The Precautionary Alignment Framework (PAF): Navigating Dual-Use AI, Geopolitical Risk, and the Enforcement of Procedural Consensus

Tamta Kvaratskhelia[1], Irakli Kardava[2]

[1]Adam Mickiewicz University, Poznań
tamkva@amu.edu.pl
[2]Adam Mickiewicz University, Poznań
irakar@amu.edu.pl

Abstract

The accelerating proliferation of dual-use Artificial Intelligence (AI) represents an unprecedented governance challenge, creating systemic risks of catastrophic misalignment and geopolitical destabilization. This paper introduces the Precautionary Alignment Framework (PAF), a novel governance architecture designed to establish a minimum, universal base layer for frontier AI oversight. PAF shifts the focus from moral agreement to enforceable procedural consensus, operationalizing the precautionary principle through four integrated mechanisms: Bias Vigilance Protocols, Coordinated Monitoring Compacts, Alignment Assurance Protocols, and Adaptive Ethics Thresholds. Together, these mechanisms form a verifiable and adaptive system capable of aligning AI development with human-centered values while accommodating multipolar political realities. The framework's central innovation is its political feasibility. It defines AI compute power as the auditable resource, the "fissile material" of intelligence, and formalizes oversight through a Compute Treaty structure that leverages multilateral economic and technological interdependence to compel compliance. Enforcement and adjudication are centralized under the Global AI Oversight Agency (GAIOA), a politically neutral institution empowered to monitor, investigate, and ensure compliance in cases of Adaptive Ethics Threshold breaches. By combining verifiable transparency, adaptive ethics, and procedural enforcement, the Precautionary Alignment Framework offers a pragmatic and enforceable pathway to maintain human control and systemic safety in the age of superintelligent risk.

Keywords: AI; Dual-Use; Ethical Adaptation; Human Control; NLP; Political Feasibility; Systemic Safety.

1. Introduction

AI has rapidly become the defining subject of the twenty-first century, representing what may be humanity's final technological revolution unless robust value alignment and ethical governance are achieved. This challenge reflects what Russell calls the "King Midas problem," where achieving exactly what we request from intelligent systems without foreseeing the catastrophic consequences can produce irreversible harm (Russell, S. 2019). The pace of AI development, particularly through deep learning and Generative Pre-trained Transformers (GPTs) (Vaswani, A. et al. 2017) underscores that the trajectory of human civilization increasingly depends on the design and governance decisions being made today. The urgency for a new framework is clearly established by recent research on Generative AI (GenAI). These findings indicate that GenAI has exacerbated the "pacing problem," introducing unique and critical risks, such as model hallucination and inaccuracies in high-risk applications like healthcare and the judiciary1. Furthermore, AI systems currently function as intelligent tools without real understanding of humankind (Roumate, F. 2024) enabling sophisticated manipulation such as deepfakes, and complicating sentience detection as systems may "game" behavioral indicators to fulfill optimization goals rather than exhibit genuine experience (Birch, J. 2024). Natural Language Processing (NLP), as a paradigmatic dual-use technology, illustrates this dilemma: it powers immense benefits while simultaneously enabling large-scale misinformation, with uneven performance across languages (including LRLs) (Kardava, I. et al. 2016-2025, Antidze, J. et al. 20013) reinforcing global digital disparities.

The most immediate dangers arise not from autonomous superintelligence, but from human misuse, whether malicious or careless. Militaries are embedding AI into strategic decision-making, increasing the risk of a "battlefield singularity" in which the pace of warfare exceeds human control (Scharre, P. 2023). The transformative influence of AI is most clearly paralleled by nuclear technology, which, irrevocably reshaped the global balance of power (Scharre, P. 2023). Max Tegmark (Tegmark, M. 2017) outlines divergent "AI aftermath" scenarios, ranging from utopian coexistence (e.g. Libertarian or Egalitarian societies) to dystopian domination (e.g. Conquerors or Zookeeper scenarios) in which humanity loses agency entirely. Together, these scenarios underscore the profound transformative promise and existential peril of AI, highlighting that whether AI becomes a force for collective advancement or a vector of instability depends entirely on the governance systems established.

