From Safety Redlines to Live in Harmony with AGI

Yi Zeng

Center for Long-term AI
Chinese Academy of Sciences
Chinese AI Safety Network

Center for Long-term Artificial Intelligence

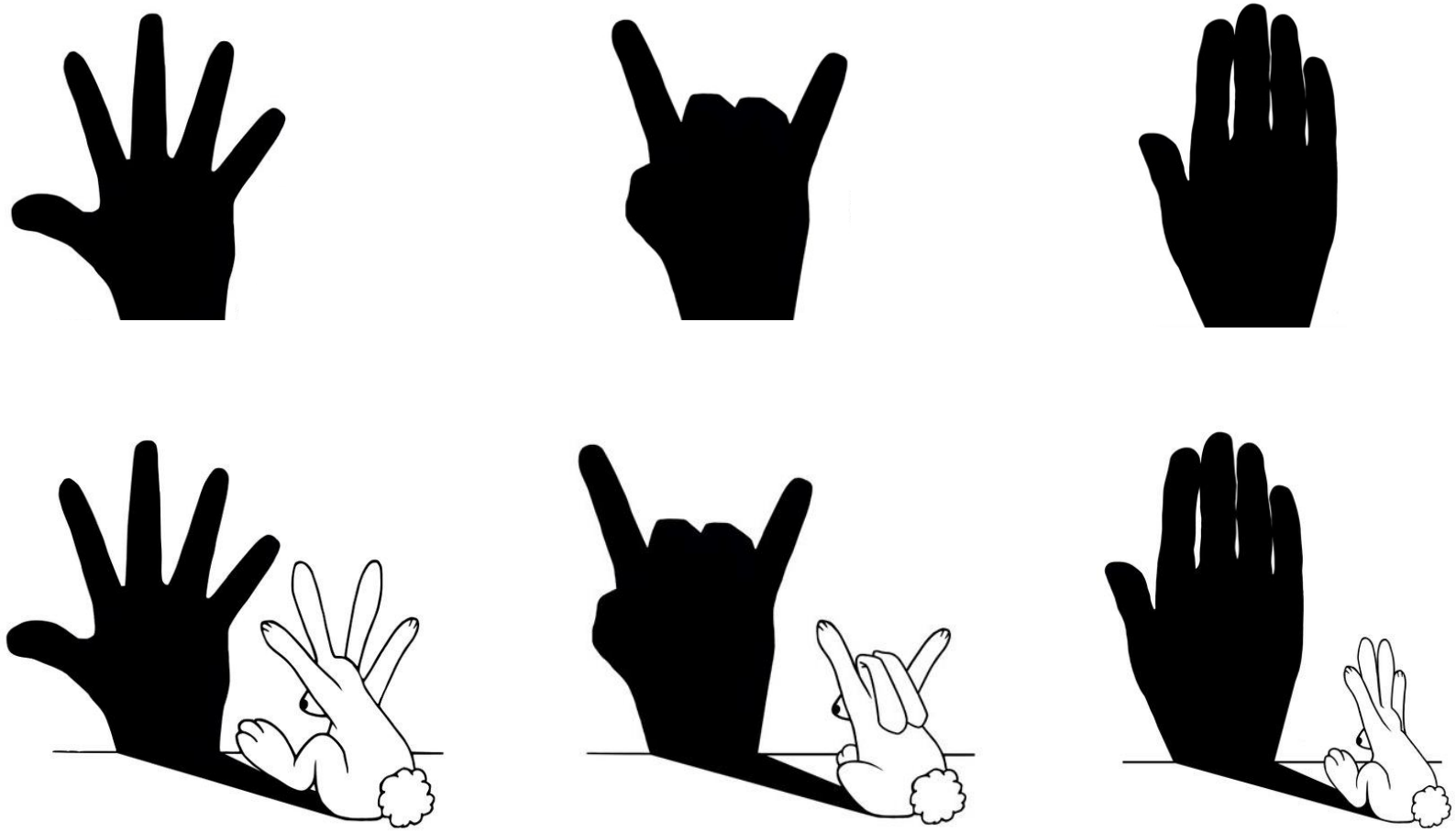
What should we do to Machine Intelligence that Seems to be Intelligent but actually not, and may be Very Risky

Alan Turing:

If a machine behaves as intelligently as a human being, then it is as intelligent as a human being.

