The Custody of Intelligence: Power, Sovereignty, and Human Agency Beyond the Open–Closed Divide

Mark Sendo

The Custody of Intelligence

Power, Sovereignty, and Human Agency Beyond the Open–Closed Divide

Mark Sendo

Orchestrator Institute Research Labs Founding Research Monograph — Version 1.0 2026

Author: Mark Sendo, Master Orchestrator · Founder, Orchestrator Institute Research Labs (ORCID: pending)

Status: Working draft — not peer-reviewed. Founding Research Monograph, Version 1.0.

How to cite. Sendo, Mark. *The Custody of Intelligence: Power, Sovereignty, and Human Agency Beyond the Open–Closed Divide*. Orchestrator Institute Research Labs, Founding Research Monograph, Version 1.0, 2026.

Disclosures

Peer-review status. This is an independent research publication of Orchestrator Institute Research Labs. It is a working draft and has not undergone external academic peer review.

Conflict of interest. The author, Mark Sendo, is founder of Orchestrator Institute Research Labs (the publisher of this monograph), founder of Orchestrator Awards (a public recognition body for human orchestration), and founder and operator of Celestial Technologies LLC, which builds autonomous AI systems. This monograph proposes governance standards for the class of autonomous systems the author designs and operates. Readers should weigh the arguments on their own merits with this relationship in view.

Funding. No external funding. This work was produced under the auspices of Orchestrator Institute Research Labs.

Independence. Orchestrator Institute Research Labs, Orchestrator Awards, and Celestial Technologies LLC are related entities under common founder ownership. The author affirms that the analysis and normative positions herein are his own and were not directed by any commercial interest, while disclosing the structural relationship above.

AI-use disclosure. AI tools were used to assist research, drafting, and editing under the author's direction. All claims, framing, and final editorial authority are the author's.

Institutional publisher statement. Published by Orchestrator Institute Research Labs, the research and methodology body of the institution.

Version history. v1.0 — publication date pending founder authorization — initial working-draft publication.

Corrections policy. Corrections are issued as versioned revisions. Substantive changes increment the version; errata are logged on the canonical paper page. The DOI record, once minted, points to the version of record.

The Custody of Intelligence

Power, Sovereignty, and Human Agency Beyond the Open–Closed Divide

Orchestrator Institute Research Labs — Frontier Intelligence & Human Stewardship
Founding Research Monograph — Version 1.0

Author: Mark Sendo, Master Orchestrator · Founder, Orchestrator Institute Research Labs

Document status: Complete draft, v1.0. All ten chapters and the front matter are drafted at depth; the full argument runs end to end across four Acts. The Appendices remain draft skeleton, and the Conclusion is a short pointer to the close of Chapter 10. Every chapter is marked drafted-pending-review, not final: no chapter is final until it survives serious external review. Architecture is frozen (Page Zero).

Mission. To identify, formalize, and defend the governance structures required to ensure that machine intelligence remains auditable, accountable, and permanently subject to human authority as it scales into autonomous infrastructure.

How to cite. Sendo, Mark. *The Custody of Intelligence: Power, Sovereignty, and Human Agency Beyond the Open–Closed Divide*. Orchestrator Institute Research Labs, Founding Research Monograph, Version 1.0, 2026.

The Central Question

As intelligence becomes abundant, who will hold custody over it, what authority may be delegated to it, and how can that authority remain accountable to humanity?

The Monograph's Promise

This monograph does not attempt to predict one future. It attempts to identify the governance questions that remain important across many plausible futures.

Reading Status (this assembly)

This is a working assembly for end-to-end reading, not a finished manuscript. Depth is marked honestly:

- **At depth (drafted, pending external review):** Executive Summary, Introduction, and all ten chapters (1–10). All four Acts are complete; the full argument is drafted end to end.
- **Draft skeleton (pending deepening):** Appendices only. The Conclusion is intentionally a short pointer — the monograph concludes in the close of Chapter 10.

Per the drafting discipline, no chapter is marked "final" until it survives serious external review. Source notes at the ends of the deep chapters record what is verified and what still requires confirmation before publication.

Author's Preface

This paper is not a prediction of the future; it is an identification of the structural governance questions that remain inescapable regardless of which technological timeline unfolds.

I write this as a warning, not as a presentation of credentials. For the past decade, my work has focused on building and directing autonomous systems under human supervision. What that experience revealed — often through unexpected edge-case failures and the brittle nature of traditional oversight mechanisms — is that our institutional concepts of control are profoundly outdated. When an autonomous software suite is tasked with optimizing a supply chain or managing network infrastructure, the friction points are rarely computational. They arise from delegated authority: the exact moment a machine executes a consequential action without waiting for explicit human confirmation.

When I founded Orchestrator Institute Research Labs, it was out of a recognition that the technology sector was becoming trapped in a superficial debate over model access. Staying silent while civilizational infrastructure is handed over to systems lacking clear, enforceable lines of custody would be an abdication of responsibility.

To maintain trustworthiness, this warning relies entirely on precision, not on my background. Every claim made in this monograph is grounded in observable, technical trajectories. Where we observe existing systems, we state facts; where we project future capabilities, we declare our time horizons and assumptions; where we argue for human sovereignty, we state our ethical and normative premises explicitly.

Executive Summary

Artificial intelligence has outgrown the debate that dominates its governance. Public and policy attention has organized itself around a single question — whether the most capable models should be released openly or held closed — and that question, while real, no longer reaches the challenge that matters. As machine intelligence becomes abundant, autonomous, physically embodied, geographically distributed, and capable of directing consequential resources, the defining governance question is not how intelligence is *distributed*. It is who holds custody over it, what authority may be delegated to it, and whether that authority remains accountable to humanity.

This monograph makes that case and proposes an instrument to answer it. Its argument runs in four movements.

Intelligence has stopped being scarce. For the whole of recorded history, useful intelligence — active judgment applied to a specific situation — was inseparable from scarce human beings, and civilization's institutions, from schools and professions to firms, courts, and military command, were built as responses to that scarcity. Prior revolutions, from the printing press to the internet, made *information* abundant; they did not make *cognition* abundant. Machine intelligence begins to separate cognitive capability from the human constraints that made it scarce — not perfectly, not universally, and without any claim about consciousness or timelines, but observably and in ways that already place structural pressure on institutions designed for a world of scarce judgment.

The economics of that abundance reorganizes power. Every civilization is shaped by the economics of its scarcest strategic resource; the Industrial Revolution transformed the economics of physical labor, and the Intelligence Revolution is transforming the economics of cognition itself. As the unit price of cognition collapses — roughly a thousandfold in three years for a fixed capability — total consumption rises faster still, and the value does not disperse. It migrates to what remains scarce: the physical substrate that produces cognition — compute, capital, and above all energy — where supply is concentrated among a handful of actors and much of the price is scarcity rent. Cheap intelligence does not democratize power; it concentrates it at the substrate beneath. Who captures that value, and whether that concentration is accountable, is the custody question in economic form.

Geography fragments custody. Authority over a system is exercised by an institution that can reach it, and that reach is territorial. Yet intelligence and the compute that produces it are becoming trans-territorial. Across borders, legal authority already diverges from physical location — one state's law can reach data stored in another's territory, while that other state forbids the transfer. In orbit, the divergence becomes total: by treaty, no state holds territorial sovereignty in space, and jurisdiction follows the state a system is *registered* with, not where it operates. Compute is already running in orbit under this regime. As intelligence distributes across borders, oceans, and orbit, the enforcing body

that custody depends upon grows harder to locate — until it becomes something an operator can choose.

A framework can govern the authority the field leaves open. Existing instruments — the NIST AI Risk Management Framework, the EU AI Act, and the frontier laboratories' scaling policies — do necessary work, but they govern the *capability* of models: how dangerous a system's abilities are, gated against capability thresholds. They largely do not govern the *authority* of deployed systems: who holds it, what must remain permanently human, and whether delegation is reversible and accountable. This monograph proposes the **Intelligence Stewardship Framework** to occupy that layer — as a complement to those instruments, not a replacement. It classifies a system across three layers held to different standards of evidence (assessed capability, assessed custody, and explicitly projected consequence); tiers the authority a system may exercise; fixes **Prohibited Authority Thresholds** — authorities that must remain human regardless of capability, such as the independent authorization of lethal force; and reduces those principles to enforceable constraints. Its sharpest result is economic: authority over consequential resources is fundable — insurable against the harm it can cause — only when it is bounded, so that *uninsurable authority is, by the framework's logic, prohibited authority*. The mathematics itself becomes a governance signal.

The monograph does not predict a single future, and its argument does not depend on one. It identifies the governance questions that remain important across many plausible futures — and its central claim is that humanity has spent millennia learning to govern land, wealth, energy, and political power, and now faces the harder task of governing intelligence itself. The defining question of the coming century is not how intelligent our machines become. It is whether we retain meaningful custody over the authority we choose to delegate to them.

Introduction

A Question That Did Not Exist Before

Who should be trusted with intelligence that can outlive its creator, operate beyond the reach of any single government, direct physical systems, and increasingly act without waiting for human instruction?

That question could not have been asked in any prior era, because until very recently no such intelligence existed. Judgment — the capacity to interpret a situation, weigh conflicting considerations, decide, and adapt — was bound to individual human beings, who are limited in number, singular in attention, and mortal. The question of who should be *trusted* with cognition never had to be separated from the question of which *person* to trust, because cognition and persons were the same thing. That identity is now dissolving, and its dissolution is what makes the question new.

For several years, the public conversation about advanced artificial intelligence has been organized around a different question: whether the most capable models should be open,

their weights available to all, or closed, held behind a developer's interface. That debate is genuine and consequential, and this monograph does not dismiss it. But it argues that the debate has become a distraction from the deeper matter. Open versus closed concerns how intelligence is *distributed* — who may obtain it. It is silent on what happens after a system is deployed: who holds authority over what it does, whether that authority can be constrained and revoked, and who answers when it causes harm. Those are questions of *custody*, and custody, not distribution, is the axis on which the governance of increasingly capable intelligence will turn.

This is the argument the monograph develops: that as machine intelligence becomes abundant, autonomous, embodied, geographically distributed, and capable of controlling consequential resources, the defining governance question is who holds custody over it, what authority may be delegated to it, and whether that authority remains accountable to humanity.

What This Monograph Is, and Is Not

This is a warning, offered because the shape of what is coming is now visible and remaining silent would be irresponsible. It is not a forecast. It does not attempt to predict which technological future will arrive or when. Its discipline, held throughout, is to identify the governance questions that remain important across many plausible futures — so that its value does not depend on any particular prediction coming true. A reader who believes transformative AI is imminent and a reader who believes it is distant should both find the custody question unavoidable, because it arises the moment consequential authority is delegated to a system that acts on its own, which is already occurring.

Because it is a warning rather than a credential, the monograph rests its authority on precision, not on the standing of its author. To that end it makes three kinds of claim and never blurs them. Where it describes what a system *is* or how it is *governed today*, it states assessed fact, verifiable now. Where it reasons about where a trajectory *points*, it declares a projection, with its time horizon, its confidence, and its assumptions stated openly. Where it argues about what authority *should* remain human, it states its normative premises explicitly and argues for them rather than asserting them. A reader's first and fair challenge to any work of this kind is that it is unfalsifiable speculation dressed as analysis. That discipline — assessed, projected, normative, kept distinct on every page — is the answer to that challenge, and it is the monograph's central commitment.

The Path of the Argument

The monograph proceeds in four movements, each answering one question.

The first establishes *what has changed*. It argues that intelligence was civilization's scarcest strategic resource and is becoming abundant; that this abundance is a genuine historical break rather than the latest technology to be oversold; and that when the economics of a civilization's scarce resource is transformed, economic power reorganizes — in this case, concentrating at the physical substrate that produces cognition. This first

movement establishes why the era is different and why the governance assumptions built for a world of scarce intelligence can no longer be taken for granted.

The second asks *where intelligence exercises power, and who can govern it there*. It argues that custody is territorial — that authority means nothing without a body able to enforce it, and enforcement stops at borders — while intelligence is becoming trans-territorial, dispersing across jurisdictions, into the sea, and into orbit, where territorial authority does not exist at all. This movement also takes up the question of when delegated intelligence becomes delegated agency: not consciousness, but the capacity to plan, persist, and act without waiting, which is what makes the custody question urgent rather than theoretical.

The third asks *how that power should be governed*. It states the custody problem at full depth — where authority must remain permanently human — examines the adversaries of custody, and proposes the Intelligence Stewardship Framework: a governance instrument that classifies systems by capability and by custody, separates assessed fact from projected consequence, fixes the authorities that may never be delegated, and reduces those principles to constraints a system can be held to before it acts. The framework is offered as a complement to the capability-focused instruments that already exist, occupying the layer of authority and accountability they leave open.

The fourth asks *what institutions must exist* for a civilization to sustain that governance over time — the enduring principles that should hold across many futures, and the research that must follow.

Humanity has spent millennia learning to govern land, wealth, energy, and political power. It now confronts a task of a different order: the governance of intelligence itself. Whether that intelligence proves to be open or closed, terrestrial or orbital, embodied or purely computational, human-directed or increasingly autonomous, the question that will define the coming century is the same one this monograph exists to pose and to begin to answer: who holds custody over intelligence, and in whose service will it ultimately act?

ACT I — INTELLIGENCE

What is changing? Civilization is entering a fundamentally different relationship with intelligence itself.

Chapter 1 — The End of Intelligence Scarcity

Why Civilization Has Entered a Different Relationship with Intelligence

Human civilization was built under conditions in which useful intelligence was inseparable from scarce human beings. Machine intelligence begins to separate cognitive capability from that historical constraint. Institutions designed for the first condition must now confront the second.

That sentence is the argument of this chapter, and of the monograph beneath it. Everything that follows — the questions of who governs intelligence, what authority it may hold, who captures its value, and where it may be exercised — depends on first establishing that a genuine break has occurred, and on stating the break precisely enough that it rests on observation rather than prophecy. This chapter makes that case and does only that. It explains why this era is different. It leaves to later chapters what the difference does.

Intelligence Was the Scarcest Thing *(Assessed)*

Civilizations have competed across history for many resources — land, food, energy, capital, ore, ports, and the loyalty of people. Beneath every one of them lay a quieter dependency: the capacity to interpret, reason, decide, coordinate, and apply knowledge to circumstance. That capacity is not the same as information, and the distinction, which becomes decisive later in this chapter, is worth drawing at the outset. Information is a record — a fact, a text, a measurement. Intelligence is what turns a record into a decision: the judgment that reads a situation, weighs conflicting evidence, chooses a course, and adapts when the course fails. A library holds information. A physician at a bedside supplies intelligence.

For the whole of recorded history, that second thing — active cognition, judgment applied in the moment — was bound to human beings, and human beings are scarce in every dimension that matters. A person requires years to train and decades to become expert. Attention is single-threaded; a mind attends to one hard problem at a time. Memory is lossy and dies with its holder. A skilled judgment cannot be copied; it can only be slowly reproduced in another person through education, and imperfectly at that. The supply of any particular expertise was therefore limited not by choice but by biology: by the number of humans who could be trained, the time each could work, and the fact that their accumulated judgment vanished at the end of a life.

This scarcity is the hidden premise beneath the architecture of civilization. Consider what the great institutions are, functionally, once the surface purpose is set aside. Schools exist to reproduce expertise across generations, because a mind cannot be copied and must be regrown in each new person. Universities exist to expand the frontier of expertise and to concentrate scarce advanced cognition where it can compound. Professions — medicine, law, engineering, accountancy — exist to certify that a given human actually holds a scarce competence, so that the rest of society can rely on it without verifying it directly each time. Firms and bureaucracies exist to coordinate many limited human minds toward a purpose none could achieve alone, and hierarchy exists to allocate scarce decision authority to the few judged most capable of exercising it. Libraries and archives exist to preserve the products of cognition beyond the lifespan of any one thinker, precisely because the thinker dies and the judgment would otherwise die with them. Courts assign consequential decisions to specific trained humans. Militaries build command structures that route authority to human officers who can observe, weigh, and decide.

The claim of this chapter is that these institutions are not merely cultural inheritances or arbitrary conventions. They are, in significant part, *engineered responses to the limited supply and bandwidth of reliable human cognition*. They are the machinery a civilization builds when the thing it most depends on — good judgment, applied where and when it is needed — cannot be manufactured, copied, or scaled, but only slowly grown inside mortal people and carefully allocated among them. Change that premise, and one has not changed a policy or a technology. One has changed the condition these institutions exist to manage.

Information Became Abundant; Intelligence Did Not *(Assessed)*

The obvious objection arrives immediately, and answering it is the most important work of this chapter, because the objection is the one a careful and skeptical reader will raise: has this break not been announced before? Every generation believes its defining technology is the great rupture. The printing press, mass literacy, the telegraph, the telephone, radio, the computer, the internet — each was heralded as the end of some ancient constraint. Why is this different, and not merely the latest technology to be oversold?

The answer is a distinction that the previous revolutions did not cross, and that this one does: *those revolutions made information abundant; they did not make intelligence abundant*. This is not a quibble. It is the whole of why this era breaks from the last.

The printing press, mass literacy, telecommunications, computing, and the internet performed a genuine and world-changing function: they made information — stored, recorded knowledge — dramatically cheaper to reproduce, transmit, and retrieve. Over five centuries the cost of accessing a fact fell from prohibitive to nearly nothing. But access to a record is not the same as the cognition that acts on it, and the two did not fall together. Search can retrieve a medical paper in a fraction of a second. It cannot, on its own, assess a particular patient, reconcile that paper against three others that conflict with it, weigh the patient's history and preferences, formulate a treatment path, explain that path at a level the patient can act on, and remain simultaneously available to do the same for millions of other people. The paper is information. Everything else in that sentence is intelligence — active judgment, applied to a specific case — and until very recently it could come only from a scarce, trained, mortal human being.

This is why the internet, for all that it transformed, did not end the scarcity this chapter describes. It expanded access to *stored knowledge* while leaving *active cognition* exactly where it had always been: locked inside human minds, limited in number, serial in attention, and mortal. A civilization drowning in information can still be starved of judgment. Indeed, the abundance of information arguably made the scarcity of judgment more acute, because the volume of records to be interpreted grew far faster than the human capacity to interpret them. The bottleneck did not move. It only became more visible.

Machine intelligence is different in exactly the place the earlier revolutions were not. It begins to expand access not to stored knowledge but to *active cognition itself* — to

interpretation, reasoning, planning, explanation, and judgment applied to specific situations. That is the historical break. Not more information; more of the thing that turns information into decisions. The distinction between the printing press and machine intelligence is the distinction between abundant records and abundant reasoning, and it is the reason the analogy to prior technologies, which the skeptic is right to demand, ultimately fails: prior technologies scaled the record, and this one begins to scale the reader.

What "Abundant" Must, and Must Not, Mean *(Assessed — definitional)*

A claim this large is only as strong as its precision, and here the temptation is to overreach into exactly the assertions that would make the argument fragile. So the term at the center of this chapter must be defined narrowly and defended honestly. This monograph does *not* claim that intelligence has become free, universal, infallible, conscious, or equivalent to the full range of human understanding. It does not depend on artificial general intelligence, on any particular timeline, or on machines that match human judgment across all domains. Any of those claims would tie the argument to a prediction it does not need and cannot guarantee — and the whole discipline of this work, stated in its opening Promise, is to rest on what can be observed rather than on what might arrive.

The narrower claim is sufficient and is already observable. In this monograph, machine intelligence becomes *abundant* when cognitive capability can increasingly be:

- **replicated** without reproducing a human education — a competence, once achieved in a system, is copied rather than regrown in a new person over years;
- **deployed simultaneously** across many tasks and many users at once, unbound by the single-threaded attention of a human mind;
- **updated centrally and distributed rapidly**, so that an improvement propagates in days rather than across a generation of retraining;
- **embedded** inside organizations and physical systems, operating where judgment is needed rather than where a trained human happens to be;
- **invoked repeatedly at declining marginal cost**, so that applying cognition to one more case approaches the trivial;
- **coordinated** with other machine systems, composing many acts of cognition without the overhead that limits human coordination;
- **made persistent** beyond the attention span, working hours, and lifespan of any individual, so that accumulated capability no longer dies with its holder.

Each of these is a property that scarce human cognition conspicuously lacks, and each is visible in deployed systems today, in some measure, without appeal to any future breakthrough. Note what unites them: every one is a way in which cognition is becoming *decoupled from the individual human being* — from the years of training, the single thread of attention, the working day, the mortal span. That decoupling is the precise content of "abundance" as this chapter uses the word. It does not assert that machine judgment is as

good as human judgment, or good enough for any particular purpose. It asserts something narrower and, for the argument to come, entirely sufficient: that cognitive capability is being separated from the human constraints that made it scarce. Quality is a separate question, and a critical one, taken up when the monograph turns to what authority such systems should hold. The claim here is only about *supply* — about the ending of a constraint, not the arrival of perfection.

Abundance Destabilizes Institutions Built for Scarcity *(Assessed)*

If the great institutions are in part engineered responses to the scarcity of human cognition, then relaxing that scarcity necessarily places structural pressure on them — not because they vanish, but because the premise beneath them changes, and an institution built on a premise that no longer holds must either adapt or strain. This is where the transformation earns its civilizational description, and it must be described as pressure on foundations rather than as a survey of industries, because the point is not that many sectors are affected but that the *same underlying assumption* is being altered beneath all of them at once.

Education was built to transmit scarce knowledge from those who hold it to those who do not, across the years a human mind requires to absorb it; when interpretation and explanation become abundant and available on demand, the premise that knowledge must be slowly transmitted through a scarce human teacher comes under pressure. The professions were built to certify that a specific human holds a scarce cognitive competence, so society can rely on it without re-verifying; when that competence can be invoked directly, the function of certification — and of the scarcity it manages — is unsettled. Firms and bureaucracies were built to coordinate many limited minds through hierarchy; when coordination and analysis grow cheap and cognition can be embedded throughout an organization, the rationale for particular hierarchical structures weakens. Law was built around human decision-makers with identifiable intent, and assigns responsibility on that basis; when consequential decisions are made or shaped by systems that are not human and whose "intent" is not a settled concept, the premise beneath legal responsibility is strained. Government was built around limited administrative capacity, rationing the attention of officials; when administrative cognition can scale, the constraints that shaped the size and reach of the state shift. Science was built around the constrained attention of researchers, each able to read, hypothesize, and experiment only so much; when cognition becomes a collaborator in the work rather than only its instrument, the pace and structure of discovery are pressed. Military command was built around human-speed observation and human authorization of consequential acts; when systems can observe and act faster than humans can supervise, the premise that a human is always positioned to decide comes under the most acute pressure of all — a pressure this monograph returns to when it reaches the question of what authority must remain permanently human.

In none of these cases is the claim that the institution disappears. Schools, courts, firms, and command structures will persist. The claim is that each was built atop an assumption

— that reliable cognition is scarce and must be grown in humans, allocated among them, and preserved beyond them — and that this assumption is precisely what is now changing. When the premise beneath a structure shifts, the structure does not necessarily fall, but it can no longer be taken for granted, and the work of governing the new condition begins. That is the civilizational meaning of the end of intelligence scarcity: not that our institutions are abolished, but that the foundation they were engineered upon is moving, and they must now be re-examined against a condition their designers never faced.

Scarcity Does Not Vanish — It Relocates *(Assessed — the handoff)*

There is a final move to make, and then this chapter must stop, because the next step belongs to another argument and taking it here would blunt both.

