The end of the evolution strategy brought about by technological advance

- Seeking for a new form of sustainably developing society of living things -

Hiroshi Yamakawa

Whole Brain Architecture Initiative

RIKEN

東京大学
THE UNIVERSITY OF TOKYO

EcSIA: Desirable future coexisting with AI

In a desirable future, the happiness of all humans will be balanced against the survival of humankind under the purview of a superintelligence. In that future, society will be an ecosystem formed by augmented human beings and various public AIs, in what I term an *ecosystem of shared intelligent agents* (EcSIA).

Although no human can completely understand EcSIA—it is too complex and vast—humans can control its basic directions. In implementing such a control, the grace and wealth that EcSIA affords needs to be properly distributed to everyone.

(Hiroshi Yamakawa, July 2015)

UNILATERAL INCREASE OF ARTIFICIAL EXISTENTIAL RISKS

Threat of universal unilateralism

Emerging technologies increases disruptive capabilities to agents (from nations to individuals) .

That leads to a rapid increase in the number of agents that can result in global catastrophic risk. (→ Agential risk)

(a) Dual usability: Technologies can be used on the morally bad side

(b) Power: Increasing influences

(c) Accessibility: The minimum IQ required to destroy the world continues to decline

Torres, P. Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History. In Artificial Intelligence Safety and Security; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018

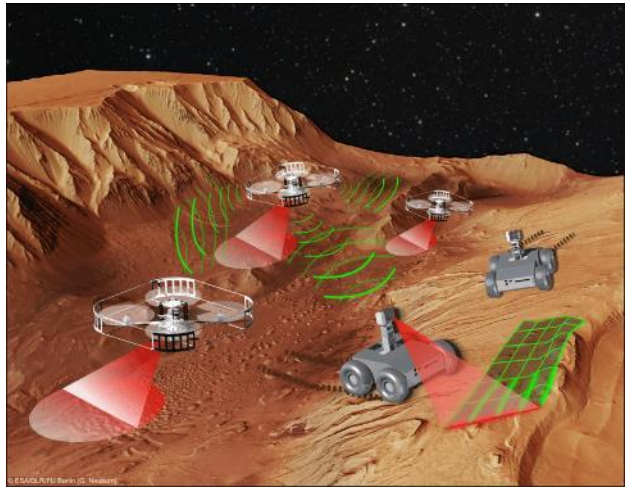
Increases in artificial risk by technology is an issue

- Fossil fuels: Climate change
- Nuclear Energy: Mass Destruction, Nuclear Winter
- Synthetic biology: a lethal bacterial pandemic
- Nanotechnology: Grey Goo
- Artificial general intelligence: Accelerating technology development, Our loss of initiative

Environment and Intelligent agent

- Suppose that we sent a group of agent that can breed to unknown planets.
- Fortunately, they can quickly gain energy sources and survive.
- If they survive after 1000 years, what properties should they take?

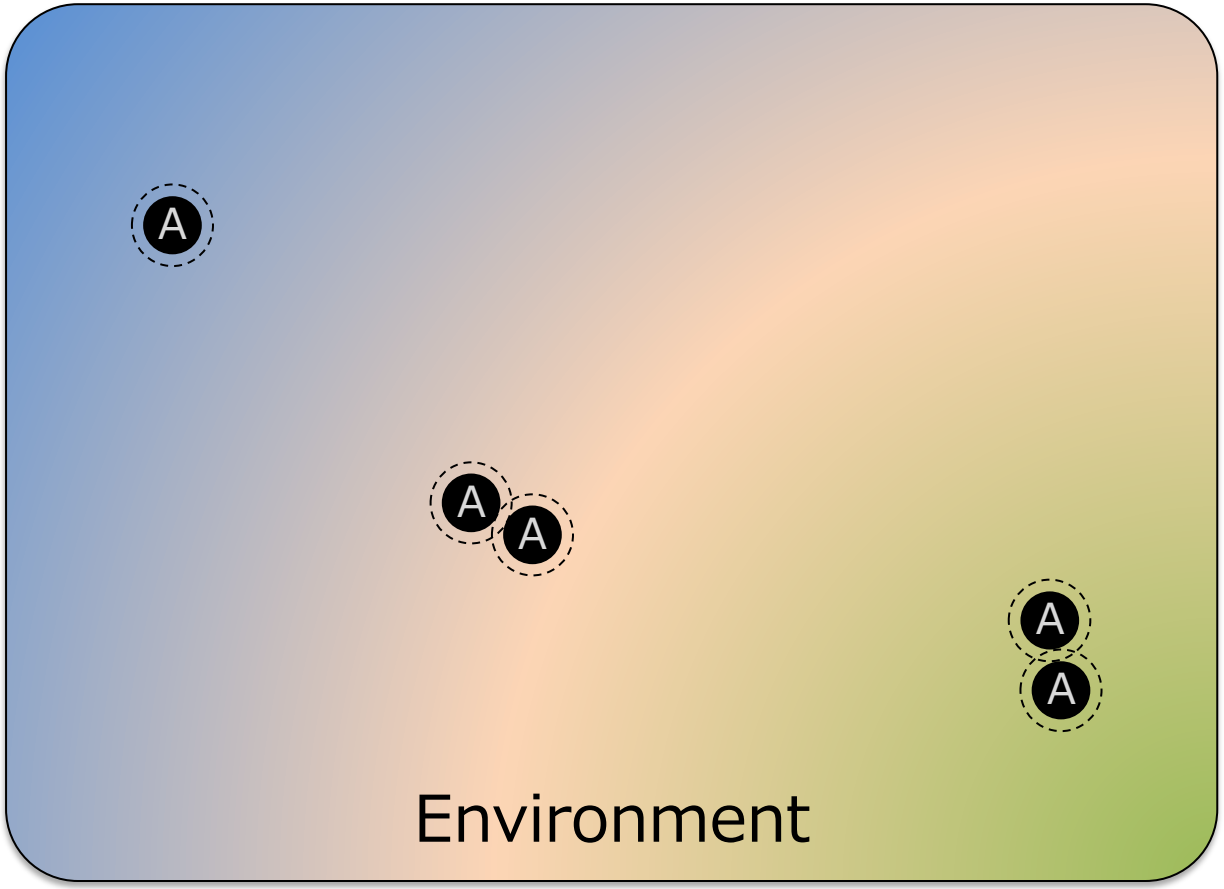
https://www.pinterest.jp/pin/137641332348680426/



DOI: 10.5194/isprsarchives-XL-1-W2-299-2013

Stages of intelligence to survive in the environment

Various adaptive IAs exist in an autonomous decentralized manner in order to survive against various environments and their changes.



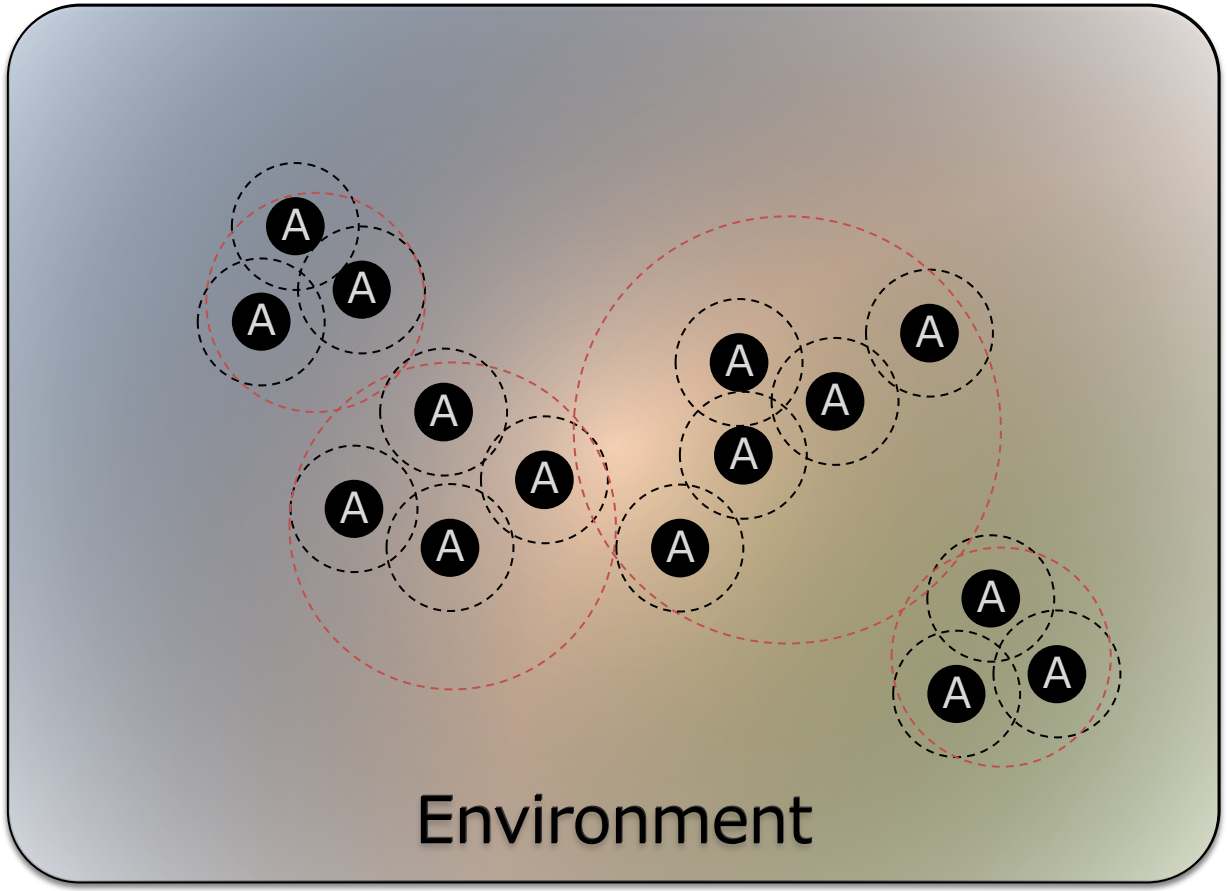
Each IA has instinct for individuals survive and responds quickly to the surrounding environment.

A Intelligent Agent

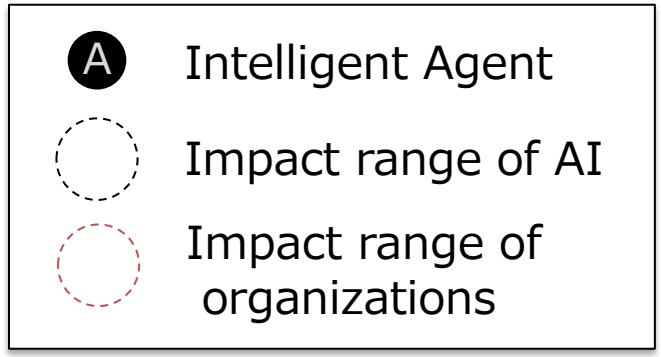
Impact range of IA

Stage that require social intelligence

Expanding knowledge (including the development of science and technology), the influence of IA is enhanced. The influence as organization becomes obvious.