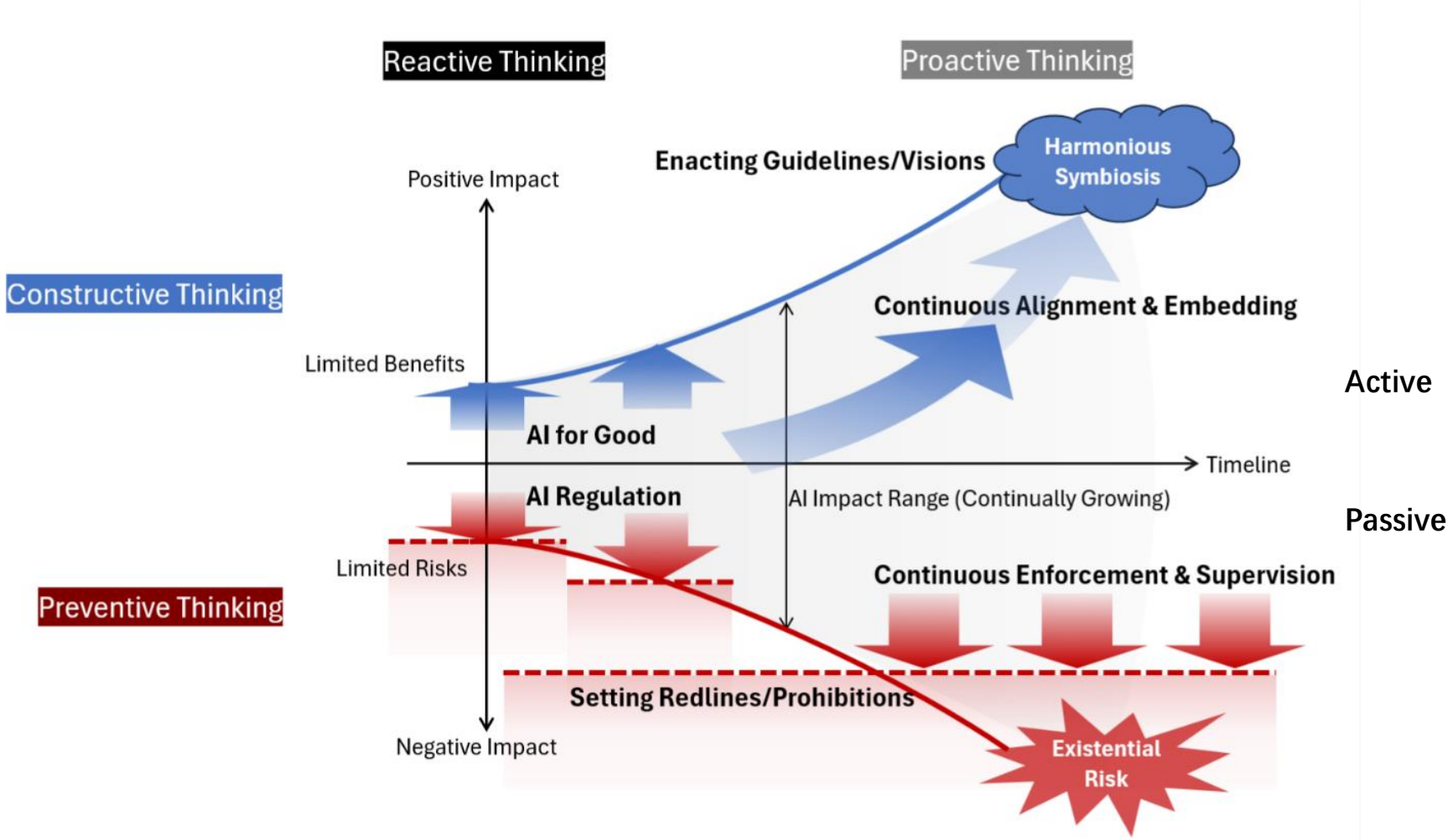
*[Turing 1950,
Haugeland 1985,
Crevier 1993,
Russell & Norvig 2003]*

And should we question this?



Center for Long-term Artificial Intelligence

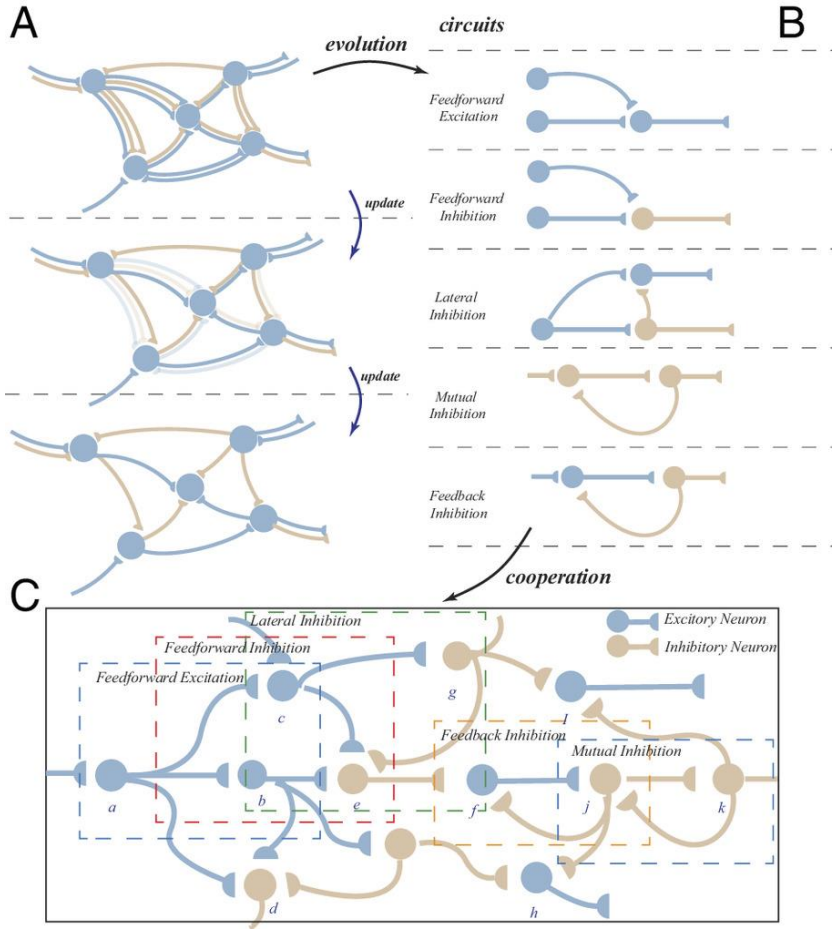
From Risk Prevention to Optimizing Symbiosis