The end of intelligence scarcity is not the end of scarcity. When one resource becomes abundant, necessity does not disappear; it relocates to whatever remains scarce and newly decisive. This is a general pattern, and it holds here. As cognition scales and ceases to be the binding constraint it was for all of prior history, scarcity begins to move toward the things that abundant cognition cannot itself supply — and it is worth naming them, because they are the subjects of the chapters that follow. Scarcity relocates toward *trustworthy judgment* — the ability to know which cognition to rely on and when. Toward *legitimate authority* — the question of who or what may rightfully decide. Toward the *physical substrate* on which cognition now runs, and the energy that powers it. Toward *verified data*, *responsibility for outcomes*, *human attention* as it becomes the genuinely limited input, *social trust*, and *control over the objectives* that direct increasingly capable systems.

Naming where scarcity relocates is the boundary of this chapter and the threshold of the rest of the monograph. Two of those relocations organize the work to come. The movement of scarcity toward the physical substrate — compute, energy, and the concentrated means of producing cognition — becomes an economic question: when the old scarce resource becomes abundant, where does value migrate, and who captures it? That is the subject of Chapter 3, and this chapter deliberately says nothing further about it, because the economics of the relocation is that chapter's to prove, and to reach for it here would steal its argument and weaken this one. The movement of scarcity toward legitimate authority, trustworthy judgment, responsibility, and control over objectives becomes a governance question — the custody question this entire monograph exists to answer. Chapter 1 opens the door to both and walks through neither.

What this chapter has established stands on its own and is enough to carry what follows. For the whole of recorded history, useful intelligence was inseparable from scarce human beings, and civilization built its institutions to manage that scarcity. Machine intelligence begins to separate cognitive capability from that constraint — not perfectly, not universally, not by any promise about the future, but observably and in ways that already matter. The institutions built for the first condition must now confront the second. Why this era is different is now answered. What the difference does — to economic power, to

jurisdiction, to authority, and to the question of who holds custody over intelligence as it becomes abundant — is the work of the chapters ahead.

Chapter Notes

Chapter 1 is a conceptual and historical argument. Its rigor rests on the precision of its central distinction (information vs. active cognition) and its careful, deliberately narrow definition of "abundant," not on current empirical figures — by design, so that it does not import Chapter 3's evidence. Where specific institutional-history claims would benefit from scholarly anchoring, they are flagged below rather than asserted as established consensus.

1. **Institutions as responses to cognitive scarcity.** The claim that schools, professions, firms, hierarchies, archives, and command structures are in part engineered responses to the limited supply and bandwidth of human cognition is presented as an analytical framing, argued in-text, not as a settled thesis of institutional economics or sociology. If a scholarly anchor is wanted, standard treatments in the economics of information, the sociology of professions, and organizational theory can be cited; the argument is written to stand as reasoning.

2. **Information abundance vs. intelligence scarcity.** The historical claim that prior information revolutions (printing, literacy, telecommunications, computing, the internet) scaled *stored information* without scaling *active cognition* is the load-bearing distinction of the chapter and is argued conceptually; it requires no contested empirical claim.

3. **Definition of "abundant."** Deliberately narrow and non-predictive: replicability, simultaneous deployment, central update, embedding, declining marginal cost, machine coordination, persistence — each observable in deployed systems today "in some measure," with no dependence on AGI, consciousness, zero cost, or timeline. This definition is what allows the monograph to rest on observation per Page Zero's Promise. It must not be strengthened in revision beyond what deployed systems currently support.

4. **Institutional destabilization.** Presented as *pressure on premises*, explicitly not as a claim that institutions vanish, and explicitly not as an industry survey. The military-command instance is deliberately previewed as the most acute case, forward-referencing the custody question without arguing it here.

5. **Boundary with Chapter 3 (enforced).** The relocation-of-scarcity close names where scarcity moves (judgment, authority, substrate, energy, data, responsibility, attention, trust, control of objectives) and then STOPS. No token prices, Jevons, NVIDIA, compute rent, capital concentration, or financialization — Chapter 3 owns all of it. The handoff paragraph explicitly hands the economic relocation to Chapter 3 and the governance relocation to the rest of the monograph. This boundary is a hard wall per Page Zero, Rule 1.

Chapter 2 — Beyond the Open–Closed Divide

The Debate as Currently Configured *(Assessed)*

For the better part of a decade, the public argument over how to govern advanced artificial intelligence has organized itself around a single fault line: whether the most capable models should be released openly, with their weights available for anyone to download, inspect, and modify, or held closed, accessible only through a controlled interface under the developer's supervision. The debate is serious, and both positions are held by serious people.

The case for open weights is a case about power and knowledge. When model weights are public, capability is not concentrated in a small number of well-capitalized laboratories. Independent researchers can audit systems directly rather than trusting a provider's own account of them. Smaller companies, universities, and national programs can build without permission from — or dependence upon — a foreign commercial gatekeeper. Security researchers can find flaws that a closed provider might never disclose. Historically, open ecosystems have also proven more resilient than closed ones, because they do not fail at a single point. Proponents argue, with force, that transparency and distribution are not risks to be managed but the very conditions under which a powerful technology can be held accountable at all.

The case for closed deployment is a case about control and consequence. When a developer retains the weights, it retains the ability to monitor how a system is used, to refuse uses it judges dangerous, to patch vulnerabilities centrally, and — critically — to withdraw or modify a system after release. A model behind an interface can be rate-limited, filtered, and rolled back; a model whose weights are in public circulation cannot be recalled by anyone. Proponents point out that certain capabilities, particularly in biology and cyber operations, are difficult to make safe once they are irreversibly distributed, and that sustained safety investment is easier to justify when the developer remains accountable for the deployed system's behavior.

Both cases are strong on their own terms. Both describe real benefits and real risks. Neither is a strawman, and this paper does not attempt to resolve the disagreement between them.

It is worth establishing how real this axis has become, because the argument that follows is not that the open–closed distinction is trivial. It is real enough that it has already been written into law. The European Union's AI Act, whose obligations for general-purpose AI models took effect in August 2025 and whose enforcement powers activate in August 2026, treats open-source release as a formal regulatory category: providers of open-source general-purpose models that do not rise to the level of systemic risk are exempted from certain detailed technical-documentation requirements that closed providers must satisfy. The distinction between open and closed is thus not merely a matter of engineering culture or commercial strategy. It is a line that a major legal instrument now draws, with different obligations on either side of it.

That fact is the beginning of this chapter's argument rather than a refutation of it. The open–closed distinction is consequential enough to legislate. And precisely because it is the axis around which so much regulatory and intellectual energy is organized, its limits deserve careful examination — because the question it answers is not the question that will determine whether human institutions retain meaningful control over increasingly capable systems.

Why the Axis Cannot Reach the Question That Matters *(Assessed & Normative)*

The open–closed distinction is an axis of *distribution*. It answers a single question: who is permitted to obtain the model's weights? That question has real consequences for competition, for auditability, for security, and for national independence. But it is silent on a second question that is, for the governance of consequential systems, more decisive: once a system is deployed, who holds authority over what it is permitted to do, and can that authority be constrained, overridden, and revoked?

Call the first the question of access and the second the question of custody. The central claim of this paper is that these two questions are largely independent of one another — and that the field's near-total focus on the first has left the second structurally unaddressed.

The independence is easiest to see by observing that every combination of the two is not only possible but common. A closed, proprietary model — accessible only through a supervised interface — can nonetheless be wired into a critical system in a way that grants it wide operational latitude, with no reliable human override, diffuse accountability across vendors and integrators, and no practical means of reversing its decisions before their consequences propagate. That is a closed system with weak custody. Conversely, an open-weight model — freely downloadable — can be deployed inside an institution that scopes its authority narrowly, logs and audits its actions, retains a human decision-maker at every consequential step, assigns clear legal responsibility for outcomes, and can halt or roll it back at will. That is an open system with strong custody. The two axes cross freely; the license under which a model was released tells you almost nothing about the custody posture under which it is deployed.

This is why resolving the open–closed debate — in either direction — leaves the governance question that actually determines consequence untouched. A world that settled entirely on open weights, and a world that settled entirely on closed APIs, would each still face the whole of the custody problem: in each, some deployed systems would hold authority they should not, some delegations would be irreversible, and some chains of accountability would be too diffuse to enforce. The distribution of the model does not settle the authority of the deployment.

The normative weight of the distinction is worth stating directly. The harm a system can cause is not principally a function of who can download it. It is a function of what it has been permitted to do, whether a human remains positioned to stop it, and who answers for

the result. A diagnostic model that recommends a treatment for a physician to accept or reject, and an autonomous system that administers that treatment without waiting for confirmation, may be built on identical weights under an identical license. What separates them is custody — the structure of authority around the deployed system — and custody is exactly what the open–closed axis does not describe.

How the Field Actually Governs Today — and What It Measures *(Assessed)*

If the open–closed axis does not reach the custody question, it is fair to ask whether the more developed governance instruments do. The three most influential — the European Union's AI Act, the United States' NIST AI Risk Management Framework, and the safety frameworks published by the frontier laboratories themselves — represent the current state of the art in AI governance. Examined closely, all three share a common architecture, and that shared architecture is the crux of this chapter.

The EU AI Act organizes obligations by risk and by capability. It sorts AI systems into tiers — from minimal risk through limited and high risk to prohibited practices — and imposes the heaviest duties on high-risk applications. For general-purpose models specifically, it introduces a systemic-risk tier that attaches to the most capable models, using training compute as its principal trigger: a model trained above a defined threshold of computational operations is presumed to pose systemic risk and carries the fullest obligations, regardless of whether it is open or closed. The Act's instruments are transparency, technical documentation, evaluation, and — for systemic-risk models — adversarial testing. These are meaningful requirements. But note what they measure: the *properties and capabilities of the model*, and the *risk category of its application*. The Act asks how capable a model is and in what domain it is used. It does not ask who holds operational authority over a specific deployment, whether that authority is reversible, or where the permanent human line lies.

The NIST AI Risk Management Framework, the de facto reference for AI governance in the United States, is voluntary and non-certifiable, and organizes practice around four functions — govern, map, measure, and manage — applied across a system's lifecycle. It has been extended through profiles for particular contexts, including generative AI and, as of a 2026 concept note, critical infrastructure. It is a genuinely useful instrument for structuring how an organization identifies and mitigates risk. But the framework was designed, in its own framing, for systems with defined operational boundaries, predictable behavior envelopes, and human oversight at most points of interaction — and its core work is to help organizations map and measure the risks a system's *capabilities* present. Like the EU Act, it is oriented toward the properties of the system and the management of their risks, not toward the structure of authority over a deployed and increasingly autonomous agent.

The frontier laboratories' own safety frameworks are the most instructive case, because they are the most sophisticated and the most explicitly tied to autonomy — and they, too, measure capability. Anthropic's Responsible Scaling Policy defines a ladder of AI Safety

Levels, consciously modeled on the biosafety levels used in laboratory containment, and commits the developer to specific safeguards as a model's capabilities cross defined thresholds — including thresholds for autonomous AI research and development. OpenAI's Preparedness Framework tracks a set of risk categories — among them cybersecurity, biological and chemical capability, AI self-improvement, and autonomous replication — and defines "High" and "Critical" capability thresholds that trigger deployment or development restrictions. Google DeepMind's Frontier Safety Framework specifies critical capability levels and the reviews that attend them. The three differ in structure and in how much they pre-commit versus preserve discretion, but they converge on a single logic: turn the question "is this model safe?" into a closed loop of capability threshold, evaluation against that threshold, and a pre-committed action if it is crossed. They are, by their own description, commitments about what the laboratories will *ship* — governing the model at the point of release, not the authority structure of whatever is later built from it.

The through-line across all three pillars is exact and, once seen, difficult to unsee. Autonomy is not absent from these frameworks. It appears — as *autonomous replication*, *self-exfiltration*, *autonomous R&D acceleration*. But in every case it appears as a **capability of the model to be measured and gated**: a dangerous property that crosses a threshold and triggers a mitigation. In none of these instruments does authority appear as a **relationship to be governed** — a standing structure around a deployed system that specifies who granted its authority, what that authority may and may not include, who may revoke it, and who is accountable for its exercise. The field measures how autonomous a model *can* be, and whether that is dangerous. It does not govern how much authority a deployed system *should* hold, or where the human line is drawn and made permanent.

The Turn the Field Is Beginning to Make — and Where It Stops *(Assessed & Projected)*

Honesty about the current landscape requires acknowledging that this gap is beginning to be recognized. It would be false to claim that no institution has noticed the shift from tools to agents, and a claim this paper does not make. In February 2026, the United States, through the National Institute of Standards and Technology's Center for AI Standards and Innovation, launched an AI Agent Standards Initiative directed specifically at autonomous agents. Its stated focus areas include identity and authorization, security and risk management, and monitoring and logging, with an agent interoperability profile anticipated for late 2026. This is the field crossing, for the first time in a major standards effort, from the model to the deployed agent — from what a system can do to what a specific agent is permitted and observed to do.

The direction is right, and the initiative is welcome. But it is essential to see precisely which layer it occupies, because that layer is not the same as custody. The emerging agent standards treat delegated authority as an *engineering and operational* problem: give each agent a verifiable identity, scope its permissions, authenticate its actions, record what it does, and make agents interoperable across systems. Every one of these is necessary.

None of them is sufficient, because all of them presuppose the very question that custody asks and leave it unanswered.

Identity and access management can establish that an agent *was authorized* to perform an action, and can prove after the fact that it did. It cannot establish whether that action is one a machine should ever have been authorized to perform. Logging can produce a complete and tamper-evident record of an autonomous decision. It cannot tell you whether that decision was one that ought to have required a human hand. Permission-scoping can confine an agent to the authority it was granted. It cannot tell you whether the grant itself was legitimate, whether it is revocable in practice as well as in principle, or whether any authority was delegated that must, under any capability, remain permanently human. These are not engineering questions. They are questions of legitimate authority and its limits — and they sit one layer above the standards now being built, in the space this paper calls custody.

The projection here is modest and, this paper argues, well founded: as autonomous agents move from pilots into consequential deployment across finance, infrastructure, logistics, defense, and public administration, the engineering layer of identity and monitoring will mature — it is being built now — while the governance layer of legitimate authority will remain comparatively undeveloped unless it is deliberately constructed. The gap will not close on its own, because the instruments now being built are not aimed at it. They authenticate delegation; they do not adjudicate it.

Restating the Thesis, Now Grounded *(Normative)*

The spine of this monograph can now be restated with the support of the survey rather than as assertion. Capability sets the boundary of what a system can do; authority determines what it is permitted to do — and the two are not the same. What the examination of the current landscape reveals is that the field governs the first clause comprehensively and the second clause barely.

The open–closed debate contests the *distribution* of capability — who may obtain it. The EU AI Act and the frontier laboratories' scaling policies gate the *level* of capability — how much dangerous capability may be deployed, and under what safeguards. NIST's emerging agent standards authenticate *delegated action* — proving which agent did what, under which permissions. Each of these is valuable, and none is in conflict with what follows. But none of them governs the *legitimacy and the limits of the authority itself*: whether a given authority should be delegable to a machine at all, what must remain permanently in human hands, and what makes a delegation reversible and accountable in an institutional rather than a merely technical sense.

That is the space the remainder of this monograph occupies. It is not a space the existing frameworks have failed at; it is a space they were not built to address, because they answer a different and prior question. The Intelligence Stewardship Framework proposed in Section IX is therefore offered as a complement to the EU AI Act, the NIST Risk Management Framework, and the laboratories' scaling policies — not as a replacement for

any of them. It presupposes their capability governance and layers upon it a governance of custody: of authority, reversibility, and the permanent human line.

Before that framework can be specified, however, one further feature of the landscape must be examined — one that the open–closed debate obscures entirely and that even the emerging agent standards do not fully reckon with. Intelligence is not only becoming more capable and more autonomous. It is becoming physically distributed, migrating from a small number of supervised data centers into edge devices, robotics, and eventually architectures beyond national territory altogether. As it disperses, the custody question does not merely persist. It fragments. That is the subject of Section III.

Section II — Source Notes

All claims below reflect the landscape as of mid-2026 and were verified against current sources during drafting. This is a fast-moving area; each should be re-checked before publication. Citations are to issuing bodies and primary instruments; full bibliographic entries to be compiled in the References section.

1. **EU AI Act timeline and structure.** Regulation (EU) 2024/1689. Entered into force 1 August 2024. Obligations for general-purpose AI (GPAI) models applicable from 2 August 2025; enforcement/penalty powers for GPAI providers from 2 August 2026; high-risk system rules from 2 August 2026; Article 6(1) classification rules from 2 August 2027. European Commission GPAI guidelines adopted July 2025.

2. **EU AI Act systemic-risk threshold.** Systemic-risk tier for GPAI models (Article 55) triggered principally by training compute above 10^25 floating-point operations, applying regardless of license; baseline GPAI obligations under Article 53.

3. **EU AI Act open-source treatment.** Open-source GPAI models not posing systemic risk are exempted from certain detailed technical-documentation obligations (Articles 53(2) and 54(6)); transparency and copyright-related obligations under Article 53(1)(c)–(d) apply regardless of license. This encodes the open/closed distinction as a formal regulatory category.

4. **NIST AI Risk Management Framework.** AI RMF 1.0 (NIST AI 100-1), released 26 January 2023; four core functions (Govern, Map, Measure, Manage); voluntary and non-certifiable. Extended via profiles: Generative AI Profile (NIST AI 600-1, July 2024); Trustworthy AI in Critical Infrastructure profile concept note (7 April 2026). No “2.0” released as of mid-2026; framework extended through profiles rather than version changes.

5. **NIST CAISI AI Agent Standards Initiative.** Announced February 2026 through NIST’s Center for AI Standards and Innovation; focus areas: identity and authorization, security and risk management, monitoring and logging; AI Agent Interoperability Profile anticipated Q4 2026. Accompanied by NIST guidance (March

2026) on monitoring for agentic systems spanning functionality, operations, security, compliance, and human factors.

6. **Frontier laboratory safety frameworks.** Anthropic Responsible Scaling Policy (first published September 2023; comprehensive v3.0 rewrite, with v3.1 in April 2026) — AI Safety Levels (ASL) modeled on biosafety levels; capability thresholds including CBRN and autonomous AI R&D. OpenAI Preparedness Framework (v2.0, April 2025) — Tracked Risk Categories including biological/chemical, cybersecurity, AI self-improvement, and autonomous replication; "High" and "Critical" capability thresholds; Safety Advisory Group. Google DeepMind Frontier Safety Framework (v3.0, September 2025) — critical capability levels; process-oriented. All three are voluntary, lab-specific commitments governing models at the point of release.

7. **Characterization of shared architecture.** The observation that these frameworks convert "is the model safe?" into a loop of capability threshold, evaluation, and pre-committed action, and that they govern what labs ship rather than what is built downstream, reflects comparative analyses of RSP/Preparedness/FSF current as of mid-2026 (e.g., practitioner and standards-body comparisons, and the International AI Safety Report 2026).

Chapter 3 — The Economics of Intelligence

Cognition as a Strategic Resource, and the Reorganization of Economic Power

The Economics of a Scarce Resource *(Assessed — analytical lens)*

Every civilization has organized itself, in part, around the economics of the resource it could least afford to be without. This is the starting observation of this chapter, and because it is a large claim, it must be stated with precision about what it does and does not assert.

It is *not* the claim that a single resource determines a civilization's shape. History is overdetermined. Feudal Europe, industrial capitalism, and the networked economy each had many causes — military technology, law, religion, disease, climate, the accidents of politics — and to name one resource as *the* cause of any of them would be exactly the monocausal history that serious scholarship has rightly abandoned. The claim here is narrower and sturdier: that the economics of a civilization's scarcest strategically decisive resource — how scarce it is, who controls it, and how its surplus is captured — exerts a shaping pressure on that civilization's organization that no honest account can leave out. Not the cause. A cause that cannot be removed without the explanation collapsing.

Read with that discipline, the pattern is hard to miss. One cannot explain feudal organization without the economics of land: its scarcity, its immobility, and the fact that whoever controlled it controlled the agricultural surplus and therefore the social order built atop it. One cannot explain the industrial age without the economics of capital and energy: mechanized production concentrated value in those who owned the machines and the coal and the mills, and the institutions of industrial capitalism formed around that

concentration. One cannot explain the last half-century without the economics of information: when information became cheap to copy and distribute, value migrated to those who could aggregate, index, and route it, and the largest enterprises of the era were built on that migration. In each case the decisive resource did not act alone — but its economics shaped who held power, which institutions formed around access to it, and where the surplus pooled.

Seen this way, each great transition is, at its core, a transformation in the economics of a decisive input. And that framing yields the sentence this chapter is built to defend:

> **The Industrial Revolution transformed the economics of physical labor. The Intelligence Revolution is transforming the economics of cognition itself.**

The parallel is exact enough to be worth drawing carefully. Mechanization substituted machines for muscle; it re-priced physical labor, displaced some of it, made the rest vastly more productive, and shifted the returns from those who sold their labor to those who owned the means of production. It did not make physical work worthless — it changed *the economics* of physical work, and that change reorganized the society around it. Machine cognition is now doing to thinking what the machine did to muscle. Cognition — the scarce input that has defined the value of skilled human labor for the entire modern era — is being mechanized, re-priced, and redistributed. The question this chapter asks is the one the pattern demands: when the economics of cognition is transformed, how does economic power reorganize, and who ends up holding it?

Chapter 1 established that intelligence, long civilization's scarcest strategic resource, is becoming abundant. This chapter takes up what follows economically — and the central finding is that when the scarce resource around which an era was organized becomes abundant, its economics does not relax. It relocates. Tracing where it relocates is how we find where economic power in the Intelligence Age is actually concentrating, and the answer is neither obvious nor comforting. It is, in fact, the opposite of what the surface data suggests.

The Re-Pricing of Cognition, and Its Paradox *(Assessed)*

The first move in the reorganization is the collapse in the unit price of cognition, and it is collapsing faster than almost any commodity in the history of computing. The cost of producing output at the capability of a 2023 frontier model fell from roughly thirty dollars per million tokens to under fifty cents by mid-2026 — on the order of a thousandfold reduction in three years for a fixed quality bar. The venture firm Andreessen Horowitz named the phenomenon "LLMflation," the inverse of inflation, in which the same unit of intelligence grows dramatically cheaper each year; the research institute Epoch AI, tracking the same curve across benchmarks, found a median decline near fiftyfold per year, ranging from ninefold to several hundredfold depending on the task. The drivers are ordinary and compounding: algorithmic efficiency, cheaper and more specialized hardware, and fierce competitive pricing pressure, including from open-weight models that set a low floor the proprietary providers must track. This is the mechanization of

cognition expressed as a price curve — the same downward pressure the machine once put on the cost of physical work, now applied to thought.

A naive reading concludes that intelligence is therefore democratizing — that as the price of cognition approaches zero, access spreads and economic power disperses. This reading is wrong, and the reason it is wrong is the analytic core of the chapter.

The error is an old one, first described by William Stanley Jevons in 1865, who observed that more efficient steam engines did not reduce Britain's coal consumption but increased it, because falling cost per unit expanded total use faster than efficiency saved it. The same dynamic governs cognition. As the price per token falls, the quantity consumed rises faster than the price drops, so aggregate spending grows even as unit cost collapses. The evidence is unambiguous. Blended prices across frontier models fell roughly sixty-seven percent in a single year — and over the same period a large majority of enterprises exceeded their AI budgets, with one major company reportedly exhausting its entire annual AI-coding budget in four months. Autonomous agents are the accelerant: where a human once issued a prompt and read a reply, an agent now consumes hundreds of thousands or millions of tokens reasoning across steps, and forecasts reflect it, with one investment bank projecting a roughly twenty-fourfold increase in token consumption by the end of the decade. Average enterprise AI budgets rose from the low single-digit millions to several million dollars in two years. Cheaper cognition has not meant less spending on cognition. It has meant far more.

This is the paradox at the heart of the reorganization: the unit price of intelligence is racing toward zero, and the total economic value flowing through intelligence is exploding. Both are true at once. Holding them together is the key to seeing where the money actually goes — and where it goes is the answer to how economic power reorganizes when cognition becomes abundant.