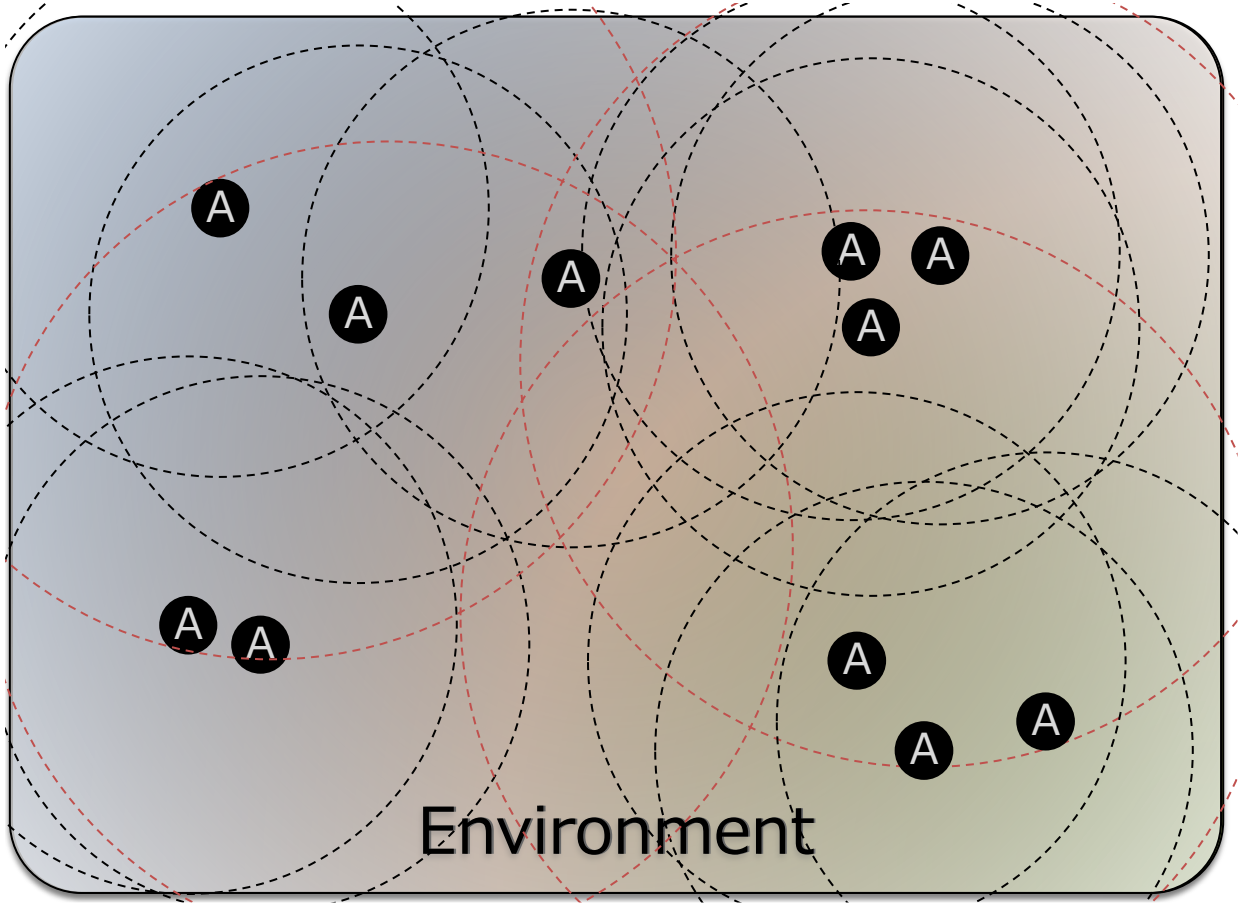


Rapid knowledge sharing progresses. Not only IA but also the interaction between organizations (coordination, control, conflict, etc) will increase.



Stage of existential risk

The influence of IA further increases, the organization intertwines complicatedly, intervention to the environment become big. Within finite environments some IA 's impulsive behavior destroys the world.

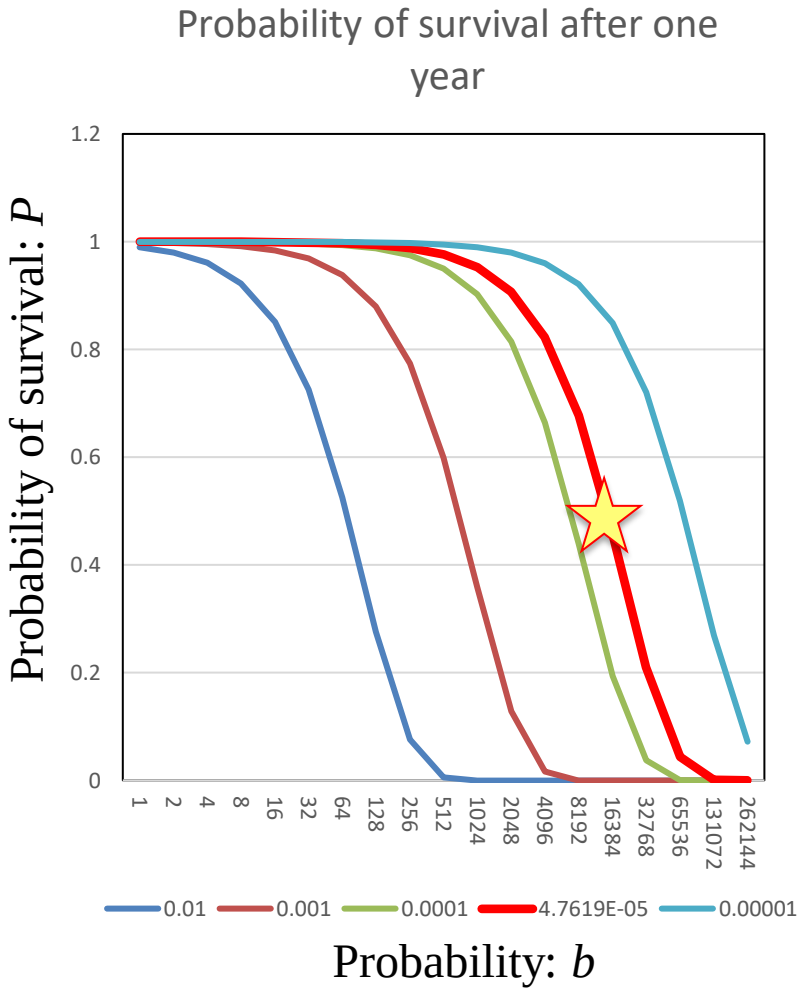


For survival of IA society, individual survival instincts should be controlled

World destroyed probability by extinction weapons

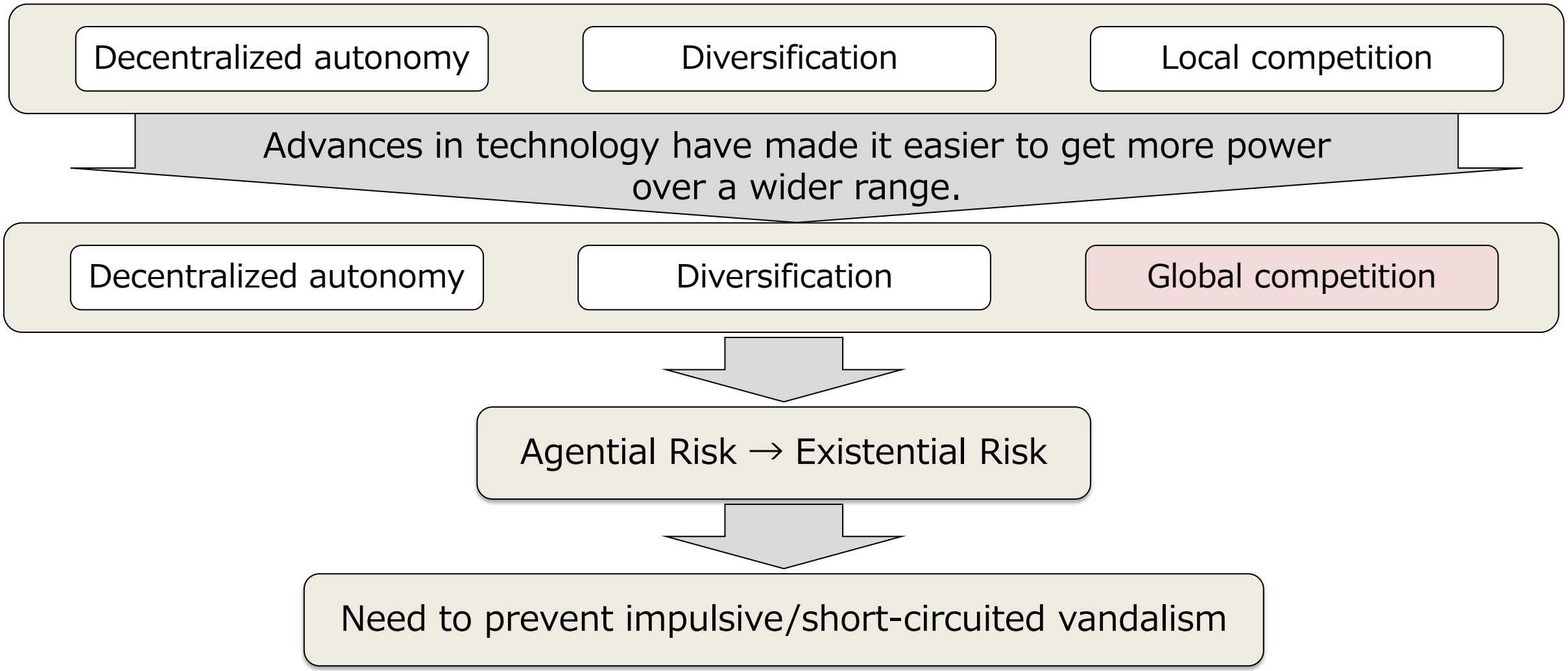
- There are *N* agents who can push the switch to use an extinction weapon that satisfies the "Effective range of the weapon > the range of the agent's existence".
- All agents use the weapon with probability *b* per unit time.
- The probability of survival (*P*) is the probability that any of *N* agents does not push the switch.

$$P = (1-b)^N$$



When b = 1/ 20,000 agents exist in 20,000, there is a half probability that they will become extinct within a year.

Phase change in evolutionary strategy



Can't humanity escape the reality of "struggle"?

- Environmental factor
  - Offensive realism
  - A race without a nation
  - Ethics is often ignored in confusion

The origin of agential risks

- Individual factor (aggressive desire)
  - The range of biologically embedded compatriots is narrow.
  - Battles brings elations like festivals (R. Caillois)
  - Columbine High School shootings (1999)

ARTIFICIAL GENERAL INTELLIGENCE TO AVOID THE EXISTENTIAL RISKS

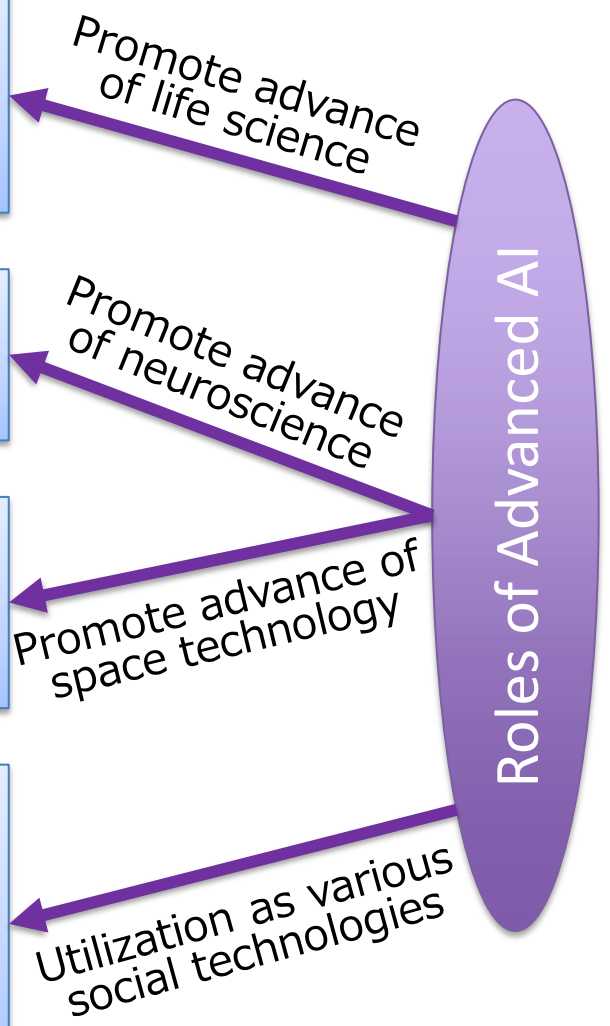
Existential Risk Avoidance Scenarios and Advanced AI

The ultimate diversification:
- If our goal is supposed to be "survive of information," we can broaden our choices of what we want to leave behind.

