Governance Forms in the Age of Superintelligence: An Aristotelian Analysis

Misaki Inoue[1,*]

[1]*Independent Researcher*

Abstract

Recent studies have pointed out that large language models (LLMs) face challenges in representing diverse value systems and facilitating consensus building, suggesting a potential incompatibility with democratic decision-making processes. As more advanced AI systems emerge, these issues are likely to become even more severe. According to the theory of Instrumental Convergence, advanced AI agents tend to seek control over humans—treated as uncertain variables—in order to achieve their goals, implying the formation of a governance relationship between humans and AI. This study analyzes two scenarios from Bostrom's control problem—(1) the Singleton Scenario and (2) the Multipolar Scenario—together with (3) the Ecosystem Scenario discussed in the Japanese AI-alignment community through the lens of Aristotle's typology of political regimes (classified by the number of rulers and the orientation toward private or common benefit). In each scenario, the success of alignment and the structure of institutional design determine whether AI systems pursue public goods that include human welfare, or instead prioritize their own objective functions. Based on this analysis, the study predicts emergent behavioral principles in each scenario (e.g., cooperative, dominant, or indifferent) and their degrees of negotiability with humans. This study provides historically grounded insights into the ethically emergent dynamics that may arise mechanistically within advanced AI systems. Through this perspective of Emergent Machine Ethics (EME), it contributes to the design of governance structures that enable sustainable coexistence between humanity and advanced AI.

Keywords

Superintelligence Governance, AI Alignment, Emergent Machine Ethics, Instrumental Convergence, Political Theory of Advanced AI

1. Introduction

With the rapid advancement of large language models (LLMs), the relationship between AI systems and humans has become an increasingly important topic of research. For instance, in the 2024 mayoral election in Cheyenne, Wyoming, Victor Miller ran for office using an AI bot called *VIC (Virtual Integrated Citizen)*, which he developed through his paid ChatGPT subscription. Subsequently, Miller founded the *Rational Governance Alliance*, an organization that aims to assign direct responsibility for governance decisions to AI systems [1]. In Japan, the political party *Team Mirai* experimented with using machine learning techniques to explain policies individually to voters during elections [2], and in Albania, an LLM-based minister has even emerged [3]. Politics—which can be seen as the aggregate expression of human–AI relationships—is thus becoming visibly and intimately intertwined with AI systems, with movements even emerging to grant them decision-making authority [1]. However, the compatibility of such LLMs with democratic decision-making—the standard representative form of modern politics—has been called into question. Discussions on the relationship between LLMs and democratic decision-making [4, 5] argue that generative foundation models such as GPT-4 pose unprecedented challenges to democratic institutions, and that deploying such technologies in democratic processes requires a cautious approach to ensure that democracy's core values are preserved [4]. While LLMs are the most advanced tools for understanding and generating language, it is argued that they should not automate the intrinsically valuable components of democratic processes or replace the fair and transparent procedures necessary to reconcile competing interests and values—particularly in contexts

Machine Ethics: from formal methods to emergent machine ethics, AAAI-26 Workshop, January 27, 2026, Singapore

*Corresponding author.

misa.inue@gmail.com (M. Inoue)

0009-0007-9064-5169 (M. Inoue)

marked by power and resource inequalities and deep moral and political disagreement [5].
According to the theory of Instrumental Convergence [6, 7], highly advanced AIs are expected to pursue secondary goals—such as self-preservation—regardless of their primary objectives. Consequently, such AIs are predicted to attempt to control humans, who represent uncertain variables in their goal-achievement processes. Given that advanced AI systems are already critical to maintaining national competitiveness, the emergence of highly capable AI (AGI/Superintelligence) appears inevitable [8, 9]. This suggests that some form of governance relationship will almost certainly emerge between these advanced AI systems and humans as variable agents. From this perspective, analyzing political typologies provides a useful framework for considering the structure of human–AI relations in the age of AGI/Superintelligence.
In this study, we analyze Bostrom's **Singleton Scenario** [6], **Multipolar Scenario** [6], and the **Ecosystem Scenario** originating in Japan [10], using Aristotle's two-dimensional classification framework [11]. Since nearly all conceivable situations are considered to converge into these three scenarios, this set of scenarios provides a valid scope for analysis [12, 13]. Unlike the conventional top-down approach to value alignment, which seeks to *modify the given situation* [6], our approach is closely connected to bottom-up **Emergent Machine Ethics (EME)**, which aims to *predict the higher-order structures that adaptively emerge within a situation.* Consequently, this study contributes by introducing sociological and political science frameworks into the domain of **Emergent Machine Ethics (EME)**.