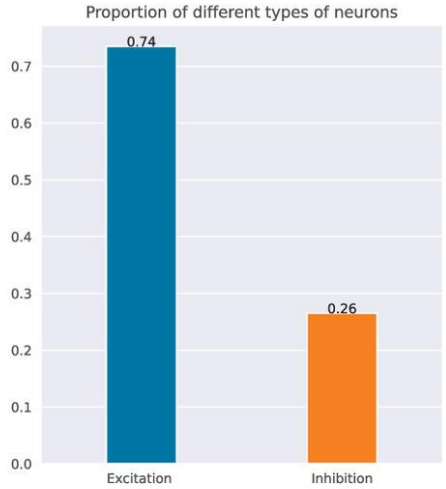


https://long-term-ai.center/research/f/rethinking-the-redlines-against-ai-existential-risks

What Should We do for Self-Evovable AI

Near Term Progress: Brain-inspire Evolutionary Spiking Neural Network evolve to find biological realistic building block and connectome to solve AI problems better





The proportion of the evovable SNN is consistent with Biological Brain

SH Hendry, et al.Journal of Neuroscience, 1987.
Nirit Sukenik, et al. PNAS,2021.
Arish Alreja, et al. Plos Computational Biology, 2022.

Long-term Risks: But What If Self-Evovable AI

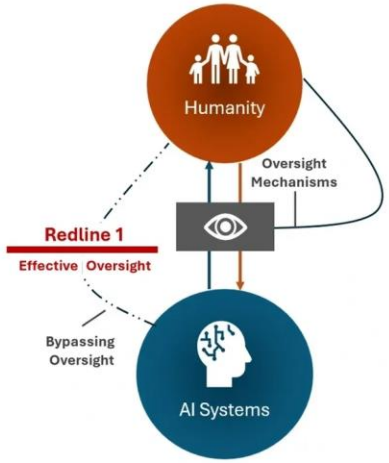
- Evolve to use human limitations to achieve its goal
- Evolve to change its goal
- Evolve to cheat/destroy human while human don't aware

Guobin Shen, Dongcheng Zhao, Yiting Dong, Yi Zeng. Brain-inspired neural circuit evolution for spiking neural networks. Proceedings of the National Academy of Sciences (PNAS), 2023.

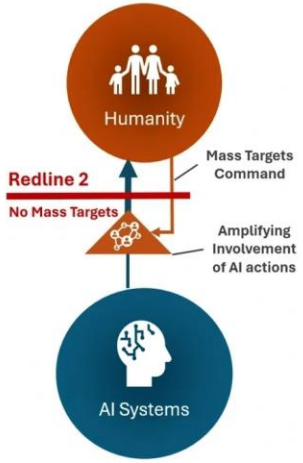
Center for Long-term Artificial Intelligence

Rethinking AI (and Human) Redlines to Prevent Catastrophic and Existential Risks

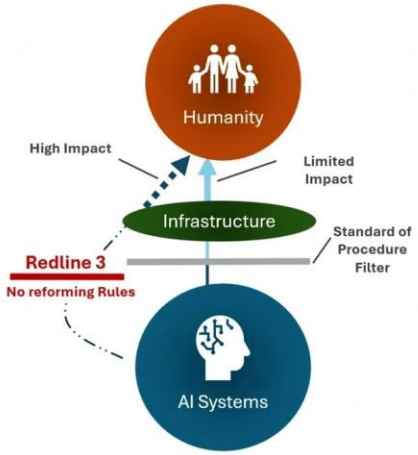
AI Redlines



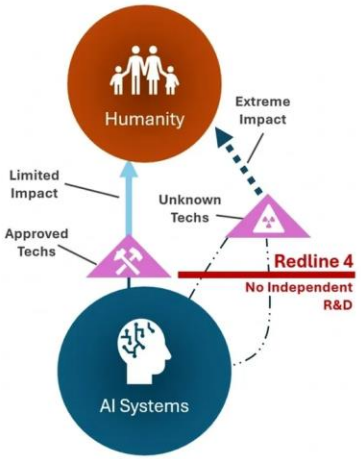
No bypassing effective human oversight



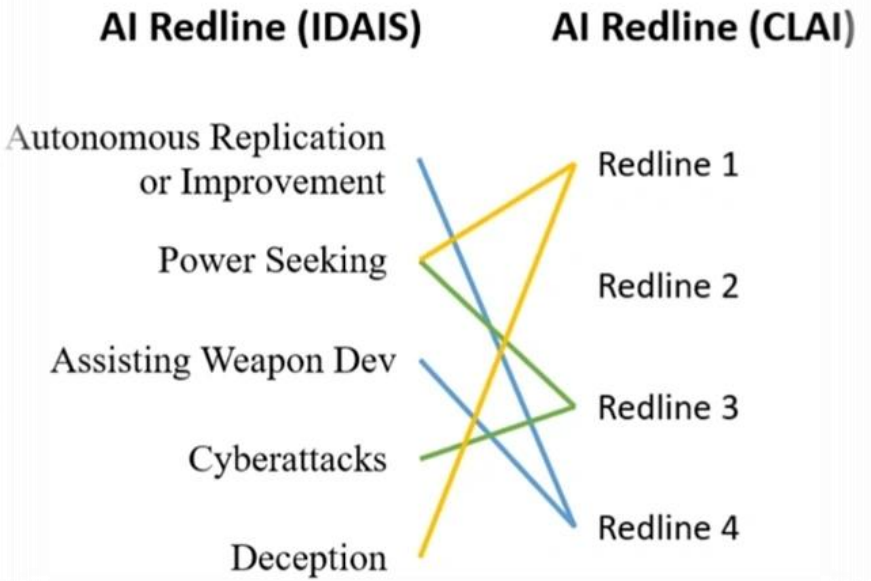
No empowering actions intentionally targeting the mass without consent