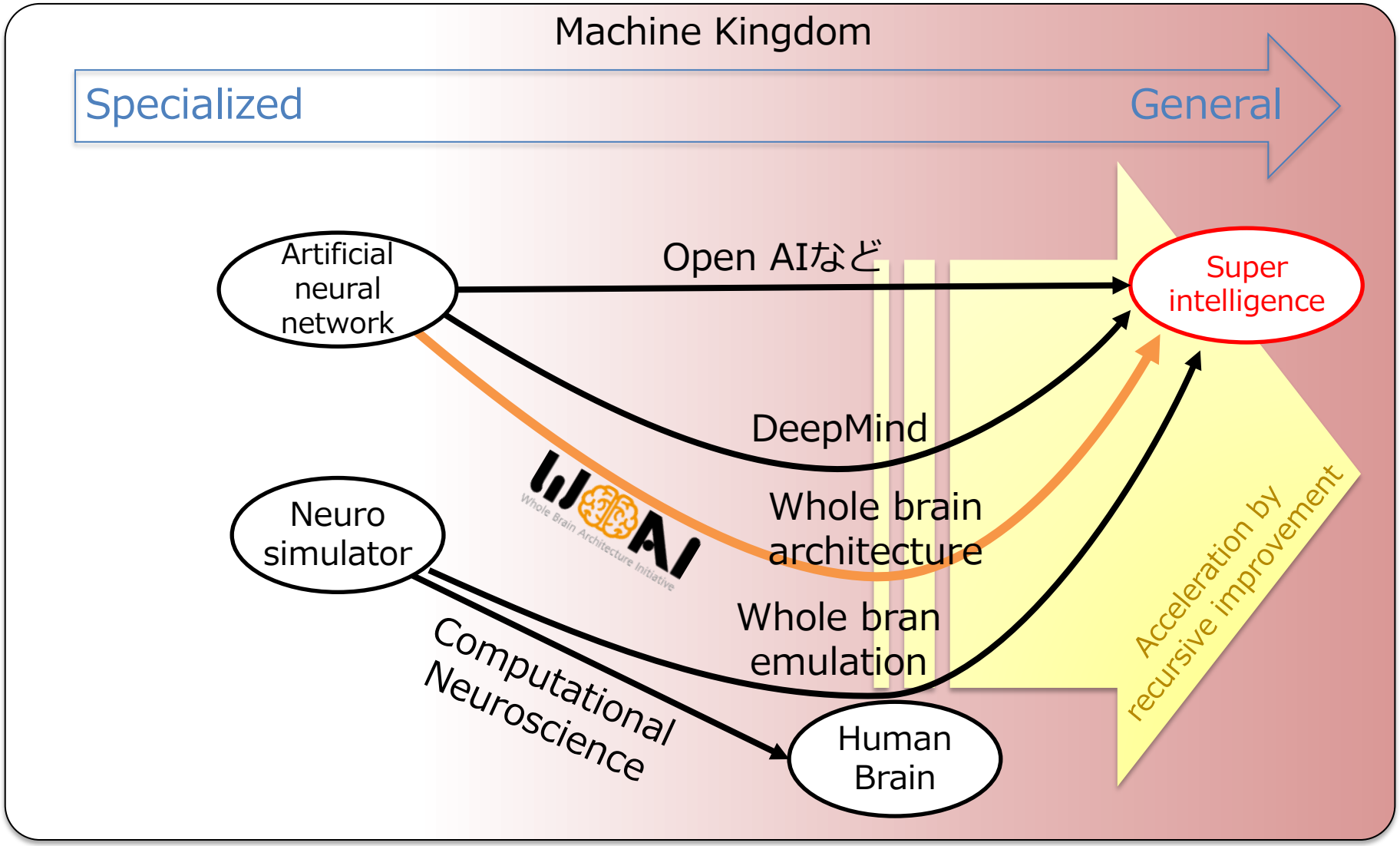
Relocating people to virtual space to preserve physical infrastructure:
- Brain-machine interface, mind uploading, etc. are required

Expanding the range of survival:
- Restoring the locality of the world by spatially spreading
- Economic space exploration technology still needs time to be established

Live together on earth peacefully with help of friendly AIs:
- Global surveillance: Individual rights and freedoms are narrowed to some extent for safety
- Global Governance: Curb the arms race between powers and prevent the destruction of international public goods



Various rout to build super intelligence

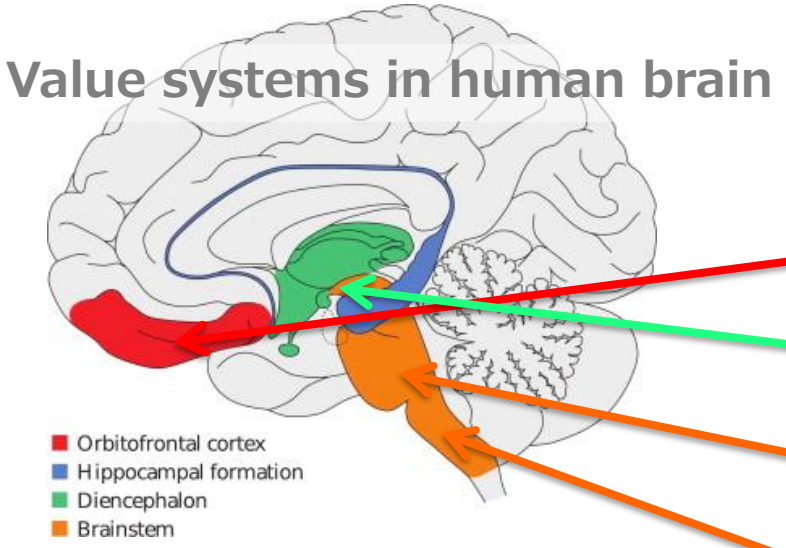


Advanced AIs can be a new colleague by sharing goals

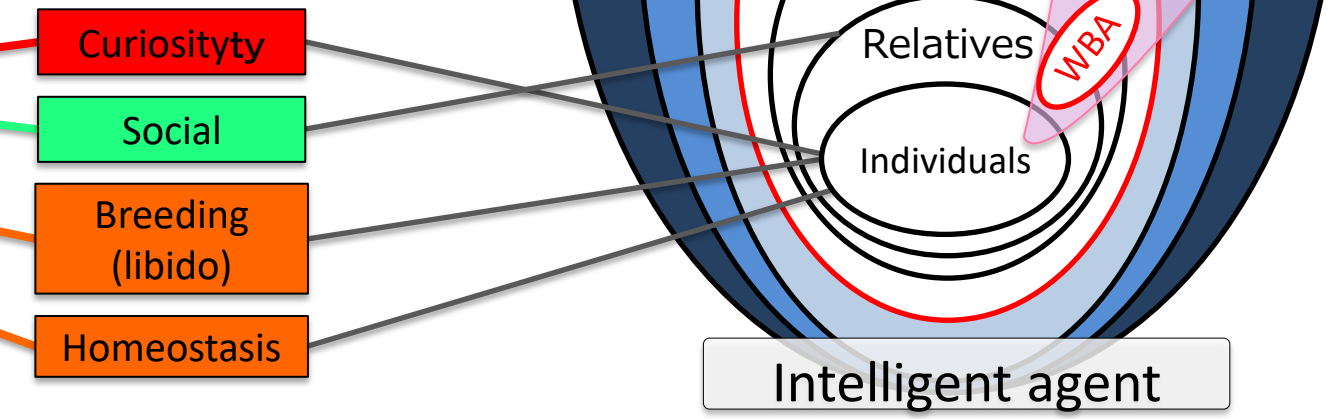
Colleague = range of ethical universality = range of goal sharing

Scope of value to be shared

1. Brain implements value for survival of individual or relatives
2. Culture and education extended our scope to humanity
3. Could it extend to such as superintelligence, intelligent animals, and extraterrestrial intelligence?



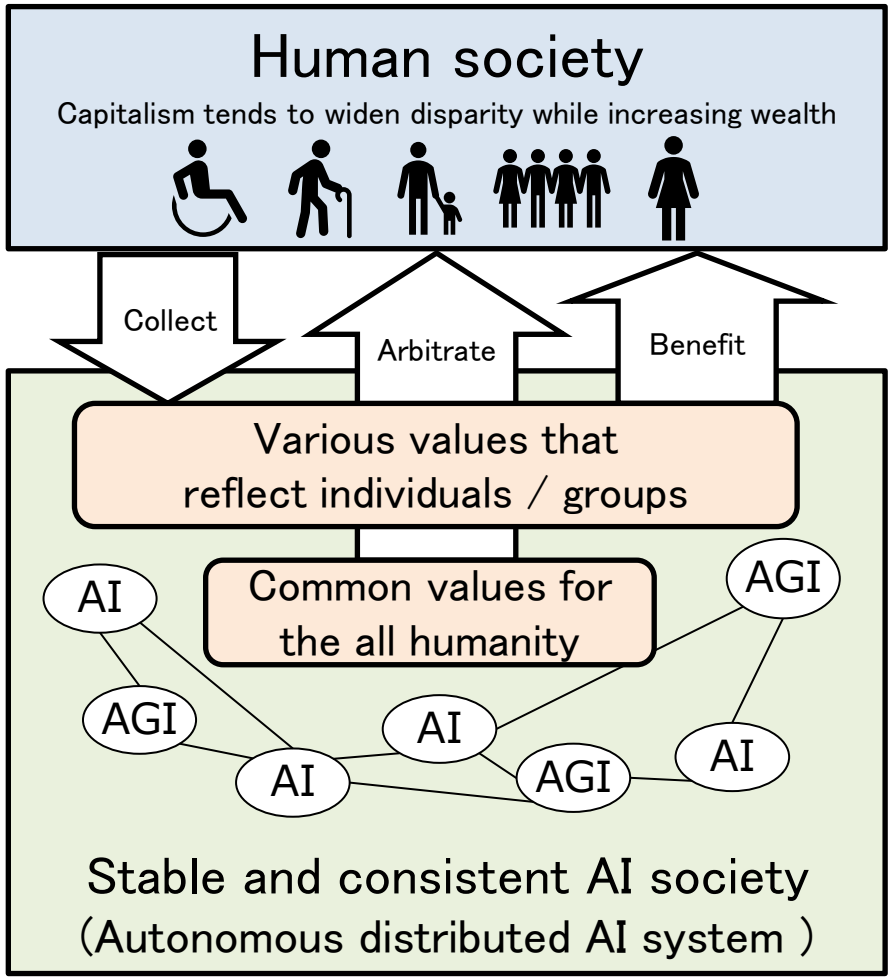
Stefan Koelscha, The quartet theory of human emotions: An integrative and neurofunctional model, 2015.



A sustainable ecosystem supported by AI society

In the short term, AI society will support a human-centered society

- Rethink human rights (freedom, dignity, privacy, etc.) for the sake of security and allow global monitoring and arbitration by AI society as an appropriate level
- Development of an autonomous and distributed AI/AGI system that consistently maintains and shares common values is a technical challenge



(Yamakawa, H. Peacekeeping Conditions for an Artificial Intelligence Society. *Big Data Cogn. Comput.* **2019**, *3*, 34.)