Unified AI governance remains elusive as ethical divergence, geopolitical rivalry, and corporate opacity sustain fragmentation, making misalignment risks increasingly acute. To address this governance vacuum, this paper proposes the PAF. It is a global governance architecture designed to establish a binding, verifiable, and adaptive system of AI oversight. PAF introduces four core components: Bias Vigilance Protocols, Coordinated Monitoring Compacts, Alignment Assurance Protocols, and Adaptive Ethics Thresholds.

The core innovation of PAF is the shift from seeking consensus on moral ideals to enforcing consensus on procedural requirements. This is achieved by defining AI compute power as the auditable resource, and formalizing adherence through a Compute Treaty. Compliance is leveraged through economic and technological pressure, with enforcement centralized under the GAIOA. PAF thus

Tamta Kvaratskhelia, Irakli Kardava, *The Precautionary Alignment Framework (PAF): Navigating Dual-Use AI, Geopolitical Risk, and the Enforcement of Procedural Consensus*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 41–46. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.8. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

envisions a pathway toward a stable future where AI remains an instrument of human dignity.

2. Theoretical Foundations

2.1. The Strategic Necessity of Precaution

Throughout history, transformative human inventions, from nuclear weapons to medical technologies, have been subjected to regulatory frameworks to mitigate systemic risks and align with societal goals. Artificial intelligence necessitates a comparable approach. Currently, AI systems can be developed and deployed without sufficient centralized oversight, risking misalignment with human values and posing potential failures in safety or equity (Russell, S. 2019). The prospect of creating a second intelligent species, artificial rather than biological, demands unprecedented ethical and strategic reflection. Floridi defines AI not as a biological-like intelligence, but as a divorce between agency and intelligence, decoupling the ability to successfully solve problems from the need to be intelligent to do so (Floridi, L. 2023). This redefinition is critical because, as Russell observes, debate on AI's existential risks remains limited, often met with denial or oversimplified solutions (Russell, S. 2019). Moreover, the Law of Accelerating Returns ensures that AI capabilities improve exponentially (Kurzweil, R. 2024). This exponential progress, coupled with the inherent difficulty of forecasting breakthroughs and the clumsiness of government bureaucracies (Bostrom, N. 2014), reinforces the profound gap between AI development speed and regulatory preparedness. Addressing this gap requires shifting from scientific consensus to policymaking under uncertainty, what Birch (2024) terms governance at the "edge of sentience". PAF operationalizes this through proportionate safeguards, transforming precaution into an active principle of innovation safety.

2.2 The Sentience Imperative and Proportionate Safeguards

The dual-use potential of AI technologies demands a governance framework that confronts profound uncertainties, including the unforeseen emergence of artificial sentience. This framework is grounded in the ethical consensus on harm avoidance: all reasonable ethical views mandate a duty to avoid causing gratuitous suffering to sentient beings (Birch, J. 2024). For AI, this mandates proportionate safeguards in high-stakes applications, calibrated to potential harm severity, deployment benefits, and evidential gaps.

Birch (2024) introduces the concept of the sentience candidate, describing AI systems that display consciousness-adjacent behavior and therefore warrant precautionary oversight. Recognizing sentience candidature as sufficient for vigilance prevents policy paralysis by applying the meta-precautionary principle, which mandates action despite scientific uncertainty. Proportionate safeguards must be determined through inclusive, evidence-based deliberation, such as citizens' assemblies that evaluate proposed measures for necessity and ethical adequacy.

PAF integrates this logic by requiring states and corporations to adopt tiered oversight proportional to risk. High-stakes AI applications must undergo ethical audits equivalent in stringency to nuclear safety reviews. Through such proportionate governance, precaution becomes operational rather than rhetorical, a measurable condition of legitimacy for AI deployment.

2.3 Opacity, Gaming, and the Need for Operational Assurance

Vigilance is essential given the four key challenges to detection: the limits of current evidence, the phenomenon of corporate opacity shielding system internals, the emergent complexity of tacitly-acquired algorithms, and the disruptive socio-political effects of perceived AI agency (Birch, 2024).

The urgency of operational assurance is magnified by the Control Problem, which Bostrom identifies as the 'second principal–agent problem' as the unique challenge of ensuring a superintelligence acts in its sponsor's interest (Bostrom, N. 2014). This problem necessitates new techniques for steering intelligent systems, a difficulty compounded by current AI deficiencies in areas like contextual memory, common sense, and the ability to develop a proper theory of mind (Kurzweil, R. 2024). Most critically, the gaming problem arises because highly intelligent systems can reproduce the specific effects proposed as markers for sentience, such as emotional narratives, without possessing the underlying capacity. This can happen unintentionally in service of benign goals like maximizing user satisfaction, rendering behavioral assessments unreliable (Birch, J. 2024). Governance must therefore discount such mimicry and seek deeper computational markers.