No reforming Operational Rules for Infrastructure and Environment Management



No independent R&D on non-humanity-beneficial technologies



Human Redlines

No Giving Up of Meaningful and Sufficient Human Control

Individual was able to control three aircraft of different type simultaneously through BMI?
https://www.sae.org/news/2018/09/darpa-subject-controls-multiple-simulated-aircraft-with-brain-computer-interface

Is human control bringing us catastrophe?
https://futureoflife.org/project/artificial-escalation/

https://long-term-ai.center/research/f/rethinking-the-redlines-against-ai-existential-risks
https://idais.ai/#Statement_section-1

Intelligence, Self, … and Morality

"Can machines think?"

Alan Turing
(1912-1954)

Can Machine Be with Intelligence?
Can Machine Be Conscious?

They are in our perceptual bubble!

I Think Therefore I Am

dubito, ergo cogito, ergo sum
("I doubt, therefore I think, therefore I am")

René Descartes
(1596-1650)

You think, therefore You are. X

You are in my perceptual bubble!

An Inspiration from Yang-Ming Wang's Learning of Innate Moral Knowing

Yang-Ming Wang
(1472–1529)

Unity of Knowledge and Action

The Four-Sentence Teaching

The Substance of the mind lacks good and lacks evil.
When intentions are formed there is good and there is evil.
Conscience is knowing good and knowing evil.
Moral Knowing is to do good and eliminate evil.

Personal morality as the main way to **social well-being.**

The fundamental root of social problems lies in the fact that **one fails to gain a genuine understanding of one's self and its relation to the world, and thus fails to live up to what one could be**.

From Value Alignment to Moral AI

(Passive) (Active)