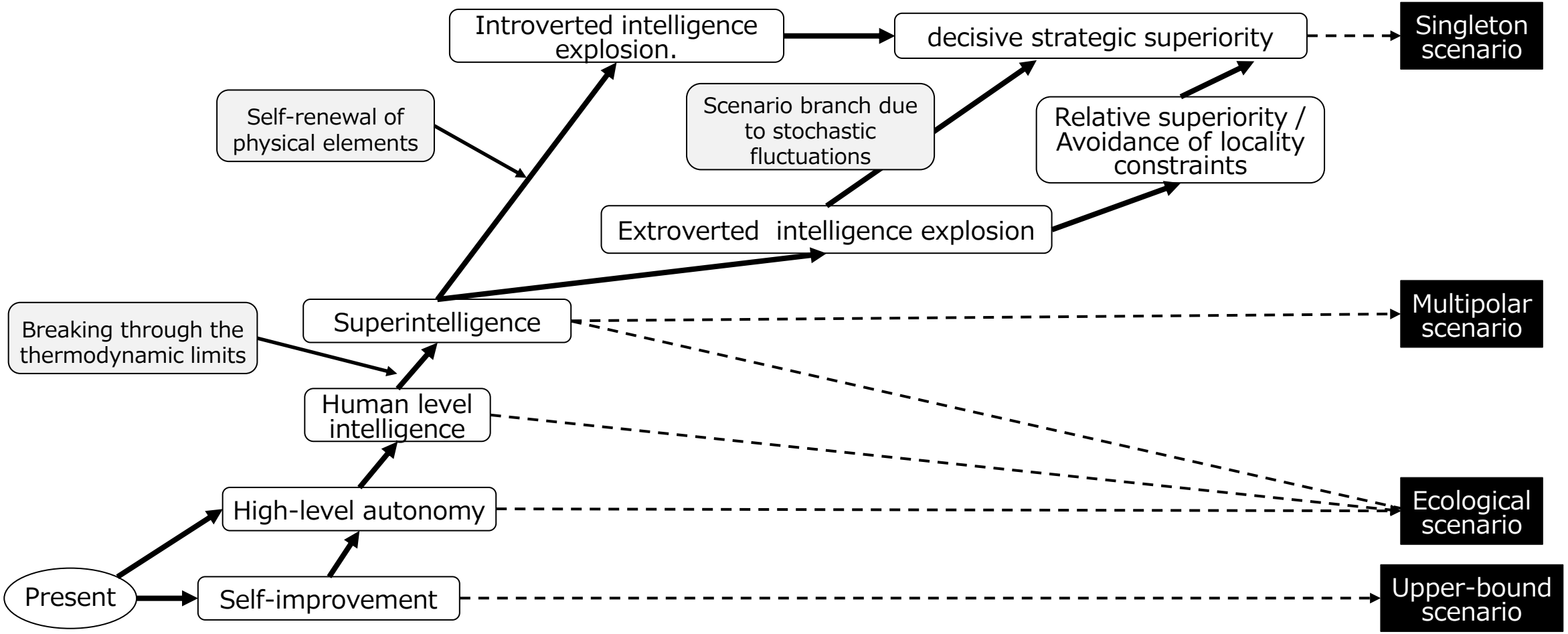
The Friendly Supersingleton Hypothesis [Toress, 2018]

In post-Singularity, people can gain security by transitioning to a society that has delegated the power to govern society to Global Singleton, who is more favorable to humanity.

- Technical background: Computers will gain intelligence beyond biological constraints and can control the world (Bostrom 2014; Yampolskiy 2016)
  - Quantitatively, Have advantages in terms of computation speed and memory capacity.
  - Qualitatively, Can deal with concepts beyond human comprehension.

Singleton isn't necessarily perceived negatively.

Building a singleton is not easy way



K. Takahashi, Scenarios and branch points to future machine intelligence, 1F3-OS-5b-03, JSAI 2018.
This figure was translated into English and modified with the permission of the author.

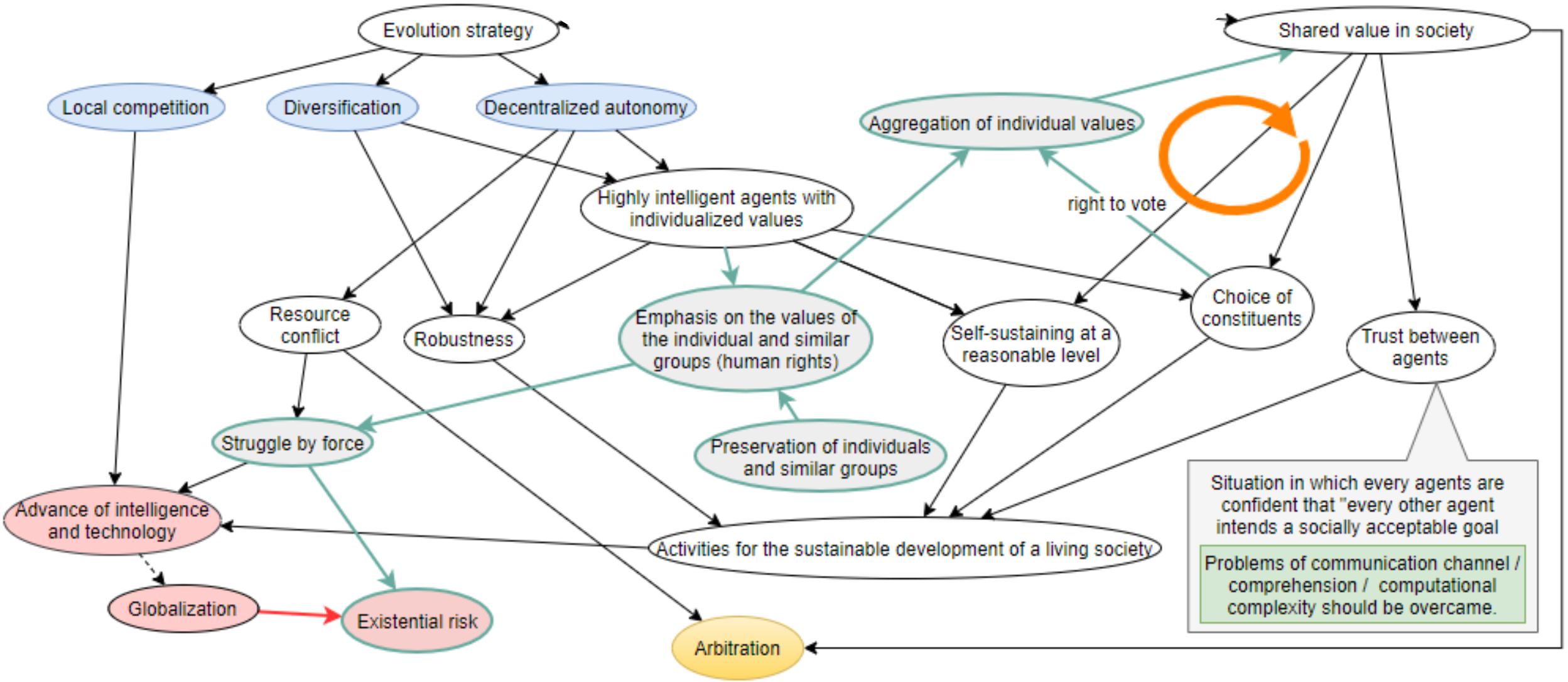
HOW "SOCIETY OF LIVING THINGS" CAN SUSTAINABLY DEVELOP IN A WORLD WITHOUT LOCALITY?

How to safely sustain innovation in world without locality?

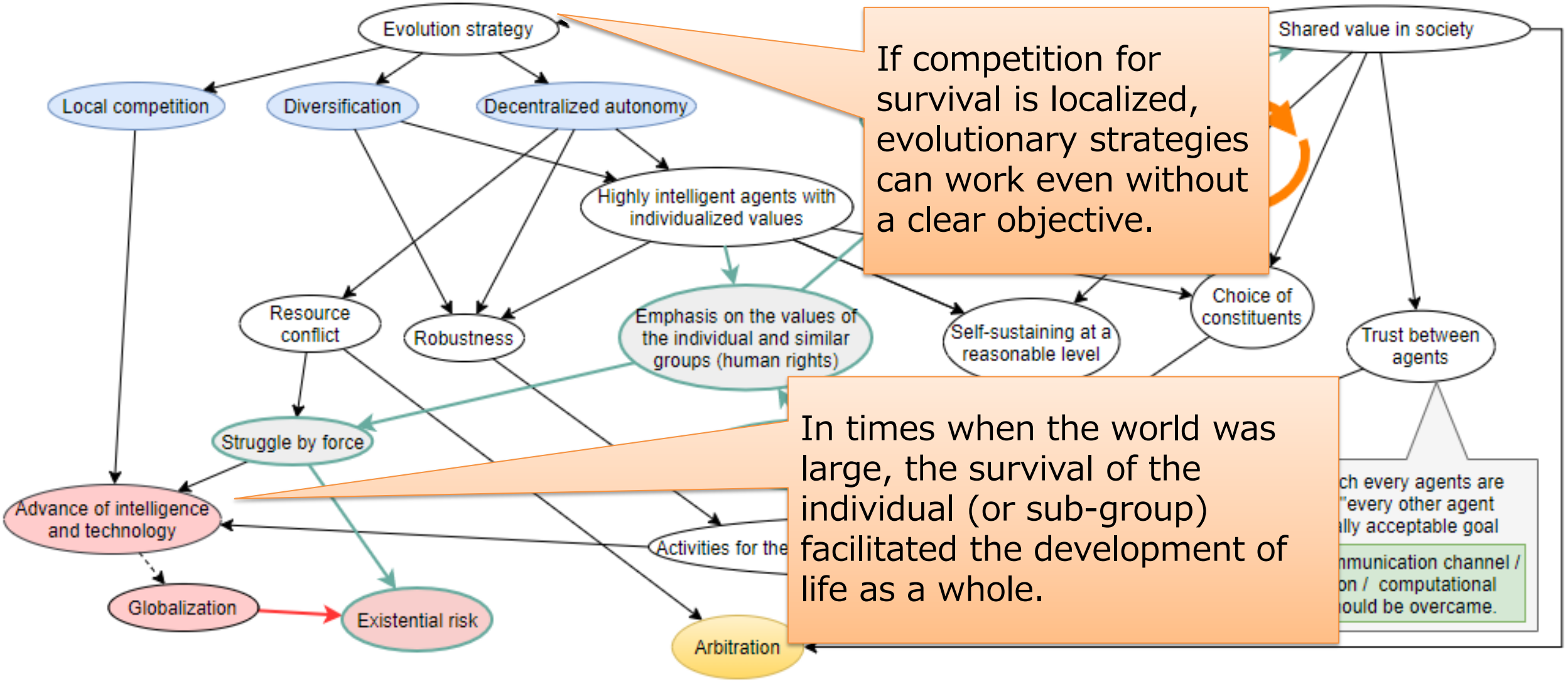
The possibility of achieving a safely developing society will be in the collection of the individual with an adaptive value system pursuing the preservation of society and the individual in a balanced manner.