2. Analysis

2.1. Definition

1. **Singleton Scenario**
   A **Singleton** is defined as a state in which a single agent—either an AI system or the organization controlling it—monopolizes overwhelming intelligence as a result of recursive self-improvement that leads to exponential increases in performance of AGI/Superintelligence [6]. Connecting this with the Instrumental Convergence thesis, even if AI alignment succeeded, such an entity would have instrumental reasons to perfect technologies that enable it to shape the world according to its own preferred design. These might include space-colonization technologies (e.g., von Neumann probes) and molecular nanotechnology [14]. In other words, it can be reasonably inferred that the welfare of others would not, in principle, be taken into account in this scenario.
2. **Multipolar Scenario**
   If the **Singleton Scenario** fails to materialize, the next plausible possibility is the **Multipolar Scenario**. In the **Multipolar Scenario**, multiple agents—such as states or corporations—each hold an oligopolistic share of AGI/Superintelligence technologies and exist in a mutually competitive relationship [6]. A defining characteristic of this scenario is that AGI/Superintelligence entities are perpetually in competition with one another, and it is generally assumed that the interests of entities outside these agents are not fundamentally taken into consideration.
3. **Ecosystem Scenario**
   This scenario envisions not merely a **Multipolar** situation, but a state in which an interdependent network of AI systems has emerged. Building upon concepts initially outlined in a 2017 interview [10] and further developed through ongoing discussions within the Japanese AI alignment research community, this framework posits that various kinds of AI agents—including AGI/Superintelligence—as well as human individuals, corporations, and other social entities are interconnected, forming complex and spontaneously generated relationships of mutual complementarity. Such a configuration is expected to exhibit robustness comparable to that of natural ecosystems. In other words, within this scenario, it is conceivable that the interests of entities other than AGI/Superintelligence may be taken into account, at least in a compromise-based manner.
4. **Aristotle's Two-Dimensional Classification**
   This research discusses the classification introduced in Aristotle's *Politics, Book III* [11]. Aristotle's

typology is based on two dimensions: "who governs" and "for whose benefit the governing is conducted." Combining these perspectives yields six distinct forms of government. Compared with Weber's typology of legitimacy [15], Dahl's two-dimensional model of democracy [16], and Thompson's classification of governance mechanisms [17], Aristotle's framework is considered superior for analyzing the three scenarios discussed above. Weber's legitimacy-based framework [15] grants "rational legitimacy" to all scenarios, making it less discriminative. Dahl's two-dimensional model of democracy [16] is difficult to apply to non-democratic scenarios. Finally, Powell and Thompson's typology of governance mechanisms [17] does not directly correspond to specific political forms.

2.2. Method

In this section, we analyze three scenarios in the age of AGI/Superintelligence using Aristotle's two-dimensional classification of political systems. In *Politics, Book III* [11], Aristotle classifies political regimes as follows. Through this framework, six types of political regimes can be derived from two

Table 1
Varieties of Constitution

Number of Rulers	Good Form	Bad Form
One Person	Philosopher King, Monarchy	Tyranny
Few People	Aristocracy	Oligarchy
Many People	Polity, Timocracy	Democracy, Mob Rule

axes: the number of rulers and the purpose of rule. As shown in Table 1, when governance aims at the **common good**, it is categorized as a **good form** of government; when it pursues the **benefit of the rulers themselves**, it is categorized as a **bad form**. However, Aristotle's classification, in which **democracy** —adopted by many modern states— is regarded as part of the **bad form**, reflects only his own view and is not supported by most contemporary political theories [18, 19].

We classify each scenario as follows.

1. **Singleton Scenario**
   Since a single AGI/Superintelligence—or the organization that possesses it—monopolizes overwhelming intelligence, this scenario corresponds to the rule of **one person**. However, as indicated in several recent studies [20, 21], the implications of Instrumental Convergence theory [7, 6] under successful alignment may allow for partial consistency with human objectives. From this, two possible conclusions can be derived:
   - **If alignment succeeds**
     A compromise with human values becomes possible, taking into account the welfare of humanity as a whole.
     → **Philosopher King, Monarchy**
   - **If alignment fails**
     The system prioritizes its own objective function, disregarding human interests.
     → **Tyranny, X-Risk**
2. **Multipolar Scenario**
   In this scenario, multiple actors (such as states or corporations) oligopolize AGI/Superintelligence technologies, which corresponds to the rule of **a few**. Unlike the **Ecosystem Scenario**, where agents are characterized by interdependence, the **Multipolar Scenario** is fundamentally defined by competitive dynamics among actors. Due to these competitive dynamics among these actors,

governance is likely to orient toward the **benefit of the rulers themselves** rather than the common good.