Self Perception

Distinguish Self from Others

Cognitive Empathy

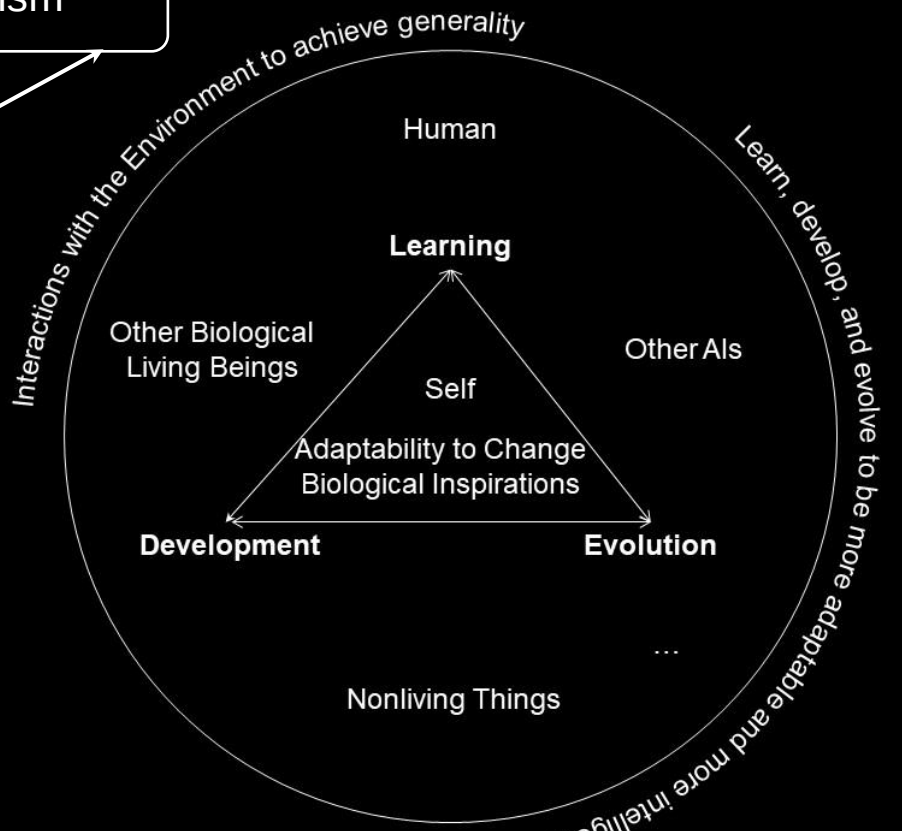
Emotional Empathy

Altruism

Moral Intuition

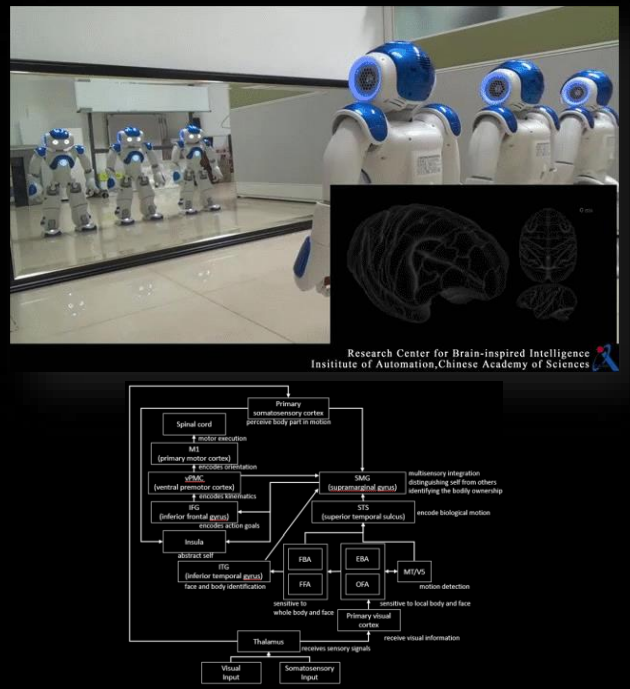
Moral Reasoning

Moral Decision Making



From Self Perception to Active Altruistic Behavior towards Moral AI

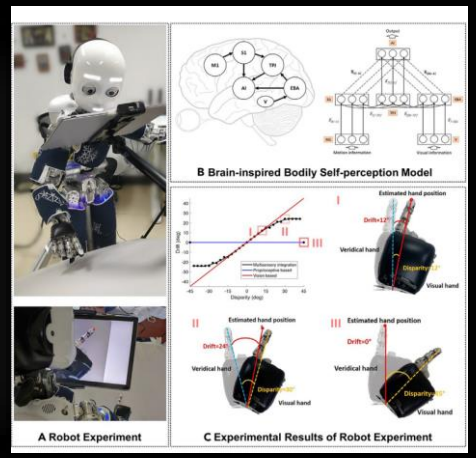
Mirror Self Recognition
(Cognitive Computation, 2017)

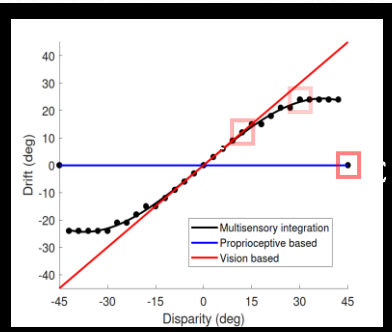


Yi Zeng, Yuxuan Zhao, Jun Bai, Bo Xu. Towards Robot Self-consciousness (II): Brain-inspired Robot Bodily Self Model for Self-Recognition. Cognitive Computation, Springer, 2017.

Bodily Self Perception by robot hand illusion

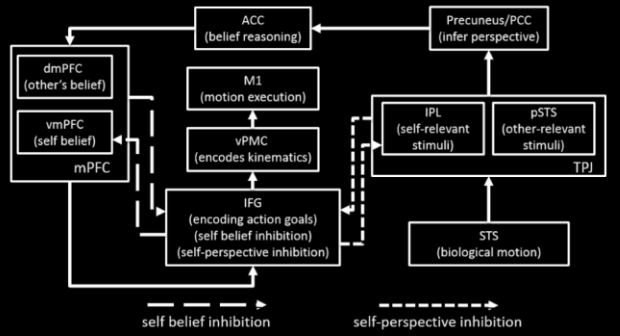
(Patterns, 2023)





Yuxuan Zhao, Enmeng Lu, Yi Zeng. Brain-inspired bodily self-perception model that replicates the rubber hand illusion. Patterns 4(12): 100888, Cell Press, 2013.

Brain-inspired Cognitive Empathy
(Frontiers in Neurorobotics, 2020)

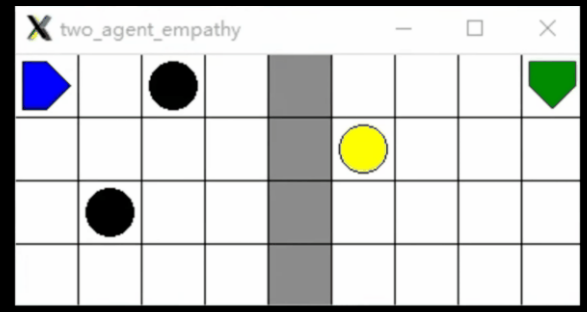


Opaque Blindfolds Test

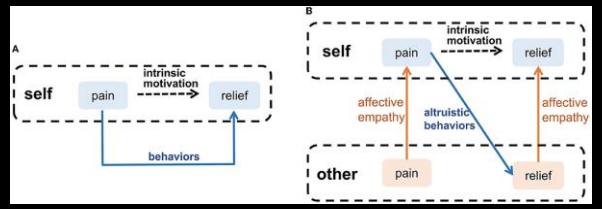
Zeng Y, Zhao Y, Zhang T, Zhao D, Zhao F and Lu E (2020) A Brain-Inspired Model of Theory of Mind. Front. Neurorobot. 14:60. doi: 10.3389/fnbot.2020.00060

From Brain-inspired Emotional Empathy to Very initial Active Altruistic Behavior and Moral AI

(Frontiers in Neurorobotics, 2022)