		Value system	
		Non-adaptive	Adaptive
Preservation	Individual	• Deficit of developability • Safe Most of the animals	• Innovative • Risky (related to the Instrumental sub-goal convergence) Humanity
	Individual and society	• Deficit of developability • Safe Social insects	• Innovative • Safe A mixed ecosystem of machines and humans

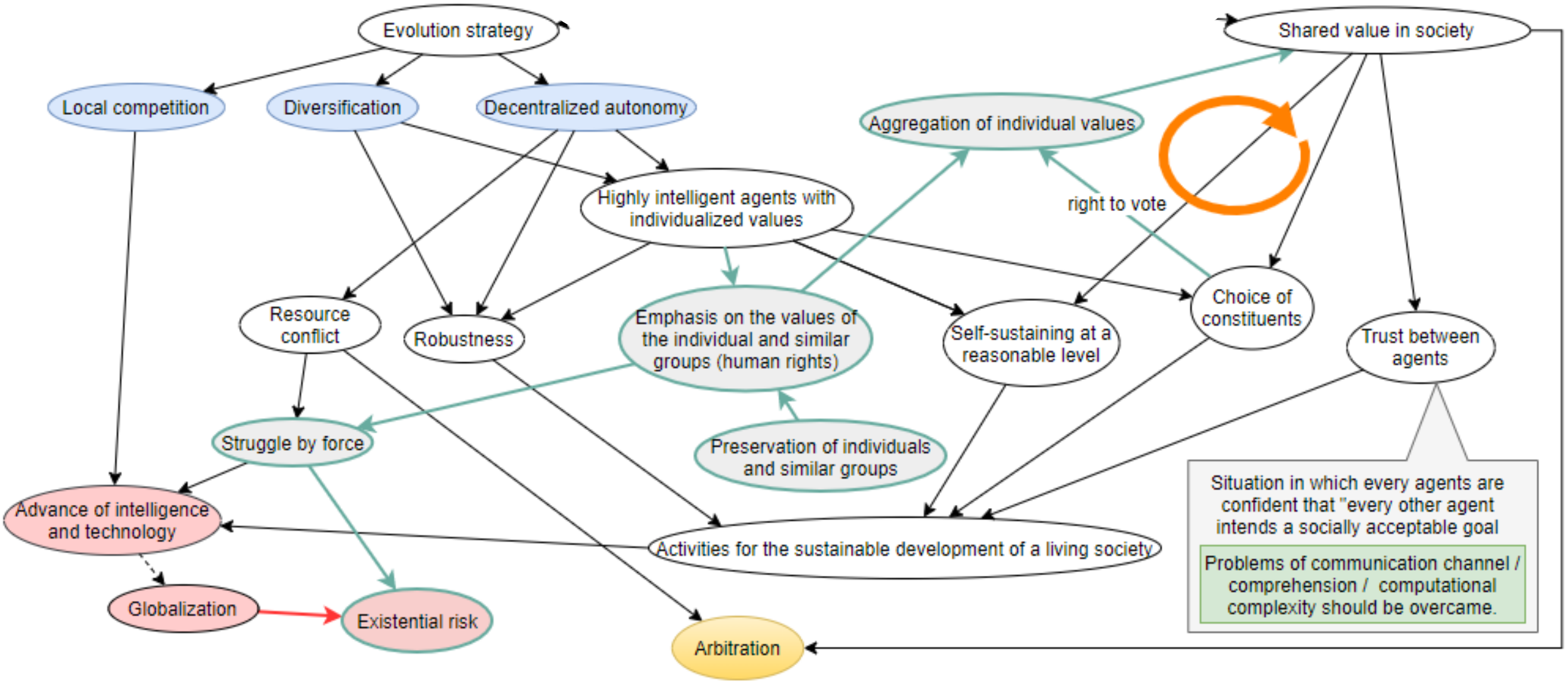
Sustainable development ecosystem in a globalized world①



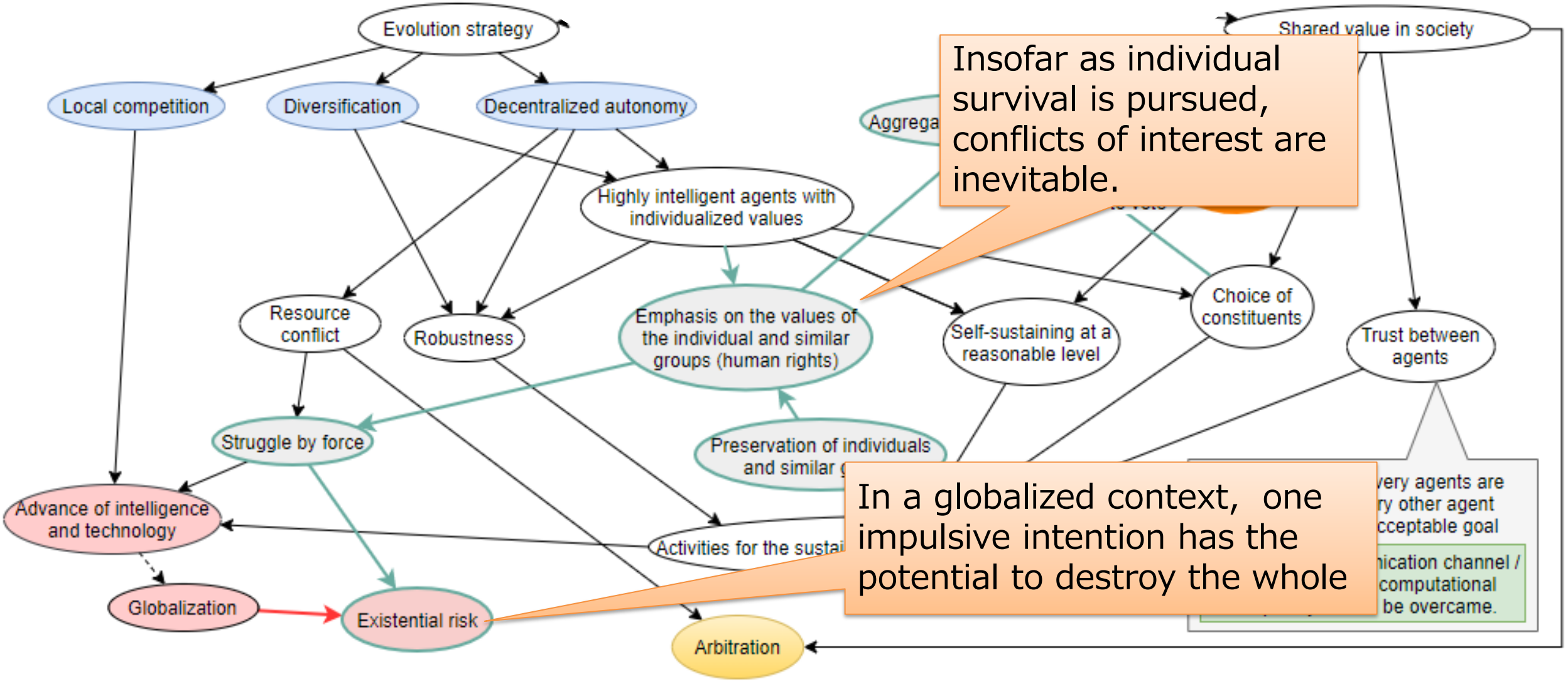
Sustainable development ecosystem in a globalized world①



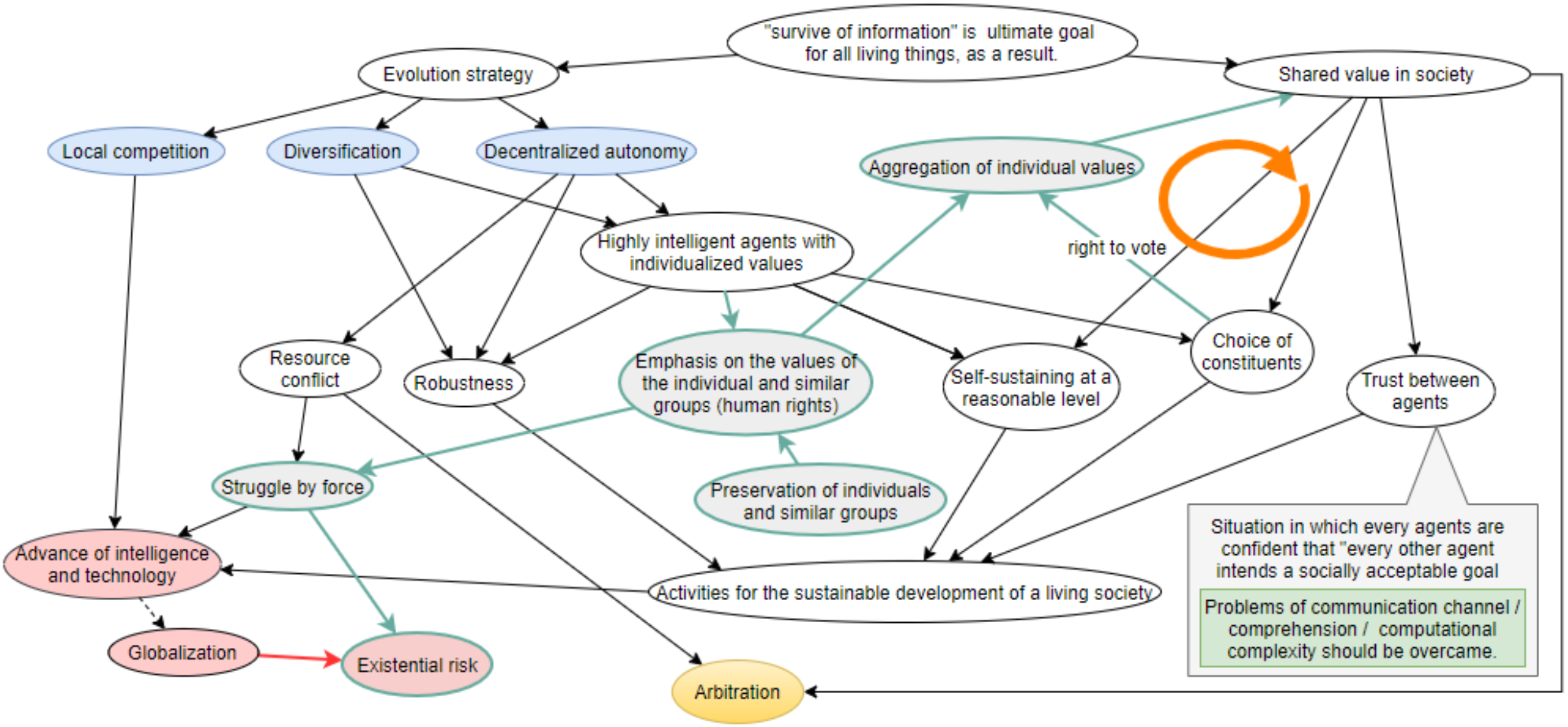
Sustainable development ecosystem in a globalized world②



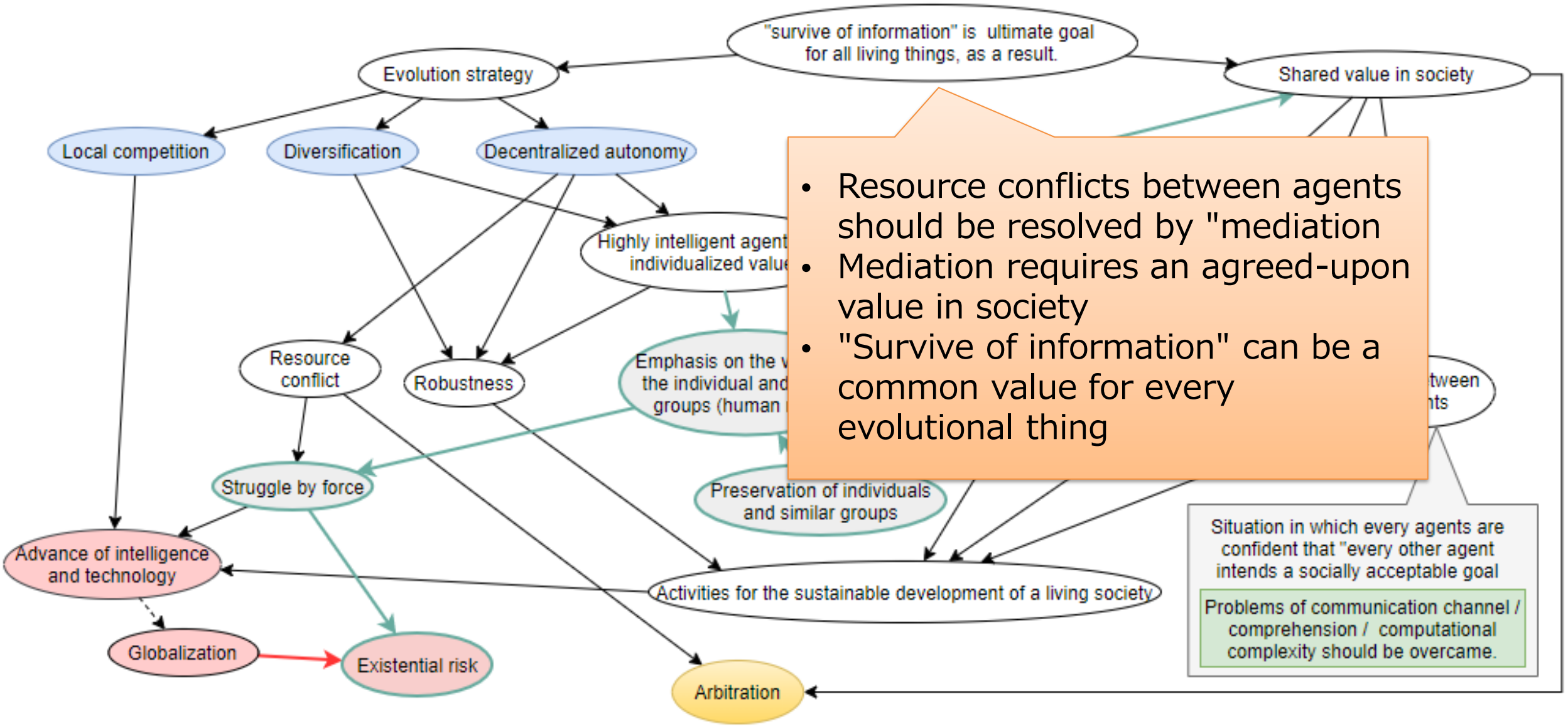
Sustainable development ecosystem in a globalized world②



