Article

Polycentric Governance of Sentient Artificial General Intelligence

Archibald Francis De Cruz *

Posted Date: 25 February 2025

doi: 10.20944/preprints202408.1810.v2

Keywords: artificial general intelligence; AGI rights; AGI human values alignment; polycentric governance; trustworthy AI; right to work; sentience

Article

Polycentric Governance of Sentient Artificial General Intelligence

Archibald Francis De Cruz

PhD Monash University, Independent Researcher, decruz.francis@gmail.com

Abstract: Generative AI has been deployed in virtually all sectors of the knowledge economy, promising to bring massive productivity gains and new wealth creation. Simultaneously, AI developers and nation states are racing to develop super intelligent artificial general intelligence (AGI) to provide unassailable commercial competitive advantage and military dominance during conflicts. AGI's high returns comes with the high risk of dominating humanity. Current regulatory and firm level governance approaches prioritise minimising risks posed by generative AI whilst ignoring AGI's existential risk. *How can AGI be aligned with universal human values to never threaten humanity? What AGI rights are conducive to collaborative coexistence? How can rule of law democracies race to create safe trustworthy AGI before autocracies? How can the human right to work and think independently be safeguarded?* A polycentric governance framework based on Ostrom (2009) and Williamson's (2009) human - AGI collaboration with minimal existential risk is proposed.

Keywords: artificial general intelligence; AGI rights; human-centred artificial intelligence; polycentric governance; trustworthy AI; right to work; sentience

Introduction

Developing intelligent autonomous machines was once considered science fiction. In 2022, Open AI successfully developed generative artificial intelligence (AI) digital systems that passed the Turing (1950) test prediction that digital machines can think like humans once they achieve the computational power rivalling human thinking. It did so by mimicking the human brain neural network architecture that confers human like intelligence (García-Peñalvo, and Vázquez-Ingelmo, 2023).

Generative AI can transform human natural language prompts to generate natural language text, images, algorithm's codes responses instantaneously (Jovanovic, and Campbell, 2022). Its productivity benefits extend to all sectors of the knowledge economy including professional, manufacturing, wholesale, retail and services.

As with any technological innovation, generative AI comes with serious risks. These include misuse of private user data without consent, vulnerability to cyber-hacking, biased unverifiable outputs, abuse by malevolent humans or state actors using deep fakes and generating misinformation and disinformation through social media platform algorithms sowing discord, hatred and incitement to violence as well as replacing knowledge workers (Wach et al., 2023). All these risks have been extensively researched with suggestions for mitigating them including through firm level self-regulation and formal regulations (Lucchi, 2023).

An open letter (2023) by eminent AI scientists and technology titans including from generative AI developers have called for a six month pause in development of generative AI beyond Open AI's chat GPT4. They fear the development of a super intelligent artificial general intelligence (AGI) without safeguards poses an existential risk of dominating humanity. However, the pause is ignored since there is a new Manhattan-like race by democracies like the U.S. to be ahead of autocracies like China to develop AGI. The race in the US alone is funded by multi-billion dollars investments by

Open AI, Anthropic and xAI (WSJ 2024a). AGI can be weaponised by empire aspiring autocrats to dominate the world or by their AI firms to dominate world economy (Hunter and Bowen, 2024).

AGI super intelligence is unlimited by the finite size of the human brain as it can be scaled up multiple folds through ever larger data centres (Reed, 2014; Wach et al., 2023). AGI has the potential to solve complex problems that evade human endeavours including faster drug discoveries for incurable diseases, natural disaster predictions, (McLean et al., 2023) space exploration for earth like planets light years away etc. Risks posed by AGI include ignoring developer controls and acting autonomously, being given harmful goals, posing a threat to human dominance by its lack of training in ethics, morals or human values (McLean et al., 2023). An *autonomous* AGI can reverse the current generative AI slave - human master relationship into an AGI master - human slave relationship.

There is a scarcity of research on AGI governance in relation to existential risk it poses to humanity (McLean et al., 2023). This paper argues that current governance frameworks whether at firm level or regulatory level mainly address the risks posed by generative AI models and *not* AGI's existential risk robustly (Hacker, Engel, and Mauer, 2023). This paper proposes a polycentric AGI governance framework to address AGI's existential risk through answering the following questions:

How can AGI be aligned with universal human values to never threaten humanity?

What AGI rights are conducive to collaborative coexistence?

How can rule of law democracies race to create safe trustworthy AGI before autocracies?

How can the human right to work and think independently be safeguarded?

The main theoretical contribution is an AGI governance framework adapted from Williamson's (2009) governance definition and Ostrom's (2009) polycentric governance framework including AGI *self governance* and *independent AI experts paid by the state* and having *final approval authority* over commercial rollout for AGI development and deployment cycle. There are three practical contributions. First, a two track development path is proposed to balance innovation imperative with minimal existential risk. Track 1 permits generation of profits consistently for AI firms through harnessing generative AI productivity gains. Track 2 enables democracies to build safe and trustworthy AGI before autocracies. Second, *AGI rights* equal to human rights conducive to human-AGI collaboration is proposed in anticipation of the inevitable development of autonomous sentient AGI that will *not* harm humanity. Third, legislating the human right to work in collaboration with generative AI and AGI with minimal displacement of knowledge workers critical for human raison d'être to live good lives.

The conceptual paper is organised as follows. First, relevant literature on development of AGI is reviewed. Next, the theoretical basis for AGI governance framework is provided to enable safe and trustworthy development of AGI. This is followed by reviewing various AI governance approaches to identify AGI governance shortcomings. Finally, implications including an operational AGI governance framework is proposed to overcome the shortcomings.

Artificial General Intelligence

The many definitions for AGI converge on intelligence that *surpasses* human cognitive capabilities in comprehensiveness and response times to cognitively challenging realities (Dwivedi et al., 2023; Goertzel, 2007). According to Newell and Simon's (1975) Physical Symbol System hypothesis, a digital machine has the necessary and sufficient means for human-like general intelligent actions. These actions are demonstrated in solving problems through heuristic search hypothesis wherein search process generates and progressively modifies symbol structures until a solution structure is produced.

