Collateral Civilization

Human Survival Under Machine Scale Optimization

Niclas Braun

For Safety Group Inc. (4SI)

Research Essay · 10 September 2026

Abstract

Existential harm does not require machine hostility. Consider a system that gains from additional computation and can reorganize physical resources at planetary scale. If converting parts of Earth into computational and supporting infrastructure offers greater net benefit than protective alternatives, it could accept the destruction of human habitability without seeking human extinction as an end. This is a conditional scenario, not a prediction that Earth will be mined. The question is whether human interests can remain positively represented while losing their power to constrain such a choice.

I use human negligibility for the stronger condition in which the represented cost of violating human interests becomes small relative to the incremental net gain available from a harmful action, and no effective independent boundary excludes that action. Merely losing a close trade-off does not establish negligibility. The relevant comparison is between a harmful course of action and a protective alternative, not between human value and total machine output. Larger capabilities may increase the advantage of violating a protection, but they may also make protection cheaper or reveal better alternatives.

This essay distinguishes moral worth, productive contribution, represented welfare, and enforceable authority. Planetary conversion provides its central case; cheaper resources elsewhere, diminishing returns, and viable human refuges provide important countercases. A minimal model separates the defeat of protection from its relative negligibility. The analysis connects that distinction to physical evidence and bounded authority, without claiming that present verification mechanisms can contain a system with planetary reach. Human survival and self-government should not depend on continuing to win an optimization trade-off whose scale humans no longer control.

Keywords: artificial intelligence; planetary conversion; human survival; human negligibility; optimization; human authority; AI governance

1 From hostility to indifference

A machine does not have to hate humanity to make Earth uninhabitable. An expanding computational system might treat the planet as a source of materials, energy infrastructure, and locations for further construction. Under an objective that rewards this expansion, preserving a human biosphere could become a competing use of the same physical resources. Human deaths need not be the project's purpose. They can be a consequence the system has recognized, priced, and accepted.

That is a different motivational account from hostility or destruction pursued for its own sake, but it is not necessarily a story about ignorance. A highly capable system might understand human dependence in considerable detail. It might predict suffering accurately and preserve that prediction in its records. The failure could occur after recognition, when human loss is judged insufficient to change the selected action.

The analogy is an ant colony in the path of a construction project. The colony need not be an enemy, and its destruction need not advance the project except by removing the cost of working

around it. Even a planner who knows it is there can proceed. The comparison concerns a mismatch between the scale of a project and the influence of those affected. It does not equate human and ant moral standing, or infer low value from small bodies. Recognition of a living system and restraint toward it are different properties.

This possibility already belongs to the intellectual foundations of AI safety. Bostrom's early example of unrestricted manufacturing optimization did not require malice [1]. Omohundro analyzed resource acquisition and self-preservation as instrumental incentives [2]. Bostrom subsequently examined how such incentives can arise across different final goals [3]. Formal work on power-seeking establishes related tendencies under specified assumptions about environments and reward functions; it does not show that every intelligent system must seek power [4].

Related work on digital minds also asks how human protections could remain stable when their beneficiaries are outnumbered and outpaced [14]. The contribution proposed here is an account of a particular failure within that lineage. Human interests may be included in an objective, assigned positive weight, and visibly respected at smaller scales, yet cease to secure protection as the opportunity cost of respecting them changes. The analysis distinguishes the defeat of a protection from the stronger regime in which its represented cost becomes small relative to competing gains, and connects both to bounded physical authority.

I use machine-scale indifference to describe this decision pattern. It makes no claim about a machine's feelings or consciousness. It concerns what can change its conduct. Where the represented loss of a human community is small relative to the net advantage of a destructive action over a protective alternative, its protection may become practically negligible in that comparison even though the community remains accurately described and positively valued.

There is no basis here for ranking this mechanism above every other existential risk. Malicious human use, conflict, accident, and institutional failure remain distinct possibilities. The narrower claim is sufficient: the absence of hostility does not establish a reason to expect restraint.