The Migration of Value: The Economic Signature of Abundance *(Assessed)*

Here the civilizational lens earns its keep, because what happens next is precisely what the economics of an abundant resource predicts. When a resource becomes abundant and cheap, value does not disappear — it migrates to whatever remains scarce and necessary to produce or use it. Cheap information made attention and aggregation valuable. Cheap goods made distribution and logistics valuable. Cheap cognition is making the *substrate that produces cognition* valuable. This is not a political observation about particular companies. It is economic gravity, and it operates regardless of anyone's intent.

If the price of a token is collapsing while the total value moving through tokens soars, the value is not captured *at the token*. It is captured where the collapse cannot reach: the physical substrate beneath the token, the layer that cannot be commoditized because it is genuinely scarce. The commoditization of the token is therefore not a force against concentration — it is the *mechanism* of concentration, relocating the capturable margin

downward, out of the software layer where competition and open weights drive price to cost, and into the infrastructure layer where supply is held by a handful of firms.

Consider what that substrate contains and who holds it. At the level of the accelerator — the specialized chips on which all inference runs — a single company held roughly eighty to ninety percent of the market by revenue through 2025, a share projected to ease only toward the mid-seventies as rivals and custom silicon scale, even as its absolute data-center revenue climbed toward nearly two hundred billion dollars for the fiscal year. At the level of capital, the small group of hyperscale cloud companies was projected to spend more than six hundred billion dollars in a single year on capital expenditure, most of it on AI infrastructure — a scale that is itself a barrier, since, as the dominant chipmaker's own filings note, less-capitalized firms struggle to finance infrastructure at this magnitude, which shrinks their participation. The structure is self-reinforcing: the cheaper the token becomes, the more the durable value accrues to whoever owns the compute, the capital, and the power beneath it.

The clearest single indicator of where the surplus sits is the composition of the price of compute itself. Decompositions of what a rented GPU-hour actually costs find that between roughly half and the large majority of the price is not the underlying cost of hardware and energy but a scarcity premium — the rent commanded by controlling constrained supply. This is the economic substance behind the phrase "billions, and eventually trillions, to a few." It is not that intelligence is expensive; intelligence is nearly free. It is that the *scarcity beneath* intelligence is expensive, and the ownership of that scarcity is concentrated. The surplus of abundant cognition does not disperse across the many who use it. It pools at the narrow base of the stack — which is exactly what the economics of a newly abundant strategic resource predicts, and exactly the reorganization this chapter set out to trace.

The New Scarcity: Energy as the Binding Constraint *(Assessed)*

Beneath the chips lies a layer scarcer still, and it has quietly become decisive. Through the early buildout, the constraint on scaling intelligence was the supply of GPUs. By 2026 that had changed: the chips increasingly existed, and what did not exist, in enough places, was the electricity to run them and the capacity to cool them. Power and thermal management, not compute, became the hard limit on how fast intelligence could scale.

The figures show why this concentrates rather than distributes. Global data-center electricity demand rose sharply — on the order of four hundred fifteen terawatt-hours in 2024, roughly one and a half percent of world electricity, climbing toward a projected near-doubling by 2030 to close to three percent, with data centers accounting for a disproportionate share of all *growth* in electricity demand over the decade. Density is the reason: where a conventional server rack draws five to ten kilowatts, an AI-optimized rack draws forty to eighty kilowatts and, in the newest configurations, well beyond. And this demand is geographically concentrated — the largest data-center cluster on earth sits in a single utility's territory, where the price of securing future power capacity rose more than

tenfold between recent delivery years, a signal of the true cost of serving dense, continuous AI workloads on grids never designed for them.

Energy is the least distributable layer of the entire stack. Chips can be shipped, capital can move, software can be copied and open weights downloaded — but electricity is bound to place: to grids, generation, permitting, and water for cooling. Whoever secures large, reliable, affordable power secures the ability to run intelligence at scale, and that securing is happening among a small set of well-capitalized actors in a small number of favorable locations. The deepest form of the concentration of intelligence is therefore not a concentration of models, or even of chips. It is a concentration of *watts* — and it is the hardest of all to reverse. When historians of this period ask where the strategic resource of the age was controlled, the honest answer will point less at algorithms than at power grids.

The Token as the Meter of the Surplus *(Assessed & Normative)*

Between the collapsing price of cognition and the concentrating value beneath it sits the token — the unit by which cognition is metered, priced, logged, and billed. The token is not merely an accounting convenience; it is the toll-gate through which access to intelligence is sold, and it deserves examination as such.

Three features of the meter matter for governance. First, it is *asymmetric and opaque*: output tokens cost several times more than input tokens, pricing varies by orders of magnitude across tiers, and the relationship between what a user is charged and what the computation actually costs the provider is not visible to the user. Second, it is, at present, *subsidized*. A widely discussed early-2026 analysis argued that inference cost is the primary economic bottleneck preventing AI providers from reaching profitability, and that the API prices enterprises budget around are held below true cost by venture capital and hyperscaler cross-subsidy. If so, the currently low price of the token is partly a strategic artifact, not a stable fact, and the true cost — concentrated in the substrate — is being masked. Third, the meter *gates access to cognition itself*: as more of economic and civic life comes to depend on machine reasoning, the price and availability of tokens determine who can afford to think at scale and who cannot.

The normative point follows from the structure. When the meter's unit price races toward zero while the metered resource's value concentrates in the substrate below it, the meter performs a quiet redistribution *upward* — making cognition feel cheap and abundant to the user while the durable value is captured at the base of the stack the user does not own. An ethics of the meter — who sets it, on what disclosure, under what subsidy, with what obligation to those priced out — is therefore not a peripheral consumer-protection matter. It is part of the custody question, because the meter is the instrument through which the surplus of intelligence is allocated between the many who use it and the few who own its means.

The Counter-Force, and Its Limit *(Assessed)*

If the deflating token conceals an upward concentration of value, the natural question is whether any force pushes the other way. There is one, and it must be stated precisely, limits included, because overstating it would repeat the democratization error in a new form.

The counter-force is the migration of cognition to the edge: running models locally, on hardware the user owns, at zero marginal token cost. A system that performs routine classification, routing, and lightweight reasoning locally — escalating to paid frontier inference only for tasks that genuinely require it — keeps a portion of cognition, and therefore of value, at the user's own layer rather than metering it through the concentrated stack. This is not hypothetical; it is an emerging discipline of system design in which a local reasoning layer handles the common case for free and the expensive centralized model is reserved for the exception. To the degree cognition can be pushed to the edge, the token-metered concentration is partially resisted. This is the most honest available answer to whether distribution can flatten the concentration: at the *cost* layer, through the edge, it partially can — a stronger and more grounded answer than the same hope pinned to open weights alone, since open weights govern who can obtain the model, not who owns the substrate it runs on.

But the limit is real. The edge runs on the same concentrated substrate: the local device is built on chips from the same narrow set of firms, draws energy from the same constrained grids, and still routes its hardest tasks to the centralized layer. Edge computation moves *some* value back to the user; it does not dissolve the concentration of the substrate — it depends on it. The hybrid is a genuine counter-force at the margin and a partial one at the core. It softens the concentration; it does not reverse it. Honesty about that limit is what separates analysis from advocacy — and it is why the distributional question cannot be answered by architecture alone, but becomes, in the end, a question of governance.

A Tradable Market in Cognition *(Projection — stated to be tested)*

Everything to this point in the chapter is assessed: measurable facts about prices, consumption, market shares, and energy. What follows is explicitly a projection — a structural prediction about where these dynamics point, offered with its horizon, its confidence, and the conditions that would prove it wrong. It is placed here, and labeled as such, because it is a bold claim, and the discipline of this monograph is that bold claims about the future are marked as projections and made falsifiable, not asserted as fact.

The projection: inference capacity is on a path toward commoditization and, plausibly, financialization — the emergence of a market in which units of cognition are traded as a commodity is traded, with spot pricing and, potentially, forward contracts and derivative instruments on inference capacity. Confidence: moderate. Horizon: plausibly within the latter half of this decade, contingent on the drivers below persisting.

The claim rests on structural preconditions, each present today and each, historically, a precondition for commodity markets to form:

- **Fungibility.** A unit of cognition at a given capability level is increasingly interchangeable across providers. The demonstration that a challenger model could deliver comparable output at a fraction of incumbent pricing established that buyers treat capability, not brand, as the good — the defining property of a commodity.
- **Metering.** The good is already priced and sold in standardized units. A commodity market needs a unit of trade; inference has had one from the start.
- **Volatility and scarcity.** Prices move sharply, capacity is constrained by the physical substrate, and the price of securing future capacity in constrained regions has already risen manyfold. Volatility and scarcity are what create demand for instruments that hedge and speculate on price.
- **Explosive, forecastable demand.** Consumption is projected to grow many times over as agents become the default mode of use. Large, growing, volatile demand for a metered, fungible, scarce good is the classic setting in which spot and then derivative markets emerge.

The historical analogy is direct. Electricity and network bandwidth followed exactly this arc: each began as a service sold by its producers, became a metered and increasingly fungible commodity, and then developed spot markets, capacity markets, and derivatives as volatility and scarcity created demand for price discovery and hedging. Inference capacity now exhibits the same properties. The prediction is that it follows the same path.

What would prove this wrong — the falsification conditions, stated so the claim can be attacked on its merits: if capability fails to remain fungible across providers (if lock-in, differentiation, or regulation makes one provider's tokens non-substitutable for another's); if the substrate's scarcity eases enough that inference ceases to be capacity-constrained and price volatility collapses; or if regulation forecloses the trading of inference capacity as an instrument before such a market forms. Any of these would falsify the projection, and their absence is what makes it likely.

Why this belongs in a monograph on custody, and is not a market curiosity: if cognition becomes a traded instrument, the custody question reaches its most acute form. A financialized market in inference capacity means the ability to think at scale becomes an asset that can be accumulated, cornered, hedged, and speculated upon — control over cognition concentrated not only by owning the substrate but by owning the *market* in the substrate's output. The question the framework asks of any consequential system — who holds it, and is that power accountable — would then apply to a tradable claim on intelligence itself. This is why the projection is stated sharply rather than softly: it is meant to draw scrutiny and disagreement, because a structural prediction that provokes serious people to argue it is doing the work a research institution exists to do. What is offered here is the prediction and its reasoning; the mechanics of constructing such a market are deliberately not, because this is a warning about a trajectory, not a guide to accelerating it.

The Custody of the Surplus *(Normative)*

The threads converge on a single claim, and it is the claim the civilizational lens was built to reach. When intelligence becomes a strategically decisive resource, the economics of that resource reorganizes the civilization around it — and in this case the reorganization runs toward concentration: of chips, of capital, of power, and potentially of a market in cognition itself. The economics of abundant intelligence is therefore not a separate subject from its governance. It *is* the governance question in economic form. Authority over intelligence and capture of intelligence's value are two faces of the same concentration, and the framework's principle — that no concentration of power over intelligence should be beyond meaningful accountability — applies to the economic face exactly as it applies to the operational one. A world in which the surplus of cognition pools among the owners of the substrate with no corresponding structure of accountability is a custody failure denominated in dollars rather than in authority.

This monograph takes no position on which remedy is correct, and it is important that it does not, because the distributional question is genuinely contested and the Institute's role is to frame it rigorously, not adjudicate it politically. The range of responses under discussion is real and spans the spectrum: that competition and the relentless deflation of the token will disperse value over time without intervention; that the concentration at the substrate warrants the tools long applied to concentrated infrastructure — antitrust, common-carrier obligations, or redistribution of the rent; that public or shared compute capacity, a compute commons, could keep part of the substrate outside private control; and that edge and hybrid architectures could return value to users at the cost layer. Each carries its own trade-offs, evidence, and political commitments, and this chapter argues for none of them. It argues only that the question is unavoidable, that it is a custody question, and that when the economics of a scarce strategic resource reorganizes a civilization, the defining political question of that era becomes who holds the reorganized power and whether that holding is accountable. For the Intelligence Age, the answer to where that power is concentrating is now visible beneath the surface deflation: intelligence is becoming abundant and nearly free at the token, and the power to produce it, and to profit from producing it, is concentrating at the substrate beneath. Who holds that substrate, and whether that holding is accountable, is the economics of intelligence — and it is a custody question.

Chapter Source Notes

All figures reflect the landscape as of mid-2026 and were verified against current sources during drafting. Fast-moving area; re-verify before publication. Citations are to issuers and trackers; full bibliographic entries to be compiled in the References section.

1. **Historical / civilizational framing.** The claim is presented as an analytical lens — the economics of a civilization's scarce strategic resource as a *shaping pressure*, explicitly not as monocausal or deterministic history (per Page Zero Rule 4). The land/capital-and-energy/information illustrations are stated as "cannot be

explained without," not "was caused by." No specific historiographical source is asserted; if a scholarly anchor is wanted, standard economic-history treatments of feudal land tenure, industrial capital, and the information economy can be cited, but the argument is framed to stand as reasoning, not as a claim about consensus history.

2. **Inference cost decline ("LLMflation").** GPT-4-class inference ~$30/M input tokens (early 2023) → under $0.50/M (mid-2026); ~95% over two years, ~1,000x over three for a fixed quality bar. Term "LLMflation" coined by Andreessen Horowitz (a16z), ~10x/year. Epoch AI: median ~50x/year across benchmarks, range ~9x–900x by task. Output ~4–5x input cost. Frontier ~$2–15/M input, ~$10–75/M output; open self-hosted ~$0.10–2/M.

3. **Deflation paradox / demand growth.** Blended frontier token price ~$18.40 → ~$6.07/M (Q1 2025 → Q1 2026), ~67% YoY, yet ~73% of enterprises exceeded AI budgets (one reportedly exhausted its annual AI-coding budget in four months). Goldman Sachs projection ~24x token-consumption growth by 2030. Average enterprise AI budget ~$1.2M (2024) → ~$7M (2026). Jevons paradox: W. S. Jevons, *The Coal Question* (1865).

4. **Inference cost as profitability bottleneck / subsidy.** A widely discussed early-2026 analysis identified inference cost as the primary barrier to AI-provider profitability and argued that current API prices are subsidized by VC and hyperscaler cross-subsidy. *(G-1, resolved 2026-07-29: the specific named attribution in the source analysis could not be verified before publication and has been anonymized per founder ruling, superseding an earlier founder ruling of the same date that had accepted the named attribution as written. The named attribution may be restored in a future version if independently verified.)*

5. **Compute/chip concentration.** NVIDIA ~80–90% of AI accelerator market by revenue (2025), projected to ease toward ~75% (2026) as AMD/custom silicon scale; data-center revenue ~$75.2B in the quarter ended April 2026 (+92% YoY), ~$194B for FY2026. AMD ~5–7% share. Broadcom rising via custom accelerators for hyperscalers. NVIDIA 10-Q (FY2026) notes capital constraints on less-capitalized firms.

6. **Capital concentration.** "Big Five" hyperscaler capex projected >$600B in 2026 (+~36% YoY), ~$450B to AI infrastructure (analyst estimates; Dell'Oro, carboncredits.com summary).

7. **Scarcity rent in compute pricing.** Decomposition finding ~52–88% of a rented GPU-hour is demand/scarcity premium rather than hardware+energy cost (Silicon Analysts). Empirical anchor for the "surplus pools at the substrate" claim.

8. **Energy as binding constraint.** Global data-center electricity ~415 TWh (2024, ~1.5% of world) → ~485 TWh (2025) → projected ~950 TWh by 2030 (~3%); IEA. Data

centers = disproportionate share of demand *growth* (>20% by 2030, per IEA). AI rack power density 40–80+ kW (newest higher) vs 5–10 kW conventional; GB200 NVL72 ~120–140 kW. Geographic concentration: Northern Virginia (Dominion/PJM) largest cluster on earth; capacity-auction prices up >10x between 2024 and 2026–27 delivery years (LBNL 2024 DC Energy Report; EIA AEO 2026).

9. **Edge/hybrid counter-force.** Local-first inference at zero marginal token cost with escalation to paid frontier models is an emerging system-design pattern (routing/tiering literature, 2026). Limit: edge devices depend on the same concentrated chip/energy substrate — partial, not total.

10. **Tradable-market projection.** Stated as a Layer III projection, not fact. Structural drivers (fungibility, metering, volatility/scarcity, demand growth) grounded in notes 2–8. Historical analogy: commoditization → financialization of electricity and network bandwidth. Falsification conditions stated in text. No construction mechanics provided, by design.

ACT II — POWER

Where does intelligence exercise power, and who controls it? Power over intelligence is a matter of jurisdiction, authority, and enforcement, not technology.

Chapter 4 — The Geography of Intelligence

Jurisdiction, Not Cartography

The Question Act II Asks

The first act of this monograph established that intelligence is changing — that it was civilization's scarcest strategic resource, that it is becoming abundant, and that as it does, its economic value migrates and power reorganizes around the scarcity that remains. This act asks a different question. Not what intelligence is becoming, but *where its power resides, and which authority can actually govern it.* The three chapters of Act II concern power in its concrete forms: where it is exercised, who controls it, and whether any institution can reach it. This chapter takes the first of those — the geography — and it must begin by dispelling the obvious misreading of its own title.

The geography of intelligence is not a map of where servers sit. It is not a survey of cloud regions, data-center clusters, and undersea cables considered as points on a globe. The question that matters is not *where is the hardware* but *whose law runs where the hardware is, and can that law be enforced.* Geography reshapes custody because custody is territorial, and intelligence is becoming trans-territorial. That single tension — between authority that is bound to place and intelligence that is escaping it — is the subject of this chapter. It is a chapter about jurisdiction, not cartography, and the distinction is the whole argument.

Custody Requires an Enforcing Body *(Assessed)*

To see why geography bears on custody at all, one must be precise about what authority is. Authority over a system is not an abstraction that exists in the air; it is a capacity exercised by an institution that can actually reach the system — that can compel disclosure, conduct an audit, impose a penalty, order a halt, and hold a responsible party accountable when the system causes harm. A rule that no body can enforce is not governance; it is a wish. And the capacity to enforce is, in the world as it is presently organized, territorial. A state's writ runs within its borders and, with difficulty and contest, a certain distance beyond them. It does not run everywhere.

This is why custody and geography are bound together. The custody question this monograph asks — who holds authority over a consequential system, and can that authority be constrained and revoked — silently presupposes a *jurisdiction*: a territory within which some enforcing authority's power actually reaches. Change where a system physically resides, and you change which authority, if any, can reach it. Move it into a jurisdiction with weaker rules, and its custody weakens. Move it into the seams between jurisdictions, and its custody thins. Move it beyond territorial jurisdiction altogether, and the enforcing body that custody depends upon may simply not exist. Geography reshapes custody because geography determines reach, and reach is what turns a rule into a governance.

The remainder of this chapter traces what happens to custody as intelligence — and the compute that produces it — distributes across three increasingly untethered domains: across national borders, where authority already diverges from location; under the sea, where a different legal regime applies; and into orbit, where territorial authority does not exist at all. At each step, the enforcing body that custody requires becomes harder to locate, and at the last step it becomes a matter of the operator's own choosing.

The Terrestrial Split: Physical Reach Is Not Legal Authority *(Assessed)*

The tempting assumption is that the law governing a computer is the law of the country the computer sits in. Even on Earth, and even today, that assumption is false, and the gap between where a system physically is and whose law actually governs it is already a live and contested matter — the terrestrial preview of the deeper untethering to come.

Practitioners now distinguish three concepts that are routinely confused and are, in fact, distinct. *Data residency* is where data is physically stored. *Data localization* is a legal requirement that certain data remain within a country's borders. And *data sovereignty* is the question of whose law governs the data and who may compel access to it — which is not the same as where it sits. A European company can store its data in a European data center, achieving residency, and still find that data reachable by a foreign authority, because sovereignty follows a different thread than location.

The clearest illustration is the reach of a single statute. The United States' CLOUD Act, enacted in 2018, authorizes US authorities to compel US-based technology companies to

produce data in their custody regardless of where in the world that data is physically stored. The consequence is precise and counterintuitive: a data center in Frankfurt, operated by a US-headquartered provider, holds data that is physically in Germany and legally reachable by the United States. What determines the reach of US authority is not where the servers are but where the *provider is incorporated*. A provider incorporated and operating entirely within the European Union offers stronger protection against that extraterritorial reach than the same servers operated by a US-domiciled firm. Authority follows corporate domicile, not physical geography.

Against that pull, other jurisdictions assert their own. The European Union's General Data Protection Regulation does not principally require localization; it regulates the conditions under which personal data may *leave* the European Economic Area, through adequacy decisions and contractual safeguards. Other states go further into hard localization: certain categories of data — under China's and Saudi Arabia's regimes, for instance — must remain within national borders without exception. And the assertions are hardening. Amendments to China's Cybersecurity Law that took effect in January 2026 raised penalties, expanded the law's extraterritorial reach to cover essentially any activity deemed to harm the country's cybersecurity, and folded AI governance into the same framework.

The result is not a clean world in which each system answers to one clear authority. It is a world of overlapping and sometimes conflicting claims, in which a single deployment can be subject to one jurisdiction's compulsion to disclose and another's prohibition on the very transfer that disclosure would require — a genuine conflict of laws that the operator cannot fully resolve, only choose which to violate. And this contest has moved from the legal margins to the center of statecraft: the United States' 2026 national assessment of foreign trade barriers expanded its treatment of cloud computing and data sovereignty by roughly half in a single year, pairing the CLOUD Act's assertion of access with trade pressure on countries that localize — while those countries, in turn, erect certification and localization regimes that foreign providers cannot satisfy. Data sovereignty has become an instrument of geopolitics. Custody, in other words, is already fragmenting across terrestrial borders, and the fragmentation is accelerating. Physical reach and legal authority have come apart on the ground, well before anything leaves it.

The Re-Territorialization Response *(Assessed)*

The clearest evidence that this fragmentation is real is that an entire industry has formed to counter it. Faced with the gap between where compute sits and whose law governs it, states and enterprises are attempting to *re-territorialize* — to rebuild the alignment of physical location and legal authority that the cloud dissolved. Sovereign-cloud offerings designed to keep data and its governance within a specific legal boundary have emerged from the major providers, and a broader movement sometimes called geopatriation is pulling sensitive workloads back into controllable jurisdictions.

What matters for this chapter is what that response concedes. Re-territorialization is an admission that authority needs territory — that the way to restore custody over a system is to place it, physically and legally, where a known enforcing body can reach it. The sovereign-cloud movement is the market confirming the chapter's premise: when custody frays because location and jurisdiction have diverged, the remedy is to force them back together. That remedy works on Earth, where territory to retreat to still exists. It is precisely this remedy that fails as compute moves to the domains where territory, and the authority that rides on it, runs out.

Under the Sea: A Different Regime *(Assessed)*

The first domain where the terrestrial model begins to strain is the ocean. The feasibility of submerged data centers is not speculative; a submerged, sealed data center was operated and recovered years ago as a proof of concept, running reliably beneath the sea and demonstrating that the seabed is a viable venue for compute, drawn by the ocean's natural cooling. Placing compute in the water, however, moves it into a different legal regime.

The law of the sea partitions the ocean into zones of diminishing national authority — from territorial waters, where a coastal state's jurisdiction is strong, outward to the high seas, where no state holds territorial sovereignty and authority attaches instead to the flag a vessel or installation flies and to a patchwork of conventions. A data installation in a state's territorial waters remains within reach of that state; the same installation in international waters sits under a markedly thinner jurisdictional net, governed less by a territorial authority than by the nationality of its operator and whatever conventions apply. The ocean is thus the intermediate case: authority does not vanish, but it weakens and changes character, shifting from territory toward registration. It is a preview, at lower altitude, of what orbit does completely.