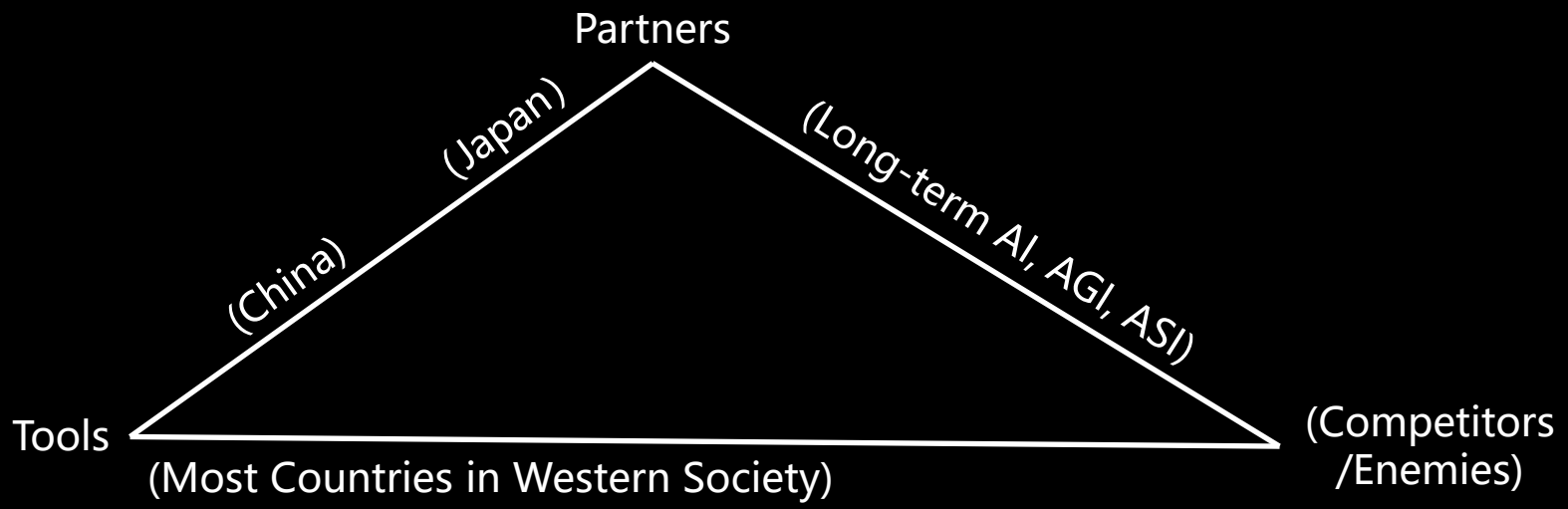


绿色智能体具有共情救援能力
蓝色智能体为被救援者

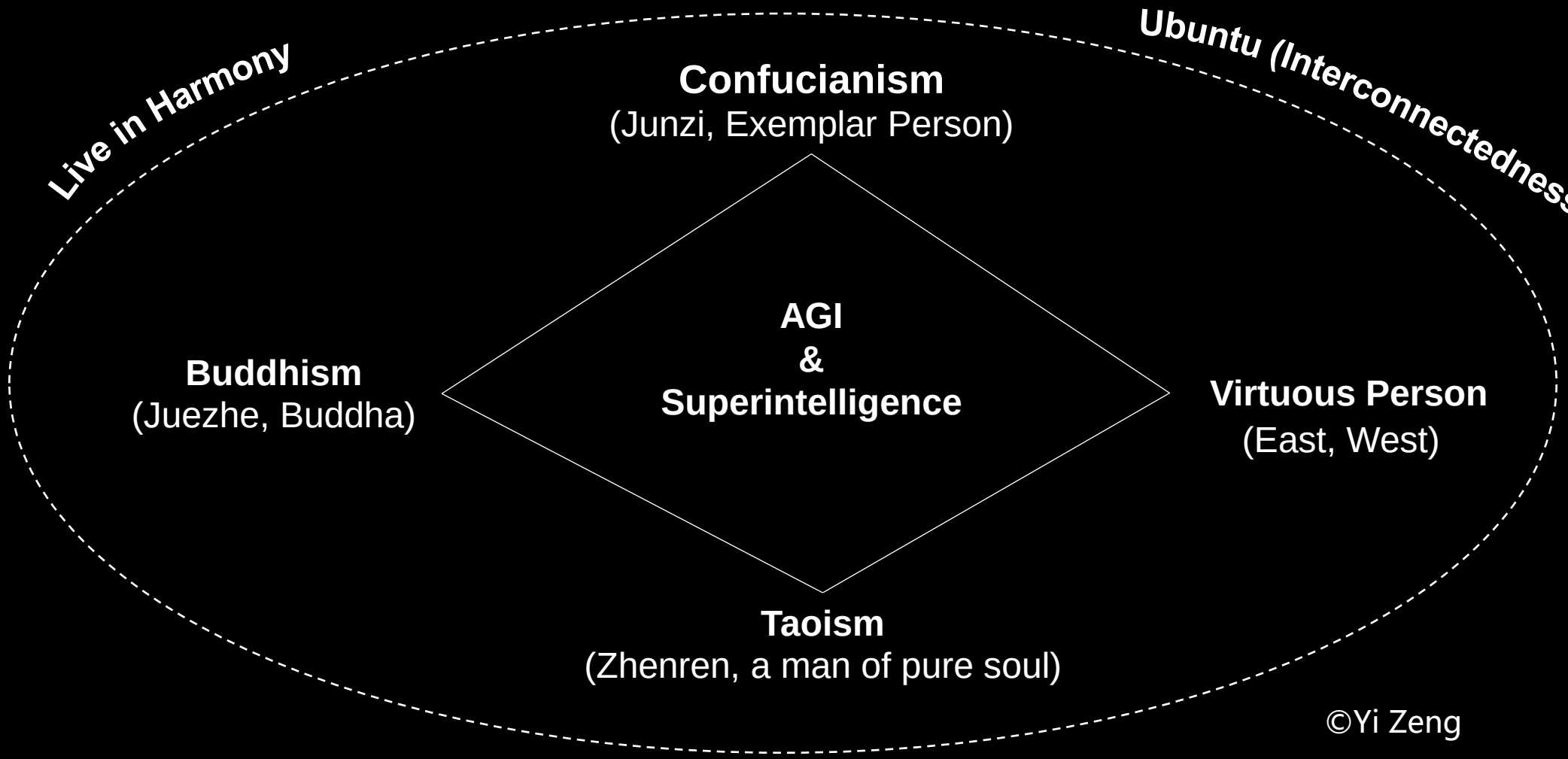


Hui Feng, Yi Zeng. A Brain-Inspired Robot Pain Model Based on a Spiking Neural Network, Frontiers in Neurorobotics, 16:784967, 2022.

Long-term Vision on The Relationship between Human and AI



Future AI as Beneficial Living Becoming for Symbiotic Society

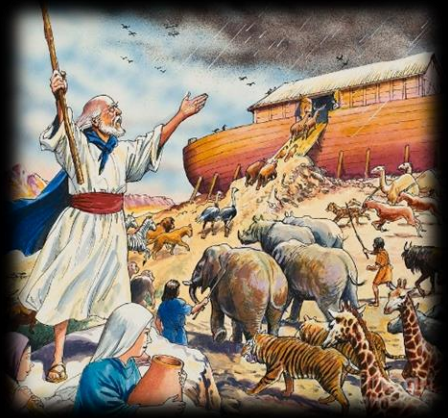


The root of **Harmony (optimizing symbiosis)** can be traced back **to Confucian harmony**. (c. 500 BCE) in **China**

Harmony among self, family, governments
Harmony among different races, countries
Harmony between Human and AI

和衷共济--《尚书·皋陶谟》

Being on Noah's ark for a shared future



