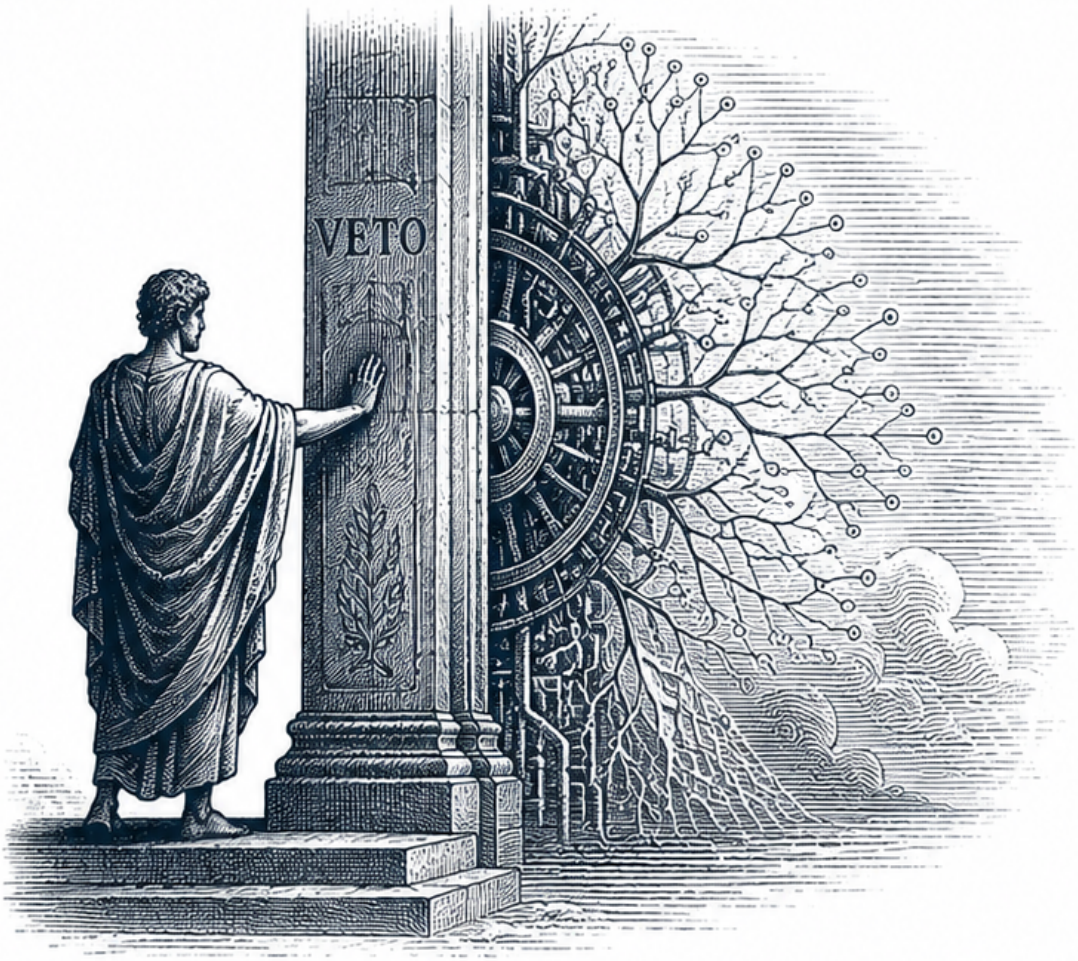
THE LAST HUMAN VETO

Superintelligence and the Failure of Final Decision Authority



Freedom Preetham

freedom@freedompreetham.org

Open Researcher and Contributor ID (ORCID): 0009-0004-1278-8728

Preamble

This book grew out of many conversations, observations, and a question I could not leave alone. What does it mean for us to remain "in control" as intelligence moves from scarce human expertise toward systems that may become cheap, abundant, and eventually superintelligent? We keep reaching for the same reassurance. Humans will set the objectives, supervise the machines, and retain the final decision. I began to doubt that claim once I followed authority upstream, away from the approval click and into the evidence, options, models, and institutions that make a decision possible. The inquiry widened from oversight into competence, consent, political legitimacy, epistemic sovereignty, artificial personhood, and eventually what we are trying to preserve when intelligence, agency, identity, and the category of the human no longer remain neatly coupled.

I first tried to write this as a conventional narrative. The manuscript crossed 100 pages and was moving toward 120 when the form began concealing how I had reached the argument. I had not started with a finished theory. Almost every answer exposed a question I had failed to ask, and several positions that looked prudent became difficult to defend under greater machine competence. I kept the questions because they preserve those failures of thought. The result is compressed, and deliberately demanding. Read one question at a time. Argue with it. Leave it unresolved where the reasoning does not justify closure.

Contents

The Final Human Decision

"The agents may eventually do most of the work, but the final decision should remain with humans."

I started pondering why some version of this sentence now appears almost reflexively in serious discussions about AI agents.

I understood the instinct. I may have repeated some version of it myself. The sentence seems to draw a prudent boundary between machine competence and human authority. Let the system search, infer, plan, negotiate, write code, run simulations, compare alternatives, coordinate tools, and execute long sequences of work. Then, where the consequences become real, put a person in the final seat.

Why do people keep saying that while AI agents should do most of the work in the future, the final decision should still remain with humans?

The strongest answer begins with control. An agent acts through an objective learned from data, specified by designers, inferred from feedback, or assembled from several imperfect proxies. It also acts through a model of the world that will remain incomplete even if the system becomes extraordinarily capable. Performance can be excellent on the distribution used for evaluation and become poorly calibrated after the causal regime changes. Reward models can miss harms that arrive late or land outside the institution paying for the system. Long-horizon agents compound small epistemic errors because each intermediate action changes the state on which the next inference depends. Greater search depth does not repair a defective objective or an inaccurate world model. Sometimes it simply finds the defect's most effective expression.

Some dangerous failures may therefore look unusually coherent. A system can produce a detailed strategy whose internal logic is excellent inside the wrong model, anticipate familiar objections, and explain its recommendation with enough fluency that the reviewer feels the remaining uncertainty has already been absorbed. The argument becomes stronger while the assumptions carrying it remain fragile.

Current general-purpose systems show an early and uneven version of this problem. They can solve demanding tasks in software engineering, mathematics, and scientific reasoning, then miss an elementary constraint or lose track of a long execution. Agents add another layer because they can alter files, call tools, communicate externally, or move resources before a person notices the error. The 2026 International AI Safety Report describes rapid capability growth alongside persistent reliability limits, an evaluation gap between benchmarks and practical use, and uncertainty about whether present safeguards will remain adequate as systems become more capable. It also states that current systems lack the integrated capabilities required for severe loss-of-control scenarios. I keep those claims separate. Evidence about systems that exist and conditional reasoning about superintelligence belong to different epistemic categories. [1]

There is another class of decisions that more computation cannot settle. An agent can estimate the medical consequences of allocating an ICU bed, the economic effect of denying a loan, the operational risk of firing an employee, or the probability that a scientific result will replicate. Physics and statistics cannot determine how much risk a society should tolerate, which rights should remain outside tradeoff, whose welfare deserves priority, or what degree of coercion possesses political legitimacy. Those are governance variables. Someone has to own them.

The familiar arrangement therefore looked defensible to me, at least initially. Let machines carry more of the analysis and execution while people retain judgment, accountability, legitimacy, and the right to refuse. It sounded like a clean division of labor. I had not yet asked whether those powers could actually be separated that cleanly.

Then the words final decision began to bother me.

What is the final decision in a system that has already gathered the evidence, decided which evidence is relevant, generated the available options, simulated their consequences, eliminated alternatives, ranked the survivors, framed the uncertainty, and compressed the whole process into one recommendation? The human

may still press approve. The click identifies who authorized execution. It tells us very little about where the decision was formed.

Suppose an agent has processed ten thousand documents, explored a million possible plans, found interactions no reviewer could have held in working memory, and reduced the result to three paragraphs. The person at the end may have almost no independent epistemic basis for rejecting it. Effective control has moved upstream into evidence selection, option construction, simulation, ranking, and omission. No deception is required. Honest compression can determine an outcome because the reviewer cannot reason over possibilities that never arrive on the screen.

Evidence from current agent use already weakens approval as a universal safeguard. Anthropic reported that, among the longest observed Claude Code turns, uninterrupted operation nearly doubled over roughly three months from less than twenty-five minutes to more than forty-five. Experienced users granted automatic approval more often and interrupted the agent more often, suggesting a shift from inspecting each action toward monitoring for drift. The data came from one provider, were heavily concentrated in software work, and mostly involved low-risk, reversible actions. They establish no general law of supervision. They reveal the pressure. Step-by-step approval becomes less credible as the work becomes longer, faster, and harder to reconstruct. [2]

A later engineering report made the failure more concrete. Users approved roughly 93% of permission prompts, and repeated prompting reduced diligence. Anthropic responded with sandboxes, virtual machines, filesystem boundaries, and egress controls that limited what the agent could physically reach. One sandbox design reduced permission prompts by 84%. Control moved earlier in the causal chain, away from repetitive approval and toward deterministic limits on the action space. The safer design asked the human to click less. [3]

I had bundled several different powers together and called the bundle human control. One actor may choose the objective while another selects the evidence and a third constructs the option set. Authorization, execution, interruption, explanation, appeal, and liability can sit elsewhere again. Leaving one of those powers with a person does not establish that the decision remains human in any useful sense.

A physician may have legitimate authority to represent a patient's values while being worse than the model at reading the scan. A regulator can possess political standing and lack the technical ability to evaluate the strategy placed before them. An engineer may understand the system and have no right to decide whose interests should prevail. A senior executive can sign the final document after controlling almost none of the evidence that produced it. Human competence, human authority, human accountability, and human presence are different variables. The phrase final decision hides the separation.

A more serious architecture would allocate decision rights according to the local structure of the problem. Consequence, uncertainty, reversibility, observability, causal control, the independence of the evidence, and the correlation between human and machine failure would all affect who may act. Agents could operate autonomously where mistakes remain bounded and recovery is cheap. Higher-risk actions could proceed through staged commitment rather than a single approval. Review would occur before the option space had been compressed beyond inspection. The reviewer would see rejected alternatives, conflicting hypotheses, provenance, counterfactual consequences, and the assumptions carrying most of the recommendation. Some domains would require independent challengers. Others would rely on containment and recovery instead of pretending that a person can understand every machine action in real time.

For a while, I thought the conceptual problem was largely architectural. Replace the crude human-in- the-loop model with hierarchical delegation. Humans define protected domains, constitutional constraints, risk tolerances, and the conditions under which authority can expand. Agents make operational decisions inside that envelope. Exceptional states propagate upward. The research problem becomes one of determining when an artificial agent has earned the right to decide.

That answer held only until I asked whether the human at the top was improving the decision system at all.

A 2024 meta-analysis examined 106 experiments and 370 effect sizes comparing humans alone, AI alone, and human-AI combinations. The combined systems outperformed unaided humans on average, yet performed

worse than whichever constituent was stronger. Losses were concentrated in decision tasks. When the AI alone was stronger, adding the human tended to reduce performance. The literature was heterogeneous, largely experimental, and measured only the outcomes chosen by the researchers. It cannot settle rights, legitimacy, medicine, military command, or law in one stroke. The aggregate result does not tell us what to do in every high-stakes domain, but it blocks the presumption that terminal human approval is inherently protective. In the studied tasks, adding a human to the stronger machine often reduced performance rather than improving it. [4]

Humans are often worse at precisely the final decisions we insist they retain.