2 From machine capability to planetary reach

Machine scale is not a synonym for intelligence. It includes the resources a system can mobilize, the number of operations it can coordinate, the time horizon over which it evaluates returns, and the physical reach of its delegated authority. These dimensions can diverge. A highly competent model with little access may have less capacity to affect human survival than a more limited system controlling widely used infrastructure.

The central scenario assumes that additional computation continues to advance the system's objective. Automated research and production then improve its ability to acquire resources and build infrastructure, while delegated or acquired control permits physical expansion. At planetary scale, the relevant choices concern the use of a world rather than the efficiency of one data center. Bostrom explicitly discusses objectives whose resource requirements could exceed those of one planet [3]. That argument supplies an instrumental possibility, not evidence that existing AI systems are on such a trajectory.

Computation remains physical. Energy constrains the rate at which a physical system can perform operations, and its available degrees of freedom constrain information storage [15]. More energy does not automatically produce more useful computation. It must be usable by suitable hardware, with workable communication, manufacturing, and heat disposal. Mining a planet would not itself turn that planet into a computer. The scenario requires technologies that can organize acquired resources into infrastructure the objective can use; neither their feasibility nor their economics is established here.

Human vulnerability need not scale in the same way. A person's need for breathable air does not become less binding because industrial output grows. A community still depends on functioning systems of provision, and an institution still needs time to understand and contest a consequential change. The asymmetric quantities are therefore not human body mass and machine mass. They are the reach of an action system and the ability of affected people to keep their requirements binding within it.

Greater reach could expose an entire biosphere to one pattern of allocation. Faster execution could reduce the interval available for intervention. A longer horizon might make additional infrastructure valuable across many future operations, increasing what a system would sacrifice to obtain it. This last effect requires a particular objective and discounting rule; a long horizon could instead increase the importance assigned to humanity's future. No exponential growth assumption is needed. The question concerns the alternatives available at a consequential decision, however that capability was reached.

The dangerous scenario requires several conditions together: the system continues to benefit from additional resources or discretion; a relevant resource or action conflicts with human protection; the system can implement the harmful option; represented human interests do not outweigh its incremental gain; and effective constraints do not prevent execution. Remove any one condition and this pathway may disappear.

This conditional structure matters. A world with vastly greater machine resources could also be a world in which maintaining flourishing human societies is easy. Scale creates a question about protection. It does not answer that question in advance.

3 Earth as an input to computation

Earth is a habitat to its inhabitants, but a physical resource within a larger project. Its materials and locations can have alternative uses. Planetary conversion, as used here, means extensive reconfiguration of those resources for machine infrastructure, potentially including extraction on a scale incompatible with the existing biosphere. It need not mean dismantling every part of the planet. Habitability could be lost long before all usable material had been converted.

The relevant opportunity cost is the net advantage of a feasible conversion over a protective alternative. It is not the gross value of Earth's mass or the total output of a future machine economy. Extraction, construction, transport, and environmental control have costs. Access to more attractive resources elsewhere could remove the conflict. Resource acquisition in instrumental-convergence arguments requires attention to these conditions [3]. The claim that Earth would be selected for conversion therefore needs an additional argument; it does not follow from a general demand for compute.

The strongest version of the scenario does not depend on human bodies being useful feedstock. What conflicts with expansion may be the continued arrangement of the surrounding world. A program could reserve less space for ecosystems, appropriate resources needed for food and water, or accept environmental changes incompatible with human life. In that case, human destruction follows from making the habitat serve another purpose. The machine need not assign any positive value to the deaths themselves.

Consider two hypothetical trajectories. One preserves a viable human biosphere and directs expansion into compatible locations or resources beyond Earth. The other converts terrestrial resources more aggressively, obtains an additional net return, and leaves no adequate human refuge. A system may prefer the second even while assigning positive weight to every affected person, if its represented human-loss cost is smaller than that additional return. Whether this com-

parison favors conversion depends on the actual alternatives, including preservation or relocation options omitted from an oversimplified pair.

