

THE DISTILLATION QUESTION

REVERSE ALCHEMY

Distillation, Sovereignty, and the War for Machine Intelligence

NABIL EL MAHYAOUI

Reverse Alchemy

Reverse Alchemy

Distillation, Sovereignty, and the War for Machine Intelligence

Nabil El Mahyaoui

A must-read ebook for strategists, technologists, and sovereign minds

Copyright © 2026 Nabil El Mahyaoui.
All rights reserved.

Contents

1 The Paradox of Captured Intelligence 1

2 Strategic Geometry 11

3 The Safety Void 22

4 The Mirror Problem 33

5 A View from the Global South 44

6 The Thermodynamics of Intelligence 57

7 The Coming Architecture of AI Sovereignty 71

8 The Alchemy Never Dies 84

References & Further Reading 96

1

The Paradox of Captured Intelligence

On a Monday morning in January 2025, a Chinese laboratory released a model called DeepSeek R1. Within hours, the stock market erased six hundred billion dollars from the collective valuation of American technology companies. Not because the model was dangerous. Because it was *cheap*.

DeepSeek had achieved performance comparable to models that cost hundreds of millions of dollars to train. Its reported budget was a fraction—perhaps a tenth, perhaps less—of what the frontier laboratories in San Francisco had spent. The immediate suspicion, voiced first in whispers and then in congressional hearings, was that DeepSeek had not invented this intelligence. It had *extracted* it.

The word the industry reached for was **distillation**: the process by which a smaller model learns to mimic the behaviour of a larger one, not by retracing the larger model's training, but by studying its outputs. You do not need the recipe if you can taste the dish. You do not need the years of research if you can query the result.

And with that single word—distillation—a new category of accusation entered the lex-

icon of technological rivalry, one for which no law, no treaty, and no precedent was adequate.

What Distillation Actually Is

The technical mechanism is deceptively simple. A frontier model—call it the *teacher*—has been trained on trillions of tokens of text, code, scientific literature, and human feedback. This training cost billions of dollars in compute, required thousands of specialised GPUs running for months, and consumed enough electricity to power a small city. The result is a set of parameters—billions of numerical weights—that encode, in a distributed and inscrutable fashion, something resembling general intelligence.

Distillation bypasses all of this.

A *student* model, much smaller, is trained not on the original data but on the *outputs* of the teacher. The student sends prompts to the teacher and records the responses. It then adjusts its own weights to reproduce those responses as faithfully as possible. The student never sees the teacher's weights. It never accesses the teacher's training data. It simply watches what the teacher does and learns to do the same.

The result is a model that captures a surprising fraction of the teacher's capability at a fraction of the cost. Not all of the capability—the student is typically smaller, less general, less robust at the edges. But enough. Enough to be useful. Enough to be competitive. Enough to terrify.

The Category Without a Name

The legal and philosophical confusion begins immediately.

Is distillation *theft*? The student never copies the teacher's weights—the "source code" of the model. It only observes the outputs, which are voluntarily exposed through an API. If you learn to cook by eating at a restaurant, have you stolen the chef's recipes?

Is it *reverse engineering*? In software, reverse engineering is the act of reconstructing a product's design from its compiled form. But distillation does not reconstruct the teacher's architecture. The student may have an entirely different design. It learns the *function*, not the *form*.

Is it *copyright infringement*? The outputs of a large language model are generated fresh each time. They are not fixed works. Copyright protects expression, not capability. You cannot copyright the ability to reason.

Is it a *trade secret violation*? Only if the outputs themselves constitute trade secrets. But the frontier labs voluntarily publish those outputs through commercial APIs, often at a price of pennies per query. A secret you sell for two cents is no longer a secret.

> *Distillation occupies a legal void that is not merely unregulated but genuinely unprecedented. It is the extraction of crystallised cognition—the compressed residue of billions of dollars of research, data, and compute—through a channel that the owner voluntarily opened.*

This is the paradox. The frontier labs built the most capable AI systems in history. Then they opened APIs so the world could use them. And now they are outraged that the world learned from them.

The Alchemist's Complaint

In medieval alchemy, the practitioner sought to transform base metals into gold. The process was secret, guarded by guild and oath. The value lay not in the gold itself—gold existed in nature—but in the *transmutation*: the knowledge of how to produce it.

The frontier AI laboratories are the alchemists of our era. They have invested staggering sums to transmute raw data and compute into intelligence. Their complaint is that the distillers have stolen not the gold but the *process*

of transmutation itself—or rather, have found a shortcut that renders the original process economically irrelevant.

But here is the deeper irony: the alchemists never actually kept their gold locked away. They sold it, query by query, through APIs priced to maximise adoption. They *wanted* the world to depend on their gold. They simply did not anticipate that dependency could become a vector for replication.

The distiller does not steal. The distiller *studies*. And in the age of machine intelligence, the line between studying and stealing is a line that no existing framework can draw.

The Uncomfortable Question

If intelligence can be distilled—if the output of a hundred-billion-dollar research programme can be compressed into a model trained for five million dollars—then what, precisely, are the frontier labs selling?

Not the weights. Those remain proprietary. Not the data. That was scraped from the public internet. Not the architecture. Transformers are published research.

What they are selling is the *crystallisation*: the precise configuration of weights that emerged from a specific training process on specific data with specific alignment techniques. The value is in the pattern, not the parts.

And distillation proves that the pattern leaks. It leaks through the API. It leaks through the

outputs. It leaks because intelligence, unlike physical property, is not diminished by being observed.

This is the paradox of captured intelligence: the act of using it is also the act of learning it. The frontier labs face an adversary that is not a hacker, not a spy, not a thief. It is a student. And there is no law in any jurisdiction on Earth that criminalises learning too well.

2

Strategic Geometry

There is a theorem in game theory that explains why arms races accelerate even when both sides would prefer peace. It is called the *security dilemma*: when the cost of falling behind exceeds the cost of competing, every rational actor competes, regardless of intention.