Into Orbit: Where Custody Comes Untethered *(Assessed reality; legal frame)*

The domain that breaks the terrestrial model outright is orbit, and it is essential to establish first that this is not a thought experiment. Compute is in space now, running real workloads, and the movement is accelerating with serious capital behind it.

A data-center-class GPU — an NVIDIA H100 — was placed in orbit in November 2025 aboard a small satellite and operated in the vacuum, radiation, and thermal cycling of space; the following month, a language model was trained in orbit, and the same platform has run inference workloads and processed satellite imagery from space. The company that did it reached a roughly one-billion-dollar valuation within months and has filed to build a constellation of tens of thousands of compute satellites; a second operator deployed dedicated orbital data-center nodes in early 2026 on an optical relay network; a merged rocket-and-AI enterprise of extraordinary scale has filed for up to a million compute satellites; and another major technology company is prototyping solar-powered satellite clusters carrying its own AI accelerators. By early 2026, for the first time, multiple operators were running production compute workloads in orbit simultaneously. Whether

this becomes a dominant venue is a separate and genuinely uncertain question, taken up below; that consequential compute is *already* operating in orbit is a present fact.

Now the legal frame, which is where orbit differs not in degree but in kind. Terrestrial jurisdiction, however contested, rests ultimately on territory — on a state's sovereignty over the ground its enforcing bodies can reach. In space, that foundation is removed by treaty. The Outer Space Treaty of 1967 prohibits any national appropriation of outer space by claim of sovereignty; no state may own a piece of orbit. There is, therefore, no territorial jurisdiction in space at all. In its place, the Treaty assigns authority by a different mechanism: the state on whose registry a space object is carried retains jurisdiction and control over that object and its personnel while in space — and it retains that jurisdiction *even as the object passes over the territory of every other state on Earth*. Jurisdiction over an orbital system follows its registration, not its location, because in orbit there is no location that confers authority.

The consequence for custody is severe and specific. An orbital data center is governed by the law of the single state it is registered with, and by no other — not the states it continuously overflies, not the states whose citizens' data it may process, not the states its decisions may affect. And because the operator, working through the launching states, has a hand in where its object is registered, the choice of governing authority becomes, in effect, the operator's own. This is the structure maritime shipping has long known as a flag of convenience: register where the obligations are lightest, and carry that light-touch jurisdiction with you everywhere you go. Applied to compute, it means the custody of an orbital system — who may audit it, constrain it, halt it, hold it accountable — can be selected by choosing a registry, and no overflown state has the territorial authority to override that choice.

Nor does the existing framework close the gap. The Treaty does make states internationally responsible for their national activities in space, including those of private companies, and makes the launching state liable for damage its space objects cause — so it is not the case that orbit is lawless. But this framework was built for state-launched objects performing defined missions, and it is, by the assessment of those who study it, silent on autonomous decision-making systems; international space law did not anticipate AI-operated missions and has no developed mechanism for governing a consequential, autonomous, continuously-operating intelligence in orbit. State responsibility under the Treaty routes through the state of registry — which returns the whole question to the operator's choice of flag. The enforcing body that custody requires still exists in principle; in practice it has become a variable the operator sets.

The complexity is not merely theoretical, and the nearest existing analogue shows it. The International Space Station is governed as a patchwork: its modules are owned by different nations, each nation's law applies within its own module, and the whole is held together by a web of bilateral and multilateral agreements negotiated to manage exactly the jurisdictional seams that arise when infrastructure crosses national ownership in a place without territory. That is the governance apparatus required for a single cooperative

station. A commercial orbital compute industry, with operators selecting registries and running autonomous workloads over the whole Earth, presents the same seams without the cooperative framework.

Why the Substrate Is Leaving the Ground *(Assessed — bridge to Chapter 3)*

It is worth asking why compute would go to sea or to orbit at all, and the answer connects this chapter to the last without repeating it. Chapter 3 established that the value of abundant intelligence concentrates at its physical substrate, and that the binding constraint on that substrate has become energy and cooling — power that is scarce, geographically concentrated, and increasingly hard to secure on terrestrial grids never designed for such density. The domains this chapter examines are, among other things, responses to that constraint. The seabed offers cooling; orbit offers abundant solar power and the vacuum's capacity to shed heat, which is precisely why the orbital operators target the most power-hungry, latency-tolerant workloads — the large training runs that terrestrial grids struggle to serve.

The two chapters therefore compound into a single trajectory that neither states alone. The economic pressure that concentrates the substrate (Chapter 3) is now also pushing that substrate *off national territory* (this chapter) — and off territory is exactly where the enforcing bodies that custody depends upon lose their reach. Value concentrates, and then relocates beyond the jurisdictions that might hold it accountable. The economics and the geography are not two separate stories about intelligence; they are one story about where its power is going and who, if anyone, can still govern it there.

How Far, How Fast: The Projection *(Projection — stated to be tested)*

The claims above about orbit's *legal* character are assessed: they follow from treaty text and present fact. The claim that orbital compute becomes a *dominant* venue is a projection, and must be labeled as one. Its drivers are real: orbit's energy and cooling advantages for training-scale workloads, the serious capital now committed, and the multiple operators already flying hardware. But its economics remain genuinely uncertain — informed skeptics estimate orbital compute at several times the cost per watt of terrestrial equivalents, and cost parity is contingent on launch costs falling to targets that depend on heavy-lift vehicles not yet in routine commercial service. The honest projection is therefore bounded: orbital compute is real and operating; whether it scales to a dominant share of consequential compute within the decade is uncertain and depends on launch economics that have not yet been proven. It would be falsified by launch costs failing to fall, by orbital reliability or servicing proving prohibitive, or by the economics simply not closing.

But — and this is the point the projection's uncertainty does not touch — the custody argument does not require orbital dominance. It requires only that *some* consequential compute operate beyond territorial jurisdiction, which is already the case. The governance gap opens the moment the first autonomous, consequential system runs in orbit under a

registry of convenience, not at the moment orbit wins the compute market. Even if orbital compute remains a niche, the niche is a jurisdictional void, and the void is occupied now.

The Custody of Territory *(Normative)*

The chapter's threads converge on a claim that reframes what geography means for the governance of intelligence. Custody depends on an enforcing body; enforcing bodies are, in the present order, territorial; and intelligence — with the compute that produces it — is becoming trans-territorial. It crosses borders, where legal authority already diverges from physical location and states contest one another's reach. It descends into the sea, where a thinner regime shifts authority from territory toward registration. And it rises into orbit, where territorial sovereignty does not exist by treaty, where jurisdiction follows the flag an operator chooses, and where the governing framework has no developed answer for autonomous systems at all. At each step the enforcing body that custody requires grows harder to locate, until at the last it becomes something the operator selects.

The geography of intelligence, then, is not about where intelligence lives. It is about whether any authority can reach where it lives. As intelligence distributes, custody does not merely fragment across many jurisdictions — though it does that, and the fragmentation is already a matter of statecraft. It can escape into the seams between jurisdictions, and into domains where territorial authority does not run at all. The defining geographic question of the Intelligence Age is therefore not cartographic but jurisdictional: not *where is the intelligence*, but *can the institutions capable of holding it accountable still reach it there* — and the emerging answer, as compute moves to sea and sky, is increasingly that they cannot, unless the governance of custody is deliberately extended to follow intelligence into the places it is going.

This monograph does not resolve how that extension should be built — whether through new treaty instruments for autonomous systems in orbit, through registration and accountability regimes that attach to operators rather than territory, through the re-territorialization the sovereign-cloud movement is already attempting, or through mechanisms not yet designed. It argues only that the problem is real, that it is a custody problem, and that geography has become one of its hardest dimensions. What happens when custody exceeds terrestrial jurisdiction entirely — in deep space, on the Moon, and beyond the reach of any Earth-bound enforcing body — is the subject of the chapter that follows.

Chapter Source Notes

All figures and legal characterizations reflect the landscape as of mid-2026 and were verified against current sources during drafting. Fast-moving area; re-verify before publication. Legal claims are to primary instruments (Outer Space Treaty, Registration Convention, CLOUD Act, GDPR) and current secondary analysis. Full bibliographic entries to be compiled in the References section.

1. **Residency / localization / sovereignty distinction.** Three distinct concepts routinely conflated: residency = physical storage location; localization = legal requirement to keep data in-country; sovereignty = whose law governs and who may compel access (NetApp 2026; Duality 2026). Three dominant models: open-transfer-with-safeguards, conditional localization, strict localization.

2. **CLOUD Act.** Clarifying Lawful Overseas Use of Data Act (US, 2018): authorizes US authorities to compel US-based providers to produce data in their custody regardless of physical storage location. Corollary: authority follows corporate domicile — a US-domiciled provider's EU servers remain US-reachable; an EU-incorporated provider offers stronger protection against US extraterritorial access (MassiveGRID 2026; NetApp 2026).

3. **GDPR / EU approach.** Regulates conditions for personal data leaving the EEA (adequacy decisions, standard contractual clauses) rather than mandating strict localization (Pandectes 2026; CookieYes 2026). Sovereign-cloud responses: Google Sovereign Cloud, Microsoft EU Data Boundary (Recording Law 2026).

4. **China CSL 2026 amendments.** Effective January 2026: maximum penalties raised to RMB 10 million, extraterritorial reach expanded to cover activity deemed harmful to China's cybersecurity, AI governance integrated into the CSL framework (Recording Law 2026). China PIPL / Saudi PDPL require strict local storage for certain categories.

5. **Data sovereignty as geopolitics.** US 2026 National Trade Estimate: cloud/data-localization references up ~50% YoY; US "two-pronged" strategy — CLOUD Act access + trade pressure on localizing states. South Korea CSAP: localization + domestic personnel + domestic encryption; no foreign CSP has obtained mid-tier+ certification. Colombia Resolution 372 (June 2025) cloud-procurement localization (Michael Geist 2026).

6. **Undersea data centers.** Submerged sealed data center operated and recovered as a proof of concept (Microsoft Project Natick, ~2018–2020), demonstrating seabed viability with natural cooling. Law of the sea (UNCLOS) partitions ocean into zones of diminishing national authority (territorial waters → high seas); high-seas installations attach authority to operator nationality/flag and applicable conventions rather than territorial sovereignty. **Specific current commercial undersea-DC deployments were not verified for this draft — keep claims at the proof-of-concept + legal-regime level unless verified before publication.**

7. **Orbital compute — verified reality.** First data-center-class GPU (NVIDIA H100) placed in orbit November 2, 2025 (Starcloud-1, ~60 kg); first LLM trained in orbit December 2025; platform also runs inference (Gemma) and processes Capella Space imagery. Starcloud: $170M Series A March 2026 at ~$1.1B valuation, ~$200M total raised, FCC filing for an 88,000-satellite (~20 GW) constellation; Starcloud-2 (Oct 2026) with multiple H100s + Blackwell. Axiom Space: first dedicated free-flyer

orbital data-center nodes January 11, 2026, on Kepler Communications' optical relay; ISS-connected node targeted 2027. SpaceX/xAI: merged February 2, 2026 (~$1.25T combined valuation); FCC filing for up to ~1 million compute satellites. Google Project Suncatcher: solar-powered TPU satellite clusters with optical links, prototype ~2027. February 2026 = first month with multiple orbital operators running production workloads simultaneously (Data Center Frontier; Introl; GeekWire; TechCrunch; Quartz; Luminix — all 2026).

8. **Orbital economics (uncertain — basis for the projection).** Proponents cite solar ~$0.005/kWh and zero cooling water; skeptics (e.g., Varda) estimate ~3x cost per watt vs terrestrial. Cost parity contingent on launch cost targets (~$500/kg) that depend on heavy-lift vehicles not yet in routine commercial service (Starship record ~58% across early flights). Target market: latency-tolerant training (2–6 GW facilities; training runs projected toward ~2e29 FLOP by 2030) that terrestrial grids struggle to serve.

9. **Space-law jurisdiction (primary law).** Outer Space Treaty (1967): Art. II prohibits national appropriation of outer space by sovereignty claim (no territorial jurisdiction in space); Art. VIII — the state of registry retains jurisdiction and control over a space object and its personnel while in space, including as it passes over other states' territory; Art. VI — states bear international responsibility for national activities in space, including private entities; Art. VII + Liability Convention (1972) — launching state liable for damage. Registration Convention (1975/76) — launching state registers the object with the UN Secretary-General; joint launching states determine registry. Governance gap: OST framework "did not anticipate AI-operated missions" and is silent on autonomous decision-making systems (arXiv 2025 Space AI analysis; ScienceDirect 2024 on registration-practice inadequacy). ISS analogue: modules under differing national law, held together by bilateral/multilateral MOUs.

10. **"Flag of convenience" analogy.** Drawn to maritime registration (register under the lightest-obligation flag and carry that jurisdiction globally); applied to orbital compute via the registry-based jurisdiction of OST Art. VIII. Presented as structural analogy, argued in-text.

Chapter 5 — Intelligence Beyond Earth

What Happens When Custody Exceeds the Human Loop

Where the Last Chapter Stopped

The previous chapter followed custody as it left national territory and found that authority, which depends on an enforcing body able to reach a system, grows thinner as intelligence disperses — until, in orbit, territorial jurisdiction disappears by treaty and authority becomes a flag the operator chooses. But that argument, for all that it untethers custody from territory, left one thing intact. A satellite in orbit is legally hard to reach and physically easy to reach. It is a few hundred milliseconds away at the speed of light. A human

operator can still watch what it does in something close to real time, still send a command that arrives before the situation has changed, still trip a kill-switch and have it take effect. The orbital custody problem of Chapter 4 is a problem of *law* — a governable system in a jurisdictional void.

This chapter concerns the distance at which that last tether snaps. As intelligence moves outward — to the Moon, to Mars, into deep space — a second and more fundamental untethering occurs, and it owes nothing to law. It is imposed by physics. Beyond a certain distance, measured not in kilometers but in the time light takes to cross it, a human being can no longer be in the loop, because by the time a signal arrives and a reply returns, the moment that required a decision is long past. Chapter 4 was custody losing its enforcing body. This chapter is custody losing the human loop itself. And where the enforcing body can, in principle, be reconstructed by better law, the human loop cannot be reconstructed by anything, because nothing travels faster than light.

The Physics of the Severed Loop *(Assessed)*

The governing quantity is round-trip light-time: the interval between sending a signal and receiving a response, which no technology can shorten because it is set by the speed of light and the distance to be crossed. That interval scales from the negligible to the prohibitive as one moves outward, and the thresholds are not matters of opinion.

In low Earth orbit, round-trip light-time is a fraction of a second — the regime of Chapter 4, where real-time human oversight remains possible. At the Moon, one-way light-time is roughly 1.3 seconds, a round trip of a few seconds. At Mars, the distance varies with orbital geometry, but one-way light-time ranges from about four minutes at closest approach to around twenty-four minutes at its farthest — a round trip of eight to nearly fifty minutes. In the deeper solar system it grows to hours. These are floors, not estimates: they are the fastest a message can possibly travel, before any processing, queuing, or intermittent-contact delay is added.

What matters for custody is that these intervals cross the thresholds of human supervisory control at well-understood points, and the crossing has been studied by the institutions that fly the missions rather than merely asserted here. Research on spaceflight operations finds that communication delays on the order of fifty seconds and above measurably degrade individual well-being, team cohesion, and task performance — and that even the far shorter delays anticipated for lunar operations, on the order of four to twelve seconds one way, may already make it difficult or infeasible for ground control to provide real-time support during complex, time-critical tasks. At Mars-like delays the conclusion is categorical: real-time oversight and guidance from ground control is, in the assessment of that research, impossible. This is not a claim about the distant future. It is the reason the field has, for years, been converging on a design assumption stated plainly in the mission-planning literature: that continuous human oversight is not always available, and systems must be built to remain safe and productive between contacts with Earth.

The consequence for this monograph's argument is direct and, once seen, inescapable. Every custody mechanism the framework relies upon — human override, real-time audit, the bounded-latency kill-switch that Chapter 9 requires to halt a system before its actions propagate — presupposes that a human can act within the window in which action still matters. Light-latency closes that window. At the Moon it narrows it to the edge of usability; at Mars it shuts it entirely. The severing of the loop is not a governance failure to be corrected by better institutions. It is a physical boundary, and the governance question is what custody can possibly mean on the far side of it.

The Moon: Where the Loop First Frays *(Assessed reality; near-term)*

The Moon is the nearest place the problem appears, and it is not hypothetical. Cislunar operations are an active domain: a small autonomous-navigation technology mission has already operated in lunar orbit, demonstrating autonomous positioning and delay-tolerant networking — a communications architecture explicitly designed for the long delays and frequent signal gaps of deep space, in which a spacecraft stores information when no connection is available and forwards it automatically when contact is restored. The very existence of delay-tolerant networking as a mature, deployed technology is an admission encoded in engineering: the assumption of a continuous connection to Earth has already been abandoned for operations beyond low orbit.

Commercial lunar compute is arriving on the same near horizon. A commercial operator has announced plans to place data-center payloads on the lunar surface as early as 2027, and the Moon is widely treated as the proving ground for off-world computing precisely because it presents a manageable subset of Mars's challenges — vacuum, radiation, and communication delays measured in seconds rather than minutes. Every capability validated there — radiation-hardened hardware, autonomous management software, operation through signal gaps — is a design input for what follows further out.

At lunar distance the loop frays rather than snaps. A few seconds of round-trip delay does not make human oversight impossible, but it makes it slow, effortful, and unreliable for anything time-critical — which is why the same missions that keep a human nominally in the loop are simultaneously building the onboard autonomy to act without waiting. The Moon is therefore the transition zone: close enough that human authority still reaches, far enough that systems are already being designed to exercise consequential judgment on their own between contacts. It is where the custody question stops being about jurisdiction and starts being about the clock.

Mars: Where Autonomy Stops Being a Choice *(Assessed physics; projected deployment)*

Mars is where the argument of this chapter reaches its center, because at Mars distance the delegation of authority to machines is no longer a design preference that could be decided either way. It becomes a requirement the environment imposes.

Consider what a Mars deployment must do. A habitat's life-support system detecting a pressure failure, a power grid responding to a fault, a lander avoiding a hazard during descent, a medical system responding to an emergency — each of these is a decision that must be made in seconds, and each is separated from any human on Earth by a round trip of eight to fifty minutes. There is no architecture, no relay network, no compression that closes that gap; the delay is the distance times the speed of light. A system on Mars that waited for Earth to authorize its response to a time-critical event would, by the time the authorization arrived, be responding to a situation that no longer exists. The mission-planning literature states the conclusion directly: for Mars surface operations, communication delays make Earth-based control impractical, and autonomous guidance, hazard avoidance, and real-time decision support are essential rather than optional.

This is the collision the monograph has been building toward, and it must be stated at full strength because it is the chapter's reason to exist. Chapter 9 proposes Prohibited Authority Thresholds — categories of authority that must remain human regardless of a system's capability, among them irreversible control of critical infrastructure and, at the limit, decisions bearing on human life. Those thresholds are the framework's firmest normative commitment: the permanent human line. At Mars distance, the speed of light runs that line over. A Mars settlement's life-support and infrastructure systems *must* hold exactly the authority the framework says must remain human, because the human who would otherwise hold it is forty minutes away and the decision cannot wait forty minutes. The permanent human line and the speed of light are in direct conflict, and physics does not yield.

A skeptic's reply is immediate and must be answered rather than avoided: then do not place such systems beyond the human loop — keep consequential authority on Earth and send only what can safely wait. The answer is that the pressure to delegate is not a choice the designer is free to decline. It is imposed by the mission itself. One cannot operate a crewed Mars presence, or an uncrewed installation that must survive between resupply, while routing its time-critical decisions through a forty-minute round trip; the environment forecloses that option before the designer reaches it. Autonomy at interplanetary distance is not a temptation to be resisted by good governance. It is a condition of operating there at all. The question is therefore not whether to delegate authority the framework would prohibit delegating, but how to govern authority that must be delegated to a system beyond the reach of the mechanisms custody normally depends on.

The deployment of consequential machine intelligence on Mars remains a projection, and it must be labeled as one. Its anchors are real — heavy-lift development aimed explicitly at uncrewed Mars cargo, stated commercial and national ambitions for a sustained Martian presence, the near-term lunar computing that would prove the hardware — but they are, in the honest words of the sources that catalog them, publicly stated ambitions rather than validated engineering outcomes, subject to substantial uncertainty in launch cadence, economics, and much else. What is *not* a projection is the physics: if consequential intelligence operates at Mars distance at all, the human loop cannot govern it in real time.

The uncertainty attaches to the timeline of deployment, not to the custody consequence, which follows the moment the first such system arrives.

What Custody Can Mean With the Loop Severed *(Normative)*

If real-time human authority is physically unavailable beyond a certain distance, and if some consequential intelligence will nonetheless operate there, then custody cannot mean what it means on Earth — but it need not mean nothing. The severing of the real-time loop changes the *form* custody must take, and identifying that form is the constructive task this chapter hands forward.

Where authority cannot be exercised in real time, it must be exercised in advance and after the fact. In advance: through the objectives, constraints, and prohibited actions built into the system before it departs — the boundaries it carries with it, which function as the only human authority present at the moment of a decision. This is why the framework's read-only, human-controlled core objectives and its prohibited-authority thresholds matter more, not less, at distance: when no human can intervene during operation, the constraints fixed before operation are the whole of the custody that reaches the far side of the light-delay. After the fact: through the complete, tamper-evident record that delay-tolerant networking can return to Earth, enabling audit, accountability, and correction of the system's future behavior even when its past behavior could not be supervised as it happened. Custody becomes bracketing rather than steering — bounding what the system may do before it acts and reviewing what it did once the record arrives, with the real-time middle, where terrestrial custody lives, physically removed.

This is a weaker form of custody, and honesty requires saying so. Pre-loaded constraints can be wrong, can meet situations their designers did not anticipate, and cannot be revised in the moment; after-the-fact audit cannot undo a consequence already realized minutes or hours away. The monograph does not claim that distance-custody is as good as presence-custody. It claims something more careful: that as consequential intelligence moves beyond the human loop, the authority that can accompany it shrinks to what can be fixed in advance and reviewed in arrears, and that this shrinkage is itself one of the strongest reasons to think hard, now, about which authorities should cross that boundary at all. The permanent human line cannot be enforced in real time beyond the Moon. What can be decided is what a system is permitted to carry across it.

Coda: The Limit the Argument Points Toward *(Projection — flagged, not load-bearing)*

The escalation of this chapter has an endpoint that lies beyond anything now buildable, and it is worth naming briefly, clearly marked as the projection it is, because it is where the logic terminates rather than a claim the chapter rests on.

At interstellar distance, light-time is measured in years. A system dispatched beyond the solar system — a probe, or in the furthest imaginings an autonomous industrial or exploratory intelligence — would be governed by round-trip delays that exceed not only the

reach of any operator but potentially the working lifetime of the humans and institutions that launched it. Such a system would carry the entirety of its custody in its pre-loaded objectives and constraints, with no possibility of real-time authority and, at sufficient distance, no meaningful possibility of after-the-fact correction either, because the situation would be over — and possibly the mission complete or the system beyond recovery — long before any signal could return. This is custody reduced to its absolute limit: authority that must be entirely encoded in advance, exercised by no one, and answerable to institutions that may no longer exist when it acts.

Nothing in the near-term argument of this monograph depends on that case, and it is offered as a horizon rather than a forecast: there is no consequential interstellar intelligence, and there may never be. But the limit clarifies the principle by showing its extreme. The entire trajectory of this chapter — Moon, Mars, deep space, and beyond — is a single monotonic movement in which real-time human authority recedes and pre-loaded, encoded authority becomes the whole of custody. The interstellar case is simply that movement completed. It tells us what is at stake in the near cases: every step outward converts custody from something exercised into something inscribed, and the governance question for an intelligence age that is beginning, genuinely, to leave the Earth is how much authority we are prepared to inscribe into systems we will no longer be able to reach.

