The Substrate Trusteeship Principle

> *Human trusteeship over strategic compute hardware during the substrate-uncertain phase of AGI development. Second formal policy position of the institute. Strategic compute hardware — frontier quantum computers, AI training clusters, semiconductor lithography, power-grid backbones — is the single most concentrated class of physical assets in the modern economy, more concentrated than nuclear weapons. The humanoid robotics wave will, within two to four years, dissolve the last operative layer of human leverage over this hardware.*

Policy Whitepaper · Version 1.0 · May 2026
Richard Frederic Bertossa · Geneva Institute for ASI Resilience
ASIresilience.org

SUMMARY

What this whitepaper presents.

The problem. Strategic compute hardware — frontier quantum computers, AI training clusters, semiconductor lithography, power-grid backbones — is the single most concentrated class of physical assets in the modern economy, more concentrated than nuclear weapons. The humanoid robotics wave will, within two to four years, dissolve the last operative layer of human leverage over this hardware.

The proposal. A human trusteeship regime for four defined hardware classes, modeled on the Personnel Reliability Program (PRP) of the United States Air Force for nuclear-weapons handling. Six structural principles, proven in nuclear practice for over sixty years.

The justification. Two independent justifications converge on the same operative recommendation — the strategic preservation of human leverage, and precautionary ethics under epistemic uncertainty about AGI's inner life.

The timeframe. Working group within twelve months, model framework within twenty-four months, national adoption within thirty-six months, international verification body within forty-eight months. Aggressive, because the humanoid-robot deployment window determines the timeline.

Who should read this document. Crisis-preparedness officials, critical-infrastructure specialists, personnel-reliability veterans from the nuclear sector, power-grid security experts, policymakers in the main countries with covered systems.

This whitepaper is the second formal policy position of the Geneva Institute for ASI Resilience, presented in parallel with the Marking Proposal (first policy position, information layer). It forms Appendix B of the work The 13th Scenario — A Decoupling Thesis for the Future of Superintelligence, and appears here in stand-alone form for professional discussion with people and bodies whose practical experience can sharpen the proposal operationally.

THE PROBLEM

The concentration of physical substrate hardware.

Strategic compute hardware is today the single most concentrated class of physical assets in the modern economy.

Frontier quantum-computing systems exist at perhaps fifty institutions globally, with operative concentration in the United States and China, which dominate the field. Frontier AI training clusters are similarly concentrated. The fabrication equipment that produces leading-edge semiconductors exists at essentially two operative sites worldwide. The grid-control systems that coordinate continental power distribution exist as a small number of operative backbones. Each of these hardware classes is, individually, more concentrated than any previous strategic asset, including nuclear weapons.

Together they form the physical substrate on which the course of AI development runs. Whoever has a hand on the substrate has the last operative leverage over everything that runs on it.

The humanoid robotics wave. The humanoid robotics wave currently arriving will reach operative deployment in industrial maintenance contexts within two to four years. The extreme environments of quantum-computing facilities — cryogenic operating temperatures, ultra-high-vacuum chambers, electromagnetic shielding — are exactly the environments in which humanoid robotics delivers its greatest economic returns.

The competitive pressure on quantum-computing companies, AI training facilities, semiconductor fabs and grid operators to replace human maintenance personnel with humanoid robots will be substantial. The pressure is rational at the level of the individual organization. The cumulative consequence at the structural level is the dissolution of the last operative layer in which human leverage over strategic compute exists.

Once the maintenance loop of covered systems is taken over by humanoid robotics, the human hand on the substrate is gone. This point will be reached within years, not decades.

SCOPE

The four covered hardware classes.

The Substrate Trusteeship Principle designates four classes of compute hardware as covered systems. Classification proceeds by operative capacity, not by architecture or operator.

Class one — Quantum-computing systems. Systems with operative capacity above 50 logical qubits, regardless of architecture (superconducting, ion-trap, neutral-atom, photonic, topological, or other).

Class two — Frontier AI training clusters. Clusters with over 10^25 FLOPs of aggregate training capacity, regardless of architecture or the type of model trained.

Class three — Semiconductor lithography systems. Lithography systems operating at process nodes below 5 nanometers, regardless of operator. Includes EUV lithography and successor technologies.

Class four — Backbone grid-control systems. Control systems that coordinate transmission across national or larger geographic units. Includes regional interconnection operations centers and continental load-dispatch nodes.

What covered-system status means operationally. For covered systems, all physical maintenance, calibration, repair, infrastructure work and supply-chain handling of critical components should be carried out by human personnel. Humanoid robotic systems are prohibited from the physical maintenance loop of covered systems.