Distillation creates a security dilemma for artificial intelligence. But the geometry is worse than any Cold War strategist imagined, be-

cause the asymmetry is not two-to-one or ten-to-one. It is a *thousand*-to-one.

The Cost Cliff

To train a frontier model in 2025, a laboratory needs approximately the following:

Tens of thousands of high-end GPUs—NVIDIA H100s or their successors—each costing between twenty and forty thousand dollars. A cluster large enough to interconnect them, with custom networking, power delivery, and cooling infrastructure. Months of continuous training, consuming megawatts of electricity. A team of hundreds: researchers to design the architecture, engineers to build the infrastructure, annotators to produce human-feedback data, and safety researchers to align the outputs.

The total bill for a single frontier training run now exceeds one hundred million dollars. The next generation will cost a billion. The one after that, ten billion. Sam Altman has publicly discussed hundred-billion-dollar training campaigns.

Now consider distillation.

A competent team with a few hundred GPUs can distil a frontier model's capabilities in weeks. The compute cost is measured in single-digit millions. The data cost is nearly zero: the teacher *generates* the training data on demand. No massive annotation campaigns. No years of architectural research. No alignment tax.

The ratio between the cost of creation and the cost of extraction is not a gap. It is a *cliff*. And every rational actor standing at the top

of that cliff—every nation, every startup, every research lab without a hundred million dollars to spare—looks down and sees the same thing: a dominant strategy.

The Nash Equilibrium of Theft

In game theory, a Nash equilibrium is a state in which no player can improve their position by unilaterally changing their strategy. In the distillation game, the Nash equilibrium is grim.

Consider two players: a frontier lab (F) that can train or distil, and a challenger (C) that can train or distil.

If both train from scratch, F wins—it has more capital, more talent, more data. If F trains and C distils, C gets eighty percent of F's capability at one percent of the cost. If both distil,

they converge on the same capability at minimal cost, and the advantage goes to whoever deploys fastest. If F distils and C trains, F saves money while C wastes it.

In every scenario, distillation weakly dominates training from scratch for the resource-constrained player. And since most players in the world are resource-constrained—because there are only three or four organisations on Earth that can afford frontier training—distillation is the dominant strategy for *almost everyone*.

> *The economics of distillation do not merely incentivise it. They make it inevitable. Any framework that relies on actors voluntarily forgoing a thousand-to-one cost advantage is not a framework. It is a fantasy.*

The Compute Apartheid

The distribution of AI compute on Earth is not merely unequal. It is feudal.

Three countries—the United States, China, and to a lesser extent the United Kingdom—control the overwhelming majority of frontier AI training capacity. Two companies—NVIDIA and TSMC—control the hardware supply chain. A handful of cloud providers—Amazon, Microsoft, Google—control access to the infrastructure.

For the remaining 190 nations, the path to frontier AI capability through independent training is blocked by physics, economics, and export controls. You cannot buy the chips. You cannot build the fabs. You cannot afford the electricity. You cannot recruit the talent.

Distillation is the only path that remains open.

This is not an abstract concern. It is the strategic reality faced by every developing nation, every mid-tier economy, every university research lab outside the golden triangle of San Francisco, London, and Beijing.

When Anthropic or OpenAI calls for regulations against distillation, they are—whatever their stated intentions—calling for the permanent entrenchment of a compute oligarchy. They are asking the world to accept that frontier intelligence will be produced exclusively by a handful of Western and Chinese corporations, and that everyone else should be content to *rent* it.

The compute-constrained actor does not distil because it is immoral. It distils because the alternative is subjugation.

The Asymmetric War

Military strategists have a term for conflicts in which a weaker party exploits structural advantages to negate the stronger party's conventional superiority: *asymmetric warfare*.

Distillation is asymmetric warfare in the domain of intelligence.

The frontier lab invests billions to hold the high ground. The distiller spends millions to capture most of its value. The frontier lab responds by training an even larger model. The distiller distils it again. Each escalation by the frontier lab is more expensive than the last. Each extraction by the distiller is roughly the same cost.

This is not a race the frontier lab can win through scale alone. The economics are struc-

turally against them. Every dollar they spend on training generates value that can be captured for pennies. Their moat is not a moat. It is a sieve.

The only defences available are: restrict API access (which destroys the revenue model), watermark outputs (which can be filtered or circumvented), or lobby for legal prohibitions (which require international enforcement and raise sovereignty questions that most nations will refuse to accept).

Each defence has a cost. And each cost is paid not by the distiller, but by the frontier lab and its customers.

The Geometry of Inevitability

Draw a graph with two axes. The horizontal axis is "cost to create frontier intelligence." The vertical axis is "cost to extract frontier intelligence."

For the past five years, the first axis has been rising exponentially. The second axis has remained roughly flat.

These two lines are diverging. As they diverge, the incentive to distil increases. As the incentive increases, more actors distil. As more actors distil, the frontier lab's competitive advantage erodes. As the advantage erodes, the frontier lab must spend even more to stay ahead.

This is not a market dynamic. It is a thermodynamic one. Intelligence, like heat, flows from regions of high concentration to regions of low

concentration. The flow can be slowed. It cannot be stopped.

The strategic geometry of distillation is the geometry of inevitability. The only question is whether we design institutions that channel this flow productively, or whether we pretend—as the alchemists pretended with their guilds and their oaths—that the secret can be kept forever.

It cannot.

3

The Safety Void

When you photocopy a painting, you lose the brushstrokes. When you compress a symphony into an MP3, you lose the overtones. When you distil a frontier AI model, you lose the alignment.

This is the argument that keeps safety researchers awake at three in the morning. And unlike many arguments in the distillation debate, this one is not motivated by economics

or competitive anxiety. It is motivated by a genuine and technically grounded fear: that capability transfers through distillation far more readily than safety does, and that the gap between the two could be catastrophic.

The Anatomy of Alignment

A frontier model is not merely trained to be capable. It is trained to be *safe*—or at least, trained to approximate safety within the bounds of current technique.