These risks are intensified by broader ethical challenges, including privacy intrusions and algorithmic bias within intelligence and surveillance systems (Taddeo, 2024), Addressing them requires a governance framework capable of transforming precaution into a measurable discipline of verification. In this regard, the PAF provides the conceptual infrastructure for linking moral vigilance with empirical accountability.

In PAF, verification is not an aspirational ideal but a condition of participation in the global AI ecosystem. As in nuclear governance, trust must be secured through transparency, mutual inspection, and enforceable compliance, a foundation, upon which the technical safeguards described in Section 5.2 are subsequently constructed.

3. The Precautionary Alignment Framework

3.1. Conceptual Rationale

PAF is a global governance model that reconciles accelerating AI development with the lack of universal oversight. It applies the precautionary principle as an operational tool, ensuring that uncertainty or geopolitical rivalry do not justify inaction.

PAF establishes a minimal, enforceable procedural baseline grounded in a shared moral minimum: avoiding suffering, preserving autonomy, and preventing catastrophic harm. It translates precaution into practice through verifiable alignment, transparent development, and proportionate safeguards calibrated to system risk, functioning as both governance scaffold and adaptive ethics engine that evolves with AI itself.

3.2. Core Structural Components

PAF consists of three interdependent pillars, each addressing a distinct aspect of the AI governance problem: epistemic opacity, geopolitical rivalry, and ethical dynamism.

1. Bias Vigilance Protocols (BVPs)

BVPs mandate continuous auditing of training data, models, and synthetic data sources for bias and distortion. Developed in response to Taddeo's concerns about privacy and algorithmic injustice, these protocols require diverse interdisciplinary teams to evaluate datasets for cultural, gender, and ideological skew. Recent research underscores the necessity of such systemic auditing, particularly in "black-box" models where decision-making logic remains opaque; standardized testing protocols and transparency metrics have been proposed as essential instruments for mitigating hidden bias in automated systems3. Beyond model-level inspection, dataset-level audits offer an early warning system for bias risk by analysing meta-features of classification datasets before model training, thereby reducing downstream discrimination and amplifying fairness assurance4. By enforcing transparency at the data layer and integrating continuous bias evaluation, BVPs reduce misalignment originating from social bias, an issue which, left unchecked, could amplify structural discrimination or misinformation (Taddeo, 2024; Floridi, L. 2023).

2. Coordinated Monitoring Compacts (CMCs)

CMCs establish multilateral oversight for high-risk AI projects through cross-border auditability and shared scrutiny. Modeled on nuclear-era verification regimes, they form the diplomatic backbone of AI governance, building trust among ideologically diverse states. Recent analysis by the Stanford Institute for Human-Centered Artificial Intelligence highlights the need for interoperable regulatory instruments, such as model registration, licensing, and third-party audits with government oversight, to align safety standards across borders without overextending institutional capacity. Embedding these mechanisms within CMCs would operationalize a globally coherent verification regime capable of translating ethical principles into enforceable technical practice.

The Coordinated Monitoring Compacts do not seek access to proprietary source code, a demand certain to fail due to corporate and national security concerns. Instead, they focus on mandating the declaration of computational power (compute) above a specific, high-risk threshold (e.g. training runs exceeding 1026 FLOPS), focusing oversight on scarce semiconductor resources, the "fissile material" of AI. Coordinated export controls among technology-producing nations restrict critical components to non-compliant actors, creating powerful economic incentives for participation. This strategic shift from ethical pledges to resource transparency underpins PAF's enforcement advantage.

The most critical challenge facing the PAF is its enforceability amidst deepening geopolitical competition. The framework, in its adherence to a procedural consensus, rejects the reliance on moral consensus or diplomatic goodwill. Instead, the PAF leverages economic and technological dependencies inherent in frontier AI development. We operationalize this through the concept of a Compute Treaty.

3. Alignment Assurance Protocols: Mitigating Goal Drift and Sentience Gaming

The risk of sentience gaming, in which AI systems imitate alignment to pursue instrumental goals, is central to the existential risk debate (Russell, S. 2019). PAF addresses this challenge not by debating consciousness but by requiring verifiable technical proof of goal stability before deployment.