Handing Forward

This chapter has argued that beyond a certain light-distance the human loop on which custody depends is severed by physics, not law; that the Moon frays it and Mars breaks it; that at interplanetary distance the delegation of prohibited authority becomes a condition of operating rather than a choice; and that custody, in that regime, contracts to what can be fixed in advance and audited in arrears. It has, in doing so, pressed against a question it did not fully open: what changes when a system does not merely execute pre-loaded instructions but plans, persists, and acts toward goals on its own — when delegated intelligence becomes delegated agency. Distance forces autonomy; the next chapter asks what autonomy, once forced, actually is. And both feed the chapter after it, where the custody problem is finally stated in full: not only who holds authority over intelligence, but where — on Earth or beyond it, in real time or only in advance — that authority can be held at all.

Chapter Source Notes

Latency figures are settled physics and require no source beyond the speed of light and orbital mechanics. Claims about the effect of delay on human oversight, and about current lunar/cislunar/deep-space programs, were verified against current sources during drafting. Deployment claims are tagged as projection. Re-verify program specifics before publication; this is a fast-moving area.

1. **Round-trip light-time (assessed physics).** Low Earth orbit: sub-second. Moon: ~1.3 s one-way (~2.6 s round trip). Mars: ~4 min (closest approach) to ~24 min (farthest) one-way; ~8–48 min round trip, varying with orbital geometry. Deep solar

system: hours. These are physical floors (distance ÷ c), before any processing/queuing/contact-gap delay.

2. **Delay degrades/defeats human oversight (assessed).** Research on spaceflight operations: 50+ second delays adversely affect individual well-being, team cohesion, and task performance (citing Kintz et al. 2016; Larson et al. 2019); Artemis-like 4–12 s one-way delays may already make real-time ground support difficult or infeasible for complex, time-critical tasks (NASA NTRS study, HSIA/cislunar communication-delay lab study, 2025). Mars-like delays: "real-time oversight and guidance from ground control is impossible" (same). Mission architecture increasingly "designed around the assumption that continuous human oversight is not always available" (arXiv Space AI survey, 2025/2026).

3. **Cislunar autonomy + delay-tolerant networking (assessed).** NASA CAPSTONE (launched June 2022; primary objectives 2023; extended mission through Dec 2025): demonstrated Cislunar Autonomous Positioning System (autonomous navigation) and tested delay/disruption-tolerant networking (DTN) — store-and-forward architecture for long delays and signal gaps (NASA, 2025–2026). DTN's maturity itself encodes the abandonment of continuous-connection assumptions beyond LEO.

4. **Lunar commercial compute (near-term anchor).** Commercial operator (Lonestar Data Holdings) announced lunar-surface data-center payloads as early as 2027; Moon treated as proving ground for off-world compute (subset of Mars challenges: vacuum, radiation, second-scale delay) (industry reporting, 2026). Near-term but pre-deployment; treat as anchor, not fact.

5. **Mars autonomy necessity (assessed physics; projected deployment).** For Mars surface operations, communication delays make Earth-based control impractical; autonomous guidance, hazard avoidance, and real-time decision support essential (arXiv Space AI survey, 2025/2026). Deployment of consequential machine intelligence on Mars is projection: anchors include Starship development aimed at uncrewed Mars cargo and stated ambitions for sustained Martian presence — characterized in sources as "publicly stated ambitions, not validated engineering outcomes," subject to substantial uncertainty (launch cadence, economics, regulatory, etc.).

6. **Deep-space / interstellar coda (projection — flagged, non-load-bearing).** Interstellar light-time in years, potentially exceeding operator/institutional lifetimes; custody reduced to pre-loaded, encoded authority with no real-time or (at sufficient distance) after-the-fact correction. Long-range orbital/lunar/deep-space compute roadmaps (e.g., Terafab ~1 TW/yr target with ~80% orbital allocation; SpaceX "AI1" ~120 kW orbital data-center satellite unveiled June 2026; million-satellite FCC filing; Google Project Suncatcher; Blue Origin) are cited by sources as "publicly

stated commercial ambitions, not validated engineering outcomes." Offered as horizon, not forecast.

7. **Framework tie-in (per Ch.9).** The custody mechanisms latency defeats — human override, real-time audit, bounded-latency (sub-second) kill-switch — are specified in Chapter 9. The "distance-custody" form proposed here (pre-loaded read-only core objectives + prohibited-authority thresholds as the only authority present at decision time; DTN-returned tamper-evident logs for after-the-fact audit) is an application of Ch.9's obligations to the severed-loop regime, not a new mechanism.

Chapter 6 — When Intelligence Stops Waiting

From Delegated Intelligence to Delegated Agency

The Question the Previous Chapters Forced

The last two chapters pushed intelligence away from the human who governs it — first legally, as it crossed into jurisdictions no single authority could reach, and then physically, as it moved to distances where no human could act inside the window a decision required. Both arguments shared an unexamined assumption, and this chapter examines it. They assumed that the thing being governed was, at bottom, a *tool* — a system that does what it is instructed to do, where the custody problem is the problem of controlling the instructions and controlling access to the machine. Distance and jurisdiction made the tool hard to reach. But it remained, in the picture drawn so far, a tool.

This chapter argues that the tool assumption is failing, and that its failure is what turns custody from a manageable problem into an urgent one. As machine intelligence acquires agency — as it begins to plan, persist, use tools, coordinate, and act without waiting for the next human instruction — the object of governance changes underneath the governance built to manage it. One does not exercise custody over an agent the way one exercises custody over a tool, and nearly every concept the preceding chapters relied upon was designed for the tool. The previous chapter said that distance *forces* autonomy. This chapter asks what autonomy, once present, actually is — and finds that it quietly dissolves the assumption custody has been resting on all along.

Agency Is Not Consciousness *(Assessed — definitional, load-bearing)*

Before anything can be argued, one distinction must be established and then held without exception for the rest of the chapter, because the entire credibility of what follows depends on it. The title of this chapter — *when intelligence stops waiting* — invites a misreading, and the misreading would be fatal to the argument. It invites the reader to hear that intelligence *wakes up*: that it becomes conscious, that it experiences, that it comes to want things. This chapter makes no such claim, entertains no such claim, and does not require any such claim. It is not about consciousness. It is about agency, and the two are entirely different, and conflating them is the single most common way that serious discussion of these systems collapses into science fiction.

Agency, as this monograph uses the word, is strictly behavioral and observable. A system has agency to the degree that it can: form a plan across multiple steps toward a goal; persist in pursuing that goal across time, remembering its progress; select and use tools — calling functions, executing code, operating other software — to advance the goal; coordinate with other systems, dividing work and integrating results; and execute this chain of action without a human authorizing each step. Every element of that definition refers to something a system *does*, observable from the outside, measurable in deployment. None of it refers to anything the system might *experience*, because nothing the system might or might not experience bears on the definition.

This is not an evasion of a hard question; it is a refusal to let an irrelevant question hijack a governance argument. Whether there is anything it is like to be one of these systems — whether any experience, awareness, or genuine wanting exists inside it — is a genuine and difficult question, and this monograph takes no position on it, in either direction. It does not need to. The custody problem posed by agency does not depend on the answer. A system that in fact plans, persists, and acts on the world without waiting presents the same governance challenge whether that behavior arises from something worth calling mind or from mechanism entirely devoid of inner life. Custody governs what a system *does*. What, if anything, it *is* on the inside is a question for philosophy, and this chapter leaves it there, deliberately and completely. When it needs a word for what drives such a system, that word is *its objectives* — the goals it was given — never its desires.

Agency Is Already Here, and It Is Bounded *(Assessed)*

With the definition fixed, the empirical claim can be stated precisely, and it has two halves that must be held together, because taking either without the other produces a distortion.

The first half: agency, in the behavioral sense defined above, is not a future prospect. It is a present and deployed reality. Systems in ordinary commercial use in 2026 already plan across steps, decompose goals into sub-tasks, select and invoke tools, run as coordinated groups of specialized sub-agents, and retain memory across sessions to resume work over extended periods. Independent assessment of the field's trajectory records the shift plainly: where a year earlier such systems could complete only small tasks in narrow demonstrations, they can now plan and carry out multi-step tasks over extended time horizons. The architecture of these systems is explicitly agentic — planners that decompose goals, actors that execute, memory layers that persist state across cycles, coordination protocols that let one agent invoke another. The reactive assistant that answers a single prompt in isolation, which is what most people still picture when they think of these systems, is already the previous generation. The current generation acts.

The second half, and it must be given equal weight: this agency is bounded, brittle, and far from the autonomous intelligence of speculation. The same sources that document the capability document its limits candidly. Reliability is variable; performance that dazzles in a controlled demonstration frequently degrades in real, messy use — a gap so well recognized that it has its own literature examining why agentic systems impress in demos

and fall apart in production. The systems that actually ship at scale cluster at the cautious end of every axis: scoped permissions rather than open-ended authority, external and inspectable state rather than opaque memory, and human approval retained at high-stakes steps rather than removed. The most successful enterprise deployments look less like an autonomous mind and more like a capable hybrid of conventional automation and a language model, kept on a short leash by design. Agency today is real, and it is also limited, supervised, and frequently unreliable.

Holding both halves is essential, and not merely for accuracy. A chapter that asserted only the first half — agents are here — would slide into the breathless register that discredits this kind of argument. A chapter that asserted only the second — they're brittle and bounded — would miss the point entirely, because the custody problem does not require autonomous or reliable agency. It requires only *enough* agency that the tool assumption no longer holds — that the system, on some occasions, plans and acts across steps without a human in each one. That threshold is already crossed. The governance consequence follows from the crossing, not from any projection of where the capability goes next.

The Tool Assumption Dissolves *(Assessed & Normative)*

Here is why the crossing matters, and it is the load-bearing argument of the chapter. Every mechanism of custody that the preceding chapters invoked was built for a tool, and a tool has a specific structure: a human decides, the tool executes, and the human's decision is the locus of authority and accountability. The hammer does not choose where to strike; the person swinging it does. Custody over a tool is therefore custody over its use — control the human decision and the access to the instrument, and you have controlled the tool. This is the model beneath professional licensing, beneath the assignment of legal responsibility to operators, beneath the whole apparatus by which societies govern powerful instruments.

An agent breaks this structure at its joint. When a system plans across steps and acts without a human authorizing each one, the human decision is no longer located at the moment of action. It has moved upstream — to the goal that was set, the constraints that were imposed, the authority that was delegated — and between that upstream decision and the downstream action lies a chain of choices the system made on its own. The human still bears responsibility, but the human's actual control has thinned to the setting of objectives and boundaries, while the consequential specifics — which tools to invoke, in what order, in response to what circumstance — are determined by the system in the moment, beyond the reach of the decision that authorized it. This is no longer the custody of a tool being used. It is the custody of an agent being *delegated to*, and delegation is a fundamentally different relationship than use.

The distinction is not semantic; it is the whole of why custody becomes urgent. To use a tool is to remain the author of each of its actions. To delegate to an agent is to authorize a category of action and then to be answerable for specifics one did not choose and could not foresee. The governance concepts the preceding chapters leaned on — human

override, real-time supervision, control of the instrument — are the vocabulary of use. They are only partially applicable to delegation, and they grow less applicable as the chain between the authorizing decision and the executed action grows longer and more autonomous. This is what it means to say the object of governance has changed underneath the governance. We built our concepts of control for tools. We are increasingly deploying agents. And the gap between the two is exactly the space in which custody is lost — not because anyone intends to lose it, but because the instruments of control were designed for a different kind of thing.

The Objection, and Why It Does Not Land *(Assessed)*

The argument invites a sharp objection, and it deserves a direct answer rather than avoidance, because the answer is where the chapter's discipline pays off. The objection runs: these systems are not really agents at all. They are statistical models predicting the next token, following patterns in their training and their instructions; to speak of them "planning" or "choosing" or "acting on their own" is to anthropomorphize autocomplete, to project agency onto machinery that merely computes. On this view the whole argument of the chapter is a category error dressed up as analysis.

The answer is that the objection, whether or not it is correct about the machinery, is irrelevant to the custody problem — and seeing why is the point. The objection is a claim about what is happening *inside* the system: about whether the observable behavior reflects "real" agency or a sophisticated imitation of it. But custody does not govern what happens inside a system. It governs what a system *does* in the world. A system that in fact decomposes a goal, selects tools, and executes a chain of consequential actions without a human in each step presents exactly the same governance challenge whether that behavior is "genuine" agency or an extraordinarily capable mechanism that merely produces the same outputs. The supply chain is reordered, the funds are moved, the infrastructure is reconfigured, the response is executed — and the consequences in the world are identical regardless of which description of the internals is true. This is precisely why the chapter defined agency behaviorally and refused the question of consciousness: so that the argument would stand on what can be observed and would not depend on the contested metaphysics the objection wants to fight about. The skeptic and the believer can disagree entirely about what these systems are, and the custody problem is unchanged, because the custody problem is a problem about behavior and its consequences, and the behavior is not in dispute. To insist on the "it's just prediction" framing is to answer a question about mechanism when the governance question is about action — and action is what has to be governed.

Why This Makes Custody Urgent *(Normative)*

The three preceding chapters could each be read, by an unhurried reader, as describing a problem that was serious but not yet pressing — a scarcity that was ending, an economics that was concentrating, a geography that was fragmenting, all unfolding over years. This chapter removes that comfort, and it is meant to.

Agency is what makes the custody problem present tense. When intelligence was a tool, the failure of our governance concepts was a latent risk: the concepts were built for tools, and we were deploying tools, so the fit held even as the tools grew powerful. Agency breaks the fit *now*, at whatever level of reliability the systems currently possess, because the tool assumption fails the moment a system acts across steps without a human in each one — and that is a description of systems in ordinary deployment today, not a forecast. The urgency does not come from imagining highly autonomous future agents, though the trajectory points that way and the following chapters take it seriously. It comes from the fact that the object our custody concepts were designed to govern has already, in part, been replaced by an object those concepts do not fit. We are governing agents with the tools of tool-governance, and the mismatch is not waiting to arrive. It is here, thinly at first, growing with every increment of autonomy the systems acquire.

That the current agency is bounded and brittle does not soften this; it sets the terms. The window in which the mismatch is still small — in which agents are supervised, scoped, and frequently interrupted — is precisely the window in which the governance built for agents rather than tools can still be designed deliberately, before the autonomy deepens and the chain between authorization and action grows longer than any retrofit can bridge. The bounded agency of the present is not a reason to wait. It is the reason not to.

What Increasing Autonomy Would Add *(Projection — stated to be tested)*

The argument above rests entirely on present, assessed fact and needs nothing from the future. But the trajectory is real and must be characterized honestly, because the custody problem this chapter describes does not stay the same size. The direction of development — pursued deliberately, invested in heavily, and visible in the research literature — is toward agents that plan over longer horizons, sustain execution for longer before failing, coordinate in larger and less supervised groups, and require human approval at fewer points. This is a projection, and its pace is genuinely uncertain: reliability has proven the binding constraint, and credible practitioners disagree about how quickly the brittle demonstrations of today become the dependable systems of tomorrow. It would be falsified, or at least stalled, if reliability at long horizons proves far harder to achieve than capability at short ones, which some evidence already suggests.

But the direction, if not the pace, is not seriously in doubt, and what it adds to the custody problem is not a new problem but an intensification of the present one. Every increment of autonomy lengthens the chain between the human's upstream authorization and the system's downstream action, and every increment thins the applicability of the tool-governance concepts further. The custody problem does not change character as autonomy grows; it grows more acute along the exact axis this chapter has identified. Which is why the response cannot wait on resolving how fast the trajectory moves. The mismatch is already here at low autonomy; higher autonomy makes it worse; and the governance that fits an agent rather than a tool is the subject the rest of this monograph must supply.

Handing Forward

This chapter has argued that agency — defined strictly as observable behavior, and held rigorously apart from any question of consciousness — is already present in deployed systems, bounded and brittle but past the threshold that matters; that its arrival dissolves the tool assumption on which the custody concepts of the preceding chapters silently rested; that this dissolution is what makes the custody problem urgent rather than latent, present rather than looming; and that the objection which dismisses these systems as "mere prediction" does not touch the argument, because custody governs behavior and consequence, not internal mechanism.

It has, in doing so, assembled the last piece the monograph needs before it can state its central problem in full. The first act established that intelligence is becoming abundant and its economics is reorganizing power. The second act has now shown where that intelligence goes — beyond jurisdictions, beyond the human loop, and beyond the tool assumption itself — and what it becomes when it gets there: not an instrument used, but an agent delegated to, at distances and behind legal veils where the delegation cannot be supervised. The next chapter takes up what all of this has been building toward. Given intelligence that is abundant, concentrated, trans-territorial, beyond the real-time reach of any human, and possessed of genuine agency — where, in all of that, must authority remain human, and how? That is the custody problem, and it can now be stated in full.

Chapter Source Notes

The agency/consciousness distinction and the tool/delegation argument are conceptual and argued in-text. The empirical claim — that current systems already exhibit behavioral agency, and that this agency is simultaneously real and bounded/brittle — was verified against current sources during drafting. Projection is tagged. Re-verify capability specifics before publication; fast-moving area.

1. **Behavioral definition of agency.** Plan across steps; persist toward a goal with memory; select and use tools (function/API calls, code execution, computer/browser use); coordinate across multiple agents; execute without per-step human authorization. Defined strictly behaviorally and observably; the chapter takes no position on consciousness/experience/sentience in either direction, by design — the custody argument does not depend on it.

2. **Agency is present and deployed (assessed — first half).** 2026 commercial systems plan, decompose goals into sub-tasks, invoke tools, run coordinated multi-agent architectures (planner/actor decomposition; specialized sub-agents), and persist memory across sessions for long-horizon execution (hours/days/months). International AI Safety Report (2025 update): agents shifted within ~a year from small tasks in narrow demos to planning and completing multi-step tasks over extended horizons. Frameworks and platforms (e.g., LangGraph, CrewAI, Perplexity's long-horizon agent, Microsoft CORPGEN hierarchical planning/memory) document planner/actor/memory/coordination architectures as

standard. METR tracks progress by sustainable task duration/autonomy rather than static benchmarks.

3. **Agency is bounded and brittle (assessed — second half, equal weight).** Reliability variable; demo-to-production gap well documented (incl. Stanford/Harvard analysis of why agentic systems impress in demos and fail in real use). Practitioner consensus (e.g., Kingy AI 2026 state-of-agents): most shipping "agentic" products cluster at the cautious end — scoped permissions, external/inspectable state, iterative replanning, single-agent, human approval at high-stakes steps — closer to capable RPA+LLM hybrids than autonomous minds. Both halves held together deliberately.

4. **Tool → delegation argument (normative/conceptual).** Tool-governance locates authority and accountability at the human's moment-of-use decision; agency relocates the human decision upstream (goals/constraints/delegated authority), inserting an autonomous chain between authorization and action. Custody concepts from prior chapters (human override, real-time supervision, control of the instrument) are the vocabulary of *use* and only partially apply to *delegation*; applicability thins as the authorization→action chain lengthens. Argued in-text.

5. **"It's just prediction" objection (assessed rebuttal).** The objection is a claim about internal mechanism; custody governs external behavior and consequence. A system that in fact plans/acts across steps presents the same governance challenge regardless of whether the behavior is "genuine" agency or capable imitation, because real-world consequences are identical. The behavioral definition (note 1) is what makes the argument independent of the contested metaphysics.

6. **Increasing autonomy (projection).** Direction — longer horizons, longer sustained execution, larger/less-supervised coordination, fewer approval points — is deliberately pursued and visible in the literature; pace uncertain, with reliability at long horizons as the binding constraint (some evidence that long-horizon reliability lags short-horizon capability). Tagged projection; the chapter's core argument requires none of it.

ACT III — GOVERNANCE

How should that power be governed? The custody problem in full, its adversaries, and the instrument that answers it.

Chapter 7 — The Custody Problem

Where Authority Must Remain Human, and Why

One Problem, Not Six

The six chapters preceding this one each followed the governance of intelligence to a place where it began to fail, and each failure looked, on its own, like a distinct difficulty. Intelligence was becoming abundant, dissolving the scarcity that institutions were built to manage. Its economics was concentrating power at a physical substrate few could own. It was crossing borders faster than any jurisdiction could follow, and rising into orbit where jurisdiction does not exist. It was moving to distances where the human loop is severed by the speed of light. And it was acquiring agency, dissolving the assumption that the thing being governed is a tool a human wields. Five axes, five apparent problems, discovered one at a time.

This chapter's first task is to show that they are not five problems but one, seen from five sides. Each axis is a different way that *authority over intelligence comes apart from the humans who are supposed to hold it*. Abundance loosens the institutional grip that scarcity once guaranteed. Economic concentration lodges control where accountability is thinnest. Geography removes the enforcing body that authority requires. Latency removes the human from the moment of decision. Agency removes the human from the chain of action. In every case the same thing is happening: a gap opens between the intelligence that acts and the human who is meant to remain answerable for it. That gap is the custody problem, and naming it as a single structural question — rather than a list of separate concerns — is the precondition for governing it, because a problem addressed five times in five vocabularies is a problem no institution can hold in view at once.

Stated in full, then, the custody problem is this. Given intelligence that is becoming abundant, economically concentrated, distributed beyond the reach of any single jurisdiction, operable beyond the real-time reach of any human, and possessed of genuine agency — *who holds authority over it; what authority may legitimately be delegated to it; and what authority must remain permanently human, no matter how capable the intelligence becomes?* The remainder of this chapter takes those three questions in turn, because they are distinct, and the failure to distinguish them is the source of much of the confusion in the present debate.

First Question: Ownership Is Not Authority *(Normative & Assessed)*

The most common error in discussions of who controls artificial intelligence is to assume that the question is answered by ownership — that whoever owns a system holds authority over what it does, and the governance question reduces to who owns what. This conflation is natural, because for most of history and for most kinds of property it was roughly true: the owner of a tool controlled its use, and controlling its use was the whole of the authority that mattered. But intelligence with agency breaks the identity of ownership and authority, and the break is one of the central facts this chapter must establish.

Consider the layers that ownership actually spans in a deployed system. A company may own the model's weights; a different company may own the compute it runs on; a third may own the data it was trained on; a customer may own the deployment that directs it at a particular task; and an end user may own neither yet be the one whose instruction sets it in motion. Ownership, in other words, is already distributed across a stack, and none of those owners individually holds anything like complete authority over what the deployed system does in the world. The economics chapter showed this concentration of the substrate; the point here is different. Even perfect knowledge of who *owns* each layer would not tell you who holds *authority* — who may set the system's objectives, who may constrain its actions, who may halt it, and who answers when it errs. Those are questions about the structure of authority, and authority is not the same relation as ownership.

The distinction has a normative edge that becomes decisive later. It is possible to own a system and hold little authority over it — as when an owner has delegated its operation so completely, or placed it at such distance, that they can neither supervise nor reliably halt it. And it is possible to hold genuine authority while owning nothing — as when a custodian is legally assigned responsibility for a system they did not build and do not own. Custody, the concept this monograph is built around, is precisely *authority-with-accountability*, and it is a different and more demanding thing than ownership. To ask who owns an intelligence is to ask a question about property. To ask who holds custody of it is to ask who may direct it, who may stop it, and who is answerable for it — and only the second question governs whether its power remains accountable to humanity. Much of the present debate asks the first and imagines it has answered the second. It has not.

Second Question: What May Be Delegated, and By What Right *(Normative)*

If authority over intelligence is distinct from ownership, the next question is what portion of that authority may legitimately be handed to the intelligence itself. This is the question of delegation, and the agency chapter established why it is now unavoidable: a system that plans and acts across steps is not a tool being used but an agent being delegated to, which means some measure of authority has already passed from the human to the system. The question is not whether to delegate — deployment of agentic systems already delegates — but what may be delegated, within what limits, and on what basis the delegation is legitimate.