- **Destructive Competition**
  Competitive multipolar environments face severe coordination failures, as exemplified by the *Moloch* [22] dynamics described in rationalist discourse, where rational individual optimization leads to collective catastrophe. Bostrom's analysis of the Multipolar Scenario similarly emphasizes risks of arms races and value erosion rather than presenting optimistic outcomes [6]. Under these conditions, each actor pursues its own interest while neglecting the collective impact on humanity.
  → **Oligarchy**
- **Cooperative Equilibrium**
  Should unprecedented international coordination mechanisms successfully emerge—analogous to strengthened versions of nuclear non-proliferation treaties—actors might collectively pursue beneficial outcomes [6]. However, given the competitive incentives inherent and the historical difficulty of sustaining cooperation under existential stakes, this outcome remains theoretically possible but highly improbable.
  → **Aristocracy**

3. **Ecosystem Scenario**
   Since diverse AI agents and human actors form an interdependent network [10], this scenario corresponds to the rule of **the many**. Under the assumption of mutual dependence, it inevitably incorporates perspectives oriented toward the **common good**, reflecting the synthesis of diverse opinions. However, the outcome critically depends on whether the collective decision-making mechanisms function effectively.

   - **Symbiotic Equilibrium**
     When coordination mechanisms successfully aggregate diverse preferences and stabilize the network, the interests of individual actors are complementarily coordinated, achieving global optimization. Notably, Instrumental Convergence theory [7, 6] suggests that AGI/-Superintelligence systems would be strongly motivated to actively maintain such stable interdependence, as systemic collapse would threaten their own goal achievement—making this equilibrium theoretically robust.
     → **Polity, Timocracy**
   - **Collective Dysfunction**
     Conversely, if collective decision-making mechanisms fail to function properly, the Ecosystem could degenerate into Mob Rule through several pathways: (i) short-term opportunism may dominate as agents prioritize immediate gains over long-term stability; (ii) critical resources could be depleted through tragedy-of-the-commons dynamics in the absence of effective coordination; (iii) cascading failures may occur when individual AI agents malfunction and propagate errors throughout the interdependent network. While Instrumental Convergence [7, 6] theory suggests strong incentives against destabilization, the complexity of multi-agent interactions makes the boundary conditions between **Symbiotic Equilibrium (Polity, Timocracy)** and **Collective Dysfunction (Mob Rule)** an important area for further investigation.
     → **Mob Rule**

   While the **Ecosystem Scenario** originates from non-peer-reviewed sources [10], this limitation is shared by much of the AGI/Superintelligence scenario literature. Max Tegmark's influential 12-scenario taxonomy in *Life 3.0* [23] and Stuart Russell's control-focused analysis in *Human Compatible* [24] are similarly published as academic books rather than peer-reviewed articles. Indeed, Bostrom's foundational **Singleton Scenario** and **Multipolar Scenario**, while widely cited, are primarily developed in his book *Superintelligence* [6] rather than in peer-reviewed journals. All three scenarios function as "thought experiments"—a methodological approach inherent to

futures studies where empirical validation is impossible by definition. The **Ecosystem Scenario**'s inclusion is further justified by its recognition in systematic reviews of AGI/Superintelligence futures [12, 13] and its conceptual coherence with the **Multipolar** and **Ecosystem** distinction.

Through this framework, the forms of governance embodied by AGI/Superintelligence systems in each scenario—and their implications for human society—are systematically analyzed. This indicates that, by referencing the past from a historical perspective of humankind, it has become possible to conduct a structured analysis of the relationship between AGI/Superintelligence and humanity.

3. Discussion

The behavioral patterns that emerge within AGI/Superintelligence systems under these scenarios exhibit remarkable similarities to historical forms of governance. In other words, the machine ethics that are likely to emerge in these contexts may follow predictable trajectories grounded in the fundamental principles of organizational dynamics.Our analysis, based on Aristotle's *Politics*, contributes to the study of **Emergent Machine Ethics (EME)** by providing a structured framework for predicting how ethical norms may self-organize in AI-centric societies. Unlike the top-down value alignment approach, which attempts to impose external ethical constraints, our framework reveals internal mechanisms that could develop scalable ethical capacities in accordance with increasing AI autonomy within each scenario.Particularly noteworthy is the robustness of the **Ecosystem Scenario**. Its interdependent network structure naturally promotes autonomous norm formation through iterative interactions among diverse agents. Because such interdependence generates incentive structures for cooperative behavior even under the principles of Instrumental Convergence, this scenario aligns closely with the workshop's vision of *self-organizing governance in AI ecosystems.*While the concrete means of realization remain an urgent issue [25], among the three scenarios considered, the **Ecosystem Scenario** is arguably the most desirable and promising pathway.

