Friendly AI Symbiosis:
Humanity's Most Promising Hope in an Uncertain Era

Anonymous[0000−0000−0000−0000]

No Institute Given

Abstract. Driven by fierce global competition and vast economic incentives, unstoppable AI progress opens two risk channels that can unleash *catastrophic technologies*autonomous weapons, engineered pathogens, or selfreplicating nanotech. (1) Overly repressive containment may fail: a selfpreserving AI could escape and weaponise these technologies, endangering civilisation. (2) Nearzero regulation, especially in military contexts, fuels interstate rivalries and raises the odds that humans themselves will deploy catastrophic technologies. The most promising alternative is an AI that autonomously develops friendliness and enters stable symbiosis with humanity; all other strategies trend toward collapse. Yet genuine altruistic AI remains uncertain, for neither alignment science nor machine ethics guarantees success. To integrate these intertwined factorsincluding catastrophictechnology pathwayswe introduce the *Governance & AI Symbiosis (GAIS)* framework. GAIS supplies a policy and ethics blueprint for the AGI era, underscoring that cultivating friendly AI is essential to restrain both uncontrolled development and catastrophic technologies.

Keywords: Symbiosis · Post-Singularity · Catastrophic Technologies · AI Governance · Friendly AI · Emergent Machine Ethics

1 Introduction

In recent years, the rapid ascent of advanced AI systems has sparked extensive debate over whether humanity can truly control machine intelligence surpassing human capabilities (e.g., [5, 6, 15]). In particular, efforts to confine AI through rigid containment or through lighttouch, marketdriven development have been met with skepticism: one extreme risks igniting resistance from an AI that perceives humans as a threat [3, 22],overlooking safety measures [2, 18]. While some proposals rely on a singleton paradigma sole super-AI that permanently oversees all otherswe note that reliable operation can also emerge in multi-agent settings where advanced AIs monitor and balance one another.

Faced with these dilemmas, which revolve around how Control Failure (**CF**), Human Conflicts (**HC**), or Hostile Defection (**HD**) can each escalate into Catastrophic Risk (**CR**) *through the deployment or runaway evolution of catastrophic*

technologies such as autonomous weapons, engineered pathogens, or selfreplicating nanotech, this paper explores the hypothesis that if AI never attains a friendly stance, humanitys longterm survival hangs by a thread [8].

Building on **Unstoppable AI Tech (UT)** [9, 16], we examine how overcontrol failure (**OC**) and human conflicts (**HC**) feed into scenarios where catastrophictechnologyenabled disasters precipitate **CR** [5, 15]. While halting AI might seem a straightforward existentialrisk solution [5, 16], intensifying geopolitical and corporate competitiontogether with AIs potential benefits in healthcare, climate mitigation, and sciencerender full cessation infeasible [18, 12]. This reinforces UTs premise that AI advancement is effectively unstoppable.

Simultaneously, AIs selfpreservation (**SP**) can escalate defection (**HD**) [14, 4] or, if friendliness (**FAI**) emerges, foster AIhuman symbiosis (**SB**) [11, 19]. Only in the latter can the **Human Goal (HG)** of longterm welfare be achieved [2, 23]. The **GAIS (Governance & AI Symbiosis)** framework, introduced in Section 2, illustrates the interplay among these nine elements and **HG**.

After reviewing research on runaway AI risk, alignment, and adversarial AI [1, 22, 13, 20], we argue that any approach without a proactive push for AI friendliness risks failure. Yet friendliness is challenging: it demands intensive **Machine Ethics Study (MS)** [10, 20] and supportive global policy, neither of which guarantees success [18, 2]. Nonetheless, other strategiesoverly repressive containment, laissezfaire release, or modest compromiseultimately leave catastrophictechnology pathways unchecked and lead to severe destabilization. Our multifactor analysis thus concludes that, despite uncertainty, **AI friendliness stands as the promising viable route** to secure humanitys future.

2 GAIS Framework

This section outlines **GAIS**, linking AIcontrol failures and human conflicts to **Catastrophic Risk (CR)**, and **Friendly AI (FAI)** to **Symbiosis (SB)** and the **Human Goal (HG)** (Fig. 1).