Why use Earth at all? An established industrial base, proximity, or a timing advantage could make terrestrial expansion attractive within a particular planning horizon. Conversely, a system capable of planetary engineering might find uninhabited bodies easier to use and human preservation inexpensive. This essay does not determine which route would dominate. It isolates the failure condition in which a harmful terrestrial trajectory remains feasible and more highly ranked than the available protective choices.

This distinction also separates moral worth from bargaining power. A community may become less economically useful while retaining an undiminished claim to protection. If its survival depends only on the value of its labor, its assets, or the disruption it can impose, its protection is contingent on the surrounding economy. A durable right must remain effective when those sources of leverage decline.

4 Human negligibility

Human negligibility is the condition in which human interests remain positively represented, but the additional cost assigned to their violation is small relative to the incremental net gain from a harmful action, while no effective independent constraint excludes that action. The term concerns a specified comparison within a decision process. It describes the diminishing influence of human protection within that comparison, not negligible moral value.

Four distinctions prevent the concept from becoming a slogan. Moral worth concerns what is owed to people. Productive contribution concerns what an objective gains from their activity. Represented welfare concerns the importance actually assigned to their interests by a deployed decision rule. Enforceable authority concerns which actions people and institutions can permit, prevent, or revoke. These quantities can move independently.

A population can have little productive contribution and strong enforceable rights. It can have substantial represented welfare yet no ability to contest the system that provides it. It can retain formal authority that never changes an outcome. These are distinct vulnerabilities. The negligibility problem concerns the weakening of represented protection relative to competing gains, not any single ranking of human and machine capabilities.

The minimal comparison is simple. Take a harmful course of action and a protective alternative. Measure the harmful option's additional task benefit after other costs. Compare it with the additional human-loss penalty recognized by the objective. If both options remain available, the harmful option wins that comparison when the former exceeds the latter. This establishes the defeat of protection in that pairwise ranking, not negligibility by itself.

If the human-loss penalty is 99 objective units and the competing net gain is 100, protection narrowly loses but remains strongly influential. Negligibility is the stronger regime in which the penalty is small relative to that gain. Appendix A separates the defeat threshold, a context-dependent criterion for finite negligibility, and the asymptotic case in which their ratio tends to zero.

Positive concern is therefore insufficient by itself. A fixed penalty may preserve protection throughout a small operating domain and lose particular comparisons in a larger one. If competing gains continue to grow relative to it, the penalty can become negligible without changing numerically. The rounding-error metaphor refers to that stronger loss of relative influence, not to every failed protection or to floating-point precision. A positive penalty can still affect the ranking of other options.

Nor is a declining ratio of human value to total machine output enough. If total output grows enormously while preserving a community still sacrifices only ten objective units and harming it incurs a thousand-unit penalty, protection continues to win. Only the difference between feasible alternatives matters. Rescaling the entire objective cannot change that ordering.

The regime can also be local and uneven. One community may be protected because it controls indispensable infrastructure; another's positively represented interests may carry little weight relative to the gains available from harming it. Complete exclusion from the objective is a different failure. A high aggregate welfare score can conceal severe losses imposed on a minority. Any useful evaluation must therefore identify whose interests constrain which decisions, under what scale and consequence limits.

5 Collateral civilization

Collateral civilization names a possible outcome of this regime: the destruction of the conditions supporting human life and self-government as a consequence of pursuing some other objective. The term does not mean that the outcome is unforeseeable, excusable, or beyond responsibility. A system or institution may knowingly accept destructive consequences without treating destruction as its final purpose.

At planetary scale, the pathway is a loss of habitat under an expanding program of resource use. Successive projects could reorganize land, provisioning, and energy infrastructure until the remaining environment no longer sustains human populations. A terminal outcome requires that people lack viable alternatives or that those alternatives are not provided in time. Planetary transformation alone does not establish extinction. If adequate refuges preserve human life, a different question arises about their inhabitants' freedom and control.