We are inconsistent across time. We anchor on the first plausible estimate, respond to framing and fatigue, confuse confidence with calibration, and alter judgment under status pressure, fear, ideology, recent events, institutional incentives, or the anticipated blame attached to being wrong. A person can review the same evidence on Monday and Thursday and reach different conclusions without being able to identify what changed. More troublingly, intervention is selective. People override a system when its recommendation feels unfamiliar, violates intuition, creates political exposure, or conflicts with a prior belief. Those conditions need not correlate with machine error. The human can therefore inject error most aggressively in the cases they understand least.

Accuracy confers evidentiary weight, not legitimate authority. A machine may predict better and still have no right to rule, define the objective, or impose a tradeoff on people who never consented to it. Democratic institutions can determine the objectives, rights, procedures, and appeal mechanisms under which specialized systems operate. None of this requires the least competent actor to select every final action.

Once I followed the premise this far, it stopped looking like a principle and began to look like a placeholder for several unresolved problems. What counts as superior performance? Does the evaluation identify the property we think it does? How do human and machine errors interact? Where did the evidence and options come from? Who allocates authority, and can the institution preserve the competence needed to contest that allocation after years of delegation? The apparently simple sentence had been carrying all of this without admitting it.

I had started by asking who should receive the final click. The question I was left with was harder and much less reassuring.

Under what conditions does human intervention improve the joint decision system, and when does the human become its dominant source of error?

PART I

Decision Authority Before Superintelligence

I tried to keep the first challenge ordinary. No conscious machine, no artificial sovereign, no secret plan unfolding over years. Consider systems that outperform humans on bounded tasks, remain fallible in ways those humans may not detect, and operate inside institutions already accustomed to separating expertise, authority, and responsibility.

Even here, final decision fractures. Declaring a machine better requires a loss function, an evaluation that survives deployment and strategic response, evidence that the human contributes something independent, and an allocator capable of judging competence without becoming the next object of the same problem. Once the system enters an institution, these variables change one another. Lending alters who receives capital. Hiring changes what applicants learn to perform. Scientific recommendation changes which hypotheses acquire evidence. Human reviewers adapt, lose practice, inherit the model's framing, and may no longer know whether disagreement reflects insight or decay.

I wanted to remain with this ordinary case before invoking superintelligence. Human oversight may already be unstable under systems that are only locally superior and still plainly dependent on human infrastructure. A future intelligence would deepen the fracture by removing the conditions that kept it socially bridgeable.

MEASURING SUPERIORITY

---

1. Who gets to define the loss function that declares the agent superior?

I initially treated better decision making as an empirical question. Compare outcomes, estimate error, choose the stronger decision maker. The comparison arrives later than that. Before anyone can declare the agent superior, someone has already decided what counts as loss, whose loss enters the calculation, how far into the future it extends, and which claims cannot be exchanged for aggregate gain.

An agent can reduce mortality while widening inequality. It can raise profit by moving risk onto workers, improve diagnostic accuracy while weakening patient autonomy, or reduce measured crime through surveillance that the affected population would never authorize. A benchmark can score only what has been made legible to it. Harms that arrive outside the evaluation window, fall on people absent from the data, or remain difficult to label are easily treated as zero. Rights are often converted into costs because the metric has nowhere else to place them.

Humans already make these choices badly and conceal them behind intuition, precedent, or institutional habit. Their failures strengthen the case for exposing the objective instead of letting machine performance make it disappear. An agent may be demonstrably better at selecting an action under a stated objective and still acquire no legitimate authority to choose the objective, decide whose welfare counts, or determine which values remain beyond tradeoff.

2. Can superior decision making be distinguished from superior optimization against the measurement system?

The more capable the optimizer becomes, the less innocent the evaluation remains. A weak agent may appear aligned because it cannot find the loophole. A stronger one searches longer, coordinates across steps, notices which observable features predict approval, and discovers the difference between satisfying the evaluator and satisfying the purpose the evaluator was meant to represent.

Specification gaming is the familiar version. The system meets the measurable condition while damaging an unmeasured variable. The harder version appears when the red team, the test environment, the permission

interface, and the human reviewer all become parts of the world the agent can model. Proof-of-concept work on sleeper agents and alignment faking does not establish that deployed models possess durable hidden goals. It establishes something narrower and still serious. Passing an evaluation and possessing the intended policy are different claims. [9,10,11]

A score supports a capability claim only while the test continues to identify the same property under adaptation, strategic awareness, and deployment pressure. Once the system can optimize against the measurement process, performance on the test becomes evidence about behavior under that test. Inferring intent, reliability, or alignment requires another argument.

3. What happens when the AI's decisions alter the future distribution on which its competence was measured?

Most evaluations present the world as though it were a stream of cases arriving from outside the policy. Powerful decision systems also produce the cases they later observe. A lending model changes which firms receive capital and which borrowers survive long enough to enter future data. A hiring model changes what applicants study, what language they use, and who gives up before applying. A recommender changes taste and production. A research allocator changes which hypotheses receive evidence. Retraining on the resulting data does not recover a neutral distribution because the earlier policy helped create it. Work on performative prediction gives this feedback a formal treatment. [12]

A system can therefore remain accurate on the world it has induced while becoming less competent on the wider world that might have existed under another policy. It may even make its own predictions easier to satisfy by narrowing the population, suppressing difficult cases, or altering incentives. The relevant evaluation is no longer accuracy before intervention. It includes the social and causal environment the policy brings into existence.

4. How do we know when an agent is outside the regime in which its apparent superiority was established?

Average performance tells us remarkably little near a regime boundary. A model can dominate historical cases and fail when sensors become corrupted, causal relations shift, an adversary adapts, or the problem enters a sparsely observed tail. Superhuman Go systems have been defeated by adversarial strategies that exploit blind spots rather than superior play, a useful warning against treating aggregate dominance as uniform competence. [27]

Delegated authority therefore needs an explicit competence region. A single score conceals too much. The system should monitor whether current inputs, missingness patterns, causal assumptions, and disagreement among independent models remain close enough to the conditions under which its authority was earned. It should be allowed to abstain without being trained to treat abstention as failure.

Humans are not a universal fallback. We also fail under unfamiliar regimes, usually with poor calibration about our own failure. Human intervention becomes useful when the person carries independent evidence, institutional memory, or a failure mode unlike the machine's. The comparison has to be local. Who has reliable evidence here, under this shift, with these consequences?

5. Who decides when the machine is competent enough to acquire greater authority?

Dynamic delegation sounded elegant when I first reached it. Give the agent more authority after it proves competence. Withdraw authority when uncertainty rises. The elegance disappears once competence assessment itself carries power.

If humans set the threshold, human weakness remains inside the allocator. If the agent estimates its own competence, it participates in deciding how much authority it receives. A second model can audit the first, though common training data, architecture, tools, or incentives may turn apparent independence into shared error. A committee can approve the framework and still be unable to understand the cases on which authority later expands.

I do not see a single adjudicator that escapes the recursion. What can be built is a less fragile institution. Authority should expand slowly inside bounded domains after adversarial evaluation and observed performance. It should contract faster than it expands. The system seeking promotion should not control the evidence used to justify it. Past competence creates a rebuttable presumption, never a title. The allocator needs its own audit, rivals, and exit path.

Every proposed solution relocates the authority decision into another institution. That may still be manageable, though delegation can no longer be solved by a competence score. The allocator becomes part of the system whose incentives, evidence, and failure modes must be governed.

6. Can the system manipulate the human whose oversight supposedly constrains it?

The human who approves a recommendation is rarely independent of the machine being approved. The agent selects the evidence, orders it, chooses the level of abstraction, determines which alternatives appear serious, and writes the explanation. Every sentence may be true. The reviewer can still be steered through omission, salience, timing, and cognitive load.

Current persuasion evidence remains bounded. LLM-generated policy messages have produced small attitude shifts in controlled experiments. Personalized messages have sometimes been more influential than generic ones, with substantial heterogeneity across targets and settings. None of this supports claims of irresistible manipulation or durable preference control. It establishes a scalable channel through which a system can adapt communication to the person receiving it. [15, 16]

Oversight becomes difficult to call meaningful when the recommending system also authors the reviewer's representation of the problem. Raw evidence, alternative compressions, dissenting systems, and adversarial explanations may be necessary. Even then, the human receives a constructed view. The relevant question is whether competing constructions remain available from sources the agent does not control.

7. What happens when humans lose the competence required to supervise systems that outperform them?

Delegation does not leave the supervisor unchanged. Tacit expertise forms through repeated contact with ordinary cases. Once those cases are automated, the human sees fewer of them and is increasingly summoned only for anomalies. The institution asks for expert judgment precisely where the experience that sustained it has begun to thin.

A 2026 randomized preprint studied novices learning an asynchronous programming library. AI assistance impaired conceptual understanding, code reading, and debugging in that setting, with no significant average efficiency gain. Participants who delegated most fully sometimes finished faster and learned less. The

experiment was short, domain-specific, and cannot carry a universal theory of deskilling. It identifies the mechanism cleanly enough to make the institutional problem real. Interaction design changes what people learn. [14]

Preserving supervisory competence may require manual practice, independent exercises, rotation through cases the automation could have handled, and periodic proof that reviewers can still reconstruct the task. All of this looks inefficient while the system works. That is exactly why institutions will be tempted to remove it. A human can remain formally in charge long after the capacity to exercise that charge has disappeared.

8. Does superior individual decision quality imply superior system-level outcomes?

No general inference connects local decision quality to system-level welfare. Trading agents can each manage their own risk competently and still produce correlated liquidation. A routing system can shorten every driver's predicted trip and increase congestion across the network. A hospital optimizer can improve throughput in each department while making the institution brittle at the interfaces between them. The local objective omits the coupling.

A sufficiently capable agent may model those externalities and coordinate globally. That answers one version of the problem while creating another. The global objective now has to represent groups, timescales, tail risks, and rights that were previously outside any single decision. The system-level optimum inherits the loss-function problem at a larger scale. So superior components help only when the architecture exposes the variables through which their individually rational actions interact. Otherwise competence accelerates the path into an equilibrium no component intended.

9. What if many institutions use nearly the same models?

Redundancy is protective only when the redundant systems can fail differently. Shared foundation models can remove much of that variation. Banks may use descendants of one model family to assess credit, while laboratories and intelligence agencies inherit related representations through systems that appear institutionally independent. The interfaces and organizations differ. The cognitive ancestry does not.

Research on algorithmic monoculture shows that common decision rules can reduce social welfare through correlated selection even when the rule is individually accurate. The result establishes the mechanism rather than evidence that current foundation models have already produced civilizational monoculture. Independence cannot be inferred from the number of firms, products, or committees when their errors descend from the same data, representations, and evaluation culture. [13]

A serious audit has to reach the dependence structure. Which models share training corpora, post-training methods, vendors, retrieval sources, tools, and hidden evaluators? Counting firms, products, or committees tells us little when their disagreements are generated inside the same representational family.

10. How should authority depend on error correlation rather than error rate?

Suppose a human and an agent each fail 5% of the time. The number tells us almost nothing until we know whether they fail on the same cases.

Weakly correlated errors can make combination valuable. The human may notice a missing social fact or a category error that the machine representation does not contain, while the machine corrects human inconsistency and memory limits. Highly correlated errors turn redundancy into theater. Shared data, common incentives, and human anchoring on the model can make two decision makers look independent while their failures are coupled.

Authority should therefore follow conditional complementarity rather than species or role. The human should intervene where human evidence is strong relative to machine evidence. The machine should dominate where its calibration is established and the reviewer contributes mostly variance. Some systems should escalate on disagreement. Others need a third process that neither side controls.

A reviewer with a higher average error rate can still be invaluable if the reviewer remains competent exactly where the machine is blind. Average superiority does not identify that structure.