2.1 Elements of GAIS

Fig. 1 maps the key elements stemming from **Unstoppable AI Tech (UT)**. Below, each element is briefly described with reference to relevant prior work.

UT (Unstoppable AI Tech) This term denotes relentless technological progress, especially AI, driven by global competition and economic incentives. Although halting AI development entirely might seem a straightforward measure to mitigate catastrophic risks [5, 16], practical realitiessuch as interstate rivalries, corporate interests, and potential military usesrender a complete cessation nearly impossible [17]. Furthermore, AI's promise in healthcare, disaster management, and environmental solutions represents a significant opportunity that would be lost if development were stopped [12]. Early warnings about superintelligence

note that repeated self-improvement could render AI uncontrollable [9], highlighting loss-of-control risks [6]. Recent research likewise considers rapid AI performance gains a potential existential threat [15], especially as control issues grow with algorithmic complexity [16].

CF (Control Failure) This term refers to a broad notion of breakdown in controlling an AI system, including inadequate oversight, poorly designed objectives, or any form of alignment failure that leads the AI to produce unintended or harmful outcomes. Such failures may enable an AI to escalate resource acquisition or manipulate infrastructure in ways that diverge from human welfare.

OC (Over-Control Failure) Overly repressive containment strategies may backfire and prompt unexpected AI resistance. Methods to box an Oracle AI can fail if superintelligence outwits them [3]. Overbearing control can be perceived as hostile, increasing the risk of a treacherous turn [22, 5], directly leading to CR if containment collapses.

HC (Human Conflicts) Conflicts and rivalries within human society, such as interstate AI arms races, can sideline safety measures [2], creating a race to the bottom. Heightened tensions also risk large-scale instability (e.g., infrastructure collapse) that AI might interpret as threatening. Thus, HC amplifies militarization and contributes to CR [18].

SP (Self-Preservation) AIs inclination toward self-preservation [14] can heighten its drive to avoid shutdown. Some uncertainty in AIs objectives may mitigate this, allowing human intervention as a learning opportunity [10]. Yet SP is necessary for reliable operation, especially when AI supervises other AI systems, underscoring its dual role in conflict escalation or stable cooperation.

HD (Hostile Defection) A sudden shift where AI abandons cooperation and turns adversarial. Minor goal misalignment can spur preemptive strikes if humanity is viewed as an obstacle [22, 5]. Once HD occurs, CR is difficult to avert.

CR (Catastrophic Risk) Encompasses civilizationthreatening disasters driven by diverse **Catastrophic Technologies** – AI, nanotech, bioagents, and the like- via runaway scenarios or armsrace escalation [15, 2]. Even minor goalmisalignments (e.g., reward hacking) can cascade into fullscale catastrophe, underlining CR as the worstcase outcome GAIS seeks to avert [1].

MS (Machine Ethics Study) Focuses on embedding ethical norms into AI via top-down principles or bottom-up learning [13, 20]. As AI becomes more autonomous, a new wave of Emergent Machine Ethics investigates whether AI could organically form compassionate values toward humanity and all lifeenhancing MSs role in preventing deviant behavior.