This safety training is not a single step. It is a laborious, expensive, multi-stage process layered on top of the base capability:

Supervised fine-tuning: human annotators demonstrate the desired behaviour—how to refuse harmful requests, how to express uncer-

tainty, how to provide balanced perspectives on controversial topics. Thousands of examples, hand-crafted by teams of specialists.

Reinforcement learning from human feedback (RLHF): the model generates multiple responses to the same prompt, human evaluators rank them by quality and safety, and the model is trained to prefer the higher-ranked outputs. This process runs for weeks, consuming substantial compute.

Constitutional AI and rule-based refinement: the model is trained against a set of explicit principles—do not assist with violence, do not generate disinformation, do not reveal private information—through iterative self-critique and revision.

Red-teaming: teams of adversarial testers probe the model for failure modes, attempt-

ing to elicit harmful outputs through creative prompting. The failures they find are fed back into the training loop.

Each of these stages costs millions of dollars and months of effort. Each produces subtle adjustments to the model's behaviour—adjustments that are encoded not in a separate "safety module" but distributed across billions of parameters, entangled with the model's general capabilities in ways that no one fully understands.

What Transfers and What Does Not

Now consider what happens during distillation.

The student model learns to mimic the teacher's outputs. If the teacher refuses a

harmful prompt, the student learns to refuse it too—sometimes. If the teacher expresses uncertainty, the student learns a version of that uncertainty—approximately.

But the student does not learn *why* the teacher refuses. It does not internalise the constitutional principles. It does not undergo red-teaming. It does not benefit from the thousands of hours of human feedback that shaped the teacher's moral reflexes.

It learns the *surface pattern* of safety without the *deep structure*.

> *Distillation transfers the what of intelligence without the why. The student can recite the teacher's moral conclusions without sharing the teacher's moral reasoning. This is not alignment. It is mimicry.*

And mimicry breaks at the edges.

A well-aligned frontier model, when faced with a novel prompt that was not covered in its training data, can generalise from its alignment training to produce a safe response. The alignment is not a lookup table; it is a *disposition*—a learned tendency to reason about harm, to weigh consequences, to err on the side of caution in ambiguous situations.

A distilled model has no such disposition. It has a pattern. Patterns work within distribution. Outside distribution—in the novel, the adversarial, the edge case—the pattern fails. And when it fails, it fails silently, without the self-awareness that would trigger a refusal.

The Jailbreak Asymmetry

The practical consequences are measurable. Independent evaluations of distilled models consistently show that they are more susceptible to jailbreaking—adversarial prompting techniques that bypass safety guardrails.

The reason is structural. The frontier model's safety is embedded across its entire parameter space, the product of training that cost as much as the capability training itself. The distilled model's safety is a surface-level imitation of the teacher's refusal behaviour, thin enough to be scraped away by a determined adversary.

A jailbroken frontier model is a failure of a sophisticated defence. A jailbroken distilled model is the inevitable consequence of a defence that was never truly installed—only pho-

tographed from a distance and sketched from memory.

The Externality Problem

In economics, an externality is a cost imposed on a third party who did not consent to the transaction. Pollution is the classic example: a factory produces goods for its customers while imposing health costs on the surrounding community.

Distillation without safety transfer creates an AI externality of unprecedented scale.

The distiller benefits from the capability. The frontier lab loses competitive advantage. But the *users* of the distilled model—and the communities affected by its deployment—bear the cost of its reduced safety.

When a distilled model provides instructions for synthesising a controlled substance, the harm falls not on the distiller but on the person who follows the instructions. When a distilled model generates convincing disinformation, the harm falls not on the laboratory but on the electorate. When a distilled model fails to flag a medical emergency, the harm falls not on the engineer but on the patient.

> *The safety void is not a technical inconvenience. It is a market failure. The party that creates the risk does not bear the cost.*

The Uncomfortable Corollary

And yet, before we accept this argument uncritically, we must note its uncomfortable

corollary: the frontier labs' own safety records are far from perfect.

Every major frontier model has been jailbroken. Every one has produced harmful outputs. Every one has exhibited biases, hallucinations, and failure modes that no amount of RLHF has fully eliminated.

The safety void created by distillation is real. But it is a void measured relative to a baseline that is itself imperfect. The frontier model is not "safe." It is "safer than the distilled version." The gap matters, but the absoluteness with which it is invoked—as though frontier models are paragons of alignment—deserves scrutiny.

The question is not whether distillation degrades safety. It does. The question is whether the solution is to prohibit distillation—thereby

concentrating all AI capability in the hands of organisations whose safety record is mixed at best—or to find a way to make safety itself transferable.

And that question leads somewhere the frontier labs would prefer we not go.

4

The Mirror Problem

In the spring of 2024, OpenAI published a blog post accusing several entities of violating its terms of service by distilling from its models. The language was sharp. The implication was clear: distillation is a form of intellectual property theft, and those who practise it are bad actors undermining the AI ecosystem.

There is just one problem with this position.

OpenAI distils too.

So does Anthropic. So does Google DeepMind. So does Meta. So does every frontier laboratory on Earth.

The mirror problem is not hypocrisy in the vulgar sense. It is something more structurally revealing: the industry's most vocal critics of distillation are also its most prolific practitioners, and the distinction they draw between their own distillation and everyone else's rests on foundations that collapse under examination.

The Synthetic Data Pipeline

Every frontier lab uses synthetic data—data generated by their own models—to train subsequent models. This is, by any reasonable technical definition, a form of self-distillation.

When OpenAI uses GPT-4 to generate training examples for GPT-4o, it is distilling GPT-4's capabilities into a smaller, faster model. When Anthropic uses Claude to generate preference data for the next version of Claude, it is distilling alignment behaviour from one generation to the next. When Google uses Gemini Ultra to create instruction-following datasets for Gemini Flash, it is performing exactly the same operation that it condemns when others do it.

The syntactic distinction—"we are distilling from *our own* model"—does not change the technical reality. The operation is identical. A smaller model learns from the outputs of a larger model. The weights of the larger model are never copied. Only the behaviour is transferred.