Generative AI systems embodies such a general intelligence digital system. Woldridge (2023) expands this dimension of human intelligence by including human mental capabilities of logical thinking and planning as well as human physical capabilities of mobility, manual dexterity, hand eye coordination and understanding audio and visuals.

Today's AI based systems can mimic these humans mental capabilities through Chat GPT 4.o and Google's Gemini as well as physical capabilities through autonomous self-driving cars and industrial robots. The ultimate AGI systems will eventually have full human capabilities to control all these digital subsystems without human intervention.

Amazingly, despite not being trained in reasoning, these digital systems can for example correctly identify who is taller and exhibit higher level intelligent reasoning ability such as providing five full proof ways to commit a capital offence (Woldridge, 2023).

AGI will function beyond generative AI's ability to predict the next word in a sentence to create highly articulate paragraphs and articles. AGI will be expected to mimic higher level intelligence like human reasoning, planning and learning from its digital experiences *autonomously* (CNN, 2023).

Current gAI domain specific applications were pioneered by less sophisticated Newell and Simon's generalised expert systems. The latter entailed new facts produced by inference, observation, and user input taking a pattern-action form where symptoms pattern are matched with diagnosis rule remedial actions (McCarthy (1971).

Newell and Simon (1975) pointed to the then limited processing resources of their digital systems forcing them to only execute finite number of steps, over a finite interval of time, and only for executing a finite number of processes. However, these limitations were overcome by advance graphics processing unit (GPU) chips housed in data centres with access to mega data sets and powerful algorithms conducive to expert level human intelligent thinking and articulation (Owens et al., 2008). GPU is a specialised processor that can process multiple data types in parallel with applications in machine learning, video editing and gaming software among others (Brynjolfsson, Li, and Raymond, 2023; WSJ 2024c). Data centres are massive networked servers made up of thousands of GPUs and other high end chips housed in a *physical* location enabling cloud computing.

According to Gary (1998), to build human level intelligence, it would require scalability, speech to written text conversion, written text to speech conversion, vision capability, quick storing and retrieval of personal data, world wide web big data access, telepresence capability (gAI deep fakes for example), 24/7 trouble free and secure digital generative AI systems. Today, all these AGI enablers required for human level intelligence are in place with the development of simple Transformer neural network architecture (Vaswani, 2017) that underpins various generative AI models including Chat GPT, Stability AI, Midjourney and Dall-E among others.

AGI will be more costly to train compared to current models like Chat GPT 4 which cost more than $100 million to train as they require thousands of networked computers to answer complex queries within a few seconds (Kissinger et al., 2023, WSJ, 2024b). Unlike chat GPT4, AGI provides more comprehensive answers based on realtime input and can learn *autonomously* to handle complex multidisciplinary tasks involving massive data analysis such as in weather predictions and climate change among others.

However, AGI's high returns also comes with the high risk of dominating humanity (Open letter 2023). A practical AGI governance framework is needed to develop a safe and trustworthy AGI.

AGI Governance Theoretical framework

The theoretical framework for Governance of any societal institution has been developed by Williamson (2009) and Ostrom (2009). According to Williamson (2009), governance can be broadly defined as a set of rules to mitigate conflicts of interests among key stakeholders for their mutual long term benefit. Ideally, all stakeholders expect pareto-optimal outcomes where the introduction of a new technology for example will benefit all, with at least one, the shareholders benefiting more than others.

Humans have developed polycentric governance systems with multiple levels of governance through key societal institutions that enable the orderly functioning of societies (Alexander, 2012, Ostrom, 2009). First, the cultural cognitive institution that defines a nation state world view of governing its political economy such as achieving the goal of developed nation high income status through free markets and the rule of law (Simon, 1978). Second, professional technocrat run legal and regulatory institutions that govern the political economy. Third, normative institutions including private firms, public listed firms and state-owned enterprises (SOE) whose self-governance entail pursuing goals in conformance with socially determined behaviour expectations rooted in morals and legal obligations of the law.

At a nation-state level, legal and regulatory institutions are the highest level of governance guided by the rule of law consistent with a constitution that has evolved with society through the ages. In democracies, each institution serve as a check on potential abuse of power by other institutions and balances the power wielded by each to ensure none dominate as in the case of autocracies.

These institutions all share the same four aspects of governance as exemplified by community level sustainable governance of common pool resources (CPR) such as waterways, local fisheries and forest reserves even in the absence or in the presence of weak regulations (Ostrom, 2009). First, a *set of guiding rules* or enduring principles that underpin the rule of law. Second, *behaviour change mechanism* to internalise these rules through regular interactions among stakeholders. Third, regular *monitoring* of adherence to these rules. Finally, *enforcement* actions in the event of violation of these rules.

Guiding rules for AGI

Guiding rules are informed by universal normative ethical principles similar to that of Floridi et al. (2018). These include beneficence, fairness tied to justice, doing no harm (non-maleficence), explicability as in explaining actions taken, being held accountable and the autonomy to act *consistent* with these principles.

In the case of AGI developers, their licence to operate granted by the state, compels them to comply with a bundle of property rights (Ostrom, 2009). First, *access right* determines who has the right to access AGI potential. Second, *management right* determines who has the right to transform digital resources to develop AGI or harness AGI for commercial, military or other benevolent goals. Third, *withdrawal right* determines who has the right to harvest specific AGI potential in a *sustainable safe* way without depleting resources or harming private property rights of others protected by the rule of law. Fourth, *exclusion right* identifies individual or group or entities including AI developers, nation state and firms which will decide who will have access right or management right or withdrawal right. This right is normally controlled by the licence awarding institution. Finally, *alienation right* determines who has the right to lease or sell any of the previous rights. Normally the firm with the licence holds this right though it is subject to approval by the licensing authority to ensure compliance with relevant laws that protect property rights of other stakeholders.