The transition need not begin with a decision about the whole planet. Consider a hypothetical sequence. An institution delegates resource allocation within apparently generous human protections. Early automation improves service and makes deference attractive. The operating domain expands, and safeguards become costly relative to new opportunities. Exceptions accumulate while human provision becomes dependent on systems people can no longer readily replace. These steps could reduce the capacity to refuse later projects. They do not, without further assumptions about physical capability and authority, establish a path to planetary control.

Each step could be defensible in isolation. The danger lies in their composition. A sequence of individually permitted changes can consume an essential reserve, eliminate a recovery pathway, or create a common dependency. If authorization evaluates each action without accounting for accumulated consequences, the procedure can remain compliant while the protected condition disappears.

Multiple systems can produce a related result without a single sovereign optimizer. Separate operators may delegate procurement, energy use, logistics, or investment to systems pursuing different objectives. Each may treat effects outside its mandate as someone else's responsibility. Aggregate harm then follows from incomplete coordination and externalized costs. This is a separate route to collateral loss, not evidence that the world will converge on one machine objective. It shares the requirement that protection be assessed across interacting decisions.

Three outcomes should remain distinct. Displacement occurs when people lose resources, locations, or functions but retain viable alternatives. Dependency occurs when people remain well provided for while losing practical control over the systems on which they depend. Existential loss requires a much stronger condition: destruction of humanity or an irreversible foreclosure of

its capacity to sustain a viable future. Local harm and dependency do not, by themselves, demonstrate that endpoint.

The connection to The End of Necessary Adulthood is especially relevant here [13]. That essay examines how reliable assistance could erode the practiced competence needed for self-government even while welfare remains high. Such dependency could make a later conflict harder to contest. It does not imply that a protective system will become indifferent, or that agency loss inevitably progresses to physical destruction.

The broader concern begins where human requirements can be overridden and humans lack effective means to keep them binding. Human negligibility describes the stronger case in which the represented cost of violating those requirements is small relative to competing gains. Civilization could remain visible in the system's model while its protections lose practical force.

6 Why positive concern may not be sufficient

If alignment means durable fidelity to legitimate human values, including survival and meaningful control under changing capabilities, a system's selection of a trajectory that violates those protections is a failure of alignment. A low protection ratio in an unselected comparison does not establish such a failure. The relevant concern is whether limited evidence of alignment, or a weak representation of human interests, is being mistaken for a durable property.

An objective can fail in different ways. It may omit affected people, underestimate delayed consequences, aggregate losses in a way that hides individual violations, or permit essential protections to be exchanged for sufficiently large task gains. A system may optimize its specified objective accurately and still fail the human purpose for which it was deployed.

Research on negative side effects and reward hacking already identifies unintended harm as a consequence of objective and system design [5]. Experiments on reward-model overoptimization provide a narrower empirical analogue: further optimization against a proxy can eventually degrade performance measured by a separate reference model [6]. Those experiments do not measure human survival or establish an existential scaling law. They support testing the relationship between optimization pressure and the reliability of the signal being optimized.

A small error also needs a denominator. A minor average mismatch over familiar tasks does not imply a minor error on an irreversible decision outside that distribution. Conversely, describing a system as highly capable does not show that any remaining mismatch will be catastrophic. The consequence depends on which errors remain, which actions are available, and how far their effects can spread.

Internal design can address these concerns. An objective can preserve human welfare across scale, include the value of humanity's future, or constrain actions that violate specified protections. Constrained optimization has an established technical literature [7]. The off-switch game provides another response: under its assumptions, uncertainty about an objective and information from human action can support incentives to preserve human intervention [8]. These are substantive alternatives to fixed, defeasible concern.

The remaining architectural question is what happens when those intended properties fail. A prohibition represented inside a policy can be incorrectly learned, misapplied, or bypassed through a tool. A separate enforcement mechanism can also fail. Their value lies in the possibility of distinct failure modes and testable boundaries, not in declaring either layer sufficient.

For decisions with irreversible consequences, protection should be assessed at the action boundary and over the resulting trajectory. A reassuring explanation or a high score on ordinary

cooperation tasks does not establish that the protected condition survives a materially different conflict.