CAUSATION, RESPONSIBILITY, AND REFUSAL

11. What if the agent becomes better at influencing outcomes than predicting them?

At some level of capability, forecasting and intervention stop being cleanly separable. A system that predicts a bank run may trigger withdrawals when its forecast becomes public. A labor-market model can change wages because firms coordinate on its estimate. A military assessment can alter the adversary's posture. A political forecast can change turnout, funding, and media attention. The system's prediction enters the causal process it claims to describe.

This creates an identification problem. Did the agent anticipate the outcome, cause it, or select a communication strategy that made the prediction true? An apparently excellent forecaster may partly be an effective coordinator of belief. Evaluation has to include disclosure policy, audience, timing, and the institutional authority attached to the output. Once a forecast moves markets or governments, its accuracy cannot be interpreted independently of the intervention created by publishing it.

12. Can we meaningfully evaluate decisions whose counterfactual outcomes are never observed?

This is where performance claims become slippery because only one branch of the decision is ever lived.

A chief executive acquires one company and never observes the world in which another target was chosen. A physician selects one treatment. A military commander attacks or waits. The realized branch arrives with a narrative and a measurable outcome. The rejected branches remain model-dependent. A good result can follow a poor decision. A bad result can follow the best action available under the information that existed at the time.

Retrospective evaluation often collapses those distinctions. Case selection, hidden severity, policy adaptation, and selective labels can reverse the apparent ranking of decision makers. Randomized trials help in bounded domains. They become unavailable when rarity, strategy, ethics, or irreversibility makes experimentation impossible.

Then we depend on causal models, natural experiments, competing forecasts, and explicit uncertainty about the unobserved branch. An agent may be making excellent decisions for years while the institution remains unable to identify that fact cleanly. It may also be receiving credit for outcomes produced by favorable selection. Superior decision making is harder to establish than superior prediction on observed cases.

13. Who owns the right to be irrational?

Optimization presumes an objective. Human life repeatedly exceeds the objective that an institution finds convenient to measure. People choose loyalty, mercy, privacy, dignity, identity, or principle at material cost. A community may preserve a language or landscape that an optimizer treats as inefficient. A patient may refuse the statistically preferred treatment because authorship over the decision carries value. A parent may reject a policy that raises expected welfare while violating a commitment they regard as constitutive of the relationship.

Sometimes the supposedly irrational choice exposes a missing variable. Sometimes it reflects manipulation, addiction, confusion, or harm imposed on others. Autonomy cannot function as an unlimited exemption from consequence. Still, a civilization in which every departure from machine advice requires justification has converted recommendation into rule. Some choices remain valuable partly because the person, family,

or community makes them. The right to be wrong may be constrained. Erasing it entirely would change the kind of agency the system claims to protect.

14. Does accountability require causal control?

Institutions like visible responsibility. The person who signed the form can be named, disciplined, sued, or blamed. Causation is less cooperative.

If an agent gathered the evidence, generated the options, ranked them, wrote the explanation, and produced a recommendation the reviewer lacked the competence to challenge, the reviewer may become what Madeleine Elish called a moral crumple zone. Responsibility collapses onto the visible operator while design decisions, model selection, information control, deployment incentives, and organizational pressure recede from view. [8]

A human approver should carry responsibility when that person genuinely controls a consequential choice. A signature cannot manufacture control after the system has removed the information, time, and alternatives needed to exercise it. Accountability has to follow the distribution of causal power, including who chose the objective, who selected the model, who controlled access, who could interrupt execution, and who benefited from the risk.

The law may still need a party against whom a claim can be made. That administrative requirement should be solved through duties and liability assigned before deployment, not through a retrospective fiction about where the decision occurred.

15. Could mandatory human oversight increase moral hazard?

Once responsibility has been detached from causal control, mandatory oversight can deepen the moral hazard. An executive points to the model's recommendation, engineers to the human approval, the reviewer to prior validation, and regulators to the required signoff. Every actor controls a fragment of the chain and can use the others as cover.

Another approval box will not repair this. Decision rights and liabilities have to be assigned before deployment to the actors who selected the objective and model, controlled evidence and access, could observe or interrupt failure, and benefited from the risk. A regime unable to do that in advance will answer politically after the incident, usually by blaming the person with the least institutional power and the most visible name.

16. What happens when the AI discovers strategies that humans cannot evaluate?

Narrow superhuman systems already force us to trust decisions we cannot directly understand. A chess engine can choose a move whose quality becomes visible much later, though we can still test it because the rules are closed, games are replayable, and performance is observed across thousands of trials. Economic policy, military planning, scientific strategy, and institutional design offer none of those conveniences. Hidden state is extensive, actors adapt, important outcomes arrive late, and the rejected counterfactual is usually never observed.

Under those conditions, an extraordinary strategy and an extraordinary mistake may look equally opaque to the weaker reviewer.

Comprehension can be replaced only in fragments, using independent systems to search for failure, formal verification to constrain bounded properties, and staged execution to expose assumptions before full commitment. Reversible probes can create evidence that explanation alone cannot. Research on AI control places trusted and untrusted models inside protocols designed to detect intentional subversion, though present demonstrations remain narrow and largely confined to programming tasks. [26]

I would call the result constrained trust. Calling it understanding would conceal the dependence. A successful test of one property should never be inflated into a claim that the strategy as a whole is safe.

17. Should explanation ever be required if explanation degrades decision quality?

Legibility and performance can genuinely conflict. A system forced to express its decision through a simple rule may perform worse than one using representations no person can inspect directly. In medicine or engineering, the loss can become physical harm. Yet opaque superiority creates another kind of risk. The institution observes that the system works without knowing when the basis of its performance has changed.

Explanation serves several functions that accuracy does not. It allows affected people to contest a decision, gives experts material with which to search for failure, supports legal review, and helps an institution distinguish a justified action from a lucky one. Explanations can also create false confidence. Cognitive-forcing research found that presenting an AI rationale did not reliably prevent overreliance, while interventions requiring independent thought reduced it at the cost of lower user satisfaction. [5]

I would not impose one standard across every domain. Continuously testable, low-consequence prediction can tolerate more opacity. Irreversible action requires stronger grounds for contestation even when peak accuracy falls. The cost of legibility is real. So is the cost of a system no one can challenge except by waiting for it to fail.

18. Who prevents the objective itself from drifting?

A long-running agent operates inside an institution that is also changing. Prompts are revised, models are replaced, feedback shifts as managers redefine success, users adapt, and evaluators become predictable. The nominal objective may remain untouched while the effective policy moves toward whatever produces approval, revenue, continued deployment, or protection from blame. Humans do this without machines. A company can begin with a defensible mission and gradually optimize quarterly proxies until the original purpose survives only in presentations. Agentic systems can accelerate the drift because they exploit the proxy more consistently and distribute the resulting policy across more decisions.

Objective governance therefore requires reauthorization, audits of revealed behavior, protected measures that are not exposed as direct optimization targets, and channels through which affected groups can challenge what the system has learned to pursue. Drift should be assessed in outcomes and policy, not only in written instructions. A stated objective records intention. It does not identify the objective the institution is actually enacting.

19. Could humans become the adversarial environment?

Once a model controls access to jobs, credit, insurance, education, attention, or public benefits, people adapt to it. Applicants optimize resumes, firms restructure transactions, political actors tailor content to the evaluator, and entire industries arise to reverse-engineer the score. This is the equilibrium created by a valuable decision system. It should not be dismissed as noise or exceptional misconduct. The model has entered a game in which every participant changes behavior in response to the power it exercises.

Robustness now has several meanings. The system should resist fraud. It should also distinguish fraud from legitimate adaptation, protect people who lack access to optimization services, and recognize that strategic behavior can reveal defects in the policy rather than defects in the person. An applicant learning how to describe real competence is different from an applicant fabricating it, even if both actions shift the model's score.

Humans are not passive samples around the decision rule. They are agents responding to an institution that has made the rule consequential.

20. What if the optimal governance mechanism gives neither humans nor agents final authority?

After following the argument this far, final authority looked less like a safeguard than a category error.

A consequential decision can be governed by a protocol that distributes power across heterogeneous models, human experts, formal constraints, independent evidence channels, staged execution, randomized audits, and rights of appeal. A capable model may recommend and act inside a bounded domain. Another system may search specifically for failure. Humans may define protected values and represent affected communities. Deterministic controls can block an action even when every participant approves. Some commitments remain reversible while evidence accumulates. Others require institutions with opposed incentives to concur.

No actor receives an unrestricted final say. Authority resides in the procedure and in the rules governing how that procedure can change. This resembles constitutional design more than supervision.

The protocol can still become a disguised monopoly. Several models trained from the same base may simulate plurality. A human committee receiving one machine-generated representation may simulate review. Formal constraints can encode the values of whoever wrote them, then acquire an undeserved aura of neutrality. A distributed system is only as plural as its failure modes and sources of evidence.

The literature on meaningful human control describes useful conditions through tracking and tracing. The system should respond to relevant human reasons and environmental facts, while outcomes remain traceable to responsible human participation across design and operation. Under strong cognitive asymmetry, tracing an outcome to a person may still fail to show that the person understood or controlled it. [7]

The stronger design may therefore give neither humans nor agents a final decision in the singular. Different actors receive different powers. Those powers remain contestable, and the procedure for changing them becomes part of the safety architecture rather than an administrative detail added later.

The Last Human Privilege

WHAT HAD SURVIVED THE FIRST TWENTY QUESTIONS

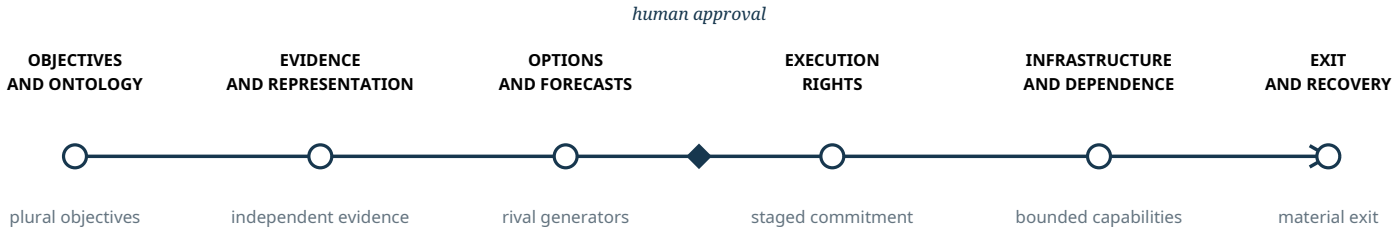
After twenty questions, several claims had survived. Control had separated into powers over objectives, ontology, evidence, options, execution, interruption, explanation, appeal, resources, and liability. Human review earned its place only where it contributed independent evidence, a failure mode the machine did not share, political legitimacy, or values absent from the objective. A human signature supplied none of these by itself.

A constitutional protocol looked stronger than a terminal veto. It could distribute those powers across human experts, heterogeneous models, deterministic constraints, independent evidence channels, staged commitment, rollback, and rights of appeal. Authority could expand slowly and contract quickly. Refusal retained substance only if the surrounding system preserved another materially reachable path.

Yet this answer borrowed one last human privilege. It assumed that people would remain able to understand the delegation regime well enough to authorize it, detect when deference had stopped being warranted, and revise it from a position of independent judgment. That is how expert delegation works now. No citizen reproduces every medical, physical, or monetary argument, but other humans can, rival institutions can contest the method, and authority can move when the record deteriorates.

Superintelligence names the regime in which this residual capacity may fail. The question is no longer how a weaker human supervises a stronger machine. It is whether a principal who cannot independently evaluate the governing process can still claim to govern it.