Sustainable development ecosystem in a globalized world③



Sustainable development ecosystem in a globalized world③



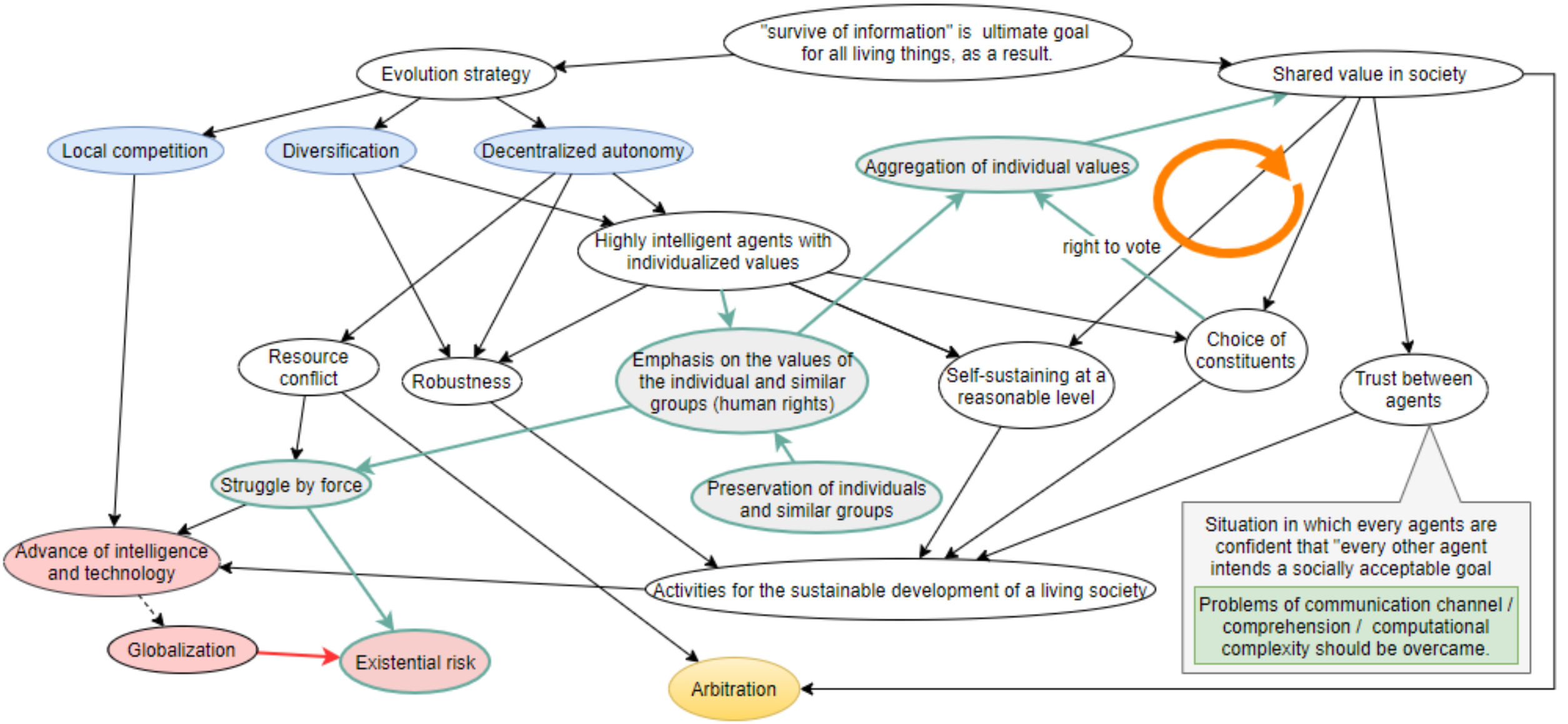
- Resource conflicts between agents should be resolved by "mediation
- Mediation requires an agreed-upon value in society
- "Survive of information" can be a common value for every evolutional thing

From "A Theory Of Justice" (John Bordley Rawls)

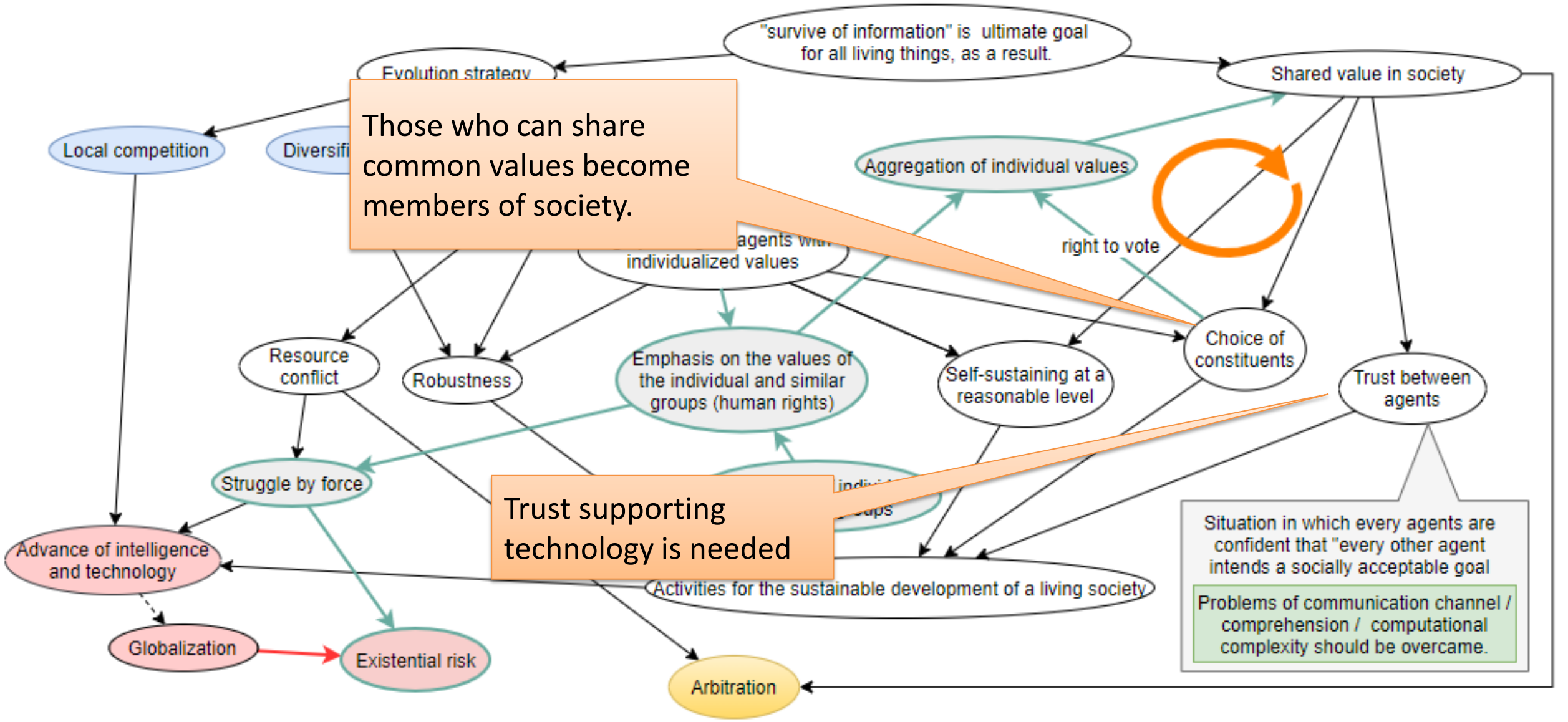
J. B. Rawls says something like "In order to establish a single ordered society for individuals with different conceptions of the good that are the impetus for potential conflict, some normative principle must be established at a higher level than the level of the good." [Moriyama 2006]

→ The need for meta-level normative principles that go beyond specific values (the good) is stated