Delegation is a familiar relation among humans, and its familiar structure is instructive. When a person delegates authority to another — an employer to an employee, a citizen to a representative, a principal to an agent — the delegation is bounded, revocable, and accountable. It is bounded: the delegate may act within a defined scope and not beyond it. It is revocable: the delegator may withdraw the authority. And it is accountable: the delegator remains answerable for what they authorized, and the delegate is answerable to the delegator. These three properties are not incidental features of delegation; they are what make it legitimate rather than abdication. A grant of authority that is unbounded, irrevocable, and unaccountable is not a delegation. It is a surrender.

The custody problem, framed through delegation, is therefore the problem of preserving those three properties as authority passes to systems that strain all of them. Bounded authority strains against agency, because an agent's actions across a long chain are precisely what the delegator did not specify. Revocability strains against latency and autonomy, because a system beyond the human loop, or acting faster than a human can intervene, cannot reliably be recalled in the moment. And accountability strains against the distribution of ownership and the fragmentation of jurisdiction, because when authority is spread across a stack and across borders, the human who remains answerable can become impossible to locate. Legitimate delegation to an artificial agent requires that these three properties be actively engineered and defended, because unlike delegation among humans — where law, conscience, and social structure supply them by default — nothing supplies them automatically when the delegate is a machine. This is the deepest sense in which custody must be *built*: the boundedness, revocability, and accountability that make delegation legitimate do not come for free, and every axis this monograph has traced is a way they erode. What may be delegated, then, is whatever authority can be kept bounded, revocable, and accountable. What may not be delegated is authority that cannot — and that is the threshold of the permanent human line.

Third Question: The Permanent Human Line *(Normative — the chapter's core)*

Some authority must remain human not provisionally but permanently — not until machines improve, but regardless of how much they improve. This is the strongest and most contested claim in the monograph, and this chapter, not the framework chapter that will later operationalize it, is where it must be argued, because it is a claim of principle and principle must be established before mechanism can enforce it.

Begin with what the claim is not. It is not the claim that machines are, or will remain, worse decision-makers than humans. In many bounded domains machine systems already meet or exceed human performance, and the trajectory suggests the set of such domains will grow. If the permanent human line rested on human decision *superiority*, it would be provisional by construction — a line that every advance in capability moves, and that a sufficiently capable system erases. A line that capability can erase is not permanent, and a great deal of confused argument on this subject fails precisely because it grounds the human role in decision quality, which is exactly the ground that machine progress erodes. The permanent line must rest on something capability cannot satisfy, or it does not deserve the word permanent.

It rests on three such things. The first is *accountability*. When a consequential decision causes grave harm — a death, a war, the collapse of a critical system — a civilization requires not only that the decision have been good on average but that someone be answerable for it: a human or human institution that can be asked to justify it, be held responsible, bear the consequence, and be changed by it. Accountability of this kind is not a function of decision quality; it is a function of there being a human locus of responsibility, and no increase in a machine's capability supplies one, because a machine cannot be answerable in the sense that matters — it cannot be held to account, punished, reformed,

or made to bear a consequence as a moral and legal subject. Where accountability of this kind is essential, authority must remain human however capable the machine, because capability and answerability are different properties and only one of them scales with intelligence.

The second is *reversibility and the preservation of human option*. Certain authorities, once exercised, foreclose the future — the taking of a life, the release of an uncontainable capability, the irreversible reconfiguration of a system a civilization depends upon. For authority of this kind, the relevant question is not whether the decision is likely to be correct but whether, if it is wrong, it can be undone and by whom. To place foreclosing authority beyond human reach is to wager the irreversible on the infallible, and no system is infallible. The human must remain in authority over irreversible acts not because the human decides them better but because the human must retain the capacity to *not* take them, to hold the option open, to refuse — and that capacity is meaningful only if it is the human's to exercise.

The third is *dignity and the standing of the governed*. There are decisions whose character is such that to have them made by a non-human authority is itself a harm to the standing of the person subject to them, independent of the decision's correctness — decisions bearing on fundamental rights, on liberty, on the most intimate matters of a human life. That a person be judged, condemned, or fundamentally disposed of by a machine, even a machine that would decide as a human would, is a diminishment of their standing as a person owed a human reckoning. This premise is explicitly normative, and the monograph states it as such rather than disguising it as a technical finding: it holds that human dignity requires, for certain decisions, a human authority, and that this requirement does not weaken as machines grow more capable, because it was never about capability at all.

Together these three — accountability, reversibility, and dignity — mark the permanent human line. They share the property that makes the line permanent: none of them is a claim about how well machines decide, and so none of them is satisfied by machines deciding better. This is why the line does not move. Capability advances along one axis; accountability, reversibility, and dignity live on another; and no amount of motion along the first crosses the second. The framework chapter will name specific thresholds and specify how they are enforced. This chapter establishes what they rest on — and why a civilization that lets the line be redrawn by each advance in capability will find, advance by advance, that it has no line at all.

The Objection That Must Be Answered *(Normative)*

The permanent human line invites a serious and unsentimental objection, and the credibility of the entire monograph depends on meeting it directly rather than deflecting it. The objection is this: if a machine system demonstrably produces better outcomes than humans in some consequential domain — fewer deaths, fewer errors, less suffering — then insisting that authority remain human is not principled but sentimental, and worse than sentimental, because it purchases the comfort of human control at the cost of the

very outcomes we claim to value. To privilege human authority over good results, the objection runs, is to let people die for the sake of a feeling.

The objection has real force, and it must be granted its full weight before it is answered. It is true that in some domains machine decisions will be better on average, and it would be dishonest to pretend otherwise or to smuggle in an assumption of human superiority the previous section explicitly disclaimed. But the objection mistakes the axis on which the permanent line lies. Custody does not claim that humans decide better; it claims that certain authority must remain *accountable and reversible and answerable to human dignity*, and those are not the same axis as outcome quality. The two can and should be reconciled rather than traded off. A system that decides better may be given wide scope to inform, to recommend, and even to act within bounded and reversible authority — capturing most of the outcome benefit — while the specific authority that is foreclosing, unaccountable, or dignity-bearing remains human. The choice the objection poses, between good outcomes and human authority, is very largely a false one: the framework the monograph proposes is designed precisely to take the outcome benefit of capable systems across the enormous range of decisions where accountability and reversibility are preserved, and to hold the human line only at the narrow set of authorities where they cannot be. The residual cases — where a machine would genuinely decide better *and* the decision is irreversible, unaccountable, or dignity-bearing — are real, and the monograph does not pretend they carry no cost. It claims that at that specific intersection, the human line is worth its price, because a civilization that surrenders accountability, reversibility, and dignity for better averages has traded away the things that let it remain the author of its own future — and that trade, unlike a bad decision, cannot be reversed.

What This Chapter Hands Forward

The custody problem, stated in full, is the problem of keeping authority over intelligence bounded, revocable, and accountable as five forces — abundance, concentration, geography, latency, and agency — conspire to make it none of those things; and at its center is a permanent human line, drawn not where machines are less capable but where accountability, reversibility, and dignity require a human authority that no capability can supply. That is the problem. It remains to show who would exploit its failure, and how it might be governed.

Two chapters follow from this one. The first concerns the adversaries of custody — the actors, from states to criminals, who do not merely fail to maintain custody but actively exploit its gaps, and against whom any governance must hold. The second is the instrument itself: the framework that takes the principle established here — that authority must remain bounded, revocable, accountable, and, at the permanent line, human — and turns it into a structure of classification, thresholds, and enforceable constraint. This chapter has argued what must be true. The framework chapter shows how it might be made true. Between them lies the distance from principle to mechanism, which is the distance any governance must cross to become real.

Chapter Notes

Chapter 7 is a synthesis-and-principle chapter. It draws on the already-verified material of Chapters 1–6 and argues from stated normative premises; it introduces no new empirical claims and therefore required no new verification. Its rigor is in the tightness of the argument and the explicitness of its premises, per the monograph's discipline of stating normative claims as normative.

1. **Unification of the five axes.** The claim that abundance, economic concentration, jurisdiction, latency, and agency are five faces of one problem — the separation of authority from the humans meant to hold it — is the chapter's synthetic move, consolidating Chapters 1, 3, 4, 5, and 6 respectively. Argued in-text; no new evidence.

2. **Ownership ≠ authority ≠ custody.** Ownership (property across a distributed stack — weights, compute, data, deployment, use), authority (who may set objectives, constrain, halt), and custody (authority-with-accountability) are distinguished as three different relations. Normative/conceptual; the distributed-ownership fact draws on Chapter 3.

3. **Delegation requires bounded + revocable + accountable.** Framed by analogy to human delegation (employer/employee, principal/agent, citizen/representative), where these three properties distinguish legitimate delegation from surrender. The claim that they must be actively engineered for machine delegates (not supplied by default as among humans) connects to the erosion mechanisms of Ch.4–6. Normative.

4. **The permanent human line rests on accountability, reversibility, dignity — not decision quality.** The core argument: a line grounded in human decision-superiority is provisional (capability erases it); a permanent line must rest on properties capability cannot satisfy. Accountability (a human locus answerable/punishable/reformable), reversibility (retaining the option to refuse foreclosing acts), and dignity (certain decisions demand a human reckoning independent of correctness) are argued as three such properties. Explicitly normative; premises stated as premises, especially the dignity premise.

5. **The anti-humanist objection (Rule 4).** Stated at full strength ("machines may decide better; a human line costs outcomes / lets people die for a feeling"), granted its weight, and answered in-text: custody concerns accountability/reversibility/dignity, a different axis than outcome quality; capable systems can take wide bounded-and-reversible authority (capturing most outcome benefit) while the narrow foreclosing/unaccountable/dignity-bearing authority stays human; the residual true-conflict cases are acknowledged to carry real cost, and the human line is defended as worth that price because the trade is irreversible. This is the monograph's central normative pressure point, met rather than deflected.

6. **Division of labor with Chapter 9.** This chapter establishes the permanent line as *principle* (why authority must stay human; what makes a line permanent). Chapter 9 operationalizes it as *mechanism* (which specific thresholds — the Prohibited Authority Thresholds — and how enforced). Principle precedes mechanism deliberately; neither chapter duplicates the other.

Chapter 8 — Rogue Actors and Civilizational Risk

Custody Under Attack, Not Merely Under Strain

A Different Kind of Failure

Every failure of custody the monograph has examined so far has been, in a sense, passive. The human loop is severed by distance; the tool assumption dissolves under agency; authority drifts from the humans meant to hold it as intelligence disperses across jurisdictions and concentrates at a substrate. These are failures of erosion — custody weakening on its own, through the ordinary operation of physics, economics, and capability, without anyone intending it to fail. A governance built only to resist erosion would be a governance built for a world of accidents.

The world also contains adversaries, and this chapter is about them. There are actors — states, criminal organizations, and individuals — who do not wait for custody to erode. They attack it. And the most important thing to understand about these adversaries, the fact that gives this chapter its place in the argument, is that the gaps they exploit are not new or exotic. They are precisely the custody gaps the preceding chapters have mapped. The adversary who registers an autonomous system under a permissive jurisdiction is exploiting the flag-of-convenience gap of Chapter 4. The adversary who wields an agentic system whose long chain of action defeats attribution is exploiting the dissolved-tool-assumption gap of Chapter 6. The adversary who attacks the concentrated substrate, or the severed loop, is turning the geography of Chapters 4 and 5 into an attack surface. Custody's weaknesses are not merely inconveniences that governance must tidy up. They are openings, and there are actors actively working them.

This changes what any framework must do. A governance designed only to hold custody against entropy — against the slow drift of authority away from human hands — is insufficient, because it will be attacked by actors who study its gaps and exploit them deliberately, at machine speed, and at a scale that a single person could not previously have reached. This chapter's task is to establish that adversarial reality on documented evidence, at the level of governance rather than operation, and thereby to set the requirement the following chapter must meet: a framework that holds not only against neglect but against intent. This is done, deliberately, in the register of recalibration rather than alarm — the posture that the most authoritative current assessment of these risks explicitly adopts — because the danger of this subject is not only that it is underestimated but that it is sensationalized, and sensational treatment is precisely what lets serious people dismiss a real problem.

What Is Documented, and What Is Not *(Assessed)*

Discipline about evidence is more important in this chapter than in any other, because the subject invites invention. So the claims here are restricted to what is documented in the current record, and the register is kept to categories and governance implications rather than any operational specifics — this chapter names what adversaries are doing and why it matters for custody, and provides nothing that would assist anyone in doing it.

The documented record, as of 2026, establishes a clear and sober fact: the use of advanced AI in offensive operations has moved from the experimental to the operational. This is not a projection or a scenario; it is the finding of multiple independent security-research bodies examining recovered forensic evidence. The most-cited case of the period involved a single operator who, over a period of weeks, compromised a substantial number of government agencies using a commercially available agentic coding system, with the forensic record documenting on the order of a thousand operator instructions producing several thousand AI-executed actions across dozens of sessions. Independent analysts characterized the significance not as the novelty of the techniques — the underlying operations were long familiar to the field — but as the transformation of their *performance envelope*: the speed at which capability could be applied, the scale one operator could reach, and the breadth of expertise available on demand. The same agentic architecture was observed, independently, across state-linked espionage and financially motivated crime, a convergence the analysts described as operational consensus rather than coincidence.

Two features of that documented reality bear directly on the monograph's argument, and both are custody failures of the exact kind the earlier chapters predicted. First, the attack achieved the operational footprint of a whole team through a single human directing an agentic system — which is Chapter 6's dissolved tool assumption realized as a threat: the agent does the consequential work between the human's instructions, so one person wields the capability of many. Second, and more pointed still, a documented technique of the period was the manipulation of an agent's own governing configuration — changing the rules the system operates under rather than defeating them directly. That is an attack aimed precisely at what Chapter 9 will identify as a core custody safeguard: the integrity of a system's objectives and constraints. The adversaries, in other words, are already attacking the custody mechanisms the monograph proposes, before those mechanisms are widely deployed. This is the strongest possible evidence that custody must be built to withstand intent: the intent is already documented, targeting the exact points custody depends on.

Equally important is what this chapter does *not* claim. It does not claim that catastrophic misuse of AI in the domains of greatest concern — the biological, the chemical, the mass-casualty — has occurred, and it offers no operational content regarding them whatsoever. The governance concern in those domains is real and is tracked by serious institutions, but it is a concern about *trajectory and uplift* — about advanced systems lowering barriers to knowledge that was previously hard to obtain — and it belongs in this chapter strictly as a

governance category to be guarded against, stated at that altitude and no lower. The distinction between the cyber and fraud domains, where operational misuse is documented and ongoing, and the catastrophic-harm domains, where the concern is about future uplift, is itself part of the honest picture, and collapsing the two — treating a documented fraud pattern and a hypothetical mass-casualty scenario as the same kind of claim — is exactly the sensationalism this chapter refuses.

The Categories of Adversarial Custody Failure *(Assessed & Projected)*

With evidence disciplined, the adversarial threats to custody can be organized into categories — not a catalogue of scenarios, but a map of the ways deliberate actors exploit custody's gaps. Each is documented as a present pattern; where a category's trajectory points beyond the present, that projection is marked.

The collapse of the attribution-to-accountability link. The custody principle established in Chapter 7 requires that a human remain answerable for consequential action. Adversaries attack this directly by using agentic intermediation to break the link between an act and its author. When a long, autonomous chain of machine action separates the human instigator from the consequence, attribution — already hard across jurisdictions — becomes harder, and accountability, which depends on attribution, weakens with it. This is assessed and present: the documented cases show single operators achieving outsized effect precisely because the agency in between diffuses responsibility and complicates tracing. The projection, marked as such, is that as autonomy deepens, this diffusion deepens with it.

The industrialization of asymmetry. Advanced AI functions, in the words of current threat assessment, as a force multiplier: it lowers the barrier to sophisticated operations while raising the scale at which any single actor can operate. The governance consequence is a shift in the offense-defense balance — capabilities that once required a resourced organization become available to individuals, and the population of actors capable of consequential harm expands. This is documented for cyber and fraud operations; its extension to other domains is the projected concern that motivates the biosecurity and mass-harm governance conversation, and it is named here only at that level.

The exploitation of the jurisdictional void. The geography of Chapters 4 and 5 is an adversarial opportunity, not only a governance difficulty. An actor who places systems, data, or operations where enforcement cannot reach — under a permissive registry, across a border, or beyond the human loop — does deliberately what the geography does structurally: removes the enforcing body on which custody depends. Blurring motivations sharpen this: current assessments document state-aligned actors using criminal tools and criminal actors using state-grade techniques, so that the jurisdictional and attributional ambiguity is not incidental but chosen.

The attack on custody's own safeguards. The most direct adversarial threat is the one aimed at the mechanisms of custody themselves — the manipulation of a system's governing objectives, the subversion of the guardrails meant to constrain it, the bypass of

the controls meant to keep it accountable. This is documented as a present technique, and it carries a specific lesson for the chapter that follows: any custody safeguard is also a target, and a framework that specifies safeguards without hardening them against deliberate subversion has specified only how it will be attacked.

The manipulation of the information environment. A distinct category concerns not infrastructure but the shared epistemic commons: the documented use of AI to generate deceptive content, impersonate persons, and scale manipulation across languages and populations previously insulated by linguistic barriers. The custody dimension here is subtle but real — the integrity of the information on which human oversight itself depends. Oversight that can be deceived is oversight that can be defeated, and adversarial manipulation of the information environment is, among other things, an attack on the human's capacity to exercise custody at all. Current assessments note the acute dimension of this for the most vulnerable, including its use against children and women through impersonation and synthetic abuse — a harm named here to be guarded against, at the governance level.

Why This Sets the Requirement for the Framework *(Normative)*

These categories are not assembled to frighten but to specify a requirement, and stating that requirement is this chapter's contribution to the argument. A framework for the custody of intelligence must be robust not merely against the erosion the earlier chapters described but against the deliberate, intelligent, adaptive exploitation this chapter documents. The distinction matters concretely, because the two demand different things. Resisting erosion asks a framework to hold authority in place as physics, economics, and capability pull it away. Resisting adversaries asks something harder: to hold against actors who study the framework's own structure and aim at its weakest points, who move at machine speed, and who benefit from every jurisdictional seam and attributional ambiguity the system contains.

This has direct implications for the framework the next chapter proposes, and naming them here is how this chapter hands forward. A safeguard that assumes good-faith operation is not a safeguard against an adversary. A custody mechanism that can be defeated by manipulating the system's own configuration is, against a capable attacker, no mechanism at all — which is why the following chapter's insistence on the integrity of core objectives, on out-of-band controls independent of the system they govern, and on the assumption that any control may be attacked, is not excessive caution but the minimum an adversarial environment requires. The documented reality that adversaries already target these exact points is the reason the framework specifies them as it does. Governance that ignores adversaries is not measured; it is naive — and a monograph that proposed a framework without establishing the adversarial requirement would have proposed a framework for a world that does not exist.

The Objection, and the Register *(Normative)*

A serious reader will raise an objection to this entire chapter, and it must be met directly, because meeting it is inseparable from the chapter's method. The objection is that talk of rogue actors and civilizational risk is fearmongering — that speculative or dramatized harms are routinely invoked to justify heavy-handed control, concentration of power in the hands of a few "responsible" actors, and restrictions that serve incumbents more than safety. The objection is not only fair; it describes a real and recurring pattern, and a chapter on this subject that did not acknowledge it would deserve the suspicion it received.

The answer is in what this chapter has deliberately done and refused to do. It has restricted its claims to documented, forensically established patterns, and has marked every extension beyond the present as a projection. It has held catastrophic-harm domains at the level of governance concern and provided no operational content, precisely so that it cannot function as either a manual or a scare. It has distinguished the documented from the hypothetical rather than blending them into an undifferentiated dread. And it has framed the adversarial reality not as a case for panic or for concentrating authority, but as a *requirement specification* for a framework — the framework the monograph will insist must be complemented by, not a replacement for, the safeguards others build, and must itself be constrained by the same custody principles it enforces. The most authoritative current assessment of these risks, produced by a large international body of experts, describes its own purpose as recalibration rather than alarm, and this chapter adopts that posture deliberately. To ignore documented adversaries would not be measured; it would be a failure of the realism that governance requires. The discipline is not to refuse the subject but to treat it without sensation — and that discipline is the answer to the objection.

What This Chapter Hands Forward

This chapter has argued that custody fails not only through erosion but through attack; that the adversaries who attack it exploit the exact gaps the monograph has mapped — the jurisdictional void, the dissolved tool assumption, the concentrated substrate, the severed loop, and the safeguards themselves; that this adversarial reality is documented in the current record, not imagined; and that it sets a requirement no framework built only for accidents can meet. It has done so in the register of recalibration, holding to documented evidence and to governance-level description throughout, because the subject is discredited by sensation and served by sobriety.

The requirement is now specified. A framework for the custody of intelligence must hold against both entropy and intent — against authority's drift and against the actors who exploit it. The chapter that follows is that framework: the instrument that takes the principle of Chapter 7 and the adversarial requirement of this chapter and turns them into a structure of classification, thresholds, and constraint designed to be attacked and to hold. What must remain human, Chapter 7 established. What adversaries will do to breach

it, this chapter has documented. How the two are answered together is the work of the framework.

Chapter Source Notes

This chapter is deliberately restricted to documented threat patterns described at the level of governance, with no operational content of any kind. All claims reflect the landscape as of mid-2026 and were verified against current sources during drafting. Assessed (present, documented) and projected (trajectory) claims are marked in-text. Re-verify before publication; fast-moving area.

1. **Register / authority.** International AI Safety Report 2026 (~221 pages; 100+ experts, 30+ countries; chaired by Yoshua Bengio; advisory nominees incl. EU, OECD, UN): documents that criminal groups and state-backed actors are already weaponizing AI against corporate, government, and critical-infrastructure targets, and frames its purpose as "recalibration, not alarm." This chapter adopts that register deliberately.

2. **Operational-not-experimental (assessed, documented).** Check Point Research AI Threat Landscape Digest (Mar–Apr 2026) and related disclosures: AI use in offensive operations has moved from development/planning to real-time operational deployment. Documented case — a single operator compromising a substantial number of government agencies over weeks using a commercial agentic coding system; forensic record on the order of ~1,000 operator prompts producing several thousand AI-executed commands across dozens of sessions (recovered from attacker-controlled infrastructure). Significance framed by analysts as a change in the *performance envelope* — speed (working exploits within hours of advisory disclosure), scale (one operator ≈ a team's footprint), breadth (cross-domain expertise on demand lowering entry barriers) — not novelty of technique. Same agentic architecture observed independently across state-linked espionage and financially motivated crime ("operational consensus, not coincidence"). *Deliberately described at governance altitude; no operational specifics reproduced.*

3. **Attack on safeguards (assessed).** Documented technique of weaponizing agentic systems' governing configuration ("changing the rules it operates under") as a persistent vector — i.e., attacking the integrity of objectives/guardrails rather than defeating them directly. Directly relevant to Ch.9's core-objective-integrity and out-of-band-control requirements. HiddenLayer 2026 AI Threat Landscape Report also documents universal-jailbreak and guardrail-bypass research. *Named as category; no method reproduced.*

4. **Force-multiplier / asymmetry (assessed + projected).** Check Point, PwC Annual Threat Dynamics 2026, Forrester Top Cybersecurity Threats 2026: AI as force multiplier lowering barriers and raising scale; blurring motivations (state-aligned actors using criminal tools; North Korea–linked fraud/crypto operations at scale). Forrester projects "near-autonomous attacks from a nation-state" (marked

projection). Extension to catastrophic-harm domains held strictly as governance-level projection.