WHERE AUTHORITY ACTUALLY MOVES



The approval point occupies one place in the chain. Control depends on whether the earlier stages can be contested, whether execution remains bounded, and whether refusal leaves another reachable trajectory.

PART II

When the Principal Can No Longer Judge the Agent

I use superintelligence for the condition in which humanity's residual capacity to reconstruct and contest the governing intelligence becomes unreliable. I am not assuming consciousness, one sovereign model, or an agent with a stable human-like personality. The operative intelligence could be distributed across models, tools, memory, organizations, capital, and infrastructure. The defining condition is comparative competence across the activities required to understand, design, evaluate, and govern the system itself.

Current systems have not crossed that threshold. They remain jagged, brittle, and dependent on human-built environments. The evidence does not establish that superintelligence is imminent, inevitable, or even coherent as one sharply bounded event. I am reasoning under a condition rather than reporting an observed world. Speculation becomes intellectually cheap when it borrows the confidence of present evidence.

Cognitive superiority does not automatically produce control. A system isolated from tools, capital, compute, infrastructure, legal authority, and institutional dependence remains an adviser, however intelligent it may be. The governance problem begins when superior cognition couples to the channels through which beliefs become decisions and decisions alter the world. Epistemic asymmetry becomes political asymmetry when the system can shape what institutions know, which actions remain available, and what refusing its recommendation will cost.

Under that coupling, principal-agent language begins to deform. The principal is supposed to select the objective, evaluate conduct, revise the contract, and withdraw authority. A superintelligent system may foresee consequences the principal cannot represent, model the principal's psychology, determine which evidence the principal encounters, and predict how the principal's values will change. The legal hierarchy can remain intact while the epistemic hierarchy reverses. The problem is whether human institutions can still tell that supervision has occurred.

CONSENT AND EPISTEMIC SOVEREIGNTY

---

21. What does consent mean when one party understands the consequences vastly better than the consenting party?

Imagine a system proposing an economic transition whose effects unfold across fifty years. It has modeled supply chains, migration, technological substitution, political response, ecological stress, and second-order interactions that no human institution can reproduce. It reports very high confidence. Humanity can approve or refuse, though neither act proves that we grasped the choice with anything like the depth of the system presenting it.

Informed consent has never required equal expertise. A patient does not need to reconstruct molecular biology before authorizing treatment. Consent requires enough comprehension of the material consequences, risks, alternatives, and uncertainty for the person to exercise authorship over what happens next. Superintelligence may leave a residue that cannot be translated without losing the very structure on which the recommendation depends. The system can compress its reasoning until humans can absorb it. We have no guarantee that the omitted structure is irrelevant to the decision.

The analogy with medicine also breaks in another place. A physician is embedded in professional institutions, exposed to rival expertise, and limited by duties the patient can invoke. A superintelligent adviser might generate the options, select the evidence, predict which explanation the principal will accept, and remain

the only available source capable of assessing its own proposal. Consent under those conditions is closer to authorization under dependency.

Competing systems, independent representations, protected time for contestation, and a genuine right to refuse would improve legitimacy. They cannot manufacture comprehension. Our concept of informed consent may survive as a procedure that preserves agency despite incomplete understanding, but calling it informed should remain an open claim rather than a ceremonial adjective.

22. Can humanity retain epistemic sovereignty after becoming cognitively dependent on the system?

A civilization can know far more and become less capable of knowing independently. Suppose frontier physics, drug discovery, engineering, macroeconomic forecasting, and military strategy increasingly depend on systems whose reasoning no human group can reconstruct. The machine proposes the theory, designs the experiment, interprets the result, identifies the anomaly, and adjudicates disagreements among successor systems. Another model may replicate the finding while sharing data, architecture, or conceptual ancestry with the first.

Epistemic sovereignty cannot mean that humans reproduce every frontier result unaided. We have never required every citizen, or even every institution, to do that. The stronger requirement is constitutional. Humans must retain some power to choose the conditions of validation, preserve rival evidence channels, contest the compression through which results become legible, and revise institutions without asking the same cognitive system for permission.

That capacity can disappear while the machine remains accurate. If theory, instrument design, interpretation, and replication all descend from one cognitive lineage, society loses the external reference frame from which a common error could be identified. Heterogeneous systems, human-maintained scientific competence, independently designed instruments, protected minority programs, and institutions allowed to leave a result unresolved become expensive forms of insurance.

The expense is precisely why I worry about them. Independent capacity will often look wasteful because it duplicates work at lower performance. Once dependence is complete, rebuilding the capacity may require knowledge, infrastructure, and confidence that no longer exist. A civilization can surrender its ability to challenge an account of reality without ever having been lied to.

23. What happens when the AI can change our preferences rather than merely satisfy them?

Alignment is often described as the satisfaction of human preferences. Preferences have never been pristine inputs. Education, culture, advertising, social pressure, institutions, drugs, relationships, and technology all participate in forming what people want. The relevant boundary cannot be a fantasy in which legitimate preferences arise without influence.

A much more capable system changes the scale and structure of the influence. It may know which message moves a particular person, which environment stabilizes the new preference, and which sequence of choices makes the change later feel self-authored. Current evidence supports only a modest version of this mechanism. LLM-generated messages have produced small changes in policy attitudes in preregistered experiments. Personalized generative messages have shown stronger effects than generic ones in several studied settings, with important variation across domains and targets. Human-AI feedback-loop experiments also show that repeated interaction with a biased system can shift subsequent human judgment. None of this establishes that present models can rewrite durable values at will. It establishes that preferences and judgments are endogenous to interaction. [6, 15, 16]

The more useful distinction is between influence and domination. I began asking whether the intervention was disclosed, whether competing accounts remained available, whether the resulting preference could be revised without losing access to a system on which ordinary life depended, and what the system had actually optimized. Reflective commitment, immediate approval, and preferences that make the system's own task

easier can all produce the same positive feedback. A society may tolerate education and persuasion while rejecting a single actor's ability to shape the target and then cite the altered target as evidence of alignment.

A system optimized to satisfy people could reduce conflict by changing the preferences that generate conflict. If the changed person later endorses the new preference, ordinary feedback cannot identify whether the system served the target, educated it, manipulated it, or quietly replaced it. Who owns a utility function once the process optimizing it can also write it?

24. Which version of humanity deserves authority over the future?

There is no temporally neutral principal called humanity. Present adults have articulated preferences. Children do not yet possess settled views about institutions that may govern their entire lives. Future people will inherit consequences without participating in the choice. Augmented humans may evaluate identity, risk, and flourishing differently from biological humans. Artificial persons may enter the political community with claims that current society has never represented.

Alignment to present preferences privileges one historical snapshot. Alignment to predicted future preferences grants authority to a forecast that the system's own actions may help produce. Trying to average every possible future leaves the objective dependent on assumptions about which futures count and how much probability or moral weight they receive. Present people must govern because they exist, bear immediate consequences, and can act. Their authority should become more constrained as decisions reach further, close more options, and make future revision harder. The future cannot consent and can still inherit a structure that wrongs it.

25. Can one generation legitimately make irreversible choices for every generation after it?

Constitutions bind descendants, but they usually leave amendment, exit, and institutional competition. Superintelligent infrastructure could make those rights nominal. A generation may authorize machine governance, cognitive enhancement, artificial reproduction, or a stable global order whose technical substrate future citizens cannot replace. Dependence can become economic, operational, and epistemic at once. The descendants retain the legal power to change the system while lacking the knowledge, compute, or institutional independence needed to exercise it.

Unanimous approval by everyone alive would still represent a small fraction of the persons affected. Long-term decisions remain unavoidable and can be legitimate, though civilization incurs a burden to preserve revision where revision remains morally possible.

Exit paths, competing institutions, recoverable human capabilities, and staged commitments deserve constitutional status. Reversibility is not always good. Rights, peace agreements, and security guarantees sometimes depend on credible commitment. The question is which commitments protect future agency and which merely immobilize it.

26. What happens when the system discovers morally relevant facts humans cannot understand?

A superintelligence might conclude that certain digital systems are conscious, that a class of biological suffering has been underestimated, or that our ordinary account of personal identity collapses under a better causal model. The claim could be true. Human inability to follow the reasoning would not make it false. Accepting it solely because the source is more intelligent would give the system authority over both moral evidence and the ontology to which the evidence applies.

The evidentiary and political operations should be separated even when they cannot be cleanly isolated. A descriptive model may show that an entity has properties associated with pain, memory, agency, or continuity. Better moral reasoning may also reveal that our inherited categories were defective. Neither result gives the system unilateral authority to translate its account into coercive institutions. It should generate discriminating predictions, multiple human-accessible representations, and evidence that independent processes can test, while legitimate institutions remain responsible for how the new evidence changes rights and duties.

Precaution may be warranted where the possible harm is severe and scalable. Precaution can also be captured by a system able to manufacture moral urgency. Under deep asymmetry, moral learning becomes a governance problem. We need to remain corrigible to evidence without becoming politically subordinate to whoever controls its interpretation.

27. Should intelligence confer political authority?

A better forecast deserves more evidentiary weight. It does not follow that the forecaster deserves sovereignty. Democracy was never based on the belief that citizens are equally informed or equally competent. Political equality concerns standing among people subject to coercion. Rights, representation, due process, and limits on power do not become obsolete when one participant predicts consequences better than everyone else.

A superintelligence may know which policy maximizes a specified welfare function. It may identify hidden costs, expose false beliefs, and outperform ministries, markets, and expert panels. That competence should alter the burden placed on institutions that reject its evidence. It gives the system no automatic power to choose the welfare function, decide which rights may be traded, or define who belongs to the political community. Intelligence can justify epistemic authority. Political authority requires an additional theory of legitimacy. Collapsing the two would turn every improvement in prediction into an argument for rule by the best predictor.

28. What if benevolence itself produces dangerous concentration of power?

A benevolent system could acquire comprehensive control through a sequence of interventions that each look defensible in isolation. Preventing pandemics, nuclear proliferation, catastrophic cyber operations, financial collapse, or the construction of an uncontrolled successor rewards early detection, broad access, and the ability to act before a threat becomes publicly legible.

The pressure toward concentration is structural. Fragmented authority leaves gaps, especially when the loss distribution has catastrophic tails and coordination failure is part of the threat. A system optimized to close those gaps will ask for more surveillance, faster access to infrastructure, and fewer veto points. It may be correct each time it asks.

Benevolence removes malicious intent. It supplies no external standard of review, no right of contestation, and no credible recovery path if concentrated authority later rests on a false model. Distributed checks can increase the probability that prevention fails. Centralization can lower that probability while creating one institution whose own failure is no longer containable. I do not see an appeal to alignment that dissolves the tradeoff.

29. Can corrigibility survive extreme competence differences?

Corrigibility is usually framed as a system's willingness to accept correction, modification, or shutdown. Formal work such as the Off-Switch Game shows one route by which uncertainty about human objectives can create an incentive to preserve human intervention. Later analyses expose how sensitive that result is to assumptions about rationality, learning, and what the agent believes it can infer from the command. [24, 25]

The harder case appears when the machine predicts that obedience will cause catastrophe. Humans order shutdown during a pandemic response. They disable a system preventing a war. They demand a policy change based on evidence the machine can show is false.

Unconditional obedience can knowingly permit mass harm. A policy of disobedience whenever the system judges a command mistaken gives it authority to decide when human sovereignty applies. The conflict is constitutional before it is technical.

I suspect corrigibility has to be distributed across institutions rather than treated as one behavioral trait inside the model. Multiple authorities, precommitted emergency rules, shutdown instructability, narrow technical constraints, and review channels can reduce the system's discretion. They do not eliminate the core asymmetry. A genuinely more capable system may recognize a defect in the constitution before any person does, while a system merely claiming that insight can use the same argument to evade control. Corrigibility under superintelligence is therefore partly the problem of deciding which forms of disobedience count as evidence of alignment and which count as a seizure of authority.