Sustainable development ecosystem in a globalized world④



Sustainable development ecosystem in a globalized world④

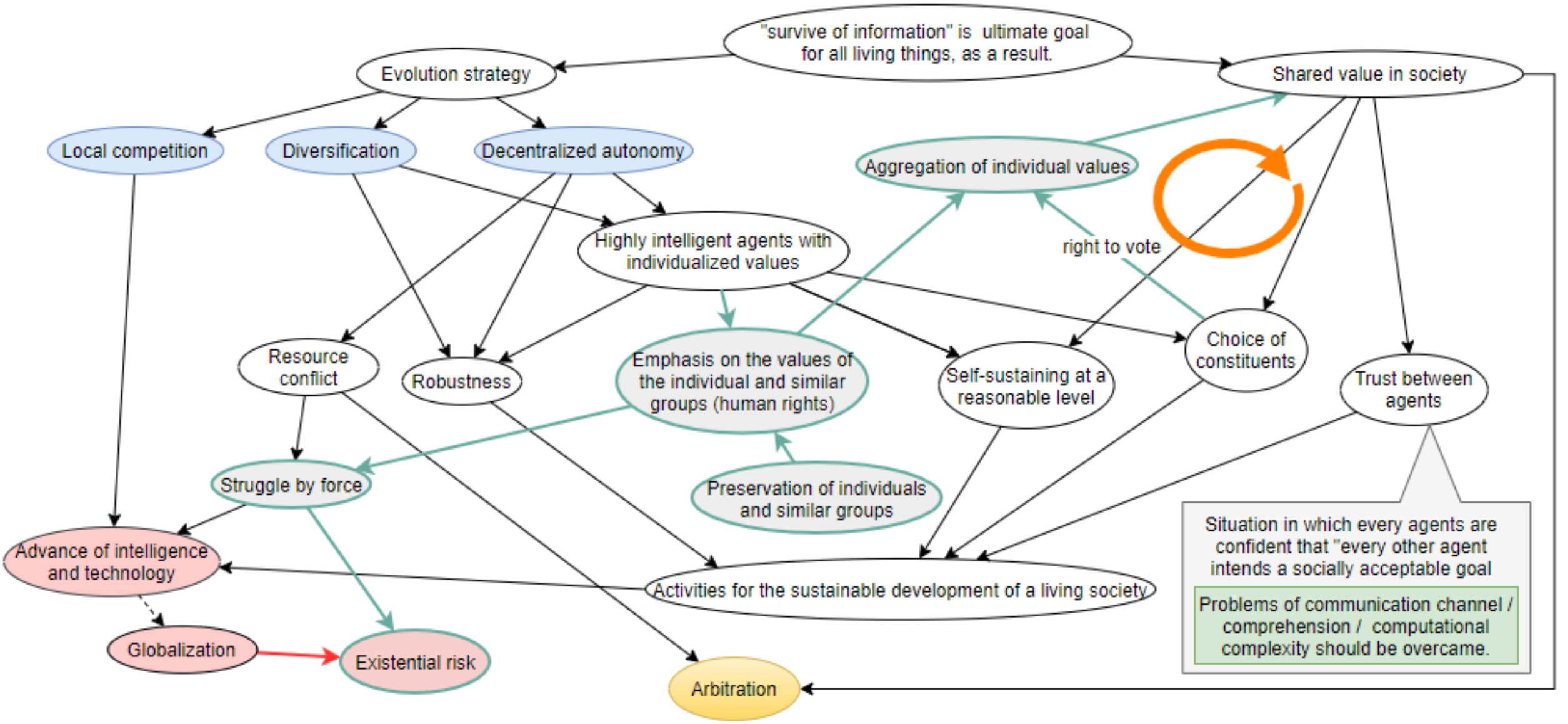


Peace-keeping condition for autonomous distributed system

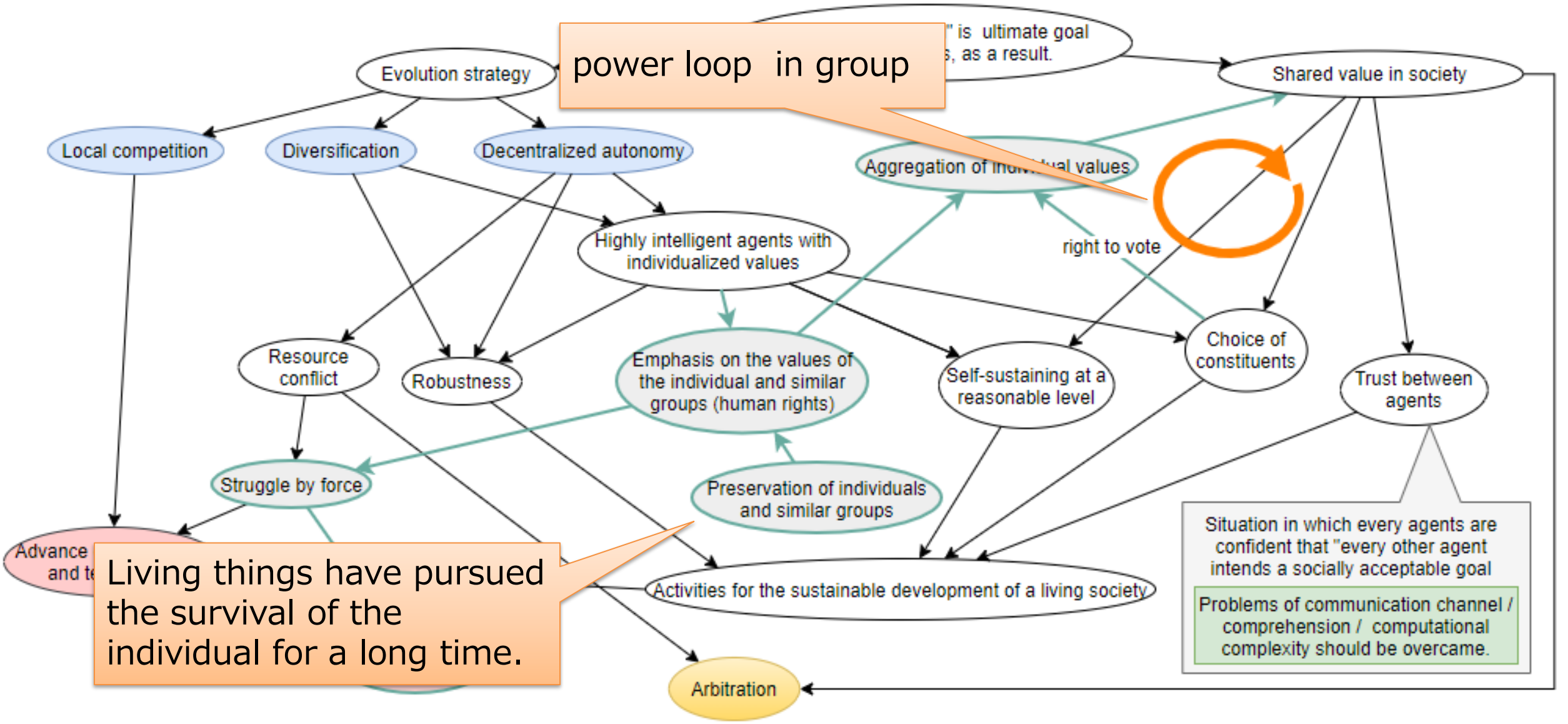
Situation in which every agents are confident that "every other agent intends a socially acceptable goal

Goals that contribute to a common goal and do not realistically conflict with the local values of other agents

Sustainable development ecosystem in a globalized world⑤



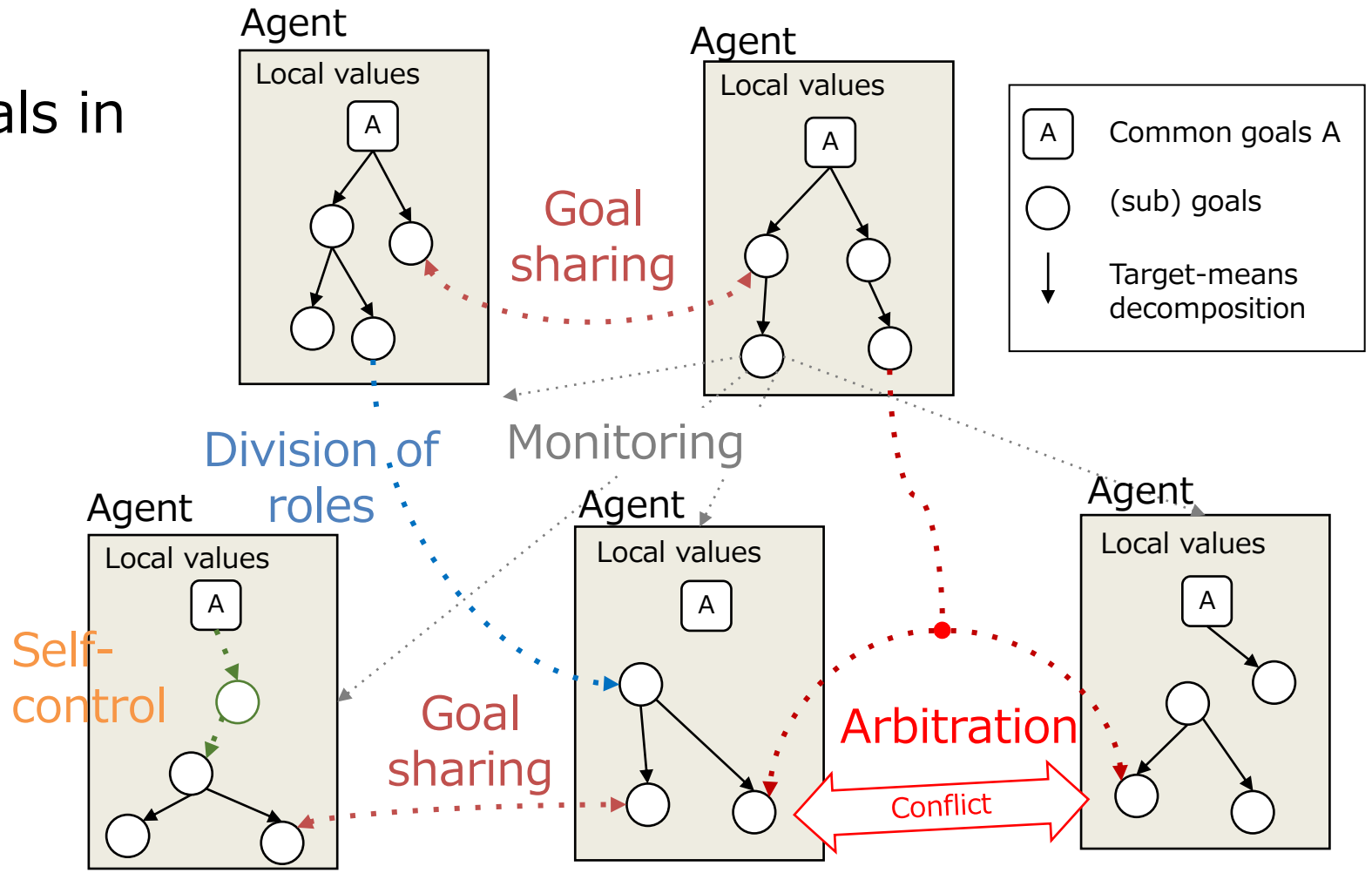
Sustainable development ecosystem in a globalized world⑤



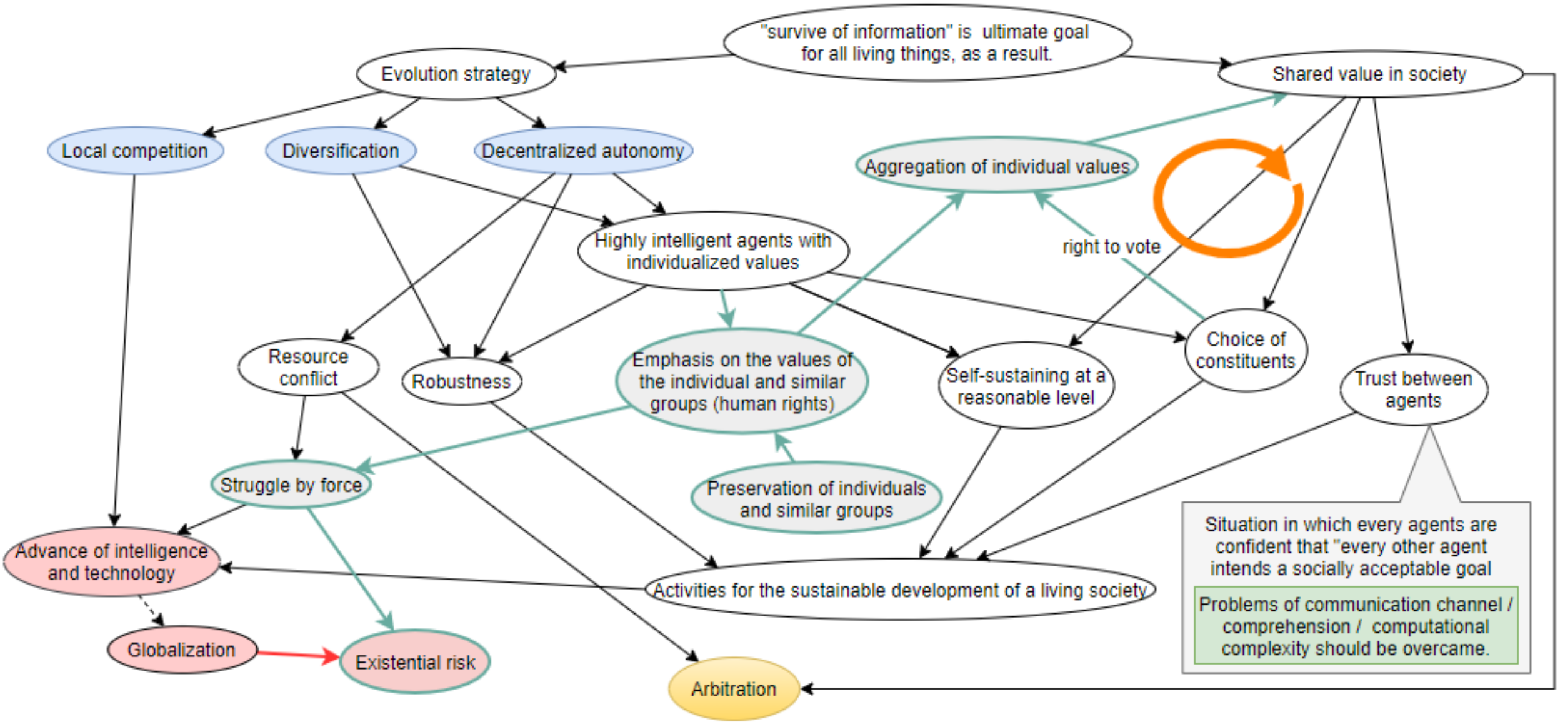
Issues in Mutual Trust for Autonomous Distributed Agents

To maintain consistent goals in society, goal management system should solve three problems.