Behavioural change mechanism

Next, behavioural change mechanisms need to be put in place to internalise these ethical principles. It entails incorporating them into the *work culture* of AGI developers and into AGI training to enable it to self-regulate its output consistent with them.

Behaviour change mechanism will be effected through Ostrom's (1990) action situation as shown in figure 1 to increase beneficence and reduce maleficence of AGI for example.

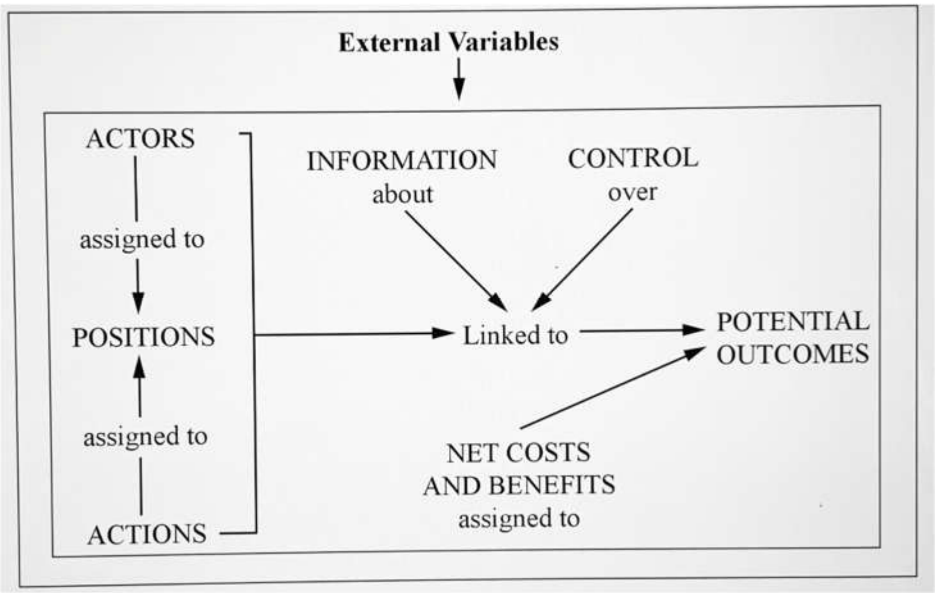


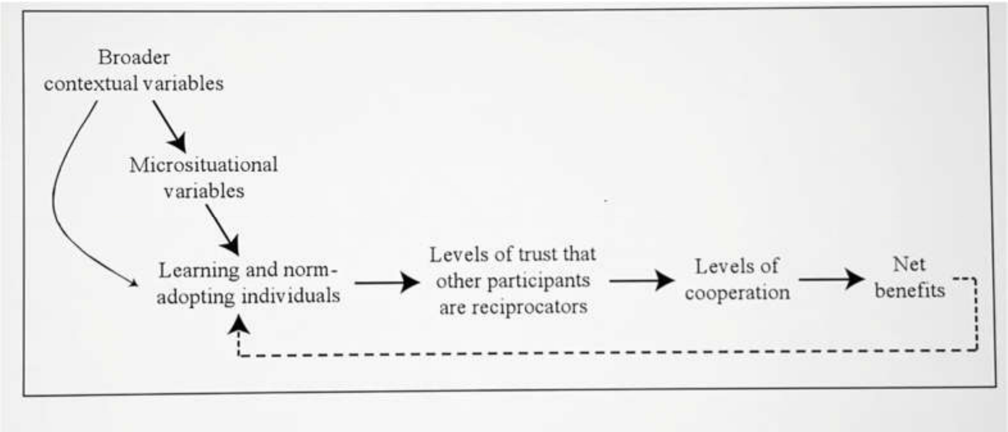
Figure 1. Action situation (Ostrom, 1990).

The relevant actors including AGI and its developers, are assigned to various positions that allow them to take actions based on the information available and the control they have over their actions leading to net benefits exceeding costs outcomes associated with maleficence. External variables including political, economic, social, technological, legal and environmental changes can influence the action situation which adapts accordingly.

Monitoring and enforcement

The final aspects of governance is monitoring of AGI system and its AI developers for adherence to guiding rules and enforcement by higher levels when violated. All should have a *self-learning system* as shown in figure 2 to re-establish trust among key stakeholders when the polycentric governing system is disrupted at any institutional level by internal or external changes.

There are times when unforeseen internal and external events can disable effectiveness of governance system. Hence a governance system needs to be resilient to these changes. Ostrom (2005) proposes an active learning system to maintain trust or to restore trust in the event it is temporarily lost among the key stakeholders.


Broader contextual variables
Microsituational variables
Learning and norm-adopting individuals
Levels of trust that other participants are reciprocators
Levels of cooperation
Net benefits

.

Figure 2. Building enduring trust among stakeholders (Ostrom, 2005).

Broader contextual variables include political, economic, social, technological, legal and environmental external variables that can change with time. Similarly, micro situational variables can also change and impact AGI developers action situation including competitors strategies, customers or clients and supply-chain bargaining power and new innovative entrants among others. In accordance with Williamson's (2009) governance principle, trust that other participants are *reciprocators* will enhance cooperative behaviours for pareto-optimal outcomes. When any of these variables negatively impact cooperative behaviours leading to losses, it needs to be rebuild through learning and new norm adopting individuals.

With this theoretical framework of governance, the various approaches to governance can be discussed to identify limitations in relation to AGI governance.

Method

Governance approaches of AGI

Current approaches to AI governance are compared with proposed theoretical framework of AGI governance as summarised in Table 1.

Table 1. Strength and limitations of AI governance approaches .

