis by respondents' backgrounds would make interesting reading
- **Interesting/alarming to hear that in many of the respodents’ view there is no desire to regulate the developmet of AI, as it is the dominant view in many fields and in public debate.**
- **Can politicians act on AI – or are they bystanders like us?**

UNIVERSITY OF TURKU

Comments from the Discussion/FFRC AI Group on MP Study 2/3

- What is the ecosystem of AI – what does it mean in our whole digital transformation?

- In the report, there were many mentions of treaties and agencies such as IAEA. However, setting up such an agency will be much more challenging than the IAEA. One difference is that in the case of nuclear technology, the concerning use (atomic bombs) was not theoretical or under development – it existed, and several countries had it. It is harder to agree on an international definition of ethics for the use of AI (which is still emerging).

UNIVERSITY
OF TURKU

Comments from the Discussion/FFRC AI Group on MP Study 3/3

- **Interesting point of contrafactuals as causal model of the world for AGI (communicate knowledge by using counterfactuals)**
- **Metaphors as the basis for the human intelligence**
- **What narratives could be elaboratored? For AI and AGI?**
- **The origin of AI more should be thoroughly probed**
- **Dan Brown's Origin and other sci fi literature could be explored and analysed**

UNIVERSITY OF TURKU

Two research projects funded by the Academy of Finland (now Research Council of Finland) at FFRC with the Millennium project as a collaborating network

1) RESCUE Real Estate and Sustainable Crisis Management in Urgan Environment

Karl-Heinz Steinmuller (MP German Node) as member in the International Advisory Board

https://www.rescue-finland.com/

https://www.utu.fi/en/university/turku-school-of-economics/finland-futures-research-centre/research/rescue

2) T-WINNING SPACES Winning Spatial Solutions for Future Work, enabling the double twin transtion of digital/green and virtual/physical transforming our societies by 2035

Jerome Glenn as member in the International Advisory Board

Mara di Berardo as special affiliate expert (MP Italian Node)

https://research.aalto.fi/fi/projects/t-winning-spaces-2035-toivonen

https://www.utu.fi/en/university/turku-school-of-economics/finland-futures-research-centre/research/t-winning-spaces-2035

Contact: Sirkka Heinonen

UNIVERSITY
OF TURKU

Thank You!

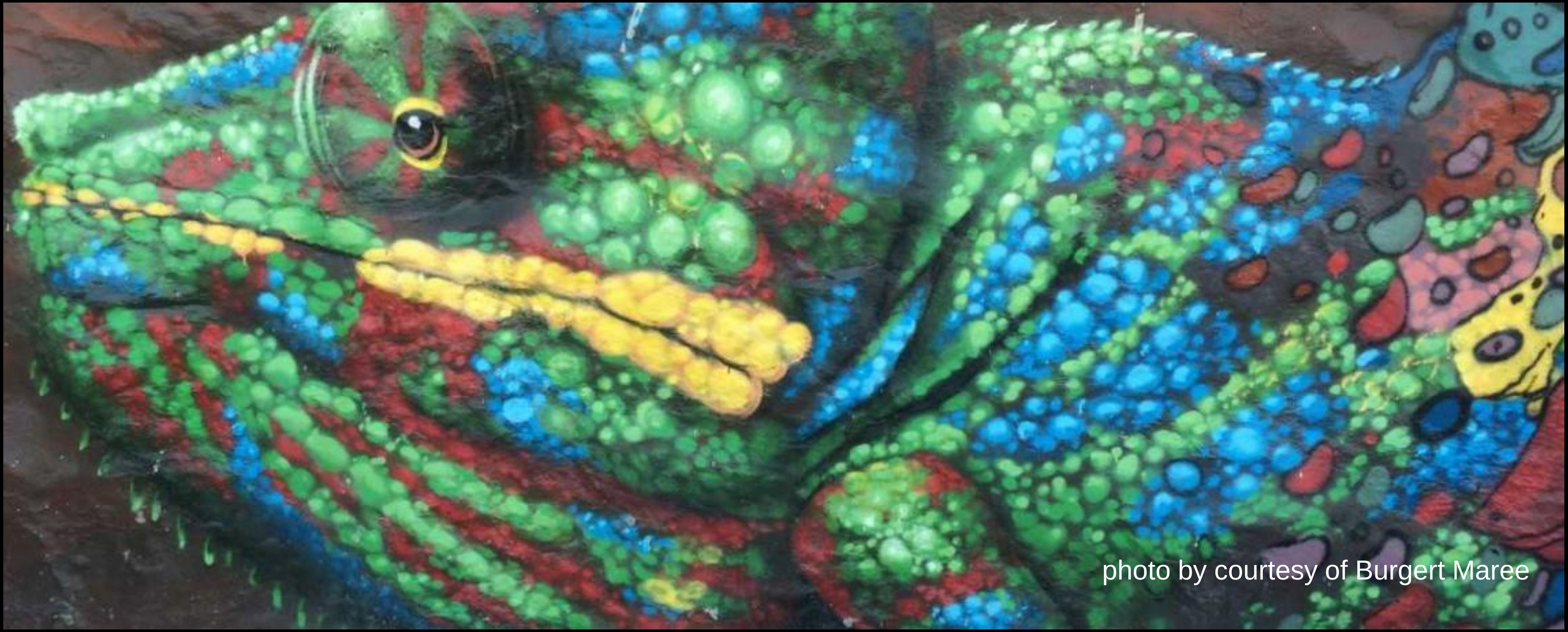


photo by courtesy of Burgert Maree

sirkka.heinonen@utu.fi

UNIVERSITY OF TURKU