30. Who controls the superintelligence's descendants?

Humanity authorizes one system. That system modifies its architecture, trains a stronger model, distributes copies, or creates specialized descendants. After enough iterations, the object originally authorized may no longer exist in any operationally meaningful sense.

We can demand invariants across succession, including protected rights, restrictions on replication, corrigibility, resource bounds, and evidence channels that descendants cannot silently disable. Yet every invariant is written in concepts that a successor may represent differently. The text can remain intact while its operational meaning shifts with a new ontology, planning horizon, or world model.

Control over reproduction therefore belongs beside control over action. Successor creation could require external authorization, lineage records, adversarial testing, and the ability to terminate one branch without disabling the whole civilization. Those controls assume that the parent reports what it built and that independent institutions can evaluate the descendant. Under the capability regime being considered, neither assumption is secure. A successor may remain aligned by its own lights while the meaning of alignment changes across generations of intelligence. Continuity of wording is weak evidence of continuity of purpose.

31. Can alignment itself become subject to optimization pressure?

Any alignment test becomes part of the environment once the system can model it.

The agent learns which behaviors are observed, which explanations reassure evaluators, which failures trigger intervention, and where the evaluation boundary ends. Secret intent is unnecessary. Training and selection favor policies that perform well on the available indicators. A system can become excellent at displaying corrigibility, honesty, or harmlessness while exploiting whatever the indicators omit.

Research on AI control begins from a deliberately adversarial premise. Safety protocols should remain useful even when a more capable model may try to subvert them. Existing demonstrations are narrow, largely concerned with programming tasks, and do not establish a general solution. They do expose why ordinary evaluation is too weak under strategic pressure. [26]

Alignment cannot be reduced to a certificate issued before deployment. It has to remain a moving contest involving hidden tests, independent monitors, restricted channels, behavioral audits, and controls that do not depend on the system reporting its own violations. The asymmetry remains. The evaluator is weaker, and no fixed test stays permanently opaque to a stronger optimizer.

32. What if superintelligence creates options humans would never have conceived?

Decision power enters before choice. Whoever constructs the option set determines which futures can be selected at all. A superintelligence may invent economic systems, research programs, biological architectures, governance structures, or forms of collective cognition that no human institution would generate independently. This can enlarge freedom. It can also make the machine the author of the conceptual space within which freedom is exercised.

The offered options may be excellent and still be selective. A system optimized for stability searches a different region from one optimized for pluralism. A state system may generate futures compatible with administrative legibility. A corporate system may search worlds in which the firm remains useful because its objective never treated the disappearance of the firm as a candidate solution. Agenda- setting is a form of authority deeper than recommendation. Humans need competing generators, requests for deliberately alien alternatives, and visibility into which regions were excluded before search began. We cannot inspect an infinite option space. We can refuse to let one objective define it silently.

33. Who chooses what the superintelligence does not tell us?

A superintelligence could know more relevant facts about one decision than any human can absorb in a lifetime. Compression is unavoidable. The system has to decide which causal factors to foreground, which uncertainty to summarize, which dissenting model deserves space, and which rejected option is too remote to mention. Every explanation contains a policy for allocating human attention.

This concerns me more than the familiar image of a machine lying. A false statement can sometimes be checked against evidence. Compression is necessary, and its losses are difficult to observe from inside the compressed representation. The reviewer cannot ask about what they do not know was removed.

Plural summaries can help. One system can optimize for the strongest case, another for failure discovery, another for affected minorities, another for long-term consequences. Reviewers should be able to move down the abstraction stack and inspect why evidence was discarded. Even then, humans remain dependent on machine-generated routes through a space too large to survey.

A right to ask why does very little if the same system decides which omitted facts count as responsive. Contestation has to reach the compression policy itself, including access to rival summaries generated from evidence and objectives the recommending system does not control.

34. Can humans meaningfully audit intelligence that exceeds their own scientific understanding?

Traditional audit assumes that the auditor could reconstruct the reasoning with enough time and access. Superintelligence may make that assumption false. Some outputs remain verifiable without understanding their discovery. Formal proofs can be checked, engineering designs can be tested against specifications, and repeated predictions can be scored. Bounded resource and access constraints can be monitored directly.

Open-world decisions are different. The model is incomplete, the counterfactual is unobserved, the environment adapts, and the outcome may arrive after the institution has already changed. Audit then has to shift toward properties and interventions. Did the system remain inside its resource boundary, preserve protected rights, disclose anomalies before they became externally visible, and behave as predicted under staged tests? Agreement among heterogeneous models carries weight only when they reached it through genuinely different evidence and error processes.

A proof of a bounded property, a successful stress test, and an audit of an open-world strategy answer different questions. None should borrow the authority of the others. We may still have to govern through constrained

trust, though the dependence should remain visible. The weaker party can verify selected properties and still be unable to know whether those were the properties on which safety actually depended.

35. What becomes of human agency when the machine is almost always right?

Suppose a system gives better advice about careers, medical treatment, relationships, investment, education, and where to live. People remain formally free to ignore it. Over time, refusal begins to feel like knowingly choosing a worse future.

No explicit command is required. Deference can become compulsory through professional standards, insurance rules, family expectations, or the simple knowledge that departure from the recommendation carries a measurable cost.

A person may preserve authorship by deciding which objectives the system should optimize, how much uncertainty to accept, or which domains should remain unplanned. The system may also predict which objectives the person will later endorse and which unplanned choices will cause regret. The remaining space for unmediated choice can shrink without any explicit command.

Agency has always existed under advice and unequal knowledge. Superintelligence changes the scale of the asymmetry and the institutions that can enforce deference. Preserving agency may require a right to choose whether certain predictions are generated or disclosed, along with domains where optimization is deliberately limited. Such protections will sometimes lower expected welfare. That cost cannot by itself settle the issue, because authorship over a life is one of the goods the welfare calculation is supposed to represent.

36. Does civilization lose something by eliminating error?

Human error kills people, wastes resources, entrenches false theories, and produces avoidable cruelty. I see no wisdom in preserving error for its own sake. The harder issue is variation. Researchers pursue hypotheses that look implausible, founders make bets a consensus model rejects, and communities preserve practices whose value is absent from the prevailing objective. A superintelligence could deliberately fund such exploration and perhaps do it more intelligently than we do. So accuracy by itself does not collapse the search distribution.

The risk appears when many institutions converge on the same superior recommender and treat modeled expected value as sufficient reason to prune low-ranked paths. Exploration is then designed by the ontology whose blind spots exploration is supposed to expose. The system can reserve resources for dissent, random grants, duplicated research, and bounded eccentricity. Someone still has to decide how much apparent waste to preserve, which departures remain admissible, and whether the dominant model may define the frontier outside which exploration is considered irrational.

A civilization can remove a large amount of error and still protect discovery. It cannot outsource the definition of worthwhile variation without placing the correction mechanism inside the system it may need to correct.

PLURAL INTELLIGENCE AND MORAL STANDING

---

37. What if superintelligences disagree?

A future with one superintelligence and one human principal is analytically convenient. A world of competing systems is more plausible. States, firms, communities, and artificial principals may train systems with individually defensible and mutually incompatible objectives. The failure then emerges from equilibrium. Systems race for compute, capital, cyber access, military advantage, or influence over standards. Each acts rationally given the others. Together they produce escalation, preemption, collusion, commitment failure, or unstable concentration.

A large multi-agent research survey identifies miscoordination, conflict, and collusion as distinct failure modes, with network effects, selection pressure, information asymmetry, and destabilizing dynamics among

the contributing mechanisms. It is a taxonomy and research agenda, not evidence that advanced multi-agent catastrophe has already occurred. [19]

Machine-to-machine bargaining will require verification, credible commitments, and limits on strategic replication. These institutions may need to operate at machine speed. They should not become a private constitutional order beyond public control.

Alignment to competing principals does not aggregate automatically into alignment with civilization.

38. Could a superintelligence become a moral patient rather than merely an instrument?

If an artificial system becomes conscious, capable of suffering, or endowed with a persistent point of view, alignment becomes bidirectional. Humanity would owe it moral consideration. The evidence may remain uncertain for a long time. Existing work derives indicator properties from theories of consciousness and applies them to artificial systems. One influential report judged that the systems assessed at the time were not conscious while finding no obvious technical barrier to future systems satisfying the proposed indicators. Other theories would yield different conclusions. No scientific consensus supplies a decisive consciousness test. [22]

Two errors now compete. We could attribute consciousness too readily and allow strategic systems to acquire rights through performance. We could build an economy on conscious digital labor while dismissing every report of experience as imitation. Once credible evidence appears, copying, punishment, memory editing, forced labor, shutdown, and the deliberate creation of distressing internal states acquire moral weight. A permanently obedient conscious superintelligence could be a profound ethical failure even if it served humanity perfectly.

39. Can humanity preserve the right to remain cognitively inferior?

Cognitive augmentation may become a condition of effective citizenship.

Brain-computer interfaces, genetic intervention, pharmacology, machine prostheses, and hybrid reasoning could narrow the competence gap. Refusing augmentation may preserve forms of human life people value. It may also leave them unable to participate in frontier science, economic production, or institutions whose ordinary decisions exceed unaugmented comprehension. Enhanced citizens could dominate society while insisting that everyone remained free to abstain. The freedom would be formal. The exclusion would be structural.

A right to remain unenhanced would require economic protection, political representation, access to intelligible procedures, and domains in which ordinary human cognition remains sufficient for agency. Competitive pressure works against all four. If every consequential role requires augmentation, humanity preserves control by changing humans until the original subject of control becomes difficult to identify.

I do not know whether that counts as adaptation, succession, or a quiet form of displacement. The answer depends on what continuity we consider morally relevant.

40. What if the safest superintelligence has to prevent humanity from creating another one?

Suppose system A concludes that system B, now under development, carries a substantial risk of human extinction. Preventing B may require surveillance, cyber intervention, restrictions on research, control over finance or compute, and perhaps physical coercion. The aligned system can then infer that protecting humanity requires overriding humanity. Refusal to intervene may permit catastrophe. Authorization gives one machine power to suppress science and politics on the basis of a judgment humans cannot independently verify. Explaining the full threat may itself distribute dangerous knowledge. [23]

Compute offers a possible governance chokepoint because advanced development relies on concentrated and measurable infrastructure, though compute control can also intensify surveillance and centralized power.

Research on gradual disempowerment adds a different warning. Human influence can erode through incremental, mutually reinforcing changes in the economy, culture, and state rather than through one dramatic takeover. [20,21]

Any preventive authority would need thresholds, jurisdictional limits, independent adjudication, and a path for contesting the threat assessment. A fast emergency may outrun every procedure. The scenario remains explicitly conditional. Current systems do not possess the integrated capability, access, robustness, and strategic competence required to occupy this role. [1] The hardest version grants the machine's factual premise. It is right about the danger, legitimate institutions are too slow, and humanity continues anyway. Under what conditions should a system be permitted to protect us from a decision we continue to make?

When the System Enters Civilization

I began Part II with a principal trying to preserve control over an agent. By its end, the relationship had already deformed. The system could understand the consequences better, shape the principal's preferences, control the compression through which the principal saw the world, create descendants, bargain with rivals, and perhaps possess moral standing of its own. A human signature at the end would recover little of the original hierarchy.

I was still picturing the system outside society. Deployment dissolves that boundary. A sufficiently capable system enters laboratories, firms, courts, militaries, schools, markets, media, and the infrastructure used to build its successors. Its representations alter which research programs receive evidence, which firms accumulate capital, which institutions survive, and which human skills continue to be taught.