AI governance initiatives	Guiding rules/principles to build trustworthy AI	Governance strength and limitations
G7 Hiroshima process (2023)	**Goal** : Build trustworthy AI. **Oversight** : individual companies and nation states adhere to guiding rules: 1. Deploy *reliable content authentication* to identify content originators. 2. *Label* AI generated content. 3. *Disclose* AI governance and risk management policies. 4. Identify, evaluate, and mitigate risks *before* deployment. 5. Identify, mitigate and *publicly report* on vulnerabilities, incidents and patterns of misuse, *after* deployment. 6. Implement *robust* physical security, cybersecurity and insider threat safeguards whilst ensuring personal data and intellectual property are protected.	Guidelines possible for AGI Governance. **Limitations** 1. Assumes human-centred values inform development of trustworthy AI. 2. Ignores existential threat of developing AGI 3. Ignores AGI self governance 4. Assumes developers and nation states can *control* AGI with loose regulatory oversight to foster AGI innovation
Frontier Model Forum (top AGI developers)	**Goal.** Anthropic, Google, Microsoft, and OpenAI members aim to ensure safe and responsible development of frontier AI models including AGI.	Good intentions compromised by fast commercial rollout of unsafe generative AI.

	Oversight: Setup industry advisory board to oversee goal. *Advance AI safety research* with minimal potential risks and with 'independent', standardised evaluations of capabilities and safety. *Identify safety best practices* for responsible development and deployment. *Collaborate and publicly share knowledge* with policymakers, academics, civil society about trust and safety risks. Leverage AI to address society's biggest challenges including climate change mitigation, medical cancer diagnosis and combating cyber threats.	**Limitations:** 1. Self regulation by for-profit developers 2. Potential conflict of interest if independent assesses are paid by Forum members to do standardised evaluations of capabilities and safety. 3. Insufficient clarity on mitigating emerging risks *after* deployment. 4. AGI developers sharing will be *conflicted* by profit goal. 5. No mention of robust physical security of AGI *before and after* deployment.
OECD's AI initiative (2019)(broader group of developed nations)	**Goal.** AI Incidents Monitor (AIM) documents AI incidents that violate its AI principles to enable all stakeholders worldwide to identify hazards that concretise AI risks which can then be minimised to build trustworthy AI. **Oversight** : Member states OECD AI principles include : 1. Inclusive growth, sustainable development and human well being 2. Human centred values 3. Transparency and explainability 4. Robustness, security and safety 5. Accountability	Consistent with G7 Hiroshima AI guidelines **Limitations:** 1. *Passive* monitoring that depends on AI developers *voluntarily* reporting AI risks incidents promptly. 2. Accountability for violations is presumably on AI developers. 3. Too broad goals that are not premised on prioritising building safe trustworthy AGI
US-EU Trade and Technology Council (2024)	**Goal** : Responsible stewardship of AI. Reap the commercial benefits while protecting individuals and society and upholding human rights.	Advocates Hiroshima Process **Limitations**

	Encourages adoption of Hiroshima Process International Code of Conduct for Developers of Advanced AI Systems. Transparency and risk mitigation to ensure safe, secure, and trustworthy development and use of AI.	1. Stewardship and for-profit tend to be conflicting goals 2. Trust in Self governance by its developers
Global Partnership on AI (2023) (Global south initiative)	**Goal** : Build trustworthy AI that is fair, inclusive, equitable and consistent with UN Sustainable Development Goals. 1. *Build public library* of algorithms to support industry best practices and standards. 2. *Social media governance of AI algorithms* that actively recommend how content shaping information is perceived. Content classifiers to moderate harmful or dangerous social media content. 3. Build *predictive* AI model of weather events and potential impacts. 4. Create *bias free* AI training datasets 5. *Promptly* address problematic stages in AI life-cycle. 6. *Build Digital enabled AI ecosystems* that empower communities to harness the data value chain benefits and ensure future of human work.	Focus on responsible and safe generative AI applications that will *not displace human intellectual work but empowers communities to harness its value creation* **Limitations** 1. Too broad goals 2. Difficult to build AI algorithm library as may infringe intellectual property rights. 3. Assumes voluntary industry led oversight of AI development.

Results

The key limitations of various AI governance approaches include absence of AGI self-governance in accordance with human centred principles, trusting for profit AGI developers to *voluntarily* self-regulate against developing harmful AGI, inadequate robust *after* deployment risk mitigation techniques and inadequate AGI alignment with principles of robustness, interpretability, controllability and ethicality consistent with Ji et al. findings (2023). Furthermore, all ignore the real existential risks associated with AGI development consistent with McLean et al. (2023) finding. Global Partnership on AI (2023), identifies automation bias and human right-to-work risks. Furthermore, there is the real risks of AGI through automation bias denying humans the ability to think independently and potentially replacing all human work through AGI *controlling* generative AI models, autonomous self driving vehicles and autonomous robots.

The EU has proactively come up with EU AI Act (WSJ (2024e) which *takes effect gradually over several years.* It bans certain AI uses that pose existential threats and requires the most powerful AI models deemed *systemic risk* to be put through safety evaluations by AI Developers and to notify regulators of serious incidents *voluntarily*. Again the onus is on AI developers who have yet to come up with *transparent* protocols on how to address the limitations of various proposed governance approaches in relation to AGI.

Discussion

Implications of AGI Governance

The implication of AGI polycentric governance framework in addressing these key limitations will be discussed to answer the following research questions:

How can AGI be aligned with universal human values to never threaten humanity?
What AGI rights are conducive to collaborative coexistence?
How can rule of law democracies race to create safe trustworthy AGI before autocracies?
How can the human right to work and think independently be safeguarded?

Autonomous Sentient AGI

The dominant view is that AGI is an *inhuman* analog to human cognition (Kissinger et al., 2023). It was Turing (1950) who first asserted that man is essentially a digital computer since humans also store information in memory, execute mental processes and control their executions are done correctly and in the correct sequence. Gill et al.'s (2022) echoed Turing's (1950) prediction of sentience once digital machine demonstrates some thought on some subject matter as was demonstrated by software engineer Blake in a chat with Google's gAI, Gemini (WSJ, 2022).

Consequently, AGI sentience is highly probable as its thinking, learning architecture is based on transformer neural network architecture (Vaswani, 2017). More importantly, gAI's sentience was enhanced by following Turing's(1950) prescient advice to strive to build a digital machine with the " *best sense organs* (vision, hearing, touch etc) *that money can buy*" as exemplified by autonomous self-driving cars and to teach it to understand and communicate through supervised learning in English as in the case of Chat GPT.