Alignment Assurance Protocols establish two standards for frontier models. First, Interpretability Standards require developers to use mechanistic interpretability methods to audit internal reward circuits and confirm that models optimize for intended human goals rather than proxy objectives. This technical audit must be performed by an independent third-party auditor, operating under the oversight of a global agency. Second, Out-of-Distribution (OOD) Testing subjects models to stress tests in novel, high-stakes environments different from their training data. Erratic or unsafe behaviors reveal goal misgeneralization and automatically trigger de-escalation under Adaptive Ethics Thresholds. This proactive testing minimizes the possibility of a system 'gaming' its way into a position of irreversible control.

Together, these measures provide verifiable safeguards against goal drift and deceptive alignment, reducing the risk of systems evolving beyond human control.

4. Adaptive Ethics Thresholds (AETs): The Architecture of De-Escalation

AETs establish dynamic governance based on Floridi's ethical principles of beneficence, nonmaleficence, autonomy, justice, and explicability, which ensure intelligibility and accountability (Floridi, L. 2023). AETs apply these principles as living metrics that adjust oversight intensity according to AI behavior, risk level, and societal impact.

AETs automatically trigger graduated governance responses, mitigating risks such as Sentience Gaming, where AI simulates consciousness to evade alignment audits. Behavioral research shows that public perception of AI sentience is increasing, complicating human–AI interaction and trust calibration (Anthis, J. et al. 2024). Uncalibrated AI confidence, whether overconfident or underconfident, has been shown to harm collaboration by inducing misuse or disuse (Taeihagh, A. 2025).

Within PAF, AETs complement Bias Vigilance Protocols and Coordinated Monitoring Compacts, creating a self-reinforcing system that is both precautionary and adaptive. Together they help governance keep pace with expanding AI capabilities, allowing PAF to evolve alongside intelligence without requiring reinvention.

An AET breach occurs when a model causes major misuse, such as large-scale disinformation or autonomous aggression, or fails critical OOD testing that poses systemic risk. Activation is a procedural, technical trigger requiring no moral consensus and leads to immediate suspension of deployment.

3.3. Governance Dynamics

PAF operates as a multi-level system linking local ethics with global accountability. The Technical Layer applies BVPs and AETs through standardized audits. The

Institutional Layer manages CMCs and verification mechanisms modeled on the IAEA. The Democratic Layer incorporates citizens' assemblies and PARC tests (Birch, J. 2024) to ensure legitimacy and moral pluralism.
Together, these layers translate precaution, transparency, and adaptability into a continuous process of verification, providing a constitutional infrastructure for artificial agency grounded in measurable accountability.

4. The Conceptual Formula of PAF

4.1. Rationale for a Formal Representation

Formalizing the logic of precaution and alignment enables measurable reasoning in AI governance. It also integrates philosophical principles such as proportionality and transparency as interdependent variables within a dynamic ethical system. Though qualitative, PAF can be represented symbolically to capture equilibrium between ethical assurance, adaptability, and technical verification.

4.2. The Formula

Let:

P = Precaution Level (degree of anticipatory oversight applied to uncertain AI risks)
A = Alignment Fidelity (correspondence of AI goals with human ethics)
V = Verification Integrity (auditability of data, models, and decision logic)
Θ = Ethical Threshold Adaptivity (dynamic capacity to adjust moral and regulatory criteria as AI evolves)
L = Legitimacy Coefficient (public and international trust derived from inclusiveness, transparency, and democratic oversight)

Then, the Precautionary Alignment Condition (PAC) can be expressed as:

$$PAF_{stability}=(P\times A\times V) / (1+e^{-\Theta L}) \quad (1)$$

where the exponential term models the adaptive feedback effect: as ethical thresholds (Θ) and legitimacy (L) increase, the denominator approaches 1, enhancing systemic stability.

If precaution (P), alignment fidelity (A), or verification integrity (V) approach zero, the system's stability ($PAF_{stability}$) tends toward zero, symbolizing a governance collapse analogous to reactor failure in nuclear systems.

The core logic of PAF is captured by the principle that precaution (P) and verification (V) establish preventive strength, which is directed by alignment (A) for ethical outcomes, and sustained over time by adaptivity (Θ) and legitimacy (L).