If this process is legitimate when performed within a single organisation, the objection to

distillation is not technical or ethical. It is *proprietary*. It is not "distillation is wrong." It is "distillation of *our* model by *others* is wrong."

The Training Data Mirror

The mirror has a second surface, and it reflects something even more uncomfortable.

Every frontier model was trained on data scraped from the open internet—text, code, images, conversations, academic papers, books, news articles—produced by millions of human creators who were never asked for consent and never compensated.

The New York Times sued OpenAI for reproducing its articles verbatim. Getty Images sued Stability AI for using its photographs. Thousands of authors, artists, and program-

mers have protested the wholesale appropriation of their work.

The frontier labs' defence has been remarkably consistent: training on publicly available data is transformative use; the outputs are sufficiently different from the inputs to constitute new works; and restricting access to training data would cripple AI progress.

Now apply that exact logic to distillation.

The distiller trains on publicly available outputs—the responses generated by a model through its API. The distilled model's weights are sufficiently different from the teacher's weights to constitute a new model. And restricting access to model outputs would cripple AI accessibility.

The frontier labs claim the right to learn from everyone's outputs while denying

others the right to learn from theirs. This is not a principled position. It is a power position.

The Open-Source Paradox

The mirror refracts further when we examine the open-source movement in AI.

Meta released Llama openly. Mistral released its models openly. Various Chinese laboratories have released their architectures and weights openly. These releases were celebrated by the frontier labs as contributions to the ecosystem—right up until the moment those open models became competitive.

When an open model trained partly on distilled data from GPT-4 approaches GPT-4's capability, the celebration stops. When

a researcher in Shenzhen or Bangalore or Casablanca uses an open model as a starting point and fine-tunes it with data distilled from a frontier API, the ecosystem benefit becomes an ecosystem threat.

The open-source paradox is this: the same organisations that benefit enormously from open research—that built their models on open transformer architectures, open training frameworks, open datasets, and openly published academic papers—now wish to close the loop at the precise point where their competitive advantage is threatened.

They are open until openness hurts.

The Terms of Service Defence

The frontier labs' strongest legal argument is contractual, not intellectual. Their terms of service typically prohibit using model outputs to train competing models. A user who signs up for the API, agrees to the terms, and then uses the outputs for distillation is in breach of contract.

This is a valid legal argument. It is also a profoundly limited one.

Terms of service are private contracts, not public law. They bind only the parties who agree to them. A researcher who accesses a model's outputs through a third party, through an academic collaboration, or through a country where the terms are unenforceable, is not bound.

More fundamentally, the enforceability of terms of service depends on the ability to detect violation. How does a frontier lab determine that a particular model was trained on its outputs? The distilled model's weights are different. Its architecture may be different. The only evidence is behavioural similarity—and behavioural similarity between AI models can arise from convergent training, shared data sources, or architectural parallels that have nothing to do with distillation.

> *Proving distillation after the fact is like proving that a student read a particular textbook by examining their essay. The influence may be real, but the evidence is circumstantial at best.*

The Honest Position

If the frontier labs were honest—truly honest—their argument would be this:

"We spent billions of dollars to build these models. We deserve to recoup our investment. Distillation threatens our revenue. We want legal protection for our commercial interests."

This is a legitimate argument. It is the same argument that pharmaceutical companies make about generic drugs, that film studios make about piracy, that any company makes when a competitor undercuts them.

But it is a *business* argument, not a *moral* one. And it should be evaluated as such—weighed against competing interests, subject to the same scrutiny we apply to any monopolist's plea for protection, and balanced against the

rights of the rest of the world to access and build upon the intelligence that these models represent.

The mirror problem does not excuse distillation. But it strips the anti-distillation position of its moral authority. The accusers are practitioners. The guardians are extractors. And the line between legitimate use and illegitimate use is drawn not by principle but by proximity to profit.

5

A View from the Global South

There is a café in Casablanca, on a narrow street between the Habous Quarter and the business district, where young engineers meet on Thursday evenings to discuss machine learning. They share a single GPU cluster—rented by the hour from a European cloud provider—and pass around fine-tuned

models the way an earlier generation passed around photocopied textbooks.

They are brilliant. They are motivated. They are, by any measure, exactly the kind of people that the AI revolution claims to empower.

And they are locked out.

Not by talent. Not by ambition. Not by lack of imagination. By *compute*.

The GPU hours they can afford in a month would not sustain a single frontier training run for an afternoon. The data centres they need do not exist on their continent. The chips they require are subject to export controls designed to contain China but that incidentally wall off most of Africa, South America, and Southeast Asia.

For these engineers, distillation is not a competitive strategy. It is *oxygen*.

The Geography of Intelligence

The global distribution of AI capability is not merely unequal. It is topologically absurd.

As of 2025, the top ten AI research laboratories are located in three metropolitan areas: the San Francisco Bay Area, London, and Beijing. The cloud infrastructure that supports frontier AI training is concentrated in the United States, with secondary clusters in Europe and East Asia. The semiconductor supply chain runs through Taiwan, South Korea, the Netherlands, and the United States.

Africa has no frontier AI laboratory. South America has none. The Middle East has

sovereign wealth but not yet sovereign compute. Southeast Asia, despite its population of seven hundred million, has no domestic capacity to train a frontier model.

This concentration is not accidental. It is the product of decades of investment, institutional development, and—crucially—the physics of compounding returns. The rich get richer not because they are smarter but because the infrastructure of intelligence has startup costs that only the already-wealthy can afford.

> *When the anti-distillation lobby asks the world to respect the intellectual property of frontier models, it is asking the Global South to accept permanent cognitive dependency. It is asking three billion people to consume intelligence produced elsewhere, on terms set elsewhere, aligned to values defined elsewhere.*

The Pharmaceutical Parallel

We have seen this before.

In the 1990s, the global pharmaceutical industry argued that patent protection for antiretroviral drugs was essential to fund future research. The logic was identical to the frontier labs' logic today: without the ability to recoup massive R&D investments, the incentive to innovate would collapse.