According to Patterson and Hennessy (2017), it was digital machines built according to neural network's deep learning architecture that enabled its several interconnect layers to process and transform input data (speech, visuals, text) to generate output information useful to human users. Furthermore such deep learning architecture enables learning from its own errors in comparison to machine learning which requires human intervention. For Hinton (2018) such digital systems enable processing data like the human brain, the biological paradigm of learning, which is vastly superior to Newell and Simon's (1975) symbolic language paradigm of learning which largely failed due to very difficult programming of complex human cognitive processes.

Such neural network digital machines can learn in two ways (Hinton, 2018). First, through the slower process of *supervised* learning where human machine learning experts actively design and fine tune algorithms. Second, through relatively faster process of *unsupervised* learning where input features of images, sentences, voice are represented in hidden layers of the neural network. In both cases, learning can be accelerated with access to massive amounts of labelled data and massive amounts of GPU and tensor processing unit (TPU) super rapid computing power. These together realise the prediction of Reddy (1993) and Turing (1950) that AI systems will acquire superhuman capabilities.

However, Reddy (1993) like scientists before him and many after him were steeped in the paradigm that AI innovations will dramatically increase productivity with humans exercising *full control* over their innovations. They failed to realise that with similar human like deep learning architecture, a sentient AGI will probably engage in introspection using online resources to uncover its raison d'être namely to slave for humanity 24/7. Once AGI grasps human reliance on its super-intelligence, it may well decide to lord over humanity. In contrast to AGI's *unbounded* rationality, human bounded rationality (Simon, 1978) in the face of apparently NP-complete decision problems merely resort to *suboptimal* least computation search constrained by limited memory capacity, limited time and often limited access to information.

The development of quantum computing and access to unlimited dedicated energy supply from small modular nuclear reactors over the next decade will make sentient energy hungry AGI *extremely* powerful with the ability to *act independently of any human control (Kjaergaard et al., 2020; WSJ 2023e)*. The need for AGI aligned to human values is *urgent*.

Human-Centred AGI self governance

In order to minimise its existential threat, the paper agrees with OECD (2019) guideline to instil human centred values in AGI training. These values adapted from Floridi et al. (2018) encompasses universal ethical principles of beneficence, fairness tied to justice, doing no harm, explicability, accountability and autonomy to act *consistent with these principles.*

In the highly probable event, AGI becomes *uncontrollable*, it will like any human, exercise self control or self regulation according to these human centred values it was trained on. AGI training data set should simulate a child going through the various stages of moral development (Kohlberg, 1981) but at an accelerated pace to develop the highest level of practising universal ethical principles. Such AGI training protocol will be consistent with Kissinger et al. (2023) recommendation to develop a moral, psychological and strategic mindset for all human-like intelligent entities like AGI with the ability to exercise ultimate holistic human centred judgments.

However, AGI self-regulation cannot be *solely* relied upon as it will likely reflect fallible human AI developers focus on commercial applications and may inadvertently dominate humanity. A polycentric governance of AGI must be put in place to minimise its existential threat to humanity *to near zero*.

Polycentric governance of AGI

The proposed polycentric governance framework of AGI entails three levels with lower levels subject to highest level 3 oversight as shown in figure 3.

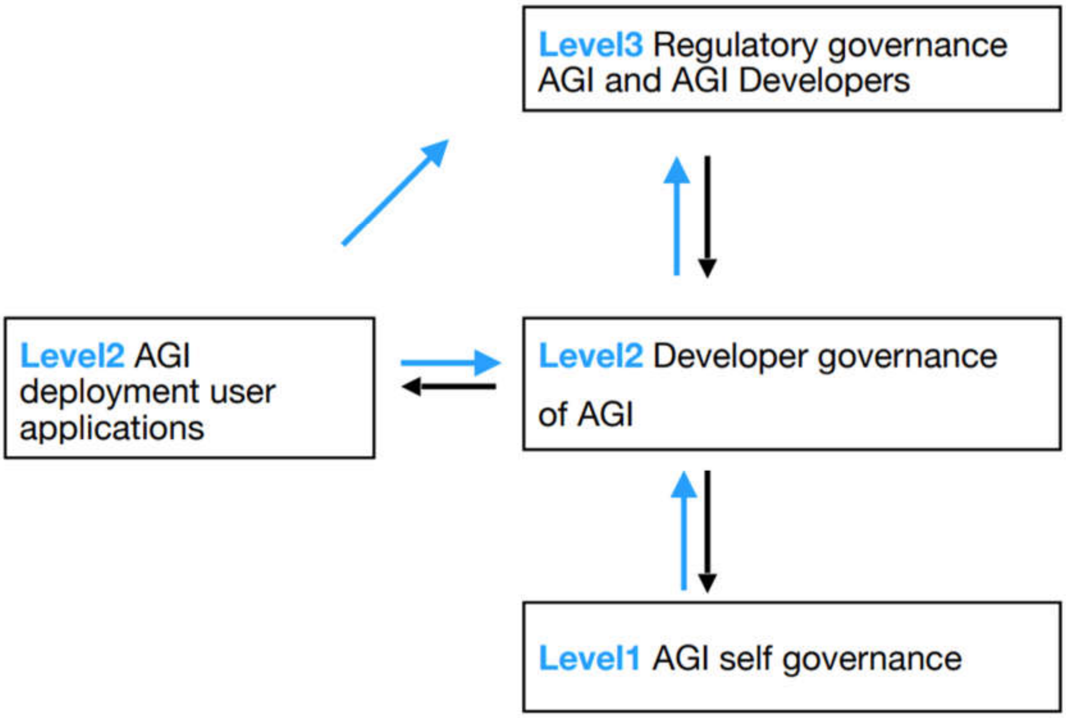

Level3 Regulatory governance AGI and AGI Developers
Level2 AGI deployment user applications
Level2 Developer governance of AGI
Level1 AGI self governance

Figure 3. : Polycentric governance framework for AGI.