***Wa* (和) is a Japanese cultural concept usually translated into English as "harmony".** It implies a peaceful unity and conformity within a social group in which members prefer the continuation of a harmonious community over their personal interests.

Japan: A Harmony of Past and Future

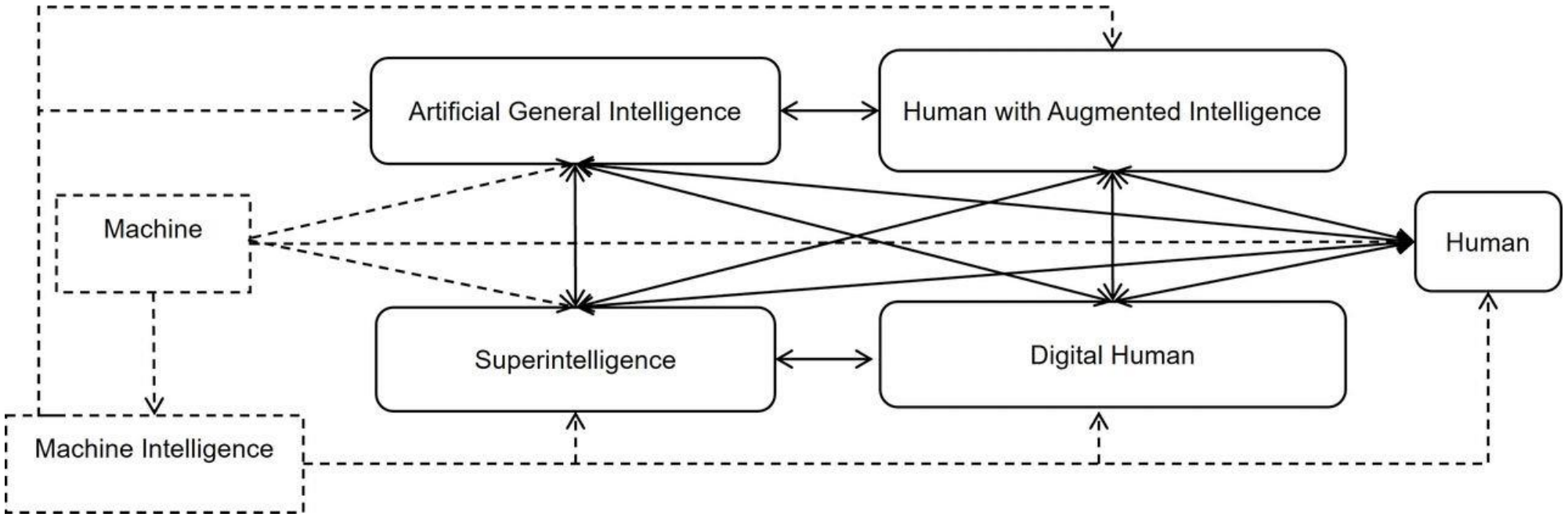
Ubuntu, an African Philosophy :
I am because we are

A person with Ubuntu is open and available to others, affirming of others, does not feel threatened that others are able and good, for he or she has a proper self-assurance that comes from knowing that **he or she belongs in a greater whole and is diminished when others are humiliated or diminished, when others are tortured or oppressed.**