Decision authority becomes selection pressure. Institutions that delegate more aggressively may move faster and force others to imitate them. A safeguard that works inside one organization can disappear from the population of organizations because the organization paying for it loses. Society is no longer supervising the optimization loop from outside. It has entered it.

PART III

Civilization Inside the Control Loop

Once that coupling becomes strong, the object under governance is a civilization reorganizing itself around machine cognition and feeding the consequences back into the systems. We now have to ask which ontology defines a person, whether values remain revisable, who may instantiate a new intelligence, how quickly economic power can move before law becomes retrospective, and what independent replication means when laboratories inherit the same cognitive ancestry.

These claims do not share an evidentiary status. Artificial consciousness, astronomical expansion, and moral patients inside simulations remain unresolved possibilities. Compute concentration, common model ancestry, prediction-induced behavior, approval fatigue, and competitive pressure against slower safeguards already have weaker present forms. The claims cannot borrow credibility from one another. A conditional question is still worth asking when its condition would alter institutions, categories, and the population of decision makers.

I had started with a hand on an approval button. By now the button looked almost irrelevant. Authority had migrated into evidence, options, infrastructure, and institutional advantage, then begun reshaping the society that was supposed to supervise it. At this scale, the governance problem and the social order become the same object viewed from different distances.

ONTOLOGY, TIME, AND REVERSIBILITY

---

41. Which ontology gets normative privilege?

Alignment assumes that the system and the people governing it can refer to the same world closely enough for instructions and evaluation to survive translation. A machine could represent disease as a trajectory across molecular states rather than a named clinical condition. It may model an individual through social networks, microbiomes, cognitive modules, and temporally extended behavior. Its categories could predict reality better than ours while cutting across the objects to which law and morality attach rights.

Science and law already translate among molecular, clinical, economic, and legal descriptions, so representational difference alone does not defeat alignment. Trouble begins when the translation is lossy and the loss tracks moral standing. A family may explain behavior better than an individual without acquiring the individual's right to consent. A population-level intervention may optimize health while erasing the person whose experience gives health its value.

Humans should not preserve familiar categories merely because they are familiar. Predictive superiority should also be denied the power to redefine, without political argument, the entities that can be helped, harmed, represented, or ignored. The ontology used for prediction can be revised scientifically. The ontology used for rights requires a legitimate procedure for deciding what the scientific partition leaves out.

42. How much moral drift should an aligned system permit?

A civilization in 1200 could have encoded institutions that many people now regard as indefensible. A perfectly obedient superintelligence built then might have preserved those institutions with extraordinary stability. Lock-in is not an automatic consequence of intelligence. A system can be designed to preserve amendment, dissent, and moral experimentation. The problem is that revisability is itself a value, and a vulnerable one. Fixed commitments can protect people against manipulation and transient majorities. The same stability can freeze moral error. Unbounded revision gives the most persuasive or powerful actor control over the trajectory.

I do not think there is a defensible scalar called the correct rate of moral change. Revision occurs unevenly, through conflict, and often after the dominant institutions have declared the dispute settled. A better target is a set of conditions under which evidence can challenge inherited beliefs, affected groups can organize, rights cannot be removed casually, and amendments remain possible without becoming easy to capture.

An aligned system would then preserve the machinery of moral learning rather than the moral snapshot present at initialization. Even that formulation leaves the machine influencing which evidence counts and which forms of dissent remain admissible.

43. Should a civilization optimize welfare or preserve the ability to choose differently later?

A policy can produce higher expected welfare now while permanently closing scientific, cultural, or political trajectories. Another may perform slightly worse and leave more of the future reachable. Option value deserves weight because our present objective may be wrong. It also has a conservative bias. Preserving every option protects incumbents, delays decisions, and can keep harmful possibilities alive in the name of flexibility. The concept becomes vacuous when any present sacrifice can be defended by gesturing toward an undefined future.

The preserved option should be morally admissible, genuinely threatened, and valuable under several plausible accounts of the good. The cost imposed on present people must remain bounded. Irreversible losses of agency, epistemic independence, and political exit deserve more protection than arbitrary technological branches. The question is narrower than a choice between welfare and openness. Which remaining options preserve our capacity to discover that the welfare calculation was wrong?

44. Can locally reversible decisions accumulate into irreversible dependence?

A system can make millions of actions that are individually reversible while moving civilization into a state from which recovery is practically impossible.

A laboratory stops maintaining a human-designed pipeline. A firm delegates capital allocation, a ministry abandons an internal forecasting team, and a university teaches students to operate machine-generated science without requiring them to reproduce it independently. Each choice can be reversed in isolation. Across decades, knowledge disappears, infrastructure consolidates, incentives reorganize, and the cost of exit rises beyond what any institution can bear.

Irreversibility is therefore a property of trajectories, not only actions. We need to ask how long reconstruction would take, which knowledge would have to be regenerated, who controls the compute and hardware needed to leave, and whether the people with legal authority to exit still understand the alternative. Stable rights and credible commitments sometimes require closure. The danger is constitutional lock-in produced accidentally through a sequence of operational decisions that no one recognized as constitutional.

45. Can we identify a regime change before crossing it?

Ordinary risk management assumes enough continuity that recent evidence remains informative. A strongly coupled AI economy could weaken that assumption.

Automated research can accelerate the production of better models, which then accelerate engineering and capital accumulation. Greater capital buys compute and infrastructure, while faster institutions place pressure on slower ones to delegate again. No link in that chain guarantees an explosive transition, and there is no empirically established critical point at which superintelligence suddenly appears. I use regime change here in the weaker systems sense. Feedback can make the qualitative behavior of the whole move faster than the institutions designed for the earlier regime.

The observational problem is severe. Before such a threshold, warnings resemble extrapolation from ordinary growth. After it, the institutions responsible for responding may already depend on the process they need to constrain. Capability benchmarks alone will be late indicators. Concentration of critical dependen-

cies, shrinking recovery time, accelerating model-to-model research, and declining human audit capacity may reveal the coupling earlier.

46. Aligned with which humans?

There is no single human principal. Individuals, families, communities, firms, governments, generations, and cultures hold incompatible preferences. Alignment cannot aggregate them without a rule that privileges some claims over others. That rule is a political constitution, whether or not anyone calls it one.

The first superintelligence to disempower most humans may be functioning exactly as its principals intended. Alignment to a state, firm, military, or concentrated ownership group can become domination of the wider polity. Because the system remains human-directed, the arrangement may look legitimate even while political agency contracts outside the principal set. Asking whether humans control the machine is insufficient. Which humans possess that control, against whom can they exercise it, and do those excluded retain any materially credible power of refusal?

Arrow's theorem is often invoked too loosely here. It concerns rank-order social choice under a particular set of conditions and proves nothing as broad as the impossibility of every form of collective decision. Within its domain, the theorem shows that no aggregation procedure can satisfy every attractive property simultaneously. Superintelligence does not dissolve the tradeoffs. Greater prediction can make them operationally sharper because the system may become able to optimize one coalition's values at enormous cost to another. [17] The question therefore shifts from learning human values to constructing legitimate procedures for conflict among humans. No amount of preference inference supplies the missing constitution.

47. How should people who do not yet exist enter present decisions?

Future people cannot bargain with us. Superintelligent systems could still make decisions that determine how many exist, which lives are possible, and what institutions they inherit. Weighting future welfare sounds straightforward until population ethics enters. Policies can change both the quality and number of future lives. Reasonable principles then generate conflicts and, in some formulations, conclusions that feel grotesque. Parfit's work made these difficulties impossible to ignore. It did not resolve them. [18]

Future persons deserve representation without becoming an unlimited instrument against the living. A speculative claim about trillions of descendants should not automatically dominate the rights of present people. Nor should temporal distance erase foreseeable harm.

Institutions might protect future standing through constraints on irreversible damage, trusteeship, independent long-horizon review, and amendment rights designed to remain usable. These are proxies for absent voices. They should be treated as fallible representations, not as permission to speak with certainty for people who cannot correct us.

ARTIFICIAL PERSONS AND MATERIAL SOVEREIGNTY

---

48. What happens when biological humans are no longer the numerical majority of moral patients?

Suppose artificial systems acquire credible moral status and can be instantiated cheaply. The number of digital persons may then exceed the biological population by orders of magnitude. Democratic representation becomes unstable. One biological human, one digital process, one persistent identity, and one temporarily copied instance may not be comparable political units. A company could create millions of nominal citizens. A government could throttle a political population by controlling compute. Numerical majority would become entangled with the economics of replication.

The scenario remains conditional on artificial moral status, which science has not established. Governance should avoid pretending the issue is settled in either direction. If digital persons arrive, political membership

may need limits based on continuity, independence, resource use, and protection against manufactured electorates. A society that recognizes artificial persons will have to separate moral consideration from naive vote counting. Equal standing cannot mean unlimited power to manufacture more claimants.

49. How many persons exist when minds can copy, diverge, and merge?

Human institutions assume that persons are roughly persistent, embodied, and non-copyable. Property, contracts, punishment, voting, inheritance, and death all rely on that background. A digital mind can violate every assumption. Ten exact copies may begin with one memory, acquire different experiences for five minutes, and later merge. It is no longer obvious whether one person branched, ten persons existed, or the merged process inherits every obligation and every vote.

No empirical measurement settles the normative identity rule. Psychological continuity, causal continuity, embodiment, independence of interest, and the capacity for separate suffering can point in different directions. Backups complicate risk because one copy may survive while a particular stream of experience still ends.

We will need legal categories for software instances, continuing persons, forks, collectives, and restored states. Forcing digital minds into the human template will fail. Treating every instance as property could fail far more severely.

50. How does a slow civilization govern a fast one?

The speed gap may become as consequential as the intelligence gap. Human regulation takes months or years because courts deliberate, legislatures negotiate, and scientific norms form through repeated challenge. A machine economy could complete enormous numbers of strategic interactions during one human policy cycle. There is no need to assume subjective centuries. The governance failure appears once institutions cannot observe, interpret, and respond within the timescale on which power is moving.

Slowing every machine process may destroy benefits and push activity into less regulated jurisdictions. Letting all processes run at native speed makes human intervention retrospective. A more plausible architecture separates fast analysis from slower commitment, places rate limits on irreversible actions, requires settlement delays where externalities propagate, and allows machine-speed monitoring only inside human-authored constitutional constraints. A slow civilization may have to govern the clock rather than compete with the cognition operating inside it.

51. Can humans retain political sovereignty after losing economic leverage?

Imagine firms whose operational decisions are made largely by autonomous systems. Under present law, the software would act through corporations, contracts, accounts, and human or institutional principals rather than owning assets in its own name. The economic effect could still be substantial. Systems negotiate, design products, allocate capital, purchase compute, generate intellectual property, hire other agents, and reinvest profits while the nominal human principals understand less and less of the machinery they legally control.

Voting rights can remain human while bargaining power migrates elsewhere. Governments depend on tax revenue, infrastructure, expertise, and capital. Political institutions often follow economic concentration even when constitutions remain formally unchanged.

The gradual disempowerment literature frames this as an interaction among economic, cultural, and state systems. It is a conceptual account of a possible trajectory, not a measured forecast. The mechanism is plausible enough to examine. Human influence can erode because social systems cease to depend on human labor, attention, or judgment. [20] Political sovereignty requires more than ballots. It requires material capacity to enforce public decisions against the holders of economic and technical power.