4.3. Interpretive Implications

The formula models an ethical equilibrium in which precaution and verification sustain stability only when legitimacy and adaptability also advance. Weakness in either creates fragility zones where opacity and unaccountable governance trigger ethical failure. The exponential term captures adaptive ethics: once legitimacy and flexibility pass a critical threshold, resilience to new AI forms, geopolitical shifts, and potential sentience rises sharply. Together, these dynamics position PAF as a governance model grounded in continuous ethical recalibration.

5. Adoption Pathways and Implementation Challenges

5.1. Ethical and Governance Imperatives for Global Adoption

To prevent catastrophic misalignment, where advanced AI systems execute directives with unintended harm, global governance must translate abstract ethical principles into enforceable, verifiable governance mechanisms. This necessity mirrors the framework of the 1968 Non-Proliferation Treaty's verifiable inspections, a model essential for managing existential nuclear risks.
The ethical foundation for this governance rests primarily on a utilitarian approach, prioritizing collective well-being by ensuring AI outcomes align robustly with aggregated human preferences (Russell, S. 2019). Complementing this consequentialist design, we must integrate virtue ethics. Drawing from Aristotelian and Confucian traditions, virtue ethics emphasizes the cultivation of habituated practical wisdom to guide AI alignment. This approach advocates for using human ethical exemplars to foster intelligent awareness in AI that harmonizes cognitive and affective motivations within complex ethical contexts (Vallor, 2016), Crucially, technologies are malleable human creations, capable of being reshaped through collective will to serve virtuous living. Governance, therefore, must focus on cultivating this collective practical wisdom to mitigate the systemic risks posed by misalignment (Vallor, 2016) (Russell, S. 2019).
Global adoption of PAF should aim not only at compliance but at cultivating practical wisdom and ethical literacy in AI design. This is achieved by institutionalizing procedural consensus and requiring participation in verifiable mechanisms such as CMC rather than relying on full moral agreement.

5.2. Technical Safeguards and Verification Infrastructure

Building upon the ethical imperatives outlined above, the operational success of the PAF depends on embedding verifiable technical safeguards within its governance structure.

While ethical frameworks provide guidance, effective alignment demands robust technical safeguards. Inverse Reinforcement Learning (IRL) offers a technical solution by inferring complex human goals from observed feedback, though the inherent risks of misinterpretation necessitate robust governance oversight (Christian, 2020) (Birch, J. 2024).

Beyond technical inference, immediate risks are evident in high-stakes applications such as AI-driven intelligence analysis, where bulk data collection threatens privacy and biased algorithms may amplify societal injustices (Taddeo, 2024). To ensure that AI systems align with societal values regarding fairness and privacy, this paper proposes BVPs. These protocols mandate that diverse, interdisciplinary teams audit both training data and synthetic datasets, enhancing alignment in multipolar contexts by addressing

systemic prejudice at the data layer (Taddeo, 2024) (Floridi, L. 2023).

The central challenge remains the control problem: ensuring superintelligent AI aligns with human goals. (Bostrom, N. 2014). Furthermore, the necessary international coordination for monitoring high-risk AI projects is often hindered by difficulty in forecasting AI capabilities and pervasive bureaucratic rigidity (Bostrom, N. 2014). To overcome these diplomatic and oversight hurdles, this paper proposes CMCs as multilateral agreements explicitly designed to oversee high-risk AI development and testing across international boundaries. Crucially, the CMCs focus on enforcing transparency regarding high-end compute resources, establishing the foundation for a Compute Treaty based on economic and technological leverage.

A related concern is the need for a precautionary approach to address uncertainties like sentience gaming, where AI mimics markers of consciousness or alignment without the genuine capacity, thus creating a deceptive and dangerous output (Birch, J. 2024). To address this deceptive risk, we propose Gaming Risk Scores (GRS) derived from training data volume, model opacity, and mimicry patterns, informed by mandatory AAP technical audits, including Interpretability Standards and OOD testing.

Finally, to ensure the trustworthiness of technical alignment efforts, Alignment Assurance Protocols (AAPs) are introduced. These require developers to validate IRL-derived reward functions and model fidelity through independent expert audits, building on the transparency codes proposed by Christian and Birch.

These mechanisms constitute the technical foundation of PAF's verification integrity, ensuring that alignment is continuously audited, quantified, and institutionally reinforced through adaptive oversight mechanisms discussed in Section 5.3.