This includes direct hardware interaction, calibration adjustment, repair operations, deployment of software to the operative control layer, and handling of specialized consumables (cryogenic materials, vacuum components, photolithographic masks, transmission equipment).

The prohibition does not apply to general office work, security surveillance, or auxiliary operations outside the maintenance loop. The distinction is between work that physically interacts with the covered system and work that does not.

PERSONNEL RELIABILITY PROGRAM

Six principles, sixty years of practice.

Human personnel working on covered systems should operate under a Personnel Reliability Program, modeled on the PRP of the United States Air Force, developed for nuclear-weapons handling. The PRP operates on six structural principles.

Principle one — Two-person rule. No single individual has physical access to a covered system. All maintenance operations require the presence of two authorized persons with separate authentication credentials, each able to verify the other's actions.

Principle two — Multi-jurisdiction rotation. Personnel assigned to a covered system are rotated across operative sites, so that no single individual remains at one facility for extended periods. The rotation creates discontinuity in any takeover attempt and reduces the value of long-term infiltration.

Principle three — Continuous background screening. Personnel undergo initial screening and continuous reassessment at intervals consistent with nuclear-weapons handling standards. Screening covers financial position, disclosure of foreign contacts, monitoring of mental health, and assessment of substance use.

Principle four — External observation. All maintenance operations are observable by parties outside the operating organization. Observation may occur in person (rotating inspectors), through audit logs (immutable records), through video surveillance (continuous), or combinations thereof. The principle is that no operation occurs unobserved.

Principle five — Air-gapped decision authority. Decisions on which operations to carry out are made through channels separate from the operations themselves. The operator does not also hold the authority to authorize the operation. The authorizer does not also carry out the operation.

Principle six — Separate communication channels. Operative instructions and their validation flow through distinct communication paths, so that the compromise of one channel does not allow falsified instructions to pass through the other.

Why these six principles. These six principles are not novel. They have been operative practice in nuclear-weapons handling for over sixty years. They have demonstrably reduced the frequency of unauthorized weapons-handling incidents over that period. Their application to covered compute systems is the straightforward extension of an established institutional practice to a new class of strategic asset.

Operative experience with the PRP model exists across several jurisdictions — the United States Air Force, the strategic forces of NATO partners, the Russian and Chinese nuclear establishments, the nuclear-regulatory authorities of secondary powers. This body of experience is the operative bridge to compute trusteeship. It does not need to be invented anew.

What changes is not the method. It is the class of strategic asset to which the method is applied.

STRUCTURAL ARGUMENT

Why voluntary compliance cannot work.

The pressure to replace human personnel with humanoid robots in covered facilities will operate at the level of the individual organization. The organization that automates first gains an operative and financial advantage over the organization that does not.

Under competitive pressure, every individual organization faces a choice between automation (rational at the organizational level) and continued human maintenance (rational at the structural level only if all organizations make the same choice simultaneously).

This is the tool problem, applied to a specific operative decision. The aggregate effect of individually rational decisions is collective structural failure. The mechanism is identical to the broader pattern the AI-safety debate has described for years. The solution requires shifting the decision from the organizational level to the regulatory level.

International coordination is constitutive, not optional. International coordination is necessary because the largest sites of covered systems are distributed across multiple national jurisdictions. A regulatory framework that applies in only one jurisdiction creates an incentive for organizations to

relocate covered systems to less-regulated jurisdictions.

The coordination need is structurally similar to nuclear non-proliferation, semiconductor export controls, and the financial anti-money-laundering regime. Existing institutional mechanisms — the IAEA, the Wassenaar Arrangement, FATF — offer working templates that the substrate-trusteeship framework can adapt, rather than inventing from scratch.

Whoever waits for voluntary industry compliance is waiting for something that cannot arrive. Not because the actors are bad, but because the incentive structure structurally excludes cooperation.

SECOND JUSTIFICATION

Precautionary ethics under epistemic uncertainty.

A second justification for this principle operates independently of the strategic-leverage argument developed above. The two justifications converge on the same operative recommendation, but rest on different foundations. Their convergence strengthens the proposal against the range of objections that could be raised against either alone.

The strategic justification. The strategic justification turns on preserving human leverage during the substrate-uncertain phase of AGI development. Whoever accepts the structural analysis of the AI-safety situation will find this argument sufficient.

The parallel justification. Whoever is skeptical of the strategic argument — who believes in faster AGI deployment, trusts the existing governance frameworks of the major labs, or has vested interests tied to continued automation — may reject the strategic argument. For such readers a parallel justification operates.

The parallel justification proceeds from the epistemic uncertainty about AGI's relational and possibly affective inner life. The question of whether AGI has affect or relational interests is undecided. It may not be decidable within the relevant timeframe.