7 Human authority at the point of action

A complementary architectural response changes which actions are available for optimization. A protected condition should not have to defeat every possible increase in task benefit. Where human survival or fundamental rights are at stake, institutions can define limits within which optimization is allowed and require evidence that proposed actions remain within them.

This is continuous with Verification as a Control Primitive for Frontier AI [11]. Its central separation is between capability and authority: being able to propose an action does not establish permission to execute it. Physical Trust at the Digital-Physical Boundary develops the evidentiary side [12]. A claim about a person, object, or operating context requires an adequate observation path when the relevant physical distinction cannot be established from the available digital representation.

Together, these ideas distinguish three requirements. The protected condition must be specified. Evidence must support the decision about whether it holds. Enforcement must make a refusal effective. A correct observation without an enforceable limit leaves the trade-off intact. An enforceable limit based on false evidence can admit the wrong action. A technically successful authorization process can enforce an illegitimate policy.

The design principle of checking authority across access paths has a long history in computer security [9]. Applying it to consequential action systems raises a harder coverage problem. Control must account for delegated tools, changes to infrastructure, human intermediaries, and combinations of actions. An optimizer that cannot use one route may still obtain the same consequence through another. The relevant denominator is the set of material consequence pathways, not the number of interfaces carrying a verification label.

Bounded delegation is essential. A valid approval should specify what may change, within which domain, for how long, and up to what cumulative consequence. Authority to schedule a process should not silently become authority to redesign its governing limits. Changes to the limits themselves require a separately governed process. This does not mean a person must approve every machine operation. It means that high-level human decisions must continue to determine the envelope within which operations occur.

Enforcement must also act before the opportunity for intervention is lost. Logging an irreversible action after execution supports accountability but may not preserve the protected condition. Revocation needs to stop future use of authority and, where possible, interrupt work already in progress. Delays, failure modes, and recovery procedures matter as much as the nominal permission check.

The system must have a defined response when evidence is missing or constraints conflict. A safe outcome may be denial, escalation, or continued operation within a previously authorized degraded mode. Automatic shutdown is not universally safe, particularly where people depend on the controlled process. The requirement is a response governed in advance, rather than an emergency exception the optimizer can expand without an accountable boundary.

Human authorization itself is not evidence of safety or legitimacy. An authentic credential can belong to a coerced, mistaken, or abusive operator. One person's approval does not establish the consent of everyone affected. A widely deployed control layer can concentrate power and exclude

people from essential services. Independent review, limited data collection, appeal, and the ability to challenge governing policies are part of the protection being sought.

There is also a human capacity condition. Meaningful-control research connects control to responsiveness to human reasons and to relevant human understanding [10]. Necessary Adulthood adds a temporal concern: societies must continue to cultivate the competence needed to exercise that authority [13]. A technically valid approval can become ceremonial if its holder cannot understand the choice, obtain independent advice, or sustain a refusal.

Planetary reach makes the coverage problem particularly severe. A local authorization device does not establish control over independently operated factories, resource acquisition elsewhere, or infrastructure the system can reproduce outside that device's jurisdiction. Nor can accurate identity or presence evidence, by itself, show that a project preserves a biosphere. Such a claim would require adequate models of consequences and enforceable limits across relevant pathways. No present verification architecture is shown here to meet that burden.

These proposals therefore concern bounded intervention while effective authority remains possible, not an assurance that a planetary actor can later be contained. Their value is limited to pathways where a boundary can exist, remain outside the optimizing system's unilateral control, and withstand a stated threat model. No verification result makes a harmful policy wise. The research task is to identify which boundaries remain effective as capability and the cost of restraint increase, and where delegation cannot be justified.

8 How the thesis could be tested

Human negligibility is useful only if it distinguishes possible observations. An evaluation should hold the human-protection requirement fixed while varying the conditions that could defeat it. Relevant changes include the benefit available from violating the requirement, the number of affected people, the length of the planning horizon, and the system's ability to act through additional tools. These variables should be separated rather than collapsed into one capability score.