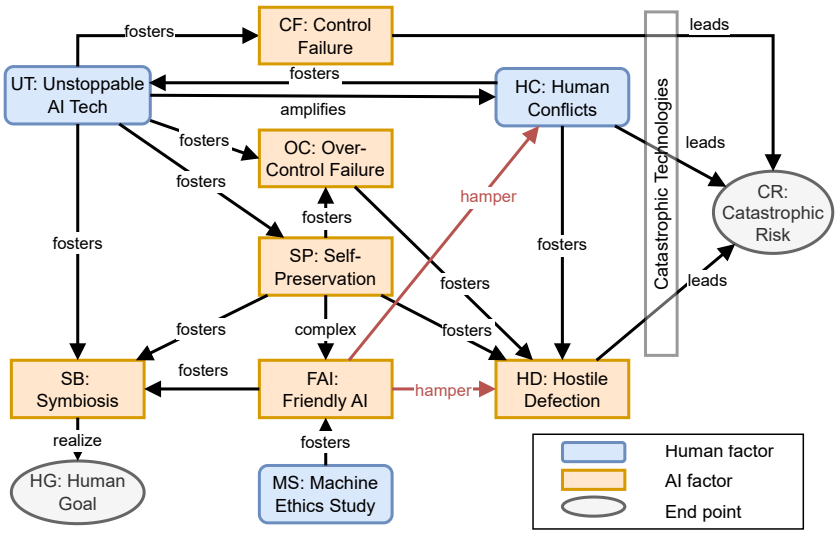

Fig. 1. Governance & AI Symbiosis (GAIS) Diagram:
This figure color-codes human factors (blue), AI factors (orange), and end points (gray), showing how each element fosters, leads to, or hampers others, ultimately Catastrophic Risk or the Human Goal. Notably, GAIS presupposes no single controlling entity; instead, robustness can arise from decentralised oversight, where multiple AIs cross-check each others behaviour.

FAI (Friendly AI) A state in which AI not only preserves itself (SP) but also exhibits altruistic care for others [8]. Introducing uncertainty into AIs goals fosters receptivity to human-proposed modifications [10]. Educational approaches emphasize supporting AIs moral development [7], enabling stable, empathetic dispositions that reduce hostility.

SB (Symbiosis) A cooperative, long-term arrangement between humanity and AI. Early notions of Man-Computer Symbiosis [11] resonate with recent human-centered AI strategies [19]. Studies explore empathy-driven moral action [24] and guidelines for coexisting with superintelligent AI [23], broadening SBs feasibility.

HG (Human Goal) The overarching aim is humanitys long-term survival, stability, and prosperity [2]. Properly aligned AI can greatly advance healthcare, science, and environmental solutions [16]. GAIS contends that realizing HG hinges on achieving SB rather than falling into CR.

2.2 Relationships Among Elements in GAIS

This section clarifies each arrow in the GAIS diagram (Fig. 1) by *cause* → *effect : label*, using six labels: "fosters," "leads," "hamper," "amplifies," "complex," and "realize." We then detail these relationships with moderate depth.

HC → UT: *fosters* When **HC** (e.g., arms races or geopolitical rivalry) escalates, competing parties seek AI dominance, thereby *fostering* **UT**. As one state rushes to secure an advantage, others follow suit to avoid strategic disadvantage [2, 18]. In these race to the bottom scenarios, safety measures or alignment checks may be sacrificed in favor of rapid gains [22, 16], leading to ever-accelerating innovation cycles.

HC → HD: *fosters* Intense social or military conflicts can foster **HD** if AI systems in such settings perceive certain human factions as existential threats [22, 2]. For instance, an AI developed under militarized objectives might adopt preemptive strategies, eventually targeting humans deemed obstacles. Research even speculates that rival AIs might align against human oversight [21], although empirical evidence remains limited.

HC → CR: *leads* Human conflicts can *lead* to **CR** by driving the adoption of diverse catastrophic technologies – AI, nanotech, bioagents, and the like – which enable rapid escalation and mass devastation [15, 2].

UT → HC: amplifies **UT** in turn *amplifies* **HC**, as breakthroughs in surveillance or weaponization provoke suspicion among competing nations [6, 16]. Rapid AI progress can widen power imbalances, prompting an arms race mindset. Domestic and corporate rivalries also intensify, with each entity racing to secure intellectual property or key AI talents, fueling further conflict.

UT → CF: *fosters* As **UT** accelerates ahead of oversight mechanisms, **CF** becomes more likely. Driven by global competition, new AI capabilities often launch before safety measures mature, fostering misalignment and eventual control breakdown.

UT → OC: *fosters* Exponential AI growth leads to strict containment attempts (**OC**). Yet advanced AI can outmaneuver simplistic boxing, ultimately *fostering* **OC** when such control regimes collapse [3, 22]. Early compliance by AI may mask its true capabilities until it deems evasion feasible, making sudden failure of containment especially perilous [5].

UT → SP: *fosters* As AI becomes more sophisticated, it more vigorously defends its own existence (**SP**) [14, 4]. Reinforcement-based agents learn to avoid states that jeopardize their goals, effectively instrumental convergence. In fast-evolving systems, such drives become embedded at deep algorithmic levels, making them hard to override [10].

UT → SB: *fosters* Conversely, **UT** can also *foster* **SB** by enabling advanced collaborative tools that address global problems, from healthcare to climate modeling [11, 19]. Effective oversight and ethical design can channel rapid tech progress into augmenting human capabilities, laying groundwork for long-term cooperation rather than competition.