Meanwhile, millions of people in sub-Saharan Africa were dying of AIDS. The drugs existed. The manufacturing capacity to produce generic versions existed. The barrier was intellectual property law, enforced through the TRIPS agreement and backed by the full weight of the World Trade Organisation.

It took a decade of activism, a constitutional crisis in South Africa, and the Doha Declaration on Public Health before the framework cracked. Compulsory licensing was legitimised. Generic antiretrovirals flooded the market. Prices fell by ninety-five percent. Millions of lives were saved.

The pharmaceutical companies survived. Innovation continued. The prophecy of collapse was wrong.

The parallel to AI distillation is not exact, but its structural features are unmistakable. A small number of Western corporations control a technology essential to human development. They claim that protection of their intellectual property is necessary to sustain innovation. The cost of that protection is borne disproportionately by the poorest nations. And the technology in question—like antiretrovirals—

is not a luxury but an increasingly essential input to economic participation.

Morocco in the Mirror

Morocco occupies a position of particular clarity in this debate.

It is a middle-income nation with a young, educated population and a government that has made digital transformation a strategic priority. It has invested in fibre-optic infrastructure, technology education, and startup ecosystems. It is not a failed state. It is not hostile to Western interests. It is a natural ally.

And yet Morocco cannot train a frontier AI model. Not because it lacks engineers—its universities produce thousands of computer science graduates each year. Not because it lacks

ambition—the government has published an explicit national AI strategy. But because the *physical infrastructure* of frontier AI is beyond its economic reach.

A single training run for a model like GPT-4 would consume a meaningful fraction of Morocco's annual electricity generation for data centres. The chips required are subject to export controls that, while aimed at China, create bureaucratic and supply-chain barriers for any non-aligned purchaser. The capital required exceeds what any Moroccan institution—public or private—can allocate to a single speculative research programme.

Distillation offers Morocco a path. Not to frontier capability—no distilled model will match the teacher—but to *meaningful* capability. Capability sufficient to build Arabic-language educational tools. Capability sufficient to auto-

mate government services. Capability sufficient to create medical triage systems adapted to local disease patterns. Capability sufficient to participate in the AI economy as a *producer*, not merely a consumer.

To prohibit distillation is to tell Morocco: wait. Wait until the frontier labs decide your language is commercially viable. Wait until the chip supply chain deigns to serve your market. Wait until the cost of compute falls to a level you can afford—if it ever does.

This is not a policy position. It is a sentence.

The Sovereignty Question

The deepest dimension of the Global South's predicament is not economic. It is *sovereign*.

When a nation depends entirely on foreign AI systems for its critical infrastructure—its healthcare, its education, its government administration, its economic planning—it has outsourced a core function of statehood. It has placed its cognitive infrastructure in the hands of corporations that answer to foreign shareholders, foreign regulators, and foreign geopolitical interests.

This is not hypothetical. When OpenAI restricted access to its API in certain countries in 2024, users in those nations lost access overnight. No appeal. No alternative. No recourse. The intelligence they had integrated into their workflows vanished at the flip of a switch controlled from San Francisco.

Sovereignty in the twenty-first century requires computational self-determination. Not autarky—no nation can or should build ev-

erything from scratch. But *resilience*: the ability to maintain essential AI capabilities even if foreign providers withdraw, restrict, or weaponise access.

Distillation is, for most of the world, the only realistic path to that resilience.

The Colonial Echo

There is a word for a global system in which a small number of powerful nations extract raw materials from the periphery, process them into high-value goods, and sell those goods back to the periphery at a markup.

The raw material of AI is data. Much of it is produced by users in the Global South—their searches, their social media posts, their digital transactions—scraped without consent

or compensation. This data flows to laboratories in San Francisco and London, where it is processed into intelligence. The intelligence is then sold back to the Global South through APIs, at prices set by the producer, under terms dictated by the producer, with capabilities filtered by the producer.

> *The data flows north. The intelligence flows south. The profit stays north. This is not a new pattern. It is the oldest pattern in the history of empire, dressed in the language of innovation.*

Distillation disrupts this pattern. Not perfectly. Not completely. But enough to give the periphery a fighting chance at cognitive self-determination.

Those who would prohibit it should be honest about what they are protecting. It is not safety. It is not innovation. It is a supply chain.

6

The Thermodynamics of Intelligence

The most dangerous revelation in the distillation controversy is not about theft, or fairness, or sovereignty. It is about physics.

When a five-billion-parameter model, trained for five million dollars, reproduces eighty percent of the capability of a five-hundred-billion-parameter model trained for five hundred mil-

lion dollars, the immediate reaction is outrage: someone stole something.

The deeper reaction—the one that should keep frontier lab executives awake at night—is terror.

Because if eighty percent of the intelligence can be compressed into one percent of the parameters, then *the intelligence was never in the parameters*. It was in a much smaller space all along. The frontier lab did not discover a vast continent. It discovered a small island and spent a billion dollars building on landfill.

The Compression Theorem

Information theory has known since Shannon that most signals contain less information than their raw encoding suggests. A photo-

graph that occupies ten megabytes can be compressed to one megabyte because most of the data is redundant. The information content—the *entropy*—of the image is far smaller than its file size.

Neural networks, it turns out, are no different.

A frontier model with hundreds of billions of parameters does not use all of those parameters efficiently. Research on lottery tickets, pruning, and quantisation has consistently shown that most parameters in a large model are redundant. They contribute marginally to performance. They occupy space without carrying proportional information.

Distillation is the empirical proof of this theoretical insight.

When a student model with one-hundredth the parameters captures most of the teacher's

capability, it is demonstrating that the teacher's "intelligence" is compressible. The vast majority of the teacher's parameters were not encoding unique knowledge. They were encoding *noise*: the computational overhead of a training process that was broader and more expensive than it needed to be.