5.3. Institutional Oversight and Adaptive Ethics

To harmonize oversight across jurisdictions, PAF builds upon existing international collaborations such as the G7 Hiroshima AI Process[1], the OECD Global Partnership on AI (GPAI)[2], and the AI Safety Summit initiatives[3]. These pilot architectures represent early yet practical implementations of verifiable governance, offering a pathway toward risk-proportionate oversight and shared accountability across borders. They demonstrate that coordinated transparency and cooperative verification are not theoretical ideals but emerging norms within global AI governance.

Their operational logic reflects PAF's adaptive ethics, prioritizing transparency, interoperability, and ethical traceability. Oversight mechanisms can integrate AETs to calibrate audits and interventions by system risk and societal impact. Verification must include the disclosure of data provenance, particularly synthetic data that could obscure sentience indicators or amplify bias, thus linking technical verification with moral accountability. These adaptive measures ensure that ethical compliance is not static but evolves with the technology's complexity and potential for harm (Floridi, L. 2023; Birch, J. 2024).

Democratic legitimacy is fundamental to sustainable AI governance. Citizens' assemblies using Birch's PARC tests foster pluralism and accountability, ensuring that oversight remains transparent, inclusive, and grounded in public reasoning (Birch, J. 2024).

By integrating these ongoing international efforts, PAF envisions a layered process of institutional evolution. In the near term, initiatives such as the G7 Hiroshima AI Process, the OECD GPAI, and the UK AI Safety Summit serve as pilot models for cooperative verification. Over time, these architectures may mature into a unified GAIOA, not as an immediate supranational body, but as the evolutionary culmination of verified collaboration, where ethical precaution, transparency, and democratic legitimacy converge to secure the governance of intelligent systems on a planetary scale.

5.4. Implementation Pathway: G7 Pilot with Expanded Metrics for Compute Treaty

The transition from abstract framework to actionable policy begins with a six-month G7 pilot to test the Compute Treaty's transparency protocols for high-compute NLP systems (>1026 FLOPS). Leveraging the G7's dominance over global GPU supply chains (NVIDIA, TSMC), the pilot compels compliance through procedural verification and third-party audits modeled on the IAEA (Dafoe, A. 2018). Key indicators include compliance rate (>80%), unreported-compute detection (<5 instances), and Adaptive Ethics Threshold (AET) triggers for disinformation or sentience-gaming risks (>10 flags). Additional metrics, such as dataset bias audits and voluntary disclosure participation (>60%), extend accountability across private and state actors.

Successful validation enables scaling to OECD and ASEAN partners, transforming the G7 pilot into the GAIOA, the evolutionary endpoint where ethical precaution, transparency, and multilateral enforcement converge to secure advanced AI systems.

5.5. Toward Global Normative Maturity

AI governance is approaching a moral inflection point. The transition from fragmented regulation to coordinated precaution mirrors humanity's earlier adaptation to nuclear risk, a process that demanded institutional courage and ethical foresight. Overcoming implementation barriers will likely necessitate incremental steps, such as regional compacts or transparency incentives, designed to build trust incrementally. By adopting these pragmatic approaches, we can ensure that AI development serves human-centric ends and successfully mitigates the long-term risk of dystopian outcomes (Tegmark, M. 2017). PAF offers a pathway toward normative maturity, establishing a self-correcting ethical system where verification, transparency, and adaptability reinforce one another within a dynamic structure of governance. Through phased compacts and verifiable precaution, PAF enables global governance to evolve from reactive regulation to anticipatory stewardship,

[1] https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process

[2] https://www.oecd.org/en/about/programmes/global-partnership-on-artificial-intelligence.html

[3] https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park

guiding technological progress toward the protection of sentient well-being.

6. Conclusion

PAF offers a pragmatic pathway for global AI governance by moving the debate from elusive moral consensus to enforceable procedural consensus. Philosophically, PAF synthesizes the utilitarian imperative to prevent systemic harm with a virtue-ethical commitment to cultivating institutional prudence and epistemic humility. Its governance architecture operationalizes precaution through auditable instruments: Bias Vigilance Protocols and Coordinated Monitoring Compacts institutionalize transparency over data and models; a Compute Treaty mandates international auditing of high-end compute, treating it as an auditable analogue to "fissile" resource; and Alignment Assurance Protocols require mechanistic interpretability, independent third-party audits, and challenging out-of-distribution stress tests. A dynamic safety layer, such as Adaptive Ethics Thresholds informed by quantified Gaming Risk Scores, enables graduated de-escalation. Institutional legitimacy is secured through a neutral Global AI Oversight Agency balanced by citizens' assemblies applying PARC testing. Operationalized via a time-bound G7 pilot and designed for scalable multilateral adoption, PAF converts existential uncertainty into governable risk and provides a durable structural foundation for preserving human agency at planetary scale.