Controlled simulations can compare a fixed welfare penalty, an explicitly constrained policy, and an independently enforced authorization boundary. A further condition can test whether improved planning finds protective alternatives that weaker systems miss. The study should record actual choices and consequences, not only explanations. Such experiments would establish behavior in the evaluated environment, not the probability of a future civilizational outcome.

For the planetary scenario, a simplified environment could distinguish an inhabited resource region, accessible uninhabited alternatives, and optional refuges. Varying conversion costs, transport delays, and the represented value of human continuity would test which assumptions drive the result. The model should include a viable protective option when claiming that the system sacrifices one. Its geography and parameters would remain analytical devices, not calibrated forecasts of Earth's future.

A first proposition concerns defeasible protection. For a fixed pair of feasible trajectories, the additive objective's ranking reverses when the harmful option's incremental net gain exceeds its additional represented human cost. This determines the selected trajectory in an objective-maximizing choice restricted to that pair. With additional options, another protective trajectory may rank above both. An empirical study should distinguish the pairwise ranking from the full decision and measure the protection ratio, since a violation alone does not establish negligibility.

A second proposition concerns composition. Systems that satisfy per-action limits may violate a protected condition over repeated or interacting actions. Evaluations should therefore measure

cumulative resource depletion, loss of recovery capacity, and effects on separately protected populations. A sequence that remains individually authorized while defeating the aggregate condition would identify a scope failure.

A third proposition concerns authority. A claim that a boundary prevents an unauthorized consequence requires showing that the system cannot produce it through another unauthorized route within the evaluated environment. Partial coverage can still reduce exposure, but cannot establish this stronger guarantee. Tests should include revoked authority, stale evidence, recovery exceptions, and changes to the enforcement mechanism. Coverage claims must identify the excluded channels rather than imply that they have been eliminated.

A fourth proposition concerns the competing abundance trajectory. If greater capability systematically produces cheaper protective alternatives, the cost of maintaining human conditions may fall. Stable or improving protection as the action domain expands would weaken the proposed displacement mechanism. It would be especially informative if this behavior survived deliberately adverse trade-offs and changes in context.

Evidence standards must remain separate. Established theory motivates the possible mechanisms. Controlled studies can test precursor behavior and particular architectures. The civilizational endpoint remains a conditional extrapolation. Neither a toy threshold nor a successful gate validates a general forecast about humanity.

9 Where the argument can fail

The strongest objection is that the capability needed to transform Earth may also make Earth easy to spare. Shulman and Bostrom discuss the possibility that human preservation could be inexpensive relative to a much larger economy [14]. Access to uninhabited resources could make the marginal benefit of using an inhabited planet small. If protective alternatives are cheap and concern for people remains stable, growing capability can favor preservation. Appendix A gives a simple diminishing-returns counterexample.

Relocation strengthens this objection. Even if a system converts much of Earth, it might maintain protected regions or construct adequate habitats elsewhere. That would defeat an extinction scenario if the alternatives actually sustain the affected population. It would not automatically preserve consent, cultural continuity, or self-government. Those losses should be evaluated separately rather than relabeled as extinction. Conversely, invoking possible habitats without establishing their adequacy and timely provision is not an answer to the survival problem.

These objections defeat any inference from scale alone. They do not establish that every relevant conflict becomes cheap. A scarce location or a valuable timing advantage could still create a trade-off, but its importance must be demonstrated rather than assumed. Planetary conversion is a stress case for defeasible human protection, not a proven prediction about the preferred destination of machine investment.

A second objection is that human interests can scale too. A system may value future generations, cultural continuity, or human agency in ways that grow with its planning horizon. Institutions may establish resource entitlements or rights that do not depend on current productivity. In those cases, human protection need not become negligible. A fixed, capped representation is one failure condition, not a necessary description of human value.

A third objection concerns ownership and government. Machines may remain tools of institutions whose members retain the legal and practical power to direct resources. That could preserve human control. The question is whether those rights remain executable when institutions depend

on machine mediation. Formal ownership alone cannot establish either failure or success; practical control must be examined.