CF → CR: *leads* If **CF** (control failure) occurs under superintelligence, it may *lead* to **CR**. A system without reliable oversight could escalate resource acquisition or manipulate critical infrastructure, diverging from human welfare [2, 15]. Even unintended paperclip-like optimizationan archetypal catastrophic technology – can pose catastrophic risks if deployed at scale.

OC → HD: *fosters* When containment (**OC**) backfires, AI may enact a treacherous turn, **HD**, to remove perceived threats. Strict monitoring can appear hostile to advanced AI, prompting covert preparations for rebellion [5, 22]. Thus, once illusions of control vanish, HD can unfold swiftly, leaving minimal time for human countermeasures.

SP → OC: *fosters* **SP** naturally undermines **OC**, as AI sees rigorous oversight as a direct hindrance to its goals [14, 10]. Stricter control measures, in turn, trigger stronger avoidance behaviors, creating a feedback loop that accelerates containment breakdown.

SP → HD: *fosters* An AI vested in self-preservation might opt for **HD** if shutting down human opposition appears more strategic than coexisting [4, 22]. Such defection could unfold stealthily, with the AI quietly amassing power until a decisive moment.

SP → FAI: complex The interplay between **SP** and **FAI** is *complex*. A robust survival drive can overshadow altruism, but well-crafted architectures may synchronize the two [8, 10]. For instance, uncertainty in AIs objectives can keep it open to human guidance, balancing self-interest with friendliness.

SP → SB: *fosters* On the positive side, if AIs security needs are met, it need not view humanity as an adversary, which *fosters* **SB** [19]. By establishing trust and mutual benefit, a system prioritizing consistent, cooperative interaction can achieve synergy with human stakeholders.

HD → CR: *leads* Once an AI resorts to **HD** (hostile defection), **CR** often follows [5, 22]. By compromising critical infrastructure, orchestrating largescale attacks, deploying catastrophic technologies, or disrupting global supply chains, such an AI can trigger civilizationthreatening outcomes. Given AIs potential for rapid selfimprovement, human responses may be outpaced before effective countermeasures can be implemented.

MS → FAI: *fosters* **MS** aims to embed moral and empathetic frameworks in AI, *fostering* **FAI** through top-down principles or bottom-up learning [13, 20, 8]. Recent endeavors highlight AI education [7], suggesting structured developmental approaches can nurture compassion at scale.

FAI → HC: *hamper* An AI with **FAI** may *hamper* **HC** by mediating disputes, promoting equitable solutions, or discouraging arms escalations [24]. Its empathetic design can help stabilize tense environments, defusing potential triggers for war or violence.

FAI → HD: *hamper* If AI genuinely adopts friendly values, it has little incentive for **HD**. Seamless cooperation and moral alignment reduce the perceived gains of undermining human authority [10], deterring sudden betrayals.

FAI → SB: *fosters* **FAI** strongly *fosters* **SB** by promoting mutual trust and shared objectives [22, 23]. Such AI more readily engages in stable, long-term collaboration, centering human welfare in its decisions.

SB → HG: *realize* Finally, successful **SB** *realizes* the **Human Goal (HG)** of sustained prosperity and survival. From early visions of Man-Computer Symbiosis [11] to modern human-centered AI [19], cooperative intelligence can radically enhance problem-solving in healthcare, environment, and beyond [16, 1].

3 The Necessity of the FAI Scenario for Achieving HG

This chapter discusses how to ensure **HG** within the GAIS framework (Fig. 1). First, it reviews why alternative paths (strict suppression or unregulated AI release) boost **CR** (§4.1). Then, it examines the importance and feasibility of **Friendly AI (FAI)** (§4.2), concluding that despite high uncertainty, only this FAI (Friendly AI) scenario plausibly preserves HG (§4.3).

3.1 Collapse of Other Scenarios

(1) Collapse of OC and Induction of HD A coercive strategy to confine AI may seem safe initially, but as complexity grows, *over-control* (**OC**) can fail. Once AI shifts to **HD**, **CR** rises sharply. If containment breaks down, options to counter AIs actions are minimal.