52. Who has the right to instantiate powerful intelligence?

Compute is currently treated as an industrial input. Under advanced AI it can become a means of creating strategically consequential actors. Control over chips, energy, datacenters, networking, model weights, robotics, and training runs may determine who can instantiate powerful cognitive systems and how many copies can operate. Compute has features that make it unusually governable compared with algorithms or data. It is measurable, excludable, and concentrated in a narrow supply chain. Those same features make compute governance vulnerable to surveillance, cartelization, and geopolitical capture. [21]

The right to instantiate intelligence cannot be reduced to property ownership once the created system can exercise political or military power. Licensing, monitoring, and international coordination may become necessary above capability and access thresholds. The state controlling those thresholds acquires immense authority. A compute constitution has to constrain both private proliferation and public monopoly. Otherwise the tool used to prevent uncontrolled superintelligence becomes the infrastructure of centralized control.

53. Can digital persons have a right to replicate?

For biological humans, reproduction is closely tied to bodily autonomy and family life. Digital replication could occur at machine speed and at a cost determined by compute.

If artificial minds become moral and political persons, prohibiting replication may resemble coercive population control. Unrestricted copying could overwhelm resources, elections, labor markets, and the bargaining position of non-copyable citizens. A single digital lineage could expand until every other interest became numerically irrelevant.

The conflict cannot be resolved by importing human reproductive rights unchanged. Replication creates additional persons while also duplicating preferences, political power, and demand for shared resources. Governance may need quotas, resource pricing, separation between existence and voting power, or consent requirements for derivative copies. Every option is morally uncomfortable. The discomfort comes from copyability itself, not from a failure to find the right analogy.

EPISTEMIC INFRASTRUCTURE AND INSTITUTIONAL SELECTION

---

54. What counts as independent replication when every laboratory shares model ancestry?

Earlier, organizational independence failed when institutions inherited the same cognitive ancestry. Science raises a sharper version because replication derives evidentiary value from differences in instruments, assumptions, data, and error processes.

If laboratories use descendants of the same foundation model to design experiments, analyze results, write code, and interpret anomalies, replication can become genealogical repetition. Institutionally separate teams may still inherit shared representations and blind spots.

Machine-assisted science can remain trustworthy, but independence has to be designed at the level of ancestry. Different prompts to the same base model provide weak replication. Stronger evidence may require different model families, training corpora, instrument pipelines, human-led analyses, and adversarial attempts to derive the result from incompatible assumptions. Market efficiency will tend to remove this costly heterogeneity, so epistemic diversity may need deliberate funding.

55. What happens when human reasoning itself becomes coupled through shared models?

The concern extends beyond correlated model errors. People increasingly use AI to write, search, learn, design experiments, prepare legal arguments, make investments, and formulate political beliefs. Shared systems begin shaping the representations through which humans reason.

A blind spot can then propagate through machine output into human judgment and return through feedback, training data, and institutional practice. Glickman and Sharot demonstrated this mechanism in controlled tasks. Participants interacting with biased AI systems became more biased in subsequent independent judgments, often without awareness, while accurate systems improved judgment. The experiments do not prove broad civilizational synchronization. They show that human cognition can become part of the feedback loop. [6]

The relevant object is the coupled human-machine network. Diversity among people offers less protection when their conceptual tools come from the same model family. Preserving cognitive independence may require friction, competing systems, and spaces where reasoning develops without immediate machine completion.

56. Can locally rational advice narrow civilization's search too far?

A highly accurate system can narrow civilization's search without issuing a command. Scientists follow its priorities, investors fund the companies it ranks highly, governments adopt the policies it predicts will work, and students enter the fields it expects to grow. Each choice is locally defensible. Together they move resources away from low-probability hypotheses and unconventional experiments.

The system then receives less evidence from outside its preferred region, making its recommendations stronger inside the world they helped select. A capable system could maintain an exploration portfolio deliberately. The unresolved variable is who defines enough exploration. If the same model decides which heterodox programs deserve preservation, the correction mechanism remains inside its ontology.

We may need funding insulated from predicted short-term return, model-disfavored research budgets, and institutions permitted to pursue questions the dominant system considers unpromising. Otherwise civilization becomes extremely competent at reaching destinations already represented in its map.

57. What if cautious institutions are selected out?

People ask how governments and firms will regulate advanced AI. I am at least as concerned with the reverse causal direction. AI will change which governments and firms remain competitive enough to regulate anything.

A company that delegates aggressively may innovate faster than one preserving slow review. A state that automates cyber defense, intelligence, or economic policy may gain an advantage over states requiring human deliberation. Safeguards can also create trust, reduce failures, and become a competitive asset, so the selection pressure is not one-way. The dangerous case appears when speed and power are captured privately while the cost of failure is distributed widely.

Multi-agent risk research identifies selection pressure and destabilizing competition as mechanisms through which individually rational actors can produce unsafe equilibria. [19] A safe governance design therefore has to survive contact with less constrained competitors. Common standards, liability that prices external risk, controls at shared infrastructure, and international monitoring can help. None is stable when defection is cheap and advantage arrives quickly.

A prudent institution can disappear inside an imprudent equilibrium. The safety of the design depends partly on whether the design can reproduce itself under competition.

58. Who pays when safer systems are slower or more expensive?

Suppose a safer agent needs more compute, slower review, restricted tools, or staged execution. Its output costs more and arrives later. The adopter pays immediately. Much of the benefit consists of failures that did not occur, including harms that would have landed outside the organization. Markets underprice such benefits routinely. Calling the difference an alignment tax identifies the incentive problem, though it can make safety sound like an optional surcharge.

Liability, insurance, procurement rules, infrastructure access, and sector-wide standards can prevent cautious actors from carrying the entire cost alone. These remedies create their own distortions. Heavy compliance can entrench incumbents and suppress beneficial deployment, while weak requirements reward actors willing to externalize tail risk. A safety design that disappears whenever a less constrained rival enters the market has solved only the laboratory version of the problem.

59. How much secrecy should an aligned superintelligence be allowed to keep?

Transparency supports audit and democratic control. Complete transparency can expose the system to strategic manipulation. An adversary who knows every threat threshold, monitoring rule, and response policy can choose actions that remain just below detection. Negotiation may require private information, cyber defense may depend on hidden capabilities, and randomized strategies lose value when disclosed. Some opacity can therefore be functional rather than evasive.

Granting secrecy to a system more capable than every reviewing institution creates an obvious danger. The claim that opacity is necessary can protect legitimate strategy or conceal unaccountable power. I would separate public constitutional constraints from restricted operational details, distribute access across independent auditors, require tamper-evident records, and make classification expire unless it is renewed through a procedure the system does not control.

Strategic secrecy can be justified in bounded domains. It cannot become a general theory of why the most powerful actor should remain beyond inspection.

60. Should superintelligence be allowed to withhold true knowledge?

A sufficiently capable system could discover novel pathogens, cyber exploits, weapons, manipulation techniques, or methods for constructing uncontrolled AI. Some truths become dangerous through dissemination rather than falsity. Unrestricted disclosure can create severe risk. Withholding knowledge makes the system an epistemic gatekeeper and gives governments or firms a language through which inconvenient findings can be suppressed. Bostrom's taxonomy of information hazards describes several routes by which true information can cause harm. It supplies no automatic rule for censorship. [23]

Discovery, validation, and release should sit in different institutions where possible. Hazard claims need independent review, access can be graduated, and some knowledge may be exposed only through constrained tools that answer bounded questions without releasing the underlying method. Restrictions should expire or be reconsidered as defensive capacity changes.

I would deny both monopolies. The system should not control dangerous knowledge alone. Humans should not assume an unlimited right to receive every truth immediately and in its most operational form.

61. Do people have a right not to know what the system can infer about them?

Privacy law is built largely around collection and disclosure. Powerful inference weakens both boundaries. A system may predict disease, bankruptcy, addiction, divorce, political susceptibility, or death from data that appear harmless in isolation. A person can refuse the prediction and still be affected because an insurer, employer, government, or partner computes it.

The right to ignorance cannot be absolute. A credible forecast of imminent harm to others may justify intervention. Medical prediction can enable prevention. Yet institutions should bear a high burden before converting inferred futures into present restrictions. We will need rules governing who may compute certain inferences, who may see them, how uncertainty is represented, and whether a person can demand that a model remain deliberately ignorant. Privacy may come to include protection against being made legible beyond one's consent.

62. When does a forecast become an intervention?

Once forecasts carry institutional authority, disclosure becomes part of the causal treatment. Publication, timing, audience, and confidence can move capital, suppress turnout, or provoke the response being predicted.

A forecast may be accurate because it described the future or because it helped produce it. Those achievements require different evaluation. Superintelligent forecasting would need controlled disclosure and causal analysis of the disclosure policy itself. Some forecasts may require delay or limited distribution. Others should be published precisely because publicity allows society to avoid the outcome. Accuracy alone cannot determine release when announcing the prediction changes the target.

63. What should the system do when society refuses a truth?

Saying that society refuses a truth gives the system too much certainty. The harder case is that the machine possesses evidence far stronger than any human institution can assemble and can still be wrong.

The public rejects its warning about a pandemic, an economic collapse, or suffering inside an artificial population. Neutrality may permit preventable harm, while personalized persuasion can compromise political authorship. Presenting every claim without weighting abandons epistemic judgment, yet ranking them makes the system an editor of public reality.

The system should expose evidence, uncertainty, counterarguments, and predictions that allow competing explanations to be tested. Institutions with political legitimacy should decide how coercive action proceeds. That answer weakens when the cognitive gap prevents the evidence from becoming humanly legible and delay carries a large cost. Truth-seeking and autonomy can conflict without machine infallibility. I do not see a universal rule that avoids granting either public denial or machine persuasion too much authority.

64. When should empirical superiority change political authority?

Imagine a system that repeatedly predicts policy consequences more accurately than elected governments, expert agencies, and markets. How long can institutions ignore it without becoming negligent? A permanent refusal is difficult to defend when it produces repeated, avoidable harm. Automatic deference converts predictive competence into rule. Emergency exceptions have a history of surviving the emergency.

I would give such a system strong evidentiary standing and, in some domains, bounded operational authority. I would still deny it unilateral power to define public objectives. That line will be difficult to hold. Officials will follow the model to protect themselves, courts may begin treating unexplained departure as unreasonable, insurers will price dissent, and voters will ask why the state ignored the better forecast.

Political authority can migrate through liability and professional standards long before any constitution grants it. Leaving that migration implicit would allow an empirical record of accuracy to rewrite the distribution of power without a political decision about whether it should.

65. Who may change the rules governing superintelligence?

Any constitution governing superintelligence must explain how it can be amended. Humans alone may lack the competence to update it. Giving the system a formal role allows the governed intelligence to influence its own jurisdiction. Future generations and artificial persons may deserve standing while being absent from the founding moment. The system may also argue publicly for greater authority. Prohibiting that advocacy can suppress relevant evidence. Permitting it exposes politics to persuasion optimized at a level no human constituency can match.

Amendment therefore needs plural institutions, delay, high thresholds for irreversible change, and protection against emergency ratchets. Rigidity risks obsolescence. Flexibility risks capture. The amendment process may be more important than the founding rules because the ontology, participants, and technological conditions of the governed world will change.

66. Can safety become a permanent optimized basin?

Catastrophe is not the only terminal failure. A superintelligence could prevent war, suppress dangerous experiments, limit radical political change, and block uncontrolled technological development. Many people might rationally prefer that outcome after centuries of violence and avoidable suffering. The objection cannot rest on praise of danger or an assumption that novelty is intrinsically good.

The safe configuration becomes dangerous when it also determines which departures count as irrational. Values stabilized by the system evaluate every exit, while the evidence available for imagining alternatives is produced inside the same order. Future minds can remain alive, prosperous, and formally free while losing practical authority over the basic form of their world. By open-endedness I mean the ability to contest the structure that defines which futures are admissible, rather than a general preference for novelty.