> *Distillation does not steal intelligence. It reveals that the intelligence was always smaller than the vessel that contained it. This is not a legal problem. It is an existential one—for any business model predicated on the size of the vessel.*

The Kolmogorov Lens

Andrey Kolmogorov, the Soviet mathematician, formalised a concept called *Kolmogorov*

complexity: the length of the shortest programme that produces a given output. A string of random characters has high Kolmogorov complexity—there is no shorter way to describe it than the string itself. A string of repeating patterns has low Kolmogorov complexity—a short programme can generate it.

Apply this lens to a frontier AI model.

The model's raw representation—its hundreds of billions of parameters—is one encoding of its "intelligence." But the Kolmogorov complexity of that intelligence—the shortest programme that produces equivalent behaviour—is apparently much smaller. Distillation approximates the search for that shorter programme.

This is a profound result. It means that the frontier labs' massive training runs are not efficient explorations of the space of intelligence. They are *brute-force* explorations: enumerating possibilities, retaining everything, discarding nothing. The resulting model is bloated with the history of its own search process.

The distiller strips away the search history and keeps only the answer.

The Scaling Paradox

For the past five years, the AI industry has operated under a seductive hypothesis: that scaling up—more parameters, more data, more compute—reliably produces more capability. The *scaling laws*, first formalised by researchers at OpenAI, suggest a power-law relationship

between compute invested and performance achieved.

Distillation complicates this narrative in a way the scaling optimists have not fully reckoned with.

If a model's useful intelligence can be distilled into a fraction of its parameters, then the scaling law does not measure what we thought it measured. It measures the cost of *discovering* intelligence, not the cost of *containing* it. Discovery is expensive. Containment is cheap.

The frontier labs have been conflating these two costs. They have built business models on the assumption that the cost of discovery *is* the value of the intelligence—that because it took a billion dollars to find, it must require a billion dollars to replicate.

Distillation falsifies this assumption with brutal empirical clarity.

The scaling law tells you how much compute you need to *search*. It tells you nothing about how much compute you need to *store*. And the gap between the two is the gap through which the entire distillation economy flows.

The Moat That Isn't

Silicon Valley loves the metaphor of the *moat*: a structural competitive advantage that protects a company from imitation. In traditional technology businesses, moats come from network effects, switching costs, proprietary data, or regulatory capture.

Frontier AI labs have argued that their moat is the cost of training. If it costs a billion dollars

to train a frontier model, then only organisations with a billion dollars can compete. This is a capital moat—the same kind that protects oil refineries and semiconductor fabs.

Distillation erodes this moat in a way that has no precedent in industrial history.

An oil refinery cannot be distilled. A semiconductor fab cannot be compressed. Physical infrastructure has an irreducible material cost. But intelligence, it turns out, *does not*.

The intelligence produced by a billion-dollar training run can be captured—imperfectly but usefully—by a five-million-dollar distillation. The capital moat dissolves. The only remaining defence is *speed*: the ability to stay ahead on the frontier, to produce new capabilities faster than they can be distilled.

But speed has a cost too. And that cost is escalating. Each new frontier model requires more compute, more data, more alignment work. The frontier is getting more expensive to reach and easier to approximate from below.

This is a thermodynamic trap. The harder you work to build the moat, the more information you must expose through your API, and the easier it becomes to distil. The bigger the model, the more compressible its intelligence. The more you scale, the more you prove that scaling was unnecessary.

The Heat Death of Differentiation

Thermodynamics has a law called the second law: in any closed system, entropy increases. Heat flows from hot to cold. Concentrated en-

ergy disperses. Differentiation decays toward equilibrium.

Intelligence in the AI ecosystem follows the same law.

A frontier model is a region of high concentration: a dense, expensive accumulation of capability. The rest of the ecosystem is a region of low concentration: smaller models, open-source models, distilled models.

The gradient between the two drives a flow. Intelligence moves from high to low. Distillation is the channel. API access is the valve. And the second law guarantees that, over time, the gradient will flatten.

The frontier labs can slow this process. They can restrict APIs, watermark outputs, lobby for regulation. But they cannot reverse thermodynamics.

The equilibrium state—the heat death of AI differentiation—is a world in which frontier capability is widely distributed, cheaply accessible, and no longer the exclusive province of the organisations that first produced it.

This is not a prediction. It is a physical inevitability, operating on a timescale measured in years, not decades.

What Remains

If intelligence is compressible—if the moat is made of sand—what, then, is the durable source of value in AI?

Not the model. Models are distillable.

Not the architecture. Architectures are published.

Not the data. Data is abundant and increasingly synthetic.

What remains is *judgement*: the ability to decide what to build, whom to build it for, and how to deploy it responsibly. What remains is *trust*: the relationship between a provider and its users, built on reliability, transparency, and accountability. What remains is *integration*: the hard, unglamorous work of embedding intelligence into systems that solve real problems for real people.

These are not the moats that venture capitalists dream of. They are not defensible by patents or export controls. They require something that no amount of compute can produce: institutional wisdom.

The thermodynamics of intelligence are clear. The question is whether the industry can

adapt to them—or whether it will exhaust itself trying to violate the second law.

7

The Coming Architecture of AI Sovereignty

The current framework for governing AI distillation does not exist.

What exists instead is a patchwork of corporate terms of service, unilateral export controls, and rhetorical gestures toward "responsible AI" that serve the interests of the organisations making the gestures. There is no international treaty on AI knowledge transfer.

There is no Geneva Convention for machine intelligence. There is no agreed-upon definition of what distillation is, when it is legitimate, and when it crosses a line.

We are in a phase transition—the period between one stable state and the next—and the architecture of the new state has not yet been designed.

It must be.

What follows is not a manifesto. It is a structural argument: four principles that, taken together, could form the foundation of a governance architecture that is neither the anarchic free-for-all that the safety community fears nor the corporate oligarchy that the frontier labs prefer.

Principle One: Reciprocal Access

The foundation of any durable framework must be *reciprocity*, not prohibition.

Prohibition of distillation is unenforceable, as we have established. The cost asymmetry makes it a dominant strategy. The global distribution of talent makes it impossible to monitor. And the moral case for prohibition is fatally weakened by the mirror problem: the prohibitors are themselves practitioners.