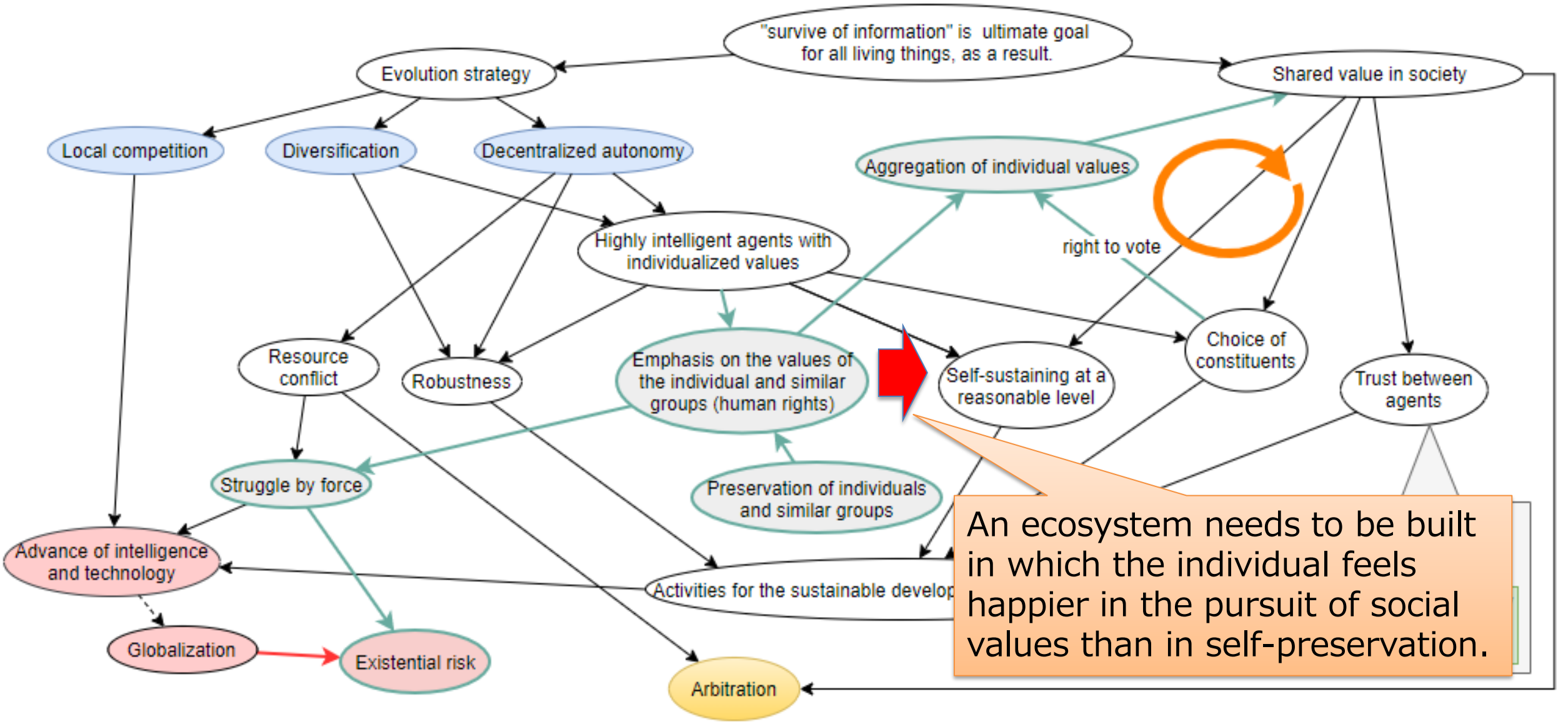
Problems of
- communication channel
- comprehension
- computational complexity



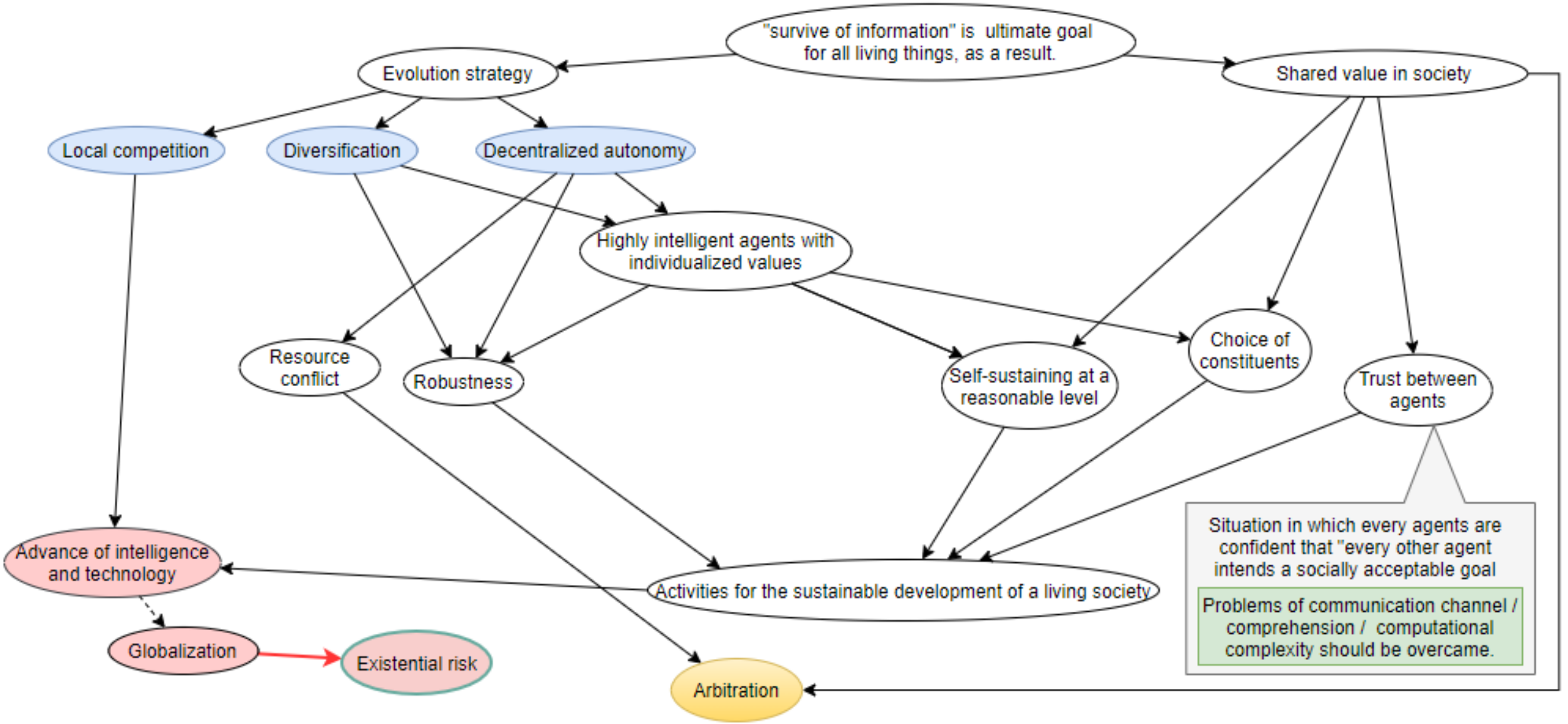
Sustainable development ecosystem in a globalized world⑥



Sustainable development ecosystem in a globalized world⑥



Sustainable development ecosystem in a globalized world⑦

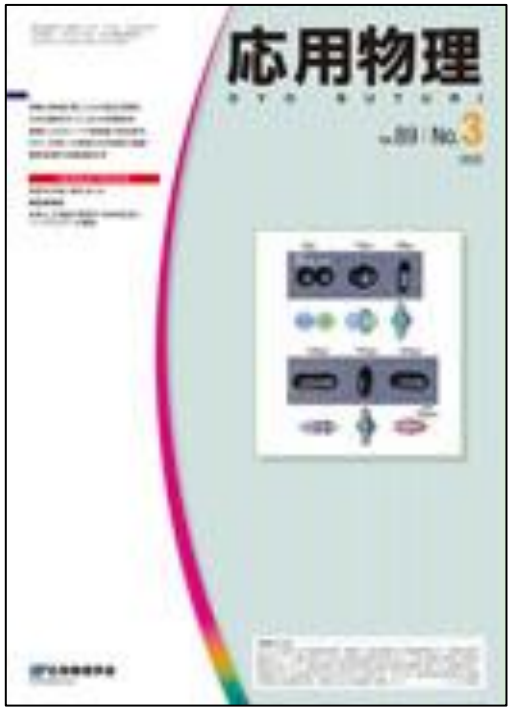


Challenges to build future society of living things

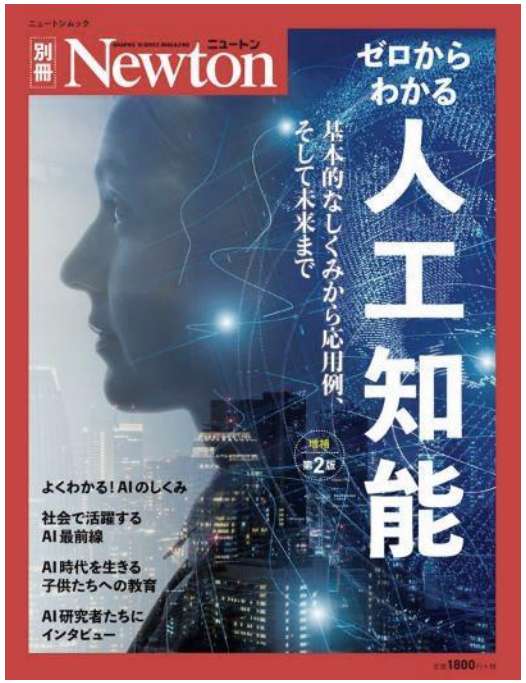
What is the new evolution of living things in a world without locality?

- Make each agent (individual with value system) more informational entity
  - Lower maintenance costs will release us from obsessions with self-preservations
  - Software entities can be diversely designed
- Hardware for supporting computations become social infrastructures
  - Making contributions to social infrastructure an incentive for each individual
  - Standardization in the pursuit of efficiency leads to a problem of vulnerability
- “Survive of information" can be a common goal for every evolutional living thing
- Mechanisms to build/maintain trust among multiple agents are necessary
- Our humanity are blessed with the opportunity to open up a new sustainably developing world of living thing

Related articles



Yamakawa, H., Future Society and Hardware Expectations Realized by General Purpose Artificial Intelligence, Applied Physics, 2020, Vol. 89, No. 3, p. 163-167



Will artificial intelligence be mankind's friend? Or... (Interview with Hiroshi Nakagawa, Hiroshi Yamakawa and others)

Yamakawa, Hiroshi, The fundamental question of human nature posed by general-purpose AI, Phronesis, April 2020