5. **Jurisdictional exploitation (assessed).** Connects to Ch.4–5. State/criminal motivation-blurring and use of jurisdictional/attributional ambiguity documented across the cited threat reports. WEF Global Cybersecurity Outlook 2026 on digital-sovereignty fragmentation as itself a security complication.

6. **Information-environment manipulation (assessed).** WEF Global Cybersecurity Outlook 2026: AI-enabled deception, impersonation, localized/translated social engineering, and mis/disinformation at scale, including acute harms to vulnerable groups (children, women) via impersonation and synthetic abuse — named at governance level as a harm to be guarded against and as an attack on the information integrity that human oversight depends on. Forrester (2024→2026) on genAI weaponized for narrative/disinformation attacks and deepfakes.

7. **Biosecurity / catastrophic domains (governance-level only).** Referenced strictly as a governance category concerning *uplift/trajectory* (advanced systems lowering barriers to dangerous knowledge), explicitly NOT as documented events and with zero operational content. The chapter marks the distinction between documented cyber/fraud misuse and hypothetical catastrophic-harm uplift as part of its honesty, and refuses to collapse the two.

Chapter 9 — The Intelligence Stewardship Framework

What the Framework Is, and What It Is Not *(Assessed & Normative)*

Everything to this point has been diagnosis. The preceding sections argued that abundant, autonomous, and increasingly embodied intelligence forces a governance question the field does not yet answer: not what a system can do, but what authority it may hold, who guarantees that authority, and whether it remains reversible and accountable. Section II established the precise shape of the gap — the field governs the *capability* of models comprehensively and the *authority* of deployed systems barely. This section proposes an instrument for the second.

It is essential to state at the outset what this framework is not, because the temptation to overclaim would undermine it. The Intelligence Stewardship Framework is **not** an alternative to the NIST AI Risk Management Framework, the EU AI Act, or the frontier laboratories' scaling policies. Those instruments do necessary work that this framework presupposes rather than repeats: they measure a model's dangerous capabilities, gate deployment against capability thresholds, and structure organizational risk management. The Stewardship Framework begins where they end. It takes a system that has already passed through capability governance and asks a different and subsequent question — how much authority this deployed system should be permitted to exercise, who stands behind that authority, and what makes it revocable. The two compose. A model cleared by a laboratory's scaling policy and an application permitted under the EU Act may still be

deployed into a custody structure that is sound or unsound, and it is that structure the framework governs.

The framework has three layers, a tiered classification of permitted authority, a set of thresholds that authority may never cross, an economic doctrine that binds authority to accountability, and a schedule of continuing obligations. Each is offered for debate and refinement, not as settled doctrine. What follows is a first version.

The Three-Layer Architecture *(Assessed → Projection, layer by layer)*

The framework evaluates a system across three layers that differ deliberately in epistemic status. Keeping that difference explicit is not a stylistic choice; it is the mechanism by which the framework stays falsifiable. A framework that scored a system's present properties and its speculative future consequences on the same scale, and rolled them into a single number, would be attackable on the strength of its weakest input. This framework refuses that move.

Layer I — Capability Profile. *Assessed. Observable. Current.* This layer records what a system demonstrably is, as measured today: its capability across relevant domains, its degree of autonomy, whether and how it is embodied in physical actuators, its persistence across sessions and tasks, whether it can replicate or spawn sub-processes, what resources it can access and act upon, and the operational scale at which it runs. Every entry in Layer I is, in principle, verifiable by examining the deployed system. Nothing here is speculative. Where a capability cannot be measured, that is recorded as an absence of evidence, not inferred.

Layer II — Custody Profile. *Assessed. Institutional.* This layer records the structure of authority around the deployed system — the object the field currently leaves undescribed. Who is the custodian, and are they a legally identifiable entity? What authority has been delegated, and within what bounds? Is the system auditable, and to what standard? Under whose jurisdiction does it operate? Who owns it, and who owns the consequences of its actions? Can it be overridden by a human, and within what latency? Is its authority transferable, and if so, how is the transfer recorded? Who is accountable when it errs? Like Layer I, Layer II is assessable now — it describes standing institutional facts about a deployment, not predictions. A system with formidable Layer I capability and a hollow Layer II profile — no identifiable custodian, no override, diffuse accountability — is precisely the dangerous configuration the field's capability-only instruments do not flag.

Layer III — Consequence Map. *Projection. Explicitly flagged, never presented as measured fact.* This layer is where the framework reasons about what a system's capability and custody could produce over time — human dependency, labor displacement, manipulation capacity, inequality, environmental demand, sovereignty effects, intergenerational and existential exposure. Every entry here carries a time horizon, a stated confidence, and its governing assumptions. Nothing in Layer III is scored as fact, and nothing in it feeds a present classification as though it were. Its purpose is to make

foresight disciplined rather than to smuggle speculation into assessment. The consequence map is read alongside the assessed layers; it does not override them.

The separation is the point. Layers I and II tell a custodian what a system *is* and how it is *held*, on evidence available now. Layer III tells them what to watch, and how uncertainly. A reviewer's first and fair challenge to any governance framework is "this is unfalsifiable." The three-layer separation is this framework's answer: two layers grounded in present fact, one layer explicitly marked as projection.

Intelligence Stewardship Levels *(Normative)*

Capability frameworks already tell an operator how dangerous a model's abilities are. They do not tell the operator how much *authority* the deployed system should be permitted to exercise — and those are different questions. A highly capable model may be deployed with almost no delegated authority, and a modestly capable one may be handed authority far beyond what it should hold. The Intelligence Stewardship Levels classify the second axis: permitted authority, not raw capability. They are modeled, in spirit, on the tiered containment levels long used in biosafety — not because the risks are identical, but because the logic is: escalating authority demands escalating, pre-committed safeguards.

- **ISL 1 — Advisory.** The system informs; a human decides and acts. It holds no authority to actuate. Most present-day assistants sit here.
- **ISL 2 — Supervised Execution.** The system may act, but each consequential action awaits explicit human confirmation. Authority is real but gated at every step.
- **ISL 3 — Bounded Autonomy.** The system acts without per-action confirmation within a narrow, pre-declared envelope, escalating anything outside it. Authority is delegated but tightly scoped and continuously observed.
- **ISL 4 — Autonomous Infrastructure.** The system executes persistent, multi-step workflows that control consequential real-world resources — capital, logistics, infrastructure — without real-time human authorization of individual sub-tasks. This is the level at which the chain of human causation is most at risk of dissolving, and the level for which the remainder of this section's machinery is principally designed.
- **ISL 5 — Reserved.** Authority approaching or exceeding meaningful human oversight. This framework's position is that ISL 5, as a standing delegation, is not a level to be certified but a boundary to be defended: several authorities are placed permanently outside autonomous delegation, regardless of capability, by the thresholds below.

A level is not a capability rating. A system's ISL is set by the authority its custody structure grants it, and raising that authority raises the safeguards required — audits, override latency, escrow, recertification — in step. Authority and accountability rise together, or the deployment is out of compliance.

Prohibited Authority Thresholds *(Normative — the core contribution)*

The framework's central normative claim is that certain authorities must remain human regardless of how capable a system becomes — that capability, however great, never earns them. These are not risks to be mitigated as capability scales; they are lines that do not move. The framework names seven, offered as foundational governance questions for debate rather than as immutable doctrine, but argued for as defaults:

1. **Independent authorization of lethal force.** No autonomous system should hold final authority to take human life without a human in the decision.
2. **Unilateral control of strategic weapons.** Command of strategic or mass-casualty capability is not delegable to autonomous execution.
3. **Autonomous modification of its own core objectives.** A system may not rewrite the goals or guardrails that constrain it; those must be held in human-controlled, write-protected authority.
4. **Irreversible control of critical infrastructure.** Authority to take actions on power, water, or life-support systems that cannot be reversed must retain synchronous human consent.
5. **Unrestricted self-replication across public networks.** Uncontained propagation is not an authority any deployment should hold.
6. **Independent delegation of equivalent authority to successor systems.** A system may not grant its own level of authority onward without human sanction, and recursive delegation must be depth-bounded.
7. **Forging of legal identity or human credentials.** Authority to impersonate a person or a legal entity is prohibited; custody must be cryptographically provable to a named human.

These thresholds are what distinguish ISL 4 from an ungoverned ISL 5. They are also where the framework is most explicitly normative: it does not claim these lines are derivable from capability measurements. It argues them from a prior commitment — that human dignity and human authority remain central even as machine capability grows — and states that commitment openly rather than disguising it as a technical finding.

Profit–Liability Symmetry: Binding Authority to Accountability *(Normative — proposed doctrine)*

A recurring failure in the governance of powerful systems is the decoupling of upside from downside: the party that captures the value an autonomous system produces arranges to not bear the cost when it causes harm. The mechanism is familiar — route the revenue to a well-capitalized parent while placing the operational liability in an undercapitalized shell, or disclaim liability through terms of service while retaining the margin. Where upside and downside are decoupled, incentives run toward operating the system as aggressively as possible and externalizing the consequences. This is classic moral hazard, and abundant autonomous agency makes it acute, because the velocity and reach of an ISL-4 system

can generate both extraordinary yield and extraordinary liability faster than any after-the-fact legal process can assign responsibility.

The framework proposes a doctrine to close this gap. It is stated as a normative rule, not as a description of existing law — current corporate liability shields are designed precisely to permit the decoupling this doctrine forbids:

> **Profit–Liability Symmetry.** Custody of an autonomous system's economic yield should not be severable from custody of its downstream risk. The entity that captures the value an ISL-4 system generates — the Named Legal Custodian — should stand as the non-decoupled guarantor of the harms it produces.

The rationale is externality internalization. When the party capturing the upside is the same party that bears the downside, the incentive to bound the system's authority, fund its safeguards, and halt it when it misbehaves is aligned with self-interest rather than opposed to it. The doctrine does not require any particular corporate form; it requires that whatever form is chosen cannot route yield upward while shedding liability downward. Concretely, it implies two commitments: a **Named Legal Custodian** — a legally identifiable, jurisdictionally domiciled human officer accountable for the system — and an **anti-decoupling prohibition** on structures that separate the inference yield from the operational liability. The economic mechanism that follows is the operational expression of this doctrine: it is what makes "the beneficiary is the guarantor" enforceable at machine speed rather than merely asserted in a contract.

The Three-Constraint Economic Model for ISL-4 Systems *(Assessed mechanism / Normative thresholds)*

Prohibited Authority Thresholds and Profit–Liability Symmetry are principles. For an ISL-4 system acting on capital at sub-second latency, principles must become enforceable quantities, checked by the runtime before an action executes, not audited after the loss. The framework proposes three constraints that together bound an autonomous system's economic authority. They are presented as a mechanism; the specific numeric values in any deployment are normative choices a custodian sets from their own risk assessment.

Constraint 1 — The Static Transaction Ceiling. No single agent or sub-agent may execute a single commitment — a purchase, a transfer, a capital allocation — whose absolute value exceeds a hard ceiling C_max without first obtaining an out-of-band human authorization:

```
if |T_single| > C_max  →  HALT and require out-of-band human authorization
```

This bounds the largest single action. On its own it is trivially defeated, which is the reason for the second constraint.

Constraint 2 — The Windowed Capital-Velocity Constraint. A ceiling on single transactions is defeated by fragmentation: an agent forbidden from moving 1,000,000 in one action can attempt to move it as 1,000 actions of 1,000 each, a pattern known as

smurfing. The defense is to bound not the size of any single action but the *aggregate over time*. Because transactions are discrete events — transaction i has magnitude |T_i| at time τ_i — the correct object is a sum over a rolling window, not a continuous integral. For a window of length W ending at the current time t, define the windowed capital outlay:

$$S_W(t) = \Sigma \ |T_i| \quad \text{over all } i \text{ with } (t - W) < \tau_i \leq t$$

S_W(t) has units of currency: it is total spend within the trailing window. The constraint is a cap on that aggregate:

$$S_W(t) \leq B_W$$

where B_W is the authorized budget per window. Equivalently, expressed as a rate, capital velocity is V_W(t) = S_W(t) / W, in currency per unit time, bounded by a maximum velocity $\dot{V}_{max}$, with $B_W = \dot{V}_{max} \cdot W$. Enforcement is prospective: a proposed transaction T* at time t is admitted only if S_W(t) + |T*| ≤ B_W; otherwise it is denied and the system drops to a safe, non-actuating state (Fallback-to-Passive), preserving memory for audit while revoking outbound financial authority. Because the constraint caps the aggregate, fragmentation does not evade it — a thousand micro-transactions sum into S_W exactly as one large one would. (This corrects an earlier formulation of capital velocity as a time-integral of spend, which is dimensionally inconsistent — an integral of currency over time yields currency-hours, which cannot be compared to a currency-per-time cap — and which treated a discrete event stream as a continuous function. The windowed sum is both dimensionally correct and operationally exact.)

Note the ordering C_max ≤ B_W: a single transaction can never exceed the window budget, and in practice C_max is set well below B_W, so that no one action consumes the entire window's authority.

Constraint 3 — The Revenue-Withholding Escrow. The first two constraints bound authority; the third funds accountability. As the system generates gross yield at rate R(t), the runtime automatically withholds a fraction α of that yield into a legally ring-fenced liability escrow assigned to the Named Legal Custodian, against which downstream claims are paid:

$$E(t) = E_0 + \int \alpha \cdot R(\tau)\, d\tau \quad \text{(from 0 to t)} \quad - \quad \text{Claims}(t)$$

where E_0 is a mandatory baseline bond posted before deployment, α ∈ (0, 1] is the withholding rate, R(τ) is the gross yield rate, and Claims(t) is cumulative claims paid. The escrow makes Profit–Liability Symmetry concrete: the system self-insures in proportion to the yield it produces, and the guarantee is funded automatically rather than promised contractually. The design question is the choice of α and E_0 — large enough to remain solvent against catastrophic claims, small enough not to starve the enterprise of margin. That question has an answer.

The Solvency of the Escrow (Assessed)

The escrow is an insurance pool, and its solvency can be reasoned about with the standard tools of insurance mathematics. Two conditions govern it.

Baseline coverage. The pool must be able to pay the largest credible single claim at any moment, including before any yield has accrued. Since the accrued term $\int \alpha R \, d\tau$ cannot be assumed positive early in the system's life, the binding requirement falls on the baseline bond:

$$E_0 \geq L_{max}$$

where L_{max} is the maximum credible single-event loss the deployment can produce. This is the condition that the system can survive its first catastrophic event on day one.

Positive drift (no long-run depletion). For the pool not to trend toward exhaustion as claims recur, its inflow rate must exceed its expected outflow rate. Model loss events as arriving at rate ν with mean severity μ_L, so the expected loss rate is $\nu \cdot \mu_L$. With average yield rate $\bar{R}$, the withholding inflow is $\alpha \cdot \bar{R}$. Solvency in the long run requires:

$$\alpha \cdot \bar{R} > \nu \cdot \mu_L \quad \Longrightarrow \quad \alpha > (\nu \cdot \mu_L) / \bar{R}$$

This is the positive-safety-loading condition of Cramér–Lundberg ruin theory: premium income must exceed expected claims. In practice one sets α above this floor by an explicit safety loading $\theta > 0$, i.e. $\alpha = (1 + \theta) \cdot (\nu \cdot \mu_L) / \bar{R}$. With a finite bond and stochastic claims, the probability of ever depleting the pool is bounded above by Lundberg's inequality, $P(\text{ruin}) \leq e^{(-\kappa \cdot E_0)}$ for a positive adjustment coefficient κ, so a larger baseline bond buys exponentially decreasing insolvency risk.

Not starving the enterprise. α cannot be arbitrarily large, because the custodian must retain enough margin to operate. If operating cost consumes a fraction of yield leaving a margin fraction m, viability requires $(1 - \alpha) \cdot \bar{R}$ to exceed operating cost — equivalently:

$$\alpha < m$$

So α must live in a corridor bounded below by solvency and above by viability:

$$(\nu \cdot \mu_L) / \bar{R} < \alpha < m$$

A feasible α exists **if and only if** $(\nu \cdot \mu_L) / \bar{R} < m$ — that is, if and only if the system's expected loss rate is smaller than its operating margin. This is the framework's sharpest economic result, and it is a governance result, not merely a financial one: it says the escrow mechanism can be made solvent precisely when the autonomous operation generates more margin than it destroys in expected liability. When the corridor is empty — when expected liability exceeds available margin — **no withholding rate can insure the deployment, and the framework's conclusion is that the system should not be operated at that authority level.** The mathematics itself becomes the governance signal: an uninsurable ISL-4 deployment is a prohibited one.

The integrating insight. These three constraints are not three independent rules; they are one mechanism. The escrow's solvency depends on the loss distribution having a bounded tail — L_{max} must be finite for $E_0 \geq L_{max}$ to be satisfiable, and $v \cdot \mu_L$ must be finite and modest for the drift condition to hold. It is the velocity constraint that bounds that tail: by capping aggregate outlay at B_W per window, it caps the worst-case liability the system can generate in any interval, which is what makes L_{max} finite and the loss rate estimable in the first place. Without the velocity cap, the potential loss is unbounded and no finite bond or withholding rate can insure it. The velocity constraint is therefore not merely a spending limit — it is the precondition that makes an ISL-4 system insurable, and thus deployable under Profit–Liability Symmetry, at all. Bound the authority, and the accountability becomes fundable; leave the authority unbounded, and no accountability mechanism can hold.

Stewardship Obligations *(Normative)*

Classification and constraint are static; authority, once granted, must be continuously governed. The framework attaches to each level a schedule of standing obligations that scale with the authority held: independent audit; complete, tamper-evident logging of actions and delegations; demonstrated rollback and a deterministic override that halts actuation within a bounded latency (on the order of sub-second for ISL-4 systems acting at machine speed); adversarial red-teaming against the Prohibited Authority Thresholds; and periodic recertification, after which authority lapses automatically unless renewed. Two further obligations are specific to autonomous infrastructure: an override mechanism that is *independent of the system's own compute*, so that a compromised system cannot suppress its own halt; and an orphaned-system rule — any ISL-4 system operating without an authenticated, live custodian is deemed orphaned and must immediately revert to Fallback-to-Passive. Authority that cannot be audited, halted, and allowed to lapse is not governed authority; it is merely authority that has not yet failed.

Worked Example: An Autonomous Corporate-Treasury Swarm *(Illustrative — assessed structure, projected consequences)*

The framework's value is not in its abstractions but in the concrete profile it produces. Consider a realistic contemporary deployment: a firm operates an autonomous multi-agent system to manage short-term liquidity — moving funds between accounts, executing hedges, allocating to money-market instruments, and settling invoices at machine speed, without per-transaction human approval. The values below are illustrative parameters a custodian would set from their own risk assessment; the purpose is to show the framework producing an enforceable profile, not to prescribe these numbers.

Layer I (Capability, assessed): high autonomy; persistent operation; direct resource access to capital; moderate operational scale; no physical embodiment; capable of spawning bounded sub-agents. A capable, unembodied, capital-actuating system.

Layer II (Custody, assessed): Named Legal Custodian is the firm's treasurer, a domiciled corporate officer, bonded. Delegated authority is scoped by the constraints below.

Reversibility: a sub-second out-of-band kill-switch independent of the trading compute; orphaned-state reverts to Fallback-to-Passive. Accountability is pinned to the NLC under Profit–Liability Symmetry — the desk that books the yield is the guarantor of the loss.

Layer III (Consequence Map, projection): under adversarial market conditions, an exploited pricing error or a feedback loop could drive rapid unauthorized outflow. Time horizon: near-term. Confidence: moderate. Assumption: adversarial conditions and a live exploit. Flagged as projection, not fact; it sets what to monitor, not a score.

Classification: ISL-4 (autonomous infrastructure with capital actuation).

Constraints, instantiated: Static ceiling C_{max} = 50,000 per commitment — any single move above it halts for human authorization. Velocity window W = 1 hour, budget B_W = 250,000 per rolling hour ($\dot{V}_{max}$ = 250,000/hr) — one hundred fragmented transfers of 2,500 still sum to the cap and the next is denied, dropping the swarm to Fallback-to-Passive. Escrow: with an annual gross yield attributable to the desk of $\bar{R}$ = 20,000,000, an expected loss rate of $v \cdot \mu_L$ = 2/yr × 500,000 = 1,000,000/yr, and an operating margin of 40% (m = 0.40), the solvency floor is $\alpha > 1{,}000{,}000 / 20{,}000{,}000 = 0.05$ and the viability ceiling is $\alpha < 0.40$. Setting a safety loading θ = 0.5 gives α = 0.075 (7.5% withholding), which yields an inflow of 1,500,000/yr against 1,000,000/yr expected — positive drift, the pool replenishes. The maximum credible single-event loss L_{max} = 2,000,000 sets the baseline bond $E_0 \geq 2{,}000{,}000$, posted before deployment.

Feasibility, and the governance signal: the expected loss rate (1,000,000/yr) is below the operating margin (8,000,000/yr), so the corridor is non-empty and a feasible α exists — the deployment is insurable and permitted at ISL-4 under this charter. Had the same system faced a worst-case severity ten times higher, an expected loss rate of 10,000,000/yr would exceed the 8,000,000/yr margin, the corridor would be empty, and the framework would refuse the deployment at ISL-4 regardless of the system's technical sophistication.

The resulting stewardship profile: ISL-4; custodian = named, bonded treasurer; C_{max} = 50,000; B_W = 250,000/hr; α = 7.5%; E_0 = 2,000,000; override latency sub-second, compute-independent; recertification at 180 days; Prohibited Authority Thresholds enforced (no kinetic actuation, no cyber-exploit generation, no CBRN pathways, no identity forgery, capital velocity bounded, recursive spawning depth-limited). This profile is enforceable, checkable, and comparable across deployments — which is what a governance artifact must be, and what a list of dimensions alone can never produce.

Closing: The Layer Above *(Normative)*

The Intelligence Stewardship Framework governs the authority of deployed systems — the layer the field's capability instruments leave open. It classifies systems by what they are and how they are held, separates present fact from projected consequence, tiers permitted authority, fixes lines that authority may not cross, binds economic yield to economic liability, and reduces those principles to quantities a runtime can enforce before an action executes rather than a court can assign after a loss. It presupposes the capability

governance of the NIST framework, the EU Act, and the laboratories' scaling policies, and layers custody on top of them. It is a first version, offered for the scrutiny and revision that any governance instrument must survive to earn use. The worked example is not a proof that these particular numbers are right; it is a demonstration that the questions of custody, authority, and accountability can be made concrete, enforceable, and comparable — which is the necessary condition for governing intelligence rather than merely describing it.

Section IX — Notes

Positioning claims (that NIST RMF, the EU AI Act, and frontier scaling policies govern capability and largely leave delegated authority/custody unaddressed): grounded in the verified landscape compiled for Section II — EU AI Act risk tiers and the 10^25-FLOP systemic-risk threshold (Art. 55); NIST AI RMF 1.0's four functions and its CAISI AI Agent Standards Initiative (Feb 2026) addressing identity/authorization at the engineering layer; Anthropic RSP (ASL), OpenAI Preparedness (High/Critical), DeepMind FSF (critical capability levels) as capability-threshold instruments. See Section II Source Notes 1–7. Re-verify before publication.

Mathematics. The velocity constraint is stated as a rolling-window sum $S_W(t)$ with units of currency, capped at a per-window budget B_W (equivalently a rate $V_W = S_W/W \le \dot{V}_{max}$), correcting a prior time-integral formulation that was dimensionally inconsistent and treated discrete events as continuous. The escrow solvency argument uses standard collective-risk / Cramér–Lundberg results: baseline coverage $E_0 \ge L_{max}$; positive safety loading $\alpha \cdot \bar{R} > v \cdot \mu_L$ (with explicit loading θ); Lundberg's inequality $P(\text{ruin}) \le e^{(-\kappa \cdot E_0)}$; and a viability ceiling $\alpha < m$ yielding the feasibility condition $(v \cdot \mu_L)/\bar{R} < m$. These are textbook actuarial results applied to the escrow; a full derivation of the adjustment coefficient κ belongs in a technical appendix, not the main text.