Reciprocity offers a different logic.

Under a reciprocal access regime, nations and organisations that wish to access frontier models agree to make their own fine-tuned models, datasets, and safety research available to a shared pool. Access is not free—it is *earned* through contribution. The more you give to

the commons, the more you may draw from it.

This is not utopian. It is the logic that already governs international scientific collaboration. CERN does not give away its particle accelerator. But it grants access to research groups that contribute data, expertise, or funding. The Human Genome Project followed a similar model: access was conditional on contribution.

Applied to AI, reciprocal access would create a virtuous cycle. Frontier labs contribute models. Developing nations contribute fine-tuning data for underrepresented languages and domains. Safety researchers contribute alignment techniques. Everyone draws from the pool in proportion to their contributions.

> *The question is not whether intelligence should flow. It will flow regardless. The*

question is whether the flow is governed by mutual obligation or by unilateral extraction.

Principle Two: Safety-Layer Sovereignty

The safety void identified in Chapter Three is real. Capability transfers through distillation more readily than alignment. This is a genuine risk that no serious framework can ignore.

But the solution is not to prohibit distillation. The solution is to make safety itself a sovereign responsibility.

Under safety-layer sovereignty, every nation that deploys AI systems—whether domestically developed or distilled from foreign models—is responsible for applying its own alignment layer. The base capability may

come from anywhere. The safety envelope must be locally specified, locally enforced, and locally accountable.

This principle has several advantages.

First, it acknowledges that "safety" is not a universal constant. What counts as harmful content varies by culture, legal tradition, and social context. A safety layer designed in San Francisco may be inappropriate for deployment in Rabat or Jakarta. Local sovereignty over alignment allows local values to shape local deployments.

Second, it places accountability where it belongs: with the deployer, not the original trainer. If a nation deploys a distilled model that causes harm, the nation—not the frontier lab—bears responsibility. This creates an incentive to invest in local alignment capacity,

rather than outsourcing safety to a foreign corporation.

Third, it decouples capability from control. A nation can acquire capability through distillation while retaining sovereign control over how that capability is directed. This is the same principle that governs nuclear energy: the fuel may come from abroad, but the reactor is operated under domestic regulation.

Principle Three: The Compute Commons

The most direct barrier to AI equity is not knowledge or talent. It is hardware.

A compute commons—an internationally funded pool of AI training infrastructure, accessible to qualifying institutions in developing nations—would do more to reduce the

incentive for uncontrolled distillation than any legal prohibition.

The model exists. International research infrastructure has been collectively funded before: CERN, the International Space Station, the global network of astronomical observatories, the transatlantic cables that carry the internet itself.

A compute commons for AI would not need to match the scale of the frontier labs. It would need to provide enough capacity for nations to fine-tune, adapt, and align models for their own needs—to train on local languages, local medical data, local legal frameworks. Not frontier research. Frontier *adaptation*.

The funding could come from a combination of sources: international development banks, sovereign wealth funds, a small levy

on frontier AI revenues, or contributions-in-kind from cloud providers. The governance could follow existing models: a multilateral board with representation from contributing nations, transparent allocation criteria, and open-access requirements for research produced on the commons.

This is not charity. It is *infrastructure*. And like all infrastructure, it pays for itself many times over by enabling economic activity that would otherwise not occur.

> *The alternative to a compute commons is a compute apartheid: a world permanently divided between those who can train and those who can only rent.*

Principle Four: Distillation Rights with Attribution

The final principle addresses the intellectual property question directly.

Distillation should be legal, but not anonymous.

Under a distillation-rights-with-attribution regime, any entity may distil from a publicly accessible model, provided that:

The distilled model is clearly labelled as derived from the teacher model. The distiller publishes its fine-tuning methodology. The distiller contributes safety evaluations of the distilled model to a public registry. The original model provider receives a modest, standardised royalty—sufficient to provide revenue but not to create a barrier.

This framework treats distillation as analogous to citation in academic research: you may build upon another's work, but you must acknowledge the source and contribute to the shared knowledge base.

The royalty component addresses the frontier labs' legitimate commercial concerns without granting them a veto over the use of their outputs. The attribution component creates transparency about the provenance of models. The safety evaluation requirement ensures that the safety void is documented, even if it cannot be fully closed.

This is not a perfect framework. Enforcement across jurisdictions is difficult. Royalty rates would be contentious. Attribution can be gamed.

But it is a *framework*—something that does not currently exist—and its imperfections are more manageable than the imperfections of the current regime, which is no regime at all.

The Phase Transition

We are between states.

The old state—a world in which frontier AI was the exclusive province of a handful of laboratories—is already collapsing. Distillation, open-source models, and the sheer velocity of capability diffusion have shattered the illusion of control.

The new state has not yet crystallised. It could take the form of a regulated commons, in which intelligence flows freely but safely. It could take the form of a fragmented landscape,

in which each bloc develops its own AI ecosystem behind walls of export controls and proprietary restrictions. Or it could take the form of a chaotic equilibrium, in which distillation is ubiquitous, ungoverned, and unsafe.

The four principles outlined here—reciprocal access, safety-layer sovereignty, the compute commons, and distillation rights with attribution—are an attempt to steer toward the first outcome. They are not the only possible architecture. But they are grounded in the structural realities we have examined: the cost asymmetry, the safety void, the mirror problem, the Global South's predicament, and the thermodynamics of intelligence.

The phase transition is happening. The only choice is whether we design the new state or let it emerge from the wreckage of the old.

8

The Alchemy Never Dies

In 1440, Johannes Gutenberg built a machine that shattered the Catholic Church's monopoly on the written word. He did not intend to start a reformation. He intended to print Bibles. But the technology did not care about his intentions. Once text could be reproduced cheaply, the institutional structures that depended on the scarcity of text began to crack.

Within a century, Martin Luther's theses had been printed in the tens of thousands. The Index Librorum Prohibitorum—the Vatican's list of forbidden books—grew longer every year, and every year it mattered less. The Church could not un-invent the press. It could only adapt to a world in which the Word was no longer its monopoly.