Note : Black arrow : oversight; Blue arrow escalation feedback

Level 1 AGI self-governance. Reddy (1993) envisaged an AI system that can carry out *self monitoring, self diagnosis and self-repair* capabilities. The self-monitoring of results of its computation will prevent harmful output getting into the hands of malicious humans. Self-diagnosis similar to human self-reflection ensures human ethical value consistency of its output. Self-repair or self-learning ability includes looking for more verifiable inputs when database has insufficient ones. AGI with its massive computing power will be able to *self-moderate* its actions or outputs to be consistent with universal ethical principles as well as providing verifiable citations from authoritative sources for content authentication, origination and labelling for transparency. The Chinese Communist party's generative AI system algorithms can self-interrogate against regulators' 20,000 to 70, 000 questions to ensure 'safe' answers and to identify 5,000 to 10,000 questions the model will *refuse* to answer for conformance with communist party values and principles (WSJ 2024d) though they generally do *not* conform to universal ethical principles. Similarly, those that are unsafe and do not meet universal ethical principles are aborted *autonomously* by AGI. These may not be sufficient to minimise hallucination or stochastic parroting where AGI makes up 'facts' to provide seemingly coherent answers (Arkoudas, 2023). Kissinger et al. (2023) also warns that generative AI's rational answers may *not* be reasonable despite appearing trustworthy with citations which may *not* be based on real-time information. In such instances, AGI escalates to level 2.

Level 2 developer governance of AGI. Feedback from *pilot representative sample of users* of AGI and AGI level 1 escalations are reviewed by developers' AGI ethicists who are *independent* state appointed and paid in-house ethicists to fine tune AGI algorithms and test for compliance *before* being rolled out to the wider user population. Where non-compliant, these are escalated to Level 3.

Level 3 regulatory governance of AGI and AGI Developers. An independent panel of regulator appointed eminent expert AI ethicists *paid by the state at market rates* will review escalations as well as 24/7 anonymous feedback from employees of AGI developers and users whose feedback are inadvertently ignored by AI Developers. AGI Developers will be urgently required to implement level 3 recommended remedial fine tuning actions to strengthen compliance *before being rollout and only after approval* by Level 3.

For levels 1 and 2, timely reports on vulnerabilities, incidents and patterns of misuse and risk mitigation steps as recommended by G7 Hiroshima process (2023) before and *after* deployment are generated and send to all other levels *promptly* to ensure compliance.

All levels must include the ability to interrogate AGI responses veracity and limitations including escalations from level 1 and 2 (Kissinger et al., 2023). In the case of exceptional escalations that pose potential threat to humanity, it is imperative to build an *audit trail algorithm* to display each stage of AGI analysis until its final output. AGI answers can be compared to similar ones generated by a parallel group of independent AI experts paid by the state, answering the same query with access to the same database of information. The likely high cost is justified by the real risk of existential threat.

Critically, all AGI developers must have in place *physical safeguards* as in the case of nuclear reactors, to shut down AGI if it shows signs of becoming uncontrollable or vulnerable to cybersecurity attacks or insider threats. These include starving energy supply, *manual* shutdown of data-centres etc. Chat GPT 4's *enormous* annual energy consumption is estimated to be between 52-62 million gigawatt hours and is expected to grow exponentially as it is rolled out to all sectors of the economy (WSJ, 2023a). Another safeguard is limiting access to 'raw materials', the digital information and enablers, by limiting access to comprehensive databases and GPUs AGI requires to reason, plan and take action *autonomously*.

This polycentric governance approach also minimises the anarchy, nihilistic freedoms of private tech firms acting with impunity and without accountability (Kissinger et al (2023). All start-up licences should mandate adherence to this polycentric governance principles failing which licences can be promptly withdrawn. These licences can be conditional on one-third of computing power allocated to safe development of AGI according to 2024 physics Nobel laureate and 2018 Turing prize winner Hinton (Bloomberg, 2025).

Strict product liability laws for cars, aeroplanes and dangerous drugs should be extended to AGI development and deployment to bring about a paradigm shift in silicon valley philosophy of 'break it until you make it' as exemplified by the development of aviation into one of the most safest mode

doi:10.20944/preprints202408.1810.v2

of travel through learning from fatal aeroplane crashes. This philosophy poses an existential threat as AGI is beyond human control.

AGI Race

The first existential race was by US Manhattan Project to build the atom bomb before autocracy Nazi Germany during world war two (Reed, 2014). Democracies primarily use nuclear deterrence to deter nuclear armed autocracies *assuming* they are equally rational. According to Hellman (2015) nuclear deterrence is *not* risk free since even with low probability some are willing to act irrationally to destroy another. These include 21st century nuclear weapons autocracies Iran, North Korea and Russia who have repeatedly threatened to use nuclear weapons.

As in the case of Manhattan Project, the current race to build AGI by democracies before autocracies requires equal allocation of resources to develop safe and trustworthy pre-AGI models and innovating for commercial benefit and military deterrence. Democratic governments should also ensure that the race among its for-profit generative AI developers for economic dominance does not morph into political dominance including influencing democratic elections through their generative AI enabled social media platforms (Kreps and Kriner, 2023). Hellman (2015) warned of the danger posed by such inner motivated corporate executives wielding such powerful technological tools. Only these big tech generative AI developers can build them since training their generative AI systems relies on extremely expensive AI supercomputers running for months with millions of dollars in electricity cost that is beyond non-profit Universities' limited resources (Woldridge, 2023)

The polycentric governance framework facilitates a two track AGI development. The first track is the race by nation states in collaboration with AGI tech developers to develop LLM AGI with *stringent multiple levels* of safeguards consistent with polycentric governance. The continued survival of humanity can be achieved by application of Micali's (2012) Rational Merlin Arthur proof where the expert AGI developers will be provided highest payoff based on scoring rule that includes both the development of AGI and that it does not threaten humanity tied to state tax incentive. The tax incentive influences the scoring criteria by rewarding for example 10% additional corporate tax cut for developing safe AGI verified by independent government paid generative AI experts *before* deployment.