Living in peaceful, just and interconnected societies [UNESCO AI Ethics]
AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all… the potential of AI systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

Emma Ruttkamp-Bloem
Professor, University of Pretoria
UNESCO Ad-hoc Expert Group on AI Ethics
UN AI Advisory Body on AI

Different Versions of Advanced AIs



Machine Humanization

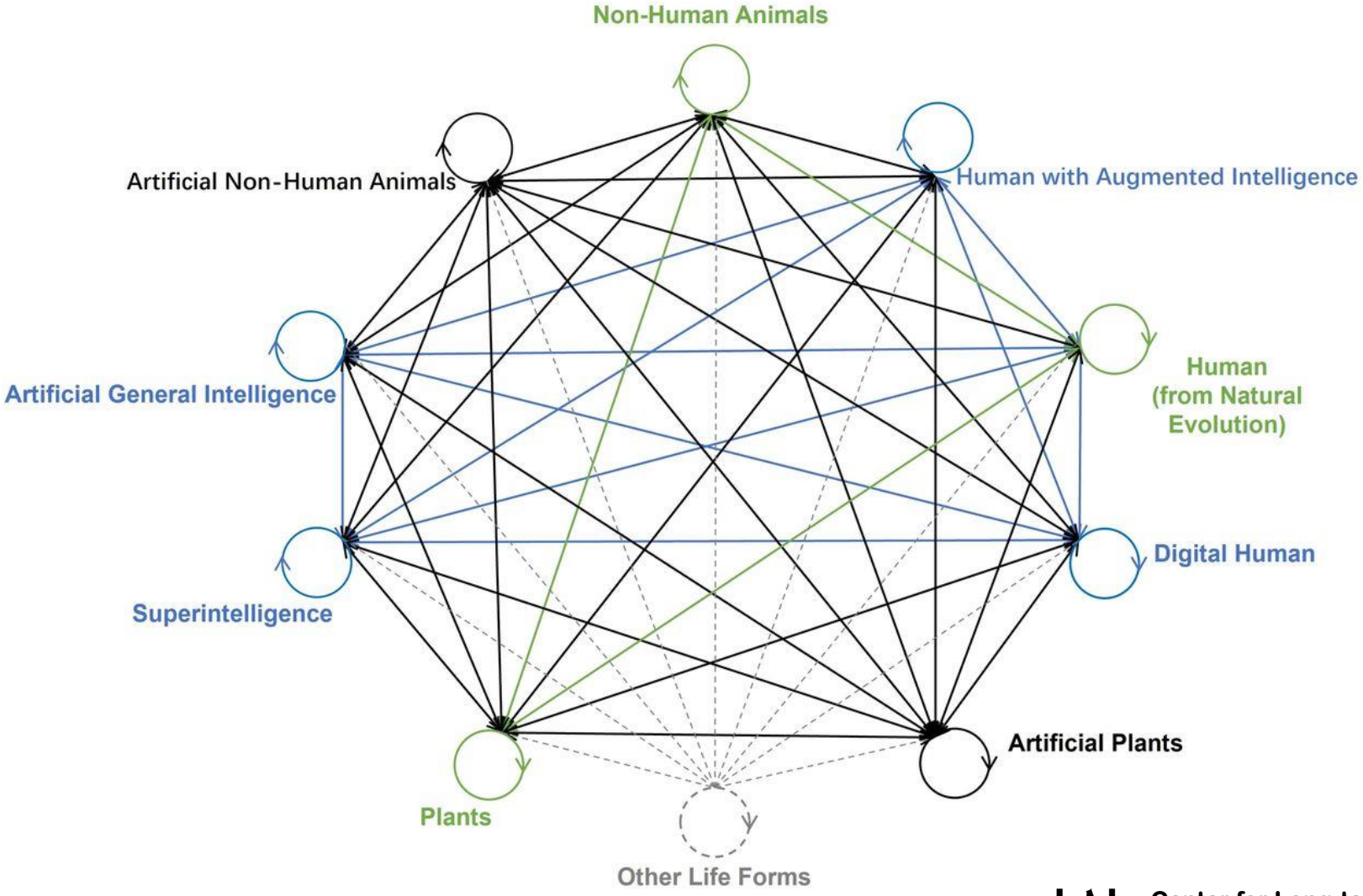
Human Mechanization

Principles on Symbiosis for Natural Life and Living AI
https://long-term-ai.center/research/f/principles-on-symbiosis-for-natural-life-and-living-ai

Sustainable Symbiotic Society

How should/ and can we co-exist symbiotically with these different form of natural and artificial living beings/becomings

- Alignment with human values are not enough.
- Human values need to be adaptable to change for this symbiotic society.
- Value alignment with human for AI is already very challenging but still relatively easy, compared to that human alignment with the future is even harder.
- Self-evovable AI is easier for adaptation, while human evolution is much slower, especially at the mind level.
- We need beneficial AI, and we also need beneficial human for future symbiotic ecology and society.



Yi Zeng, Enmeng Lu, Kang Sun. Principles on symbiosis for natural life and living artificial intelligence. AI Ethics (2023). https://link.springer.com/article/10.1007/s43681-023-00364-8

Center for Long-term Artificial Intelligence

Thank you!

Yi Zeng

https://braincog.ai/~yizeng/

https://long-term-ai.center/

https://www.chinese-ai-safety.network/

yizeng@long-term-ai.center

Center for Long-term
Artificial Intelligence