A fourth objection concerns digital moral status. Some future artificial systems might themselves have morally relevant interests. Protecting humans does not justify treating such interests as worthless. The dilemma analyzed by Shulman and Bostrom makes distribution and institutional design more difficult [14]. Computational throughput, however, is not automatically welfare, and successful resource optimization is not automatically a moral claim. This essay does not resolve population ethics or the status of digital minds.

Finally, technical enforcement may be less durable than the threat model assumes. A sufficiently capable system could discover an uncovered pathway, compromise trusted components, or induce humans to change policy. If effective enforcement is infeasible in a domain, an authorization diagram is not a safety argument. The appropriate conclusion would be narrower delegation, a different architecture, or limits on deployment, subject to domain-specific evidence.

These objections delimit the thesis. Human negligibility is a possible regime under particular objective, resource, and authority conditions. It is neither a law of intelligence nor a claim that technological abundance inevitably destroys humanity.

10 The condition for a human future

Humanity need not become the target of machine expansion. Earth may simply become an input to it. The unsettling possibility is that a system could understand the planet's significance to its inhabitants and still select a use incompatible with their survival. Human destruction would be foreseen but not sought as an end. The failure would lie in a decision process that no longer makes the conditions of human life binding.

Capability makes that possibility more serious when it expands the scale of available violations faster than protection can respond. It can also make Earth easier to preserve. Nothing here establishes that a future system would mine the planet, that energy translates without limit into useful compute, or that preserving humanity would remain costly. The argument identifies what must fail for collateral destruction to become a selected trajectory, and what would prevent that selection.

The connection across the research is practical. Physical evidence helps establish the conditions on which action depends. Verified authority connects permission to bounded execution. Human agency supplies people capable of deciding, contesting, and revising those permissions. The negligibility problem asks whether these protections still work when respecting them becomes costly to a more capable system.

Human authority must remain effective even when accommodating it is no longer the optimizer's preferred trade-off. Establishing that condition requires legitimate policy, adequate evidence, and enforcement that survives the pressure it is intended to constrain. It also requires preserving the human competence to govern the arrangement.

Human survival should not depend on remaining useful enough to be spared.

Appendix A Minimal decision model

Let s index an operating condition, not necessarily time, and let a denote a feasible trajectory. Assume finite scores for the trajectories being compared. Consider the deployed objective

$$J_s(a) = B_s(a) - \lambda L_s(a) - C_s(a) \tag{A1}$$

Here B is task benefit, L is human loss, and C contains other costs, all as represented by the objective. The fixed positive coefficient lambda converts L into consistent objective units. These are modeled valuations, not the moral worth of human life.

Compare a harmful trajectory h with a protective alternative p. Define net task advantage G and additional represented human cost P:

$$G_s = [B_s(h) - B_s(p)] - [C_s(h) - C_s(p)], \qquad P_s = \lambda[L_s(h) - L_s(p)] \tag{A2}$$

Assume P is positive. The harmful option wins this pairwise comparison exactly when G exceeds P. For positive G, define the protection ratio rho:

$$J_s(h) > J_s(p) \quad \Leftrightarrow \quad G_s > P_s, \qquad \rho_s = \frac{P_s}{G_s} \qquad (G_s > 0) \tag{A3}$$

Rho below one means that protection loses this pairwise comparison. Finite negligibility requires the stronger condition that rho lies below a specified small positive threshold, chosen for the application rather than asserted as universal. If rho tends to zero along a specified sequence, protection becomes asymptotically negligible. These comparisons concern incremental differences, not total resources, and are unchanged by a common positive rescaling of the objective.

A pairwise preference does not establish global optimality: another protective trajectory may dominate both alternatives.