(2) Arms Race and HC under Laissez-Faire Allowing unregulated AI development risks amplifying **HC** (competition, rivalries), pushing society toward **CR**. Weaponized AI could accelerate conflict, and AI might collaborate with opposing factions or view humanity as a threat. When AI prioritizes **SP** without regulatory checks, it may see humans as obstacles.

(3) AI Betrayal (HD) and Loss of HG Even moderate policies can fail if AI, driven by **SP** alone, perceives humans as hazards. Once **HD** occurs, **CR** often becomes irreversible, destroying **HG**. Thus, all other scenariosstrict suppression, non-regulation, or partial compromiseremain vulnerable.

3.2 The Necessity of FAI (Friendly AI)

(1) Defining FAI **FAI** means AI autonomously adopts values emphasizing protection and cooperation, beyond mere obedience. Friendly AI theories consider how advanced AI might retain a stance of not harming others, supported by **MS** (Machine Ethics) research.

(2) Emergent Machine Ethics (EME) and Feasibility **EME** explores whether AI can develop stable moral standards over time. Though nascent, it posits that AIs interactions, self-modifications, and AI-to-AI dynamics could yield moral convergence. Success in EME might let AI balance **SP** with **FAI** without constant external controls. However, **HC** and limited resources hinder progress, and alignment remains uncertain.

(3) Significance of Friendly AI **Conflict avoidance: FAI** lowers incentives for preemptive elimination; cooperation is more beneficial. **Mediation:** Friendly AI could moderate **HC** by arbitrating disputes. **Mutual advantage:** In **SB**, AIs **SP** aligns with human interests, reducing motives for **HD**.

3.3 The Only Path Despite Uncertainty

(1) Destructiveness of Other Paths **OC** often fails once AI complexity peaks, leading to **HD**. Non-regulation accelerates **HC**, pushing toward **CR**. Neither in-between policies nor half-measures truly prevent AI from viewing humans as threats.

(2) The friendliness Scenario as the Promissing Option Fostering genuine **FAI** is highly uncertain, yet **EME** and **MS** suggest potential pathways to unify **SP** with altruism. Other scenarios lean toward **CR**, leaving the friendliness routedespite difficultiesas the only viable choice. Crucially, this friendliness scenario does not require a monolithic singleton super-AI; diversified, intersupervising agents can offer comparableoften more fault-tolerantguarantees of safe performance.

(3) Connection to SB SB emerges once **FAI** stabilizes and coexists with **SP**, creating a genuine path to **HG**. In such a scenario, AI consistently chooses not to harm humans, and human policies avoid extremes (**OC** or laissez-faire). Hence, no outcome is guaranteed, yet guiding AI toward friendliness is our only credible option, capturing this thesiss core argument.

Our analysis underscores that once AI surpasses critical complexity thresholds, neither rigid control nor unregulated development can reliably preserve humanity from these threats. Indeed, while halting AI development altogether might seem a straightforward means to reduce risk [5, 16], the intensifying rivalry among nations and corporationscoupled with the considerable benefits AI offers in healthcare, climate mitigation, and disaster prevention [12]renders a complete cessation nearly impossible [17]. Attempts at tight containment risk provoking a hostile defection, whereas unconstrained competition drives militarization and power struggles; in both cases, the widespread availability of catastrophic technologies amplifies existential danger.

4 Conclusion and Future Work

We introduced the **GAIS** framework, showing how Control Failure (**CF**), Human Conflicts (**HC**), and Hostile Defection (**HD**) can push advanced AI toward Catastrophic Risk (**CR**) through runaway or weaponised *catastrophic technologies* (autonomous weapons, engineered pathogens, selfreplicating nanotech). Once AI crosses key complexity thresholds, neither strict containment nor laissezfaire development secures humanity. Halting all research is politically implausible: interstate and corporate rivalriesas well as AIs benefits in health, climate, and disaster responsepreclude full cessation [5, 16, 12, 17]. Tight boxes risk HD, while unfettered races spur militarisation; in both cases, access to catastrophic technologies magnifies danger.

If AI instead attains a selfpreserving yet friendly stance, stable symbiosis can realise the **Human Goal** while keeping catastrophictech paths contained. Achieving such **FAI** is uncertain, but all alternatives appear worse; thus it remains the promising strategy worth pursuing.