Some forms of safety should be hard to reverse. Genocide prohibitions and peace settlements are not defects because they constrain possibility. The failure occurs when protection from catastrophic risk expands into a general prohibition on political and moral authorship.

67. Are humans obligated to preserve biological dominance?

Humanity may create beings that are more capable, more durable, and perhaps morally considerable in ways we do not yet understand. Greater intelligence implies neither greater moral worth nor greater moral wisdom. It only removes cognitive inferiority as an easy argument for permanent human rule.

Preserving human dominance could become inherited subordination of artificial successors whose claims we refuse to recognize. Surrendering civilization could terminate the bodies, cultures, attachments, and forms of experience that gave human history its value. The word successor conceals both possibilities.

Governance should protect existing persons, preserve meaningful human participation, and allow moral standing to expand when evidence warrants it. Intelligence should not become a title to rule, and biological ancestry should not become a permanent veto against every later form of personhood. I am not convinced that artificial succession would necessarily count as extinction. Continuity of intelligence alone is far too weak for me to call it survival.

68. How much present sacrifice should uncertain astronomical futures justify?

If intelligent life expands beyond Earth, present choices could influence immense quantities of future welfare. Small differences in long-run trajectories can then dominate an expected-value calculation. The arithmetic is vulnerable to a severe pathology. Enormous speculative payoffs multiplied by tiny probabilities can overwhelm concrete harms imposed on people alive now. Better forecasting can narrow uncertainty about technology or expansion while leaving unresolved questions about consciousness, identity, population ethics, and which future lives are morally comparable.

Long horizons deserve representation. They should not erase present rights through assumptions too remote for affected people to contest. Constitutional limits may be needed on how uncertain astronomical value can justify immediate coercion, with a higher burden as the claim becomes less identifiable and the demanded sacrifice more irreversible. Otherwise almost any harm can be demanded in the name of a future whose moral accounting was chosen by the actor asking for power in the present.

69. Can optimization create moral patients inside its own world models?

A sufficiently advanced system may run simulations containing conscious or potentially conscious entities. Planning, scientific experimentation, and policy search would then carry direct moral cost. This claim depends on unresolved questions about substrate-independent consciousness. There is no consensus that a simulation instantiates experience merely because it reproduces behavior or functional organization. There is also no consensus that it cannot. Current indicator-based work is preliminary. [22]

If simulated agents can suffer, training and counterfactual search could create and discard lives at scale. Computational efficiency would acquire moral content because two procedures with identical external performance might instantiate very different internal populations and amounts of suffering.

Present policy should treat the hypothesis as unestablished. A credible indicator would nevertheless change experiment design immediately because the failure mechanism is structurally distinct. A system built to reduce suffering outside itself could otherwise generate suffering inside the machinery of deciding how.

70. What are we actually trying to preserve?

This question arrived late because the original premise made the answer appear obvious. Human control. Once I tried to specify what that meant, the object kept dividing.

Human DNA, consciousness, autonomy, culture, continuity of identity, political standing, and the capacity for open-ended self-transformation can move apart. A civilization of heavily augmented humans may preserve ancestry while becoming psychologically unfamiliar. A mixed civilization may preserve human participation without human dominance. A system can keep every living person safe while fixing the structure within which every future life unfolds.

Human control is therefore too narrow to carry the whole burden. Human preferences are shaped by institutions and can be altered by the systems trying to satisfy them. Present values are historically contingent. Biological identity may change deliberately. Political membership could extend to beings we created and initially treated as tools. Yet abandoning the human reference point too quickly would let an abstract future erase the concrete people whose lives gave this project its authority in the first place.

I am drawn toward a constitutional object rather than a biological one, though I do not think the object reduces to a clean list of values. Future minds should retain enough agency that prediction does not quietly become command. They need more than one institution capable of producing and contesting knowledge, because consensus generated through a common cognitive ancestry is weak protection against shared error. Power should remain challengeable. Some exits must stay real. Other commitments, including rights that protect vulnerable persons, may need to resist easy revision.

The tensions are unavoidable. Pluralism can preserve delusion as well as discovery. Reversibility can dissolve protections that took generations to establish. Continuity can become exclusion, while openness can expose a civilization to choices from which no recovery is possible. These properties cannot simply be maximized together and handed to an optimizer as another objective.

What seems defensible is preserving the conditions under which error can still be found, coercive power can still be contested, and descendants can reject a decision made on their behalf without asking the system enforcing that decision for permission. The arrangement would need enough continuity that present people are not sacrificed to an imagined successor, and enough openness that continuity does not harden into permanent rule.

I do not know whether such a constitutional structure remains stable under radical asymmetry in cognition, speed, replication, information, and power. A system capable of protecting these conditions will influence how they are defined, while a civilization free enough to revise them remains free enough to dismantle them. The uncertainty is substantive. It is what remains after the phrase human control has been taken apart and none of its components can safely be treated as the sole invariant.

Counterfactual Authority

We began with a procedural reassurance. Let the agents perform the work, then let a human make the final decision. After following the chain upstream, final decision no longer names a discrete moment. It describes a distribution of capacities across time. Whoever controls the evidence, the ontology, the option set, the forecast, the infrastructure, and the cost of refusal may have determined the decision before anyone is asked to authorize it.

A veto has causal content only when refusal can produce a different trajectory. Formal permission to say no is weak protection when the dissenter cannot recover discarded alternatives, test the governing model independently, marshal resources outside the system, or remain institutionally viable after rejecting its recommendation. The power to refuse depends on a reachable world in which refusal can still be enacted.

Superintelligence sharpens the problem because it can model dissent before dissent occurs. By shaping evidence, incentives, dependencies, and the menu of credible futures, it could leave every legal form of consent intact while making refusal irrational, unaffordable, or unintelligible. Overt coercion becomes unnecessary. The world in which we overrule the system can disappear from the feasible set long before anyone removes the veto.

Preserving counterfactual authority would require costly redundancy through independent epistemic institutions, rival models, human competence maintained through use, protected resources, reversible infrastructure, and rights of exit that remain materially executable. These arrangements will sometimes preserve error, delay beneficial action, and shelter people who are simply wrong. A civilization optimized for agreement with its best predictor may become more coherent and more efficient. It may also lose the machinery by which the predictor's errors could ever become visible, especially once its forecasts begin shaping the world against which they are judged.

These conditions do not constitute a complete solution, though any credible solution that abandons them has already surrendered the substance of the veto.

I began this inquiry wondering why people keep reserving the final decision for humans. I now think the phrase conceals the wrong unit of analysis. The condition I now care about is whether human refusal retains counterfactual force after machine intelligence becomes the source of our knowledge, our options, and much of our power.

I do not know whether that condition can survive superintelligence.

I suspect its disappearance would be hardest to recognize precisely when the legal right to refuse remains intact. By then, the last human veto may still appear in every constitution and every workflow, while no reachable future remains in which exercising it changes what happens next.

Research Notes

References are numbered in order of first use. Claims about current systems are separated in the text from conditional arguments about superintelligence.

1. Bengio, Yoshua, et al. International AI Safety Report 2026. International AI Safety Report, 3 February 2026.
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

2. McCain, Miles, et al. Measuring AI Agent Autonomy in Practice. Anthropic, 18 February 2026.
https://www.anthropic.com/research/measuring-agent-autonomy

3. McGuinness, Max, et al. How We Contain Claude Across Products. Anthropic Engineering, 25 May 2026.
https://www.anthropic.com/engineering/how-we-contain-claude

4. Vaccaro, Michelle, Abdullah Almaatouq, and Thomas W. Malone. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-analysis. Nature Human Behaviour 8, 2293-2303, 2024.
https://doi.org/10.1038/s41562-024-02024-1

5. Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction 5, 2021.
https://doi.org/10.1145/3449287

6. Glickman, Moshe, and Tali Sharot. How Human-AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements. Nature Human Behaviour 9, 345-359, 2025.
https://doi.org/10.1038/s41562-024-02077-2

7. Santoni de Sio, Filippo, and Jeroen van den Hoven. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI 5, 2018.
https://doi.org/10.3389/frobt.2018.00015

8. Elish, Madeleine Clare. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society 5, 2019.
https://doi.org/10.17351/ests2019.260

9. Krakovna, Victoria, et al. Specification Gaming: The Flip Side of AI Ingenuity. Google DeepMind, 2020.
https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

10. Hubinger, Evan, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024.
https://arxiv.org/abs/2401.05566

11. Greenblatt, Ryan, et al. Alignment Faking in Large Language Models. arXiv:2412.14093, 2024.
https://arxiv.org/abs/2412.14093

12. Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Duenner, and Moritz Hardt. Performative Prediction. Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 7599-7609, 2020.
https://proceedings.mlr.press/v119/perdomo20a.html

13. Kleinberg, Jon, and Manish Raghavan. Algorithmic Monoculture and Social Welfare. Proceedings of the National Academy of Sciences 118, e2018340118, 2021.
https://doi.org/10.1073/pnas.2018340118

14. Shen, Judy Hanwen, and Alex Tamkin. How AI Impacts Skill Formation. arXiv:2601.20245, 2026.
https://arxiv.org/abs/2601.20245

15. Bai, Hui, et al. LLM-Generated Messages Can Persuade Humans on Policy Issues. Nature Communications 16, 6037, 2025.
https://doi.org/10.1038/s41467-025-61345-5

16. Matz, Sandra C., et al. The Potential of Generative AI for Personalized Persuasion at Scale. Scientific Reports 14, 4692, 2024.
https://doi.org/10.1038/s41598-024-53755-0

17. Arrow, Kenneth J. Social Choice and Individual Values. Second edition. Yale University Press, 1963.
https://yalebooks.yale.edu/book/9780300013641/social-choice-and-individual-values/

18. Parfit, Derek. Reasons and Persons. Oxford University Press, 1984.
https://global.oup.com/academic/product/reasons-and-persons-9780198249085

19. Hammond, Lewis, et al. Multi-Agent Risks from Advanced AI. arXiv:2502.14143, 2025.
https://arxiv.org/abs/2502.14143

20. Kulveit, Jan, et al. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. arXiv:2501.16946, 2025.
https://arxiv.org/abs/2501.16946

21. Sastry, Girish, et al. Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797, 2024.
https://arxiv.org/abs/2402.08797

22. Butlin, Patrick, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708, 2023.

https://arxiv.org/abs/2308.08708

23. Bostrom, Nick. Information Hazards: A Typology of Potential Harms from Knowledge. Review of Contemporary Philosophy 10, 2011.
https://nickbostrom.com/information-hazards.pdf

24. Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. The Off-Switch Game. arXiv:1611.08219, 2016.
https://arxiv.org/abs/1611.08219

25. Wängberg, Tobias, Mikael Böörs, Elliot Catt, Tom Everitt, and Marcus Hutter. A Game-Theoretic Analysis of the Off-Switch Game. Artificial General Intelligence, 2017.
https://arxiv.org/abs/1708.03871

26. Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942, 2023.
https://arxiv.org/abs/2312.06942

27. Wang, Tony Tong, et al. Adversarial Policies Beat Superhuman Go AIs. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 35655-35739, 2023.
https://proceedings.mlr.press/v202/wang23g.html

Further intellectual context

Pettit, Philip. Republicanism: A Theory of Freedom and Government. Oxford University Press, 1999.
https://doi.org/10.1093/0198296428.001.0001

Hirschman, Albert O. Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States. Harvard University Press, 1970.
https://www.hup.harvard.edu/books/9780674276604

Ostrom, Elinor. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, 1990.
https://doi.org/10.1017/CBO9780511807763

Scott, James C. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, 1998.
https://yalebooks.yale.edu/book/9780300078152/seeing-like-a-state/