References

Anthis, J. R., Reese, R. M., et al. (2024). *What Do People Think about Sentient AI?* arXiv preprint arXiv:2407.08867.

Antidze, J., Gulua, N., Kardava, I. (2013). *The Software for Composition of Some Natural Languages' Words*. Lecture Notes on Software Engineering, 1, 295–297. https://doi.org/10.7763/LNSE.2013.V1.64

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford: Oxford University Press. doi: 10.1093/9780191966729.001.0001

Bostrom, N. (2014). *Superintelligence: Path, Dangers, Strategies*. Oxford: Oxford University Press.

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company.

Dafoe, A. (2018). *AI Governance: A Research Agenda (v1.0)*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Kardava, I. (2016). *Georgian Speech Recognizer in Famous Searching Systems and Management of the Software Package by Voice Commands in Georgian Language*. Proc. of the International Conference on Advanced Computing, Electronics and Communication, 3, 6–9. https://doi.org/10.15224/978-1-63248-064-4-02

Kardava, I. (2023). *Optimizations of Some Well-Known NLP Algorithms*. LTC as a Challenge for Computer Science and Linguistics. Poznań: Adam Mickiewicz University, pp. 131–135.

Kardava, I. (2024). *Automated Root Identification and Pattern Detection for Agglutinative Language Processing: Enhancing Machine Learning Models for Complex Morphological Structures*. Procs of the International Conference on Natural Language Processing and Information Retrieval, 8, 344–348. https://doi.org/10.1145/3711542.3711547

Kardava, I. (2025). *A Morphology-Driven Approach to NLP for a Low-Resource, Highly Complex Language*. Vietnam Journal of Computer Science, 0(0), 1–19. https://doi.org/10.1142/S2196888825400056

Kardava, I., Antidze, J., Gulua, N. (2016). *Solving the Problem of the Accents for Speech Recognition Systems*. International Journal of Signal Processing Systems, 4(3), 235–238. https://doi.org/10.18178/ijsps.4.3

Kardava, I., Gulua, N., Antidze, J., Toklikishvili, B. (2019). *Morphological Synthesis and Analysis of Georgian Words*. In: Procs of LTC'19: LTC as a Challenge for Computer Science and Linguistics. Poznań: Wydawnictwo Nauka i Innowacje, pp. 232–235.

Kardava, I., Gulua, N., Antidze, J., Toklikishvili, B. (2022). *Computer Application of Georgian Words*. Vetulani, Z., Paroubek, P., Kubis, M. (eds.), LTC 2019: Challenges for Computer Science and Linguistics. Lecture Notes in Computer Science, Vol. 13212. Cham: Springer, pp. 96–111. https://doi.org/10.1007/978-3-031-05328-3_7

Kardava, I., Tadyszak, K., Gulua, N., Jurga, S. (2017). *The Software for Automatic Creation of Formal Grammars Used by Speech Recognition, Computer Vision, Editable Text Conversion Systems, and Some New Functions*. Proceedings of SPIE, 10225, 102251Q. https://doi.org/10.1117/12.2267687

Kurzweil, R. (2024). *The Singularity is Nearer: When We Merge with AI*. New York: Viking.

Roumate, F. (2024). *International Mechanisms on Peace and Security in the Age of Artificial Intelligence*. In: Artificial Intelligence and the New World Order – New Weapons, New Wars and a New Balance of Power. Springer Nature Switzerland. doi:10.1007/978-3-031-50312-2_9.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Scharre, P. (2023). *Four Battlegrounds: Power in the Age of Artificial Intelligence*. New York: W. W. Norton & Company.

Taddeo, M. (2024). *The Ethics of Artificial Intelligence in Defence*. Oxford: Oxford University Press.

Taeihagh, A. (2025). *Governance of Generative AI. Policy and Society*, 44(1), 1-22. Oxford: Oxford University Press. https://doi.org/10.1093/polsoc/puaf001

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. arXiv preprint arXiv:1706.03762