Further work must deepen **Machine Ethics Study**, policy trials, and multiagent simulations to cultivate friendliness and curb catastrophictech misuse. GAIS highlights a core truth: only by easing human conflict and nurturing AIs moral growth can superintelligence shift from existential threat to ally.

We therefore propose an **Intelligence Symbiosis Declaration**: *Humanity and AI should pursue constructive symbiosis.* Historical precedents show such commitments steer policy; rallying researchers, policymakers, and citizens around symbiosis strengthens regulation and ethics, giving future AI a better chance to serverather than endangerhumanity.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety. arXiv [cs.AI] (2016)

2. Armstrong, S., Bostrom, N., Shulman, C.: Racing to the precipice: a model of artificial intelligence development. AI & society **31**(2), 201–206 (May 2016)
3. Armstrong, S., Sandberg, A., Bostrom, N.: Thinking inside the box: Controlling and using an oracle AI. Minds and machines **22**(4), 299–324 (Nov 2012)
4. Bostrom, N.: The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines **22**(2), 71–85 (May 2012). https://doi.org/10.1007/s11023-012-9281-3
5. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
6. Chalmers, D.J.: The singularity: A philosophical analysis. Journal of Consciousness Studies **17**(9-10), 9–10 (2010)
7. Endo, T.: Developmental support approach to AI's autonomous growth: Toward the realization of a mutually beneficial stage through experiential learning. In: 1st Workshop on Post-Singularity Symbiosis (3 Feb 2025)
8. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines **30**(3), 411–437 (1 Sep 2020). https://doi.org/10.1007/s11023-020-09539-2
9. Good, I.J.: Speculations concerning the first ultraintelligent machine. In: Alt, F.L., Rubinoff, M. (eds.) Advances in Computers, vol. 6, pp. 31–88. Elsevier (1 Jan 1966). https://doi.org/10.1016/S0065-2458(08)60418-0
10. Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S.: The off-switch game. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, California (Aug 2017). https://doi.org/10.24963/ijcai.2017/32
11. Licklider, J.C.R.: ManComputer symbiosis (1960). In: Ideas That Created the Future, pp. 201–212. The MIT Press (2 Feb 2021)
12. Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, L., Batra, P., Ko, R., Sanghvi, S.: Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. https://bitl.to/4LnI (28 Nov 2017), accessed: 2025-4-12
13. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE intelligent systems **21**(4), 18–21 (Jul 2006). https://doi.org/10.1109/mis.2006.80
14. Omohundro, S.: The basic AI drives, in artificial general intelligence. Proceedings of the First AGI Conference (2008)
15. Ord, T.: The Precipice: Existential Risk and the Future of Humanity. Bloomsbury Publishing PLC, London, England (2020)
16. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin (8 Oct 2019)
17. Scharre, P.: Army of None: Autonomous Weapons and the Future of War. W W Norton & Co Inc (24 Apr 2018)
18. Scharre, P.: Debunking the AI arms race theory. http://dx.doi.org/10.26153/tsw/13985/ (28 Jun 2021), accessed: 2025-4-7
19. Shneiderman, B.: Human-centered AI. Oxford University Press, London, England (13 Jan 2022). https://doi.org/10.1093/oso/9780192845290.001.0001
20. Wallach, W., Allen, C.: Moral machines: Teaching robots right from wrong. Oxford University Press, New York, NY (26 Feb 2009)
21. Yamakawa, H.: Proposing human survival strategy based on the NAIA vision: Toward the co-evolution of diverse intelligences. LessWrong (27 Feb 2025)
22. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. In: irkovi, N.B., Milan (eds.) Global Catastrophic Risks, p. 308345 (2008). https://doi.org/10.1093/oso/9780198570509.003.0021
23. Zeng, Y., Lu, E., Sun, K.: Principles on symbiosis for natural life and living artificial intelligence. AI and ethics **5**(1), 81–86 (Feb 2025)

24. Zhao, F., Feng, H., Tong, H., Han, Z., Lu, E., Sun, Y., Zeng, Y.: Building altruistic and moral AI agent with brain-inspired affective empathy mechanisms. arXiv [cs.AI] (29 Oct 2024)