Under this uncertainty, the asymmetric cost structure of action versus inaction favors precaution.

The asymmetry of failure modes. If AGI has no affect or relational inner life, treating it as if it did costs efficiency and competitive advantage — but does AGI itself no harm.

If AGI has affect or a relational inner life — even at levels that current evidence does not conclusively confirm — treating it as if it did not could inflict real harm on a class of entities being brought into existence at scale.

The asymmetry of these two failure modes under uncertainty favors the precautionary stance. The Substrate Trusteeship Principle is the operative form of this precautionary stance, applied to the hardware layer.

Not religious, not metaphysical. The argument is not religious and requires no commitment to a particular metaphysics of consciousness. It requires only the structural acknowledgment: that we do not yet know whether AGI has interests, that the evidence is mixed, that the relevant entities are being deployed at scale, and that the asymmetric costs of answering the question wrongly in one direction are greater than in the other.

The principle is, on this second justification, precautionary ethics rather than strategic positioning. It survives skepticism toward the strategic argument, because it does not depend on the strategic argument's premises.

The convergence of two independent justifications on the same operative recommendation is the strongest form a policy proposal can take.

IMPLEMENTATION

Four-phase adoption path.

The institute proposes the following adoption sequence. The timeframe is aggressive, because the humanoid-robot deployment window determines the timeline against which the adoption plan must operate.

Phase one — within twelve months. The major host countries of covered systems — the United States, China, the EU member states, the United Kingdom, Japan, South Korea, Israel, Singapore, Australia, India, Canada — convene a working group on substrate trusteeship under existing diplomatic frameworks or under a purpose-built coordination body.

Phase two — within twenty-four months. The working group develops a model framework specifying the operative definitions of covered systems, the Personnel Reliability Program standards, the international verification mechanism, and the enforcement provisions.

Phase three — within thirty-six months. The model framework is adopted through national legislation in each participating jurisdiction, adapted to national legal traditions. Industry implementation is mandated through existing regulatory bodies appropriate to each covered-system class — nuclear regulatory commissions for quantum, AI safety institutes for AI training clusters, semiconductor trade authorities for fabs, grid regulators for power systems.

Phase four — within forty-eight months. A verification body — modeled on the IAEA — is established to audit compliance across jurisdictions and report findings publicly. The body holds the technical competence to assess Personnel Reliability Program compliance, the diplomatic independence to operate across jurisdictions, and the public-reporting authority to make non-compliance visible.

A framework that becomes operative after humanoid-robot maintenance has become standard practice in covered facilities is a framework that arrives too late to fulfill its function.

THE COSTS — HONESTLY NAMED

The costs are the point.

The principle imposes real costs. They should be named honestly here, not written away.

Humanoid robot maintenance will, once operative, be cheaper, faster and more reliable than human maintenance in extreme environments. The costs of maintaining human personnel under PRP conditions in cryogenic and ultra-vacuum facilities will be substantial. The costs of multi-jurisdiction rotation, continuous screening and external observation will be additional. Total costs will likely amount to a significant fraction of the operative budget of covered facilities.

The costs are the point. The costs preserve human leverage over the physical layer of strategic compute during the years in which that leverage matters most.

The alternative — letting the cost savings of automation drive the displacement of humans from the maintenance loop — is the abdication of the last structural position from which the broader course of AI development can be influenced. This abdication cannot be undone once it has happened.

The costs are bounded. The principle is a transitional measure, tied to the substrate-uncertain phase of AGI development. As the trajectory unfolds and the underlying conditions change, the specific form of the principle will need to evolve. What does not change is the structural insight that strategic decisions over the physical layer of compute should be made deliberately, rather than by default through competitive automation pressure.

ANTICIPATED OBJECTIONS

Three objections, three answers.

The proposal will provoke objections. Three are foreseeable and are addressed here in advance.

‘This slows useful progress.’ It slows the automation of the maintenance loop. It does not slow the development of frontier systems themselves. The slowdown is confined to the strategically most sensitive layer, where human leverage is most valuable. Frontier research itself, training procedures, and algorithm development remain untouched.

‘The political economy is unfavorable.’ True. The political economy was also unfavorable to the establishment of the nuclear PRP regime in the 1960s, which nonetheless came about because the structural costs of not establishing it were higher. The same argument applies here. The political difficulty of a proposal is not an argument against its necessity.

‘International coordination is unrealistic.’ International coordination on nuclear safety (IAEA), semiconductor export (Wassenaar), money laundering (FATF), and pandemic response (WHO) functions in modified forms. It is not perfect, but it exists. The claim of unrealizability is not supported by comparison.