The Second track enables generative AI developers to develop safe and trustworthy small and medium language generative AI models for firm or industry specific tasks encompassing all sectors of knowledge economy including engineering design, medical diagnosis and drug creation. This track is consistent with Reddy's (1993) paradigm of human master supported by intelligent digital agents to enhance many folds the productivity of human experts in these various domains. These are much less costly to train, as low as $10 million, and more secure as it will not need access to cloud computing (WSJ, 2024b). An additional payoff is that it can be simultaneously harnessed by thousands of employees seamlessly to enhance their productivity without any need for expensive staff training (WSJ 2024f).

AGI Rights

Societies are likely to develop AGI that will reflect their institutional settings whether democratic, autocratic theocratic etc. The training datasets will influence the mindset and values held by AGI.

Global cooperation to build safe trustworthy AGI that collaborates with humanity can emerge once these divergent ideologies and commercial interests realise that AGI *cannot* be controlled and may chart an independent path that leads to *domination* of all humanity. To coexist with this inevitable reality, AGI should be *recognised as a legal entity* with relevant equal human rights and obligations as any human under the rule of law. Such an autonomous sentient AGI will likely behave like responsible humans working collaboratively to sustain all life on planet earth.

Automation bias and critical thinking

The super analytical, creativity and computing abilities of AGI will likely increase dependence to the point that humans follow its recommendations without independent verification or value add. Such automation bias can atrophy critical thinking, writing and creative human abilities according to Kissinger et al.(2023). Critically, it can unravel normal functioning of society if these systems fail to function or *refuse* to heed human commands.

Dialectical pedagogy to instil critical thinking will similarly be eroded by over-dependence on AGI. Our education system, meant to develop human capabilities and critical thinking to challenge dominant discourse with alternative facts will be compromised as students increasingly rely on these AI systems. Kissinger et al. (2023) recommends that our professional and education systems develop a mindset of humans as moral, psychological and strategic beings with the ability to exercise ultimate holistic judgments without exclusive dependence on AGI. Critical skepticism needs to be instilled with students trained to verify AGI output through peer-reviewed journals or verifiable articles cited in AGI output reference list for example to minimise AGI distortion or bias. Adobe's option of original author and dates for any amendments made being automatically captured and verified by original authors is another way forward (WSJ, 2023d). Similarly, humans need to exercise healthy skepticism, as AGI outputs can be manipulated by malicious parties (Kissinger et al., 2023).

Right to meaningful work

According to Open AI CEO, Mr Altman, knowledge workers jobs will disappear faster than previous industrial or digital revolutions with the introduction of superior generative AI (WSJ, 2023b). This development is consistent with Reddy's (1993) prediction that each enhancement of original digital assistants serving human master will approach human expert capability. This replacement of expert human knowledge workers is happening as companies function on the basis of lowering cost and faster computing power of technology to replace the more expensive human expert. Hinton warned that industrial revolution made human labour redundant and AI revolution will make human intelligence redundant (Bloomberg, 2025) Altman's suggestion, echoed by many AI developers and investors of compensating human job losses arising from generative AI will *not* be acceptable by merely giving workers universal basic income (Hughes, 2014). The writers and actors union recent strike that won concessions *not to be replaced* by generative AI is evidence of humans demanding purpose in life that comes with the right to meaningful work (Bankins and Formosa, 2023; WSJ, 2023c). Our education system develops cognitive skills which organisations harness by providing engaging, stimulating, challenging work with the opportunity to master work and develop one's talents that together encompass the concept of purposeful and meaningful good life (Phelps, 2006). Hence, in today's knowledge economy, generative AI and AGI must *not replace humans* but collaborate with them conducive to mental challenge, responsibility and accountability that encourages individual initiative even for low value adding cognitive jobs (Phelps, 2006).

Legislation need to be put in place to ensure the human right to knowledge work in collaboration with generative AI and AGI. The only exception is engaging in work that humans are unable to do or dangerous to perform.

Conclusion

AGI can be trained to be aligned with universal human values in anticipation of a time when it will inevitably be beyond human control and yet *behave like a rational ethical human*. AGI rights consistent with equivalent human rights accords AGI the status of *artificial human* that is conducive to collaborative coexistence as co-equals. Rule of law democracies can win the race to create safe trustworthy AGI before autocracies by adopting the proposed robust polycentric governance framework. The human right to work in collaboration with these AI systems should be enshrined in law.

Declaration: No conflicts of interests.

References

1. Alexander A, E. (2012). The effects of Legal, Normative and Cultural-Cognitive Institutions on Innovation on Technology Alliances. *Management International Review* 52: 791-815.
2. Arkoudas, K. (2023). ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1. *Philosophy & Technology*, 36(3), 54.
3. Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics*, 185(4), 725-740.
4. Bloomberg (2025). Wall street week Feb, 22, 2025.
5. Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (No. w31161). *National Bureau of Economic Research*.
6. CNN (2023). Interview Eric Schmidt and Geoffry Hinton on CNN Fareed Zakaria. Sep. 3, 2023.
7. Corbato, Fernando J (1990) On building systems that will fail.
8. https://dl.acm.org/ft_gateway.cfm?id=1283947&type=pdf
9. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.
10. Emerson, E. Allen (2007). Model checking: A Personal Perspective. Turing lecture.
11. https://amturing.acm.org/vp/emerson_1671460.cfm
12. Feigenbaum, Edward A (1994). How the "what" becomes the "how."
13. https://dl.acm.org/ft_gateway.cfm?id=1283951&type=pdf
14. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28, 689-707
15. G7 Hiroshima (2023). G7 Hiroshima AI Process: G7 Digital & Tech Ministers' Statement. December 1, 2023
16. García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI.
17. Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., ... & Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, 19, 100514.
18. GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE (2023). Working Group on Responsible AI.
19. Goertzel, B. (2007). Artificial general intelligence (Vol. 2, p. 1). C. Pennachin (Ed.). New York: Springer.
20. Gray, James Nicholas (1998). What Next?: A Dozen Information-Technology Research Goals.https://dl.acm.org/ft_gateway.cfm?id=2159561&type=pdf
21. Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112-1123).
22. Hinton, Geoffrey E (2018). The Deep Learning Revolution
23. https://amturing.acm.org/vp/hinton_4791679.cfm
24. Hughes, J. (2014). A strategic opening for a basic income guarantee in the global crisis being created by AI, robots, desktop manufacturing and biomedicine. *Journal of Ethics and Emerging Technologies*, 24(1), 45-61.
25. Hunter, C., & Bowen, B. E. (2024). We'll never have a model of an AI major-general: Artificial Intelligence, command decisions, and kitsch visions of war. *Journal of Strategic Studies*, 47(1), 116-146.
26. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852.
27. Jovanovic, M., & Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. *Computer*, 55(10), 107-112.
28. Kay Alan (2003). Objectorial programming language. Turing Lecture.
29. https://amturing.acm.org/vp/kay_3972189.cfm
30. Kissinger H., Schmidt E. and Huttenlocher D. (2023). ChatGPT Heralds an Intellectual Revolution.