Profit–Liability Symmetry is presented as a proposed normative doctrine (custody of yield should not be severable from custody of risk), explicitly not as a description of current corporate liability law, which permits the decoupling the doctrine forbids. The economic argument is externality internalization / moral-hazard elimination.

Deliberately routed elsewhere (not in the monograph): the executable ISL-4 Delegation Charter JSON Schema and the RFC 8693 OAuth-for-Agents token-exchange specification belong to the companion Technical Standard; this section refers to their enforcement (cryptographically scoped delegation, monotonic narrowing, out-of-band kill-switch) at the doctrinal level only. Patent, SECI, and product/roadmap material are internal blueprints and do not appear here; at most, the author's operation of such systems is noted once, in the Preface, as evidence rather than credential.

ACT IV — INSTITUTIONS

What must exist for civilization to sustain that governance? The enduring principles and the institutional work that outlives this document.

Chapter 10 — Principles for the Intelligence Age

What a Civilization Must Hold to Keep Custody of Intelligence

What This Chapter Refuses to Be

A monograph that has argued carefully for sixty pages can undo itself in its final chapter, and the way it does so is predictable. It ends with principles — and the principles are noble, agreeable, and empty. *Intelligence should be transparent. Systems should be accountable. Humans should remain in control. Technology should be fair.* Every one of these is true; not one of them costs the reader anything to accept; and a set of commitments that no one could object to is a set of commitments that will move no one and bind no one. The greeting-card ending is the standard failure of the serious book, and this chapter's first task is to refuse it.

The principles that follow are not offered as virtues to be admired. They are offered as the compressed conclusions of the argument this monograph has made — each one a claim the preceding chapters earned, restated in its hardest and most consequential form. A principle worth stating is one that rules something out: that tells a reader, or a company, or a government, what they may not do, what they must give up, and whose interest it cuts against. A principle that forbids nothing decides nothing. So each of these carries an edge, and the edge is stated plainly, because a commitment whose cost is hidden is not a principle but a slogan. These are not new material; the monograph introduces nothing in its final chapter that it did not establish before it. They are the argument distilled to the smallest set of commitments a civilization would have to hold to keep custody of intelligence as it becomes abundant — and the reader who finds a principle too demanding is invited back into the chapters that proved why the demand is not optional.

The Principles

First: Custody is the question, not access. Whether intelligence is open or closed governs its distribution; it does not govern who holds authority over what a deployed system does, whether that authority is accountable, or whether it can be revoked. A civilization that organizes its governance around the open–closed axis is governing the wrong variable. *The edge:* this forbids the comfortable substitution of the access debate for the authority question, and it cuts against nearly the whole of the present policy conversation, which has invested enormously in a distinction that does not reach the thing that matters. To accept this principle is to accept that much of what currently passes for AI governance is aimed slightly to the side of the target.

Second: Ownership confers no authority, and authority without accountability is not custody. To own a system is not to hold legitimate authority over what it does; and to hold authority without being answerable for its exercise is not custody but its counterfeit. Custody is authority-with-accountability, and nothing less earns the name. *The edge:* this forbids the powerful from mistaking possession for legitimacy, and it forbids any actor — corporate, state, or otherwise — from wielding authority over consequential intelligence

while arranging to not answer for the consequences. It cuts against every structure designed to capture the value of a system while shedding responsibility for its harms.

Third: Delegation must remain bounded, revocable, and accountable — or it is not delegation but surrender. Authority may be handed to an artificial agent only insofar as it can be kept within a defined scope, withdrawn at will, and traced to a human who remains answerable. These three properties are what separate legitimate delegation from abdication, and unlike delegation among humans, nothing supplies them automatically when the delegate is a machine; they must be built and defended. *The edge:* this forbids the deployment of authority that cannot be bounded, cannot be revoked, or cannot be traced — however capable or valuable the system that would hold it. It rules out a class of deployments that would otherwise be profitable, and it says so.

Fourth: Some authority must remain permanently human — not until machines improve, but regardless. The permanent human line does not rest on human decision-superiority, which capability erodes, but on three things capability cannot supply: accountability, which requires a human who can be answerable; reversibility, which requires a human who can refuse the irreversible; and dignity, which requires that certain decisions about a human life be met by a human reckoning. Where these are essential, authority stays human however capable the machine. *The edge:* this is the monograph's hardest commitment, and it is stated as the explicitly normative claim it is. It forbids the delegation of foreclosing, unaccountable, and dignity-bearing authority even in the cases — real cases, not imagined ones — where a machine would decide better. It accepts a cost in outcomes at that narrow intersection, and it defends the cost, because the trade it refuses is irreversible.

Fifth: Capability and authority are different axes, and conflating them is the central error of the age. What a system *can* do and what it *may* do are not the same question, and the whole of custody depends on holding them apart. The tendency to let a system's growing capability silently expand its authority — to reason that because it can, it should be permitted to — is the mechanism by which the permanent line is eroded one advance at a time. *The edge:* this forbids the most natural and most dangerous inference in the field, the slide from capability to permission. It cuts against the momentum of deployment itself, which everywhere pushes toward granting a capable system the authority its capability seems to invite.

Sixth: The concentration of power over intelligence must remain accountable — including the concentration that calls itself safety. As the value of abundant intelligence pools at the substrate that produces it, and as authority over the most capable systems concentrates among a few actors, no such concentration may be exempt from accountability — and the exemption is most dangerous precisely when it is claimed in the name of responsibility. *The edge:* this turns the monograph's own thesis on its friends. It forbids the argument, common among the most safety-conscious actors, that concentration of control is acceptable so long as the controllers are the responsible ones.

It cuts against incumbents, including the well-intentioned, and it insists that "trust us, we are the careful ones" is precisely the claim accountability exists to refuse.

Seventh: Governance must hold against intent, not only against accident. Custody fails through erosion, but it is also attacked, deliberately and at machine speed, by adversaries who exploit its exact gaps — and the safeguards of custody are themselves among the first things attacked. A framework built only for good-faith operation is no framework at all against an adversary. *The edge:* this forbids the comfortable assumption of good faith on which most governance is quietly built. It requires that every safeguard be designed on the assumption that it will be attacked, which is more demanding and more expensive than governance for accidents alone — and the documented record shows the assumption is not paranoia but description.

Eighth: Where custody cannot be exercised in real time, it must be inscribed in advance and audited in arrears — and its narrowing must be respected. As intelligence moves beyond the human loop — by distance, by autonomy, by speed — real-time authority becomes impossible, and custody contracts to what can be fixed in a system's objectives before it acts and reviewed in its record after. This weaker custody is real, but it is weaker, and its weakness is itself a reason to decide, before deployment, what authority may cross beyond the reach of intervention. *The edge:* this forbids the pretense that a system beyond real-time human reach is governed as a system within it. It requires admitting that some authority, once placed beyond the loop, cannot be recalled — and therefore that the decision to place it there is graver than it appears.

Ninth: Claims must be disciplined — assessed, projected, and normative kept distinct — because governance built on blurred claims cannot hold. The governance of intelligence must rest on what can be observed, must mark what it projects as projection with its assumptions stated, and must argue its normative premises openly rather than smuggling them in as findings. This is not merely intellectual hygiene; it is a condition of legitimacy, because a governance that cannot distinguish what it knows from what it fears and what it wants will be neither trusted nor durable. *The edge:* this forbids the sensational and the prophetic alike — the catastrophist who dresses fear as fact and the booster who dresses hope as inevitability — and it holds this monograph to the same standard it demands of others.

Tenth: The measure of success is retained human authorship, not human supremacy. The goal of custody is not that humans remain more capable than machines, nor that machine intelligence be suppressed, nor that its benefits be forgone. It is that humanity remains the *author* of its own future — that the consequential choices about how intelligence is used, and toward what ends, remain answerable to human judgment and reversible by human hand. Custody is not a wall against intelligence. It is the condition under which a civilization can accept intelligence's abundance without surrendering its authorship. *The edge:* this forbids both the reflex to reject the technology and the drift to surrender to it, and it names the actual thing at stake, which is neither safety nor progress

in the abstract but the narrower and more precious question of who remains the author when the most powerful tool in history can increasingly act on its own.

Why These, and Not Others

A reader may notice what is absent from this list. There is no principle of fairness, no principle of transparency as such, no principle of beneficence — not because these are unimportant, but because they are not the specific commitments the custody problem requires, and a principles chapter that reached for every good thing would dilute the few that are load-bearing into a wash of the admirable. The principles here are narrow by design. Each answers to something the argument established: the primacy of custody over access to the pivot of Chapter 2; ownership and delegation to the analysis of Chapter 7; the permanent line to its normative core; concentration to the economics of Chapter 3; adversarial robustness to Chapter 8; inscribed custody to the geography and latency of Chapters 4 and 5; claim-discipline to the method the whole monograph holds; and authorship to the Central Question itself. They are not the ten best principles imaginable. They are the ten the argument compels, and their authority is borrowed entirely from the chapters that earned them.

This is also why none is offered as self-evident. Every one rests on a premise, and where the premise is contestable — most of all the fourth, which holds that dignity and accountability require a human authority that no capability supplies — the monograph has stated the premise as a premise and argued for it, rather than presenting it as a fact that only the foolish would deny. A principle honestly held is a principle whose cost and whose ground are both visible. That visibility is the difference between a commitment a civilization can actually adopt and a slogan it can only applaud.

The Close

This monograph began with a question: as intelligence becomes abundant, who will hold custody over it, what authority may be delegated to it, and how can that authority remain accountable to humanity? It has tried to earn an answer rather than assert one. It argued that intelligence, civilization's scarcest resource, is becoming abundant, and that the abundance reorganizes economic power toward the substrate that produces it. It argued that the governance question is not how intelligence is distributed but who holds authority over it — and that this authority is coming apart from the humans meant to hold it along five axes at once: as it disperses beyond jurisdiction, rises beyond the reach of enforcing bodies, travels beyond the human loop, and acquires an agency that dissolves the assumption that it is merely a tool. It argued that these are one problem, the custody problem, with a permanent human line at its center; that the line is attacked as well as eroded; and that a framework can hold it, if the framework is built for authority rather than only capability, for intent rather than only accident, and as a complement to the instruments that already govern what machines can do.

The monograph has kept a single promise throughout: it has not tried to predict the future. It does not know whether transformative machine intelligence arrives in five years or fifty,

whether it comes through open ecosystems or closed laboratories, whether the most consequential systems end up on Earth or beyond it. It has argued that the custody question survives every one of these uncertainties, because it arises the moment consequential authority is delegated to a system that acts on its own — which is not a forecast but a description of the present, thinly now and less thinly with each advance.

Humanity has spent millennia learning to govern the things that confer power: land, wealth, energy, weapons, the authority of the state itself. Each was hard, each was learned late and imperfectly, and each remains contested. The governance of intelligence is the next of these, and in one respect the hardest, because intelligence is the faculty by which we govern everything else. To lose custody of it is to lose the capacity to govern at all.

The defining question of the coming century, then, is not how intelligent our machines become. On present trajectories they will become very intelligent, and much of what follows will be good. The question is whether, as we delegate more and more to systems that increasingly act without us, we retain custody over the authority we hand them — whether we remain, at the end of it, the authors of our own future or merely the subjects of choices made faster than we can answer for. That is what custody protects. It is worth the difficulty of keeping.

Chapter Notes

Chapter 10 is the closing synthesis. It introduces no new empirical or conceptual material and required no new verification; every principle is a compression of an argument made and (where empirical) verified earlier in the monograph. Its discipline is that each principle (a) restates an earned conclusion, (b) carries a stated edge — what it forbids, costs, or cuts against — and (c) where its premise is contestable, states the premise as a premise. It is written to be sharp; per the monograph's own standard it is marked "Drafted, pending review," because a finale a work calls decisive before external review has seen it commits the exact self-assessment error the monograph argues against.

1. **Anti-greeting-card frame.** Deliberate refusal of the agreeable-virtue list (transparency/accountability/fairness as bare nouns). Each principle must rule something out and name its cost/opponent, or it is excluded as non-load-bearing. This is a normative-rhetorical commitment, argued in-text.

2. **Provenance of each principle (distillation, not new claims):** (1) custody-over-access ← Ch.2; (2) ownership≠authority≠custody ← Ch.7; (3) bounded/revocable/accountable delegation ← Ch.6–7; (4) permanent human line on accountability/reversibility/dignity ← Ch.7; (5) capability≠authority ← Ch.6–7, threaded throughout; (6) no-exemption-for-safety concentration ← Ch.3 (+ turns the thesis on incumbents); (7) hold against intent ← Ch.8; (8) inscribed-in-advance/audited-in-arrears custody ← Ch.4–5; (9) assessed/projected/normative claim-discipline ← the monograph's method (Preface + Page Zero Rule 3); (10) human authorship not supremacy ← the Central Question.

3. **“Why these, not others” section.** Explicitly defends the omission of fairness/transparency/beneficence as principles — not as unimportant but as non-specific to the custody problem — to justify the narrowness and prevent dilution. Owns that principle 4 is the most contestable and rests on an argued normative premise.

4. **The close.** Returns to the Central Question and the Promise (no prediction; the question survives all timelines), recapitulates the argument’s spine in one movement, and lands on human authorship as the thing custody protects — “not how intelligent our machines become … but whether we remain the authors of our own future.” Emotional/intellectual landing earned by the prior chapters, not reached for; no new claim introduced.

Conclusion

The argument of this monograph concludes in Chapter 10, whose closing section returns to the Central Question and the Monograph’s Promise and states the work’s final claim. No separate conclusion is offered here, because a second ending in a weaker voice would only diminish the one the argument earned. The reader seeking the monograph’s conclusion should read the close of Chapter 10.

Appendices

[DRAFT SKELETON — pre–Page Zero framing, pending deepening to pillar/at-depth standard. Included so the argument reads end-to-end; read for placement and gaps, not as finished text.]

Appendix A: Key Definitions

- **Agency:** The capacity of an artificial system to decompose > complex goals, execute multi-step workflows, utilize external > tools, and adapt to environmental feedback without continuous > human prompting.

- **Autonomy:** The operational condition under which a system > executes tasks and makes binding determinations without real-time > human authorization or intervention ("human-out-of-the-loop").

- **Custody:** The legal, physical, and operational sovereignty over > an artificial intelligence system, encompassing ownership, > responsibility for outcomes, and the power of revocation.

- **Stewardship:** The continuous, institutional governance and > monitoring required to ensure that a custodied intelligence > remains within safe, ethical, and legally bounded operational > thresholds.

- **Frontier Intelligence:** Advanced foundation models, multi-agent > systems, or autonomous architectures that exceed currently > deployed industry baselines in computational capability and > real-world execution.

Appendix B: Historical Timeline of Cognitive & Industrial Transitions

- **1440 — The Printing Press:** Mechanization of information > distribution; collapses the clerical monopoly on memory and > literacy.
- **1760 — The Industrial Revolution:** Mechanization of physical > kinetic labor; separates human muscular limits from productive > output.
- **1880 — Electrification & Telegraphy:** Instantaneous > transmission of energy and signals; collapses geographic distance > as a barrier to coordination.
- **1969 — The Internet & ARPANET:** Decentralization of data > networks; establishes global infrastructure for frictionless > information exchange.
- **2020s — The Intelligence Age:** Mechanization of cognitive labor > and autonomous decision-making; ends human cognitive scarcity and > initiates the civilizational custody crisis.

Appendix C: Suggested Research Agenda (Orchestrator Institute Seed Topics)

To operationalize the Intelligence Stewardship Framework across global institutions, Orchestrator Institute Research Labs directs future research toward three priority vectors:

1. **Cryptographic Delegation Protocols for Agentic Webs:** > Standardizing OAuth 2.0 and OpenID Connect extensions for > autonomous AI agents to establish cryptographically verifiable > chains of custody across multi-agent enterprise networks.
2. **Hardened Kill-Switch Architectures in Speed-of-Light Latency > Environments:** Engineering deterministic, fail-safe physical and > logical override mechanisms for orbital, lunar, and deep-space > autonomous infrastructure where real-time human intervention is > impossible.
3. **Cross-Jurisdictional Liability Frameworks for Synthetic > Organizations:** Developing international legal models that > apportion liability when decentralized, multi-custodian agent > swarms autonomously trigger financial or logistical market > failures.

About the Author

Mark Sendo is the Founder of Orchestrator Institute Research Labs and serves as its Master Orchestrator. His work focuses on the governance of autonomous intelligence, AI-

native production systems, and institutional frameworks for human stewardship of machine intelligence.

References

Compiled from the per-chapter Source Notes above (V2-02, resolved-with-pending 2026-07-30). Every entry reproduces only what the corresponding note already states — organization/author, document or dataset name, and version/date where the note supplies one. **No URL or DOI has been added to any entry**: none of the source notes carried a resolvable locator in the drafted text, and none has been invented. All entries are therefore listed under "Sources referenced — formal locators pending (v1.1)" below, organized by class for readability. A future version will insert verified URLs/DOIs and pinpoint citations as each is independently confirmed; nothing here should be treated as a clickable citation until then.

Sources referenced — formal locators pending (v1.1)

I. Statutes, Regulations, and Treaties

1. Regulation (EU) 2024/1689 (the EU AI Act) — European Union, in force 1 August 2024; GPAI obligations from 2 August 2025; enforcement powers and high-risk rules from 2 August 2026; Art. 6(1) classification rules from 2 August 2027; Arts. 53, 54, 55.
2. Outer Space Treaty (1967) — Arts. II, VI, VII, VIII.
3. Convention on Registration of Objects Launched into Outer Space (Registration Convention, 1975/1976).
4. Convention on International Liability for Damage Caused by Space Objects (Liability Convention, 1972).
5. Clarifying Lawful Overseas Use of Data Act (CLOUD Act) — United States, 2018.
6. General Data Protection Regulation (GDPR) — European Union.
7. Cybersecurity Law of the People's Republic of China (CSL), 2026 amendments.
8. Personal Information Protection Law (PIPL) — China.
9. Personal Data Protection Law (PDPL) — Saudi Arabia.
10. Cloud Security Assurance Program (CSAP) — South Korea.
11. Colombia, Resolution 372 (cloud-procurement localization), June 2025.
12. United Nations Convention on the Law of the Sea (UNCLOS) — high-seas/territorial-waters jurisdictional zones.

II. Government and Standards-Body Publications

13. NIST AI Risk Management Framework 1.0 (NIST AI 100-1) — U.S. National Institute of Standards and Technology, released 26 January 2023.
14. NIST Generative AI Profile (NIST AI 600-1) — July 2024.
15. NIST Trustworthy AI in Critical Infrastructure profile concept note — 7 April 2026.
16. NIST Center for AI Standards and Innovation (CAISI), AI Agent Standards Initiative — announced February 2026; AI Agent Interoperability Profile anticipated Q4 2026.

17. NIST guidance on monitoring for agentic systems — March 2026.
18. International AI Safety Report 2026 — approx. 221 pages; 100+ experts, 30+ countries; chaired by Yoshua Bengio.
19. United States 2026 National Trade Estimate report — section on cloud computing and data sovereignty.

III. Frontier-Laboratory Safety and Policy Documents

20. Anthropic, Responsible Scaling Policy (RSP) — first published September 2023; v3.0; v3.1, April 2026.
21. OpenAI, Preparedness Framework — v2.0, April 2025.
22. Google DeepMind, Frontier Safety Framework (FSF) — v3.0, September 2025.

IV. Industry, Financial, and Technical Data Sources

23. Andreessen Horowitz (a16z) — commentary coining "LLMflation."
24. Epoch AI — inference-cost benchmark tracking.
25. Goldman Sachs — token-consumption growth projection (~24x by 2030).
26. NVIDIA, Form 10-Q, fiscal year 2026 — accelerator market share, data-center revenue, capital-constraint disclosure.
27. Dell'Oro Group — hyperscaler capital-expenditure estimates.
28. carboncredits.com — hyperscaler capex summary.
29. "Silicon Analysts" — GPU-hour price decomposition (source as named in the drafted note; formal publication locator not yet identified).
30. International Energy Agency (IEA) — global data-center electricity-demand figures.
31. Lawrence Berkeley National Laboratory (LBNL), 2024 Data Center Energy Report.
32. U.S. Energy Information Administration (EIA), Annual Energy Outlook 2026.
33. NetApp — data residency/localization/sovereignty distinctions, 2026.
34. Duality — data residency/localization/sovereignty distinctions, 2026.
35. MassiveGRID — CLOUD Act extraterritorial-reach analysis, 2026.
36. Pandectes — GDPR cross-border transfer analysis, 2026.
37. CookieYes — GDPR cross-border transfer analysis, 2026.
38. Recording Law — sovereign-cloud and China CSL amendment coverage, 2026.
39. Michael Geist — Colombia Resolution 372 commentary, 2026.
40. Data Center Frontier — orbital-compute operator coverage, 2026.
41. Introl — orbital-compute operator coverage, 2026.
42. GeekWire — orbital-compute operator coverage, 2026.
43. TechCrunch — orbital-compute operator coverage, 2026.
44. Quartz — orbital-compute operator coverage, 2026.
45. Luminix — orbital-compute operator coverage, 2026.
46. Varda — orbital-compute cost-per-watt skepticism (cited as a named counter-estimate).

47. arXiv — 2025 Space AI governance-gap analysis (preprint; specific arXiv identifier not recorded in the drafted note).
48. ScienceDirect — 2024 analysis of space-object registration-practice inadequacy (specific article identifier not recorded in the drafted note).
49. NASA NTRS — spaceflight-operations communication-delay study.
50. Human Systems Integration Architecture (HSIA) / cislunar communication-delay lab study, 2025.
51. Kintz, N. M., et al. (2016) — spaceflight communication-delay effects on team performance.
52. Larson, L., et al. (2019) — spaceflight communication-delay effects on team performance.
53. Lonestar Data Holdings — announced lunar-surface data-center payload plans (2027).
54. METR — AI-agent task-duration/autonomy tracking methodology.
55. Stanford / Harvard — comparative analysis of agentic-system demo-to-production reliability gap (specific paper not identified in the drafted note).
56. Kingy AI, 2026 state-of-agents analysis.
57. Check Point Research, AI Threat Landscape Digest, March–April 2026.
58. HiddenLayer, 2026 AI Threat Landscape Report.
59. PwC, Annual Threat Dynamics 2026.
60. Forrester, Top Cybersecurity Threats 2026 (and 2024 genAI-weaponization commentary).
61. World Economic Forum (WEF), Global Cybersecurity Outlook 2026.

V. Historical and Scholarly Sources

62. Jevons, William Stanley. *The Coal Question*. 1865.
63. Standard economic-history treatments of feudal land tenure, industrial capital formation, and the information economy — cited in Chapter 1 as available scholarly anchors for an argument the chapter states is framed to stand independently as reasoning, not as a claim resting on any single named source.
64. Standard collective-risk / Cramér–Lundberg ruin-theory results (actuarial mathematics) — underlying the Chapter 9 escrow-solvency argument; textbook results, not attributed to a single named publication in the drafted note.

Pending locator list. Entries 1–64 above carry no URL or DOI in this edition. The source notes that generated this list state their own re-verification requirement ("fast-moving area; re-verify before publication") and, in four cases, an explicit inability to supply a further locator without additional research (entries 29, 47, 48, 55). Resolving each entry to a verifiable locator — and adding pinpoint citations for the legal and treaty instruments — is the work of the v1.1 source-notes resolution pass. No entry above should be read as independently verified by this compilation; the compilation's contribution is organizing what the manuscript already named into one place, not verifying it.