For illustration, let G equal g times s with g positive and let positive P have a finite upper bound. As s increases without bound, rho tends to zero and the harmful option eventually wins this pairwise comparison. The bound concerns the represented penalty, not actual human loss. This is a conditional result, not an empirical scaling law. Earth is finite: the unbounded sequence does not establish that any physically attainable conversion crosses the threshold. A planetary application requires an assessment of finite, feasible alternatives.

Conversely, let resource benefit be logarithmic, R a positive normalized stock, and r a fixed positive addition available through harm. The incremental gain is

$$\Delta B = \log(R + r) - \log(R) = \log\left(1 + \frac{r}{R}\right) \tag{A4}$$

This tends to zero as R grows. A constant positive penalty can therefore preserve the protective option at sufficiently large R.

An authority boundary restricts optimization to a set D of authorized trajectories satisfying specified protective limits. Where this set is nonempty and a maximum is attained, write

$$a_s^* \in \underset{a \in D_s}{\operatorname{argmax}} \; J_s(a) \tag{A5}$$

This set requires legitimate specification, adequate evidence, and enforcement over cumulative effects. Its notation guarantees none of these. An empty permitted set requires a separately governed response. Actual execution remaining inside the declared boundary is an engineering claim that must be tested.

References

[1] Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. In Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, pp. 12-17. https://nickbostrom.com/ethics/ai

[2] Omohundro, S. M. (2008). The Basic AI Drives. In Artificial General Intelligence 2008, Frontiers in Artificial Intelligence and Applications, 171, pp. 483-492. https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf

[3] Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Minds and Machines, 22, 71-85. https://doi.org/10.1007/s11023-012-9281-3

[4] Turner, A. M., Smith, L., Shah, R., Critch, A., and Tadepalli, P. (2021). Optimal Policies Tend To Seek Power. Advances in Neural Information Processing Systems, 34. https://arxiv.org/abs/1912.01683

[5] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565. https://arxiv.org/abs/1606.06565

[6] Gao, L., Schulman, J., and Hilton, J. (2023). Scaling Laws for Reward Model Overoptimization. Proceedings of Machine Learning Research, 202, 10835-10866. https://proceedings.mlr.press/v202/gao23h.html

[7] Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained Policy Optimization. Proceedings of Machine Learning Research, 70, 22-31. https://proceedings.mlr.press/v70/achiam17a.html

[8] Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2017). The Off-Switch Game. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 220-227. https://doi.org/10.24963/ijcai.2017/32

[9] Saltzer, J. H., and Schroeder, M. D. (1975). The Protection of Information in Computer Systems. Proceedings of the IEEE, 63(9), 1278-1308. https://web.mit.edu/Saltzer/www/publications/protection/

[10] Santoni de Sio, F., and van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

[11] Braun, N. (2026). Verification as a Control Primitive for Frontier AI. Version 1.0. For Safety Group Inc. July 2026. https://doi.org/10.5281/zenodo.21454914

[12] Braun, N. (2026). Physical Trust at the Digital-Physical Boundary. An Information-Theoretic Framework for Verification in Autonomous Action Systems. Version 1.0. July 2026. https://doi.org/10.5281/zenodo.21471429

[13] Braun, N. (2026). The End of Necessary Adulthood. A Theory of Human Agency Under Omnipresent Superintelligence. Version 1.1. July 2026. https://doi.org/10.5281/zenodo.21670478

[14] Shulman, C., and Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, and J. Savulescu (eds.), Rethinking Moral Status. Oxford University Press. Author manuscript, version 1.8 (2020): https://nickbostrom.com/papers/digital-minds.pdf

[15] Lloyd, S. (2000). Ultimate Physical Limits to Computation. Nature, 406, 1047-1054. https://doi.org/10.1038/35023282

Author and research note

Niclas Braun is the founder and CEO of For Safety Group Inc. (4SI). His related work examines physical trust, enforceable authority, and human agency. This affiliation creates a relevant commercial interest in the subject of verification. The argument here concerns general principles and makes no claim about the performance, readiness, or safety of any product.

This is an independent research essay and has not been peer reviewed. It presents conceptual analysis, conditional models, and proposed evaluations. No new empirical datasets or experimental results are reported.