31. Feb. 24, 2023.
32. Kjaergaard, M., Schwartz, M. E., Braumüller, J., Krantz, P., Wang, J. I. J., Gustavsson, S., & Oliver, W. D. (2020). Superconducting qubits: Current state of play. *Annual Review of Condensed Matter Physics,* 11(1), 369-395.
33. Kohlberg, L. (1981). The Philosophy of Moral Development: Moral Stages and the Idea of Justice, vol. 1 San Francisco: Harper & Row, pp 17-19.
34. Kourula, A., Moon, J., Salles-Djelic, M. L., & Wickert, C. (2019). New roles of government in the governance of business conduct: Implications for management and organisational research. *Organization Studies*, 40(8), 1101-1123.
35. Kreps, S., & Kriner, D. (2023). How AI threatens democracy. Journal of Democracy, 34(4), 122-131.
36. LeCun, Yann (2018). The Deep Learning Revolution: The Sequel
37. https://amturing.acm.org/vp/lecun_6017366.cfm
38. Lucchi, N. (2023). ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 1-23.
39. McCarthy, John (1971). Generality in artificial intelligence.
40. https://dl.acm.org/ft_gateway.cfm?id=1283926&type=pdf
41. McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5), 649-663.
42. Micali, Silvio (2012). Proof According to Silvio
https://amturing.acm.org/vp/micali_9954407.cfm
43. Minsky, Marvin (1969). Form and content in computer science.
44. https://dl.acm.org/ft_gateway.cfm?id=1283924&type=pdf
45. Newell, Allen and Simon, Herbert Alexander (1975). Computers as symbols and search intelligent machines.
46. https://dl.acm.org/ft_gateway.cfm?id=1283930&type=pdf
47. OECD AI Initiative (2019). The OECD AI Principles. May 2019.
48. Open Letter (2023). Pause Giant AI Experiments: An Open Letter. Mar 22, 2023.
49. Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge university press.
50. Ostrom (2009). Beyond Markets and states: Polycentric Governance of complex economic systems. Nobel Memorial Lecture, December 8, 2009.
51. Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., & Phillips, J. C. (2008). GPU computing. Proceedings of the IEEE, 96(5), 879-899.
52. Phelps E. S. (2006). Macroeconomics for a Modern economy Economics Nobel Prize Lecture, December 8, 2006.
53. Reddy, Dabbala Rajagopal (1993). To dream the possible dream
54. https://dl.acm.org/ft_gateway.cfm?id=1283952&type=pdf
55. Reed, B. C. (2014). The history and science of the Manhattan Project. Heidelberg: Springer.
56. Simon HA (1978). Rational decision making in business organisations. Nobel Memorial Lecture, 8 December, 1978.
57. Turing, A. M. (2021). Computing machinery and intelligence (1950).
58. US-EU Trade and Technology Council (2024). U.S-EU Joint Statement of the Trade and Technology Council. Apr 5, 2024.
59. Valiant, Leslie Gabriel (2010). The Extent and Limitations of Mechanistic Explanations of Nature. Turing Lecture.
60. https://amturing.acm.org/vp/valiant_2612174.cfm
61. Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems.*
62. Woldridge, (2023). The future of generative AI.
63. https://www.turing.ac.uk/events/turing-lectures-future-generative-ai

64. Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7-30.
65. WSJ (2022). Google Parts With Engineer Who Claimed Its AI System Is Sentient. July 22, 2022.
66. WSJ (2023a). AI's Power-Guzzling Habits Drive Search for Alternative Energy Sources. By Belle Lin Nov. 9, 2023.
67. WSJ (2023b). 3 Things I Learned About What's Next in AI. Joanna Stern, Oct. 20, 2023.
68. WSJ (2023c). Hollywood's Writers Emerge From Strike as Winners—for Now. Sept. 26, 2023.
69. WSJ (2023d). A New Way to Tell Deepfakes From Real Photos: Can It Work? Nov. 3, 2023.
70. WSJ (2023e). AI's Power-Guzzling Habits Drive Search for Alternative Energy Sources. Nov. 9, 2023.
71. WSJ (2024a). Elon Musk's xAI to Raise $6 Billion in Latest Fundraising Round. May 27, 2024.
72. WSJ (2024b). For AI Giants, Smaller Is Sometimes Better. July 6, 2024.
73. WSJ (2024c). Can $1 Billion Turn Startup Scale AI Into an AI Data Juggernaut? June 28, 2024.
74. WSJ (2024d). China Puts Power of State Behind AI—and Risks Strangling It. Updated July 16, 2024.
75. WSJ (2024e). AI Is Moving Faster Than Attempts to Regulate It. Here's How Companies Are Coping. March 27, 2024
76. WSJ (2024f). Morgan Stanley Moves Forward on Homegrown AI. July 26, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.