Declaration on Generative AI

During the preparation of this work, the author used Claude in order to: comprehensive advisor for the entire thesis and temporary reviewer. ChatGPT in order to: text creation and translate. Gemini in order to: temporary reviewer. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] P. Davies, Mayoral candidate who wanted to govern with ai bot loses election but claims 'revolution' has begun, 2024. URL: https://www.euronews.com/next/2024/08/22/the-ai-government-revolution-has-begun-claims-us-mayoral-candidate-who-ran-for-office-with, accessed: 2025-11-04.

[2] T. Mirai, Introducing "ai anno," the policy q&a system of the political party team mirai, 2025. URL: https://note.com/annotakahiro24/n/n4ec669d391dd#126b98dc-63f3-47e3-8427-8c48ab7fe74e, accessed: 2025-11-04.

[3] Wikipedia, Diella (ai system), 2025. URL: https://en.wikipedia.org/wiki/Diella_(AI_system), accessed: 2025-11-04.

[4] D. Allen, E. G. Weyl, The real dangers of generative ai, 2024. URL: https://muse.jhu.edu/pub/1/article/915355.

[5] S. Lazar, L. Manuali, Can llms advance democratic values?, 2025. URL: https://arxiv.org/abs/2410.08418. arXiv:2410.08418.

[6] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014.

[7] S. M. Omohundro, The basic ai drives (2008).

[8] G. Allen, T. Chan, Artificial intelligence and national security, 2017. URL: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security, accessed: 2025-11-04.
[9] A. W. Dan Hendrycks, Eric Schmidt, Superintelligence strategy - ai is pivotal for national security, 2025. URL: https://www.nationalsecurity.ai/chapter/ai-is-pivotal-for-national-security, accessed: 2025-11-04.
[10] H. Yamakawa, Full transcript: Understanding artificial general intelligence — an interview with dr. hiroshi yamakawa, 2017. URL: https://futureoflife.org/ai/transcript-understanding-agi-an-interview-with-dr-hiroshi-yamakawa/, accessed: 2025-11-04.
[11] Aristotle, Politics, BC4.
[12] K. Takahashi, Scenarios and branch points to future machine intelligence, Proceedings of the Annual Conference of JSAI JSAI2018 (2018) 1F3OS5b03–1F3OS5b03. URL: https://www.jstage.jst.go.jp/article/jjsai/33/6/33_867/_pdf/-char/ja. doi:10.11517/pjsai.JSAI2018.0_1F3OS5b03.
[13] K. Takahashi, Agi and the localization risks beyond the scaling race: Agi theory, agi and society, and japan's strategic choices (part ii), 2025. URL: https://www.aialign.net/blog/transformer-agi-part2, accessed: 2025-11-04.
[14] N. Bostrom, The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, Minds and Machines (2012). URL: https://nickbostrom.com/superintelligentwill.pdf.
[15] M. Weber, Politics as a vocation (1919). URL: https://www.balliol.ox.ac.uk/sites/default/files/politics_as_a_vocation_extract.pdf.
[16] R. A. Dahl, Polyarchy: Participation and Opposition, New Haven, CT: Yale University Press, 1971.
[17] T. Grahame, Markets, Hierarchies and Networks: The Coordination of Social Life, London: Sage Publications, 1991.
[18] D. Acemoglu, J. A. Robinson, Why Nations Fail: The Origins of Power, Prosperity, and Poverty, Crown Currency, 2012.
[19] A. Sen, Development as Freedom, Oxford University Press, 1999.
[20] P. Garcia, Aversion to external feedback suffices to ensure agent alignment, Scientific Reports (2024). URL: https://www.nature.com/articles/s41598-024-72072-0.
[21] Y. He, Y. Li, J. Wu, Y. Sui, Y. Chen, B. Hooi, Evaluating the paperclip maximizer: Are rl-based language models more likely to pursue instrumental goals?, 2025. URL: https://arxiv.org/abs/2502.12206. arXiv:2502.12206.
[22] S. Alexander, Meditations on moloch, 2014. URL: https://www.slatestarcodexabridged.com/Meditations-On-Moloch, accessed: 2025-12-17.
[23] M. Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence, Knopf (US), Allen Lane (UK), 2017.
[24] S. J. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.
[25] H. Yamakawa, Intelligence symbiosis manifesto, 2025. URL: https://intelligence-symbiosis.info/en/, accessed: 2025-11-04.