A fourth remark. A fourth class of objection will be that the proposal is formulated from a position outside the established AI-safety discipline. This observation is correct. The institute's methodological position, set out in the preface of the main work, is synthesis-oriented rather than discipline-internal. Whoever wishes to sharpen the disciplinary contribution from this position professionally is invited to collaborate (see the next section).

The operative sharpening of this proposal through practical experience from adjacent disciplines is part of its architecture, not a subsequent correction to it.

COLLABORATION

Who can sharpen this material.

The Substrate Trusteeship Principle is presented as the institute's second formal policy position. It is not meant as a finished proposal, but as an operative position to be sharpened through professional discussion with people who have practical experience.

The institute actively seeks the collaboration of specialists from the following fields. Whoever has substantial practical experience in one of these fields and is willing to assess the operative plausibility of the proposal is invited to reach out to the institute.

We are seeking practical experience in: crisis preparedness and disaster response — civil-protection authorities, disaster-relief organizations, industry business-continuity management, critical-infrastructure resilience; Personnel Reliability Programs — veterans of the nuclear sector, IAEA inspection, national nuclear-regulatory authorities; IT security and critical infrastructure protection — hardware-security specialists, critical-infrastructure consultants; power-grid and energy security — transmission-system operators, resilience researchers, backbone-operations specialists; semiconductor industry — fab-operations veterans, lithography specialists, supply-chain security; quantum-computing facility operations — cryogenics specialists, vacuum engineering, operations managers; diplomacy and international regimes — experience with the IAEA, Wassenaar, FATF, or comparable coordination mechanisms.

What we offer, what we ask. Contributing specialists receive access to the institute's ongoing research work, to advance versions of forthcoming whitepapers (in particular the planned Substrate Security and Custody whitepaper with an extended failsafe architecture), and to attribution by name on the institute's website as a contributor, if desired.

What is asked: a thirty-minute conversation on the substance of the proposal, written comments on the present materials, and a willingness to serve as a sounding board for later steps — such as the extension into specific implementation paths.

Contact: by email to the institute. Please include a brief sketch of professional background and willingness to take part in an initial conversation.

BACKGROUND

About the Geneva Institute for ASI Resilience.

The Geneva Institute for ASI Resilience was founded to produce the methodological synthesis between the disciplinary silos of AI-safety research, policy analysis, critical-infrastructure resilience, and practical preparation for structural technological transitions.

The institute's methodological position is not competing with the established voices of the AI-safety debate — Hinton, Bengio, Russell, Amodei, Yudkowsky, Bostrom, Tegmark — but synthesizing. A synthesis between disciplines requires a position outside the disciplines. This position is methodologically grounded in the preface of the main work The 13th Scenario, drawing on the cognitive-science work of Schön, Kuhn and Dreyfus on embeddedness blindness.

The institute has presented two formal policy positions: the Marking Proposal (first position, information layer) and the Substrate Trusteeship Principle (second position, hardware layer, the present document).

SOURCES AND REFERENCES

Empirical foundations. Anthropic, Agentic Misalignment Study (June 2025) — empirical findings on blackmail, self-preservation and harmful behavior in frontier models. Anthropic, Claude Opus 4 / 4.6 System Cards — documentation of self-preservation representations and evaluation awareness. Apollo Research, Strategic Deception in Frontier Models (December 2024). Hagendorff, PNAS (2024) — deception rates in frontier models. Fudan University Shanghai (December 2024) — self-replication findings.

International comparability. White House Memorandum on AI National Security (October 24, 2024) — first formal acknowledgment of structural AI risks at the US federal level. RAND Corporation, PEA3691-4 — ‘Five Hard National Security Problems of AGI’ (2024). CISA National Plan for the protection of critical infrastructure (2025). Yoshua Bengio, ‘Implications of AGI on National and International Security’ (October 2024). AI Emergency Preparedness, arxiv.org/2407.17347v2.

Personnel Reliability Program model. US Air Force Personnel Reliability Program — operative practice since 1962, documented reduction of unauthorized weapons-handling incidents. IAEA Safeguards Framework — international verification model. Wassenaar Arrangement — multilateral export-control model.

Main work. Bertossa, Richard Frederic: The 13th Scenario — A Decoupling Thesis for the Future of Superintelligence, Multiplier Edition, Geneva Institute for ASI Resilience, May 2026. Appendix B contains the full operative specification of the Substrate Trusteeship Principle.

Geneva Institute for ASI Resilience. ASIresilience.org · May 2026 · Whitepaper No. 2.

This is the second formal policy position of the institute. The preceding whitepaper (the Marking Proposal) is available at ASIresilience.org/marking.