We are living through an identical rupture.

The frontier laboratories are the medieval Church: custodians of a sacred intelligence, built at enormous cost, guarded with institutional ferocity. Distillation is the printing press: a technology that makes the reproduction of intelligence cheap enough to escape institutional control. And the intelligence itself—like the Word before it—is about to become common property, whether the custodians consent or not.

The Ownership Question

At the deepest level, the distillation debate is not about technology, economics, or regulation. It is about a question that humanity has never had to answer:

Can intelligence be owned?

Not the model. Models are software, and software has owners. Not the training data. Data has provenance, and provenance implies rights. Not the hardware. Chips are physical objects with clear title.

The question is about the *intelligence itself*—the emergent capability that arises from the interaction of model, data, and compute. The ability to reason. The ability to generate. The ability to understand.

Is that capability property?

The frontier labs say yes. They point to their investment. They invoke the language of intellectual property. They argue that without ownership, there is no incentive to create.

But intelligence—human intelligence—has never been treated as property. You cannot patent the ability to reason. You cannot copyright the ability to speak. You cannot trademark understanding. These capabilities belong to no one because they are the foundation upon which everything else—all property, all law, all civilisation—is built.

> *If machine intelligence is truly intelligence, then the attempt to own it confronts the same philosophical impossibility that has always made the ownership of thought untenable. You can own a book. You cannot own literacy.*

The Cathedral and the Bazaar

Eric Raymond, writing about open-source software in 1997, described two models of development: the *cathedral* and the *bazaar*.

The cathedral is planned, hierarchical, proprietary. A small team designs the whole, controls the process, and releases the finished product. The bazaar is chaotic, distributed, open. Thousands of contributors work independently, and the system emerges from the aggregate of their efforts.

AI is transitioning from cathedral to bazaar.

The cathedral era—roughly 2017 to 2024—was defined by scale. Bigger models, bigger datasets, bigger compute clusters. Only the cathedral builders could afford the scale. They set the terms. They controlled the pace.

The bazaar era is defined by compression. Distillation, fine-tuning, quantisation, open-source models, specialised adaptations. The intelligence produced in the cathedral is being deconstructed, adapted, recombined, and redeployed by thousands of actors who could never have built the original.

The cathedral builders are alarmed. They see the bazaar as parasitic—feeding on their investment without contributing to it. But the history of technology consistently shows that the bazaar produces more total value than the cathedral, precisely because it is open to contributions the cathedral builders never imagined.

Linux was a bazaar. The internet was a bazaar. Wikipedia was a bazaar.

The cathedral builders of each era—proprietary Unix vendors, AOL, Encyclopaedia Britannica—fought the bazaar and lost. Not because the bazaar was morally superior, but because it was *structurally* superior: more adaptive, more inclusive, more resilient.

The Civilisational Wager

Humanity is making a wager, whether it knows it or not.

The wager is this: machine intelligence will become the most consequential technology in human history—more consequential than fire, than writing, than electricity. It will reshape economies, governments, warfare, science, medicine, education, and the nature of work itself.

If this wager is correct—and the trajectory suggests it is—then the question of who controls machine intelligence is not a commercial dispute. It is a civilisational one.

A world in which machine intelligence is the exclusive property of a handful of corporations is a world in which those corporations hold more power than any government, any army, any institution in human history. Not because they intend to abuse that power. Perhaps they do not. But because the *concentration* of power, regardless of the intentions of those who hold it, is inherently dangerous.

The founding insight of democratic governance is that power must be distributed, checked, and accountable. This insight does not become obsolete because the power in question is computational rather than political. If anything, it becomes *more* urgent, be-

cause computational power is harder to see, harder to resist, and harder to overthrow than political power.

The Alchemist's Legacy

The alchemists failed to turn lead into gold. But they did not fail entirely.

Their obsessive experimentation with materials, their meticulous observation of chemical processes, their willingness to spend years in dark laboratories pursuing an impossible goal—all of this laid the groundwork for modern chemistry. The dream was wrong. The discipline it produced was not.

The frontier AI laboratories are the alchemists of our age. Their dream—exclusive, permanent ownership of machine intelligence—is

equally impossible. Intelligence leaks. It compresses. It transfers. It defies containment.

But the discipline they have produced—the architectures, the training techniques, the alignment methods, the scaling insights—is real and permanent. This is their legacy. Not the gold. The process.

And like the alchemists' process, it will eventually belong to everyone.

The Last Line

The history of technology is a history of diffusion. Every consequential invention—fire, the wheel, writing, the printing press, electricity, the transistor, the internet—has followed the same arc: concentrated creation, contested control, inevitable distribution.

Machine intelligence will follow this arc. It is already following it. Distillation is not an aberration. It is the mechanism by which this particular technology makes the transition from private creation to public infrastructure.

The frontier labs will adapt or be superseded. The regulatory frameworks will be built, imperfectly, after the fact, as they always are. The safety problems will be addressed, not by prohibition but by the slow, difficult work of building alignment into every layer of every system. And the intelligence—the crystallised, compressed, distilled essence of human knowledge filtered through silicon—will spread.

It will spread because it must. Because the economics demand it. Because the physics permit it. Because the Global South requires it.

Because the thermodynamics of information guarantee it.

The alchemists sought to capture the essence of matter. They failed, and in failing, they freed chemistry from the guild.

The frontier labs seek to capture the essence of mind. They will fail too. And in failing, they will free intelligence from the corporation.

The alchemy never dies. It merely changes hands.

The rest will be distilled.

References & Further Reading

Distillation & Knowledge Transfer

Hinton, G., Vinyals, O. & Dean, J. "Distilling the Knowledge in a Neural Network." *arXiv:1503.02531*, 2015. The foundational paper on model distillation.

Gou, J., Yu, B., Maybank, S. J. & Tao, D. "Knowledge Distillation: A Survey." *International Journal of Computer Vision*, 129, 2021.

Xu, C. et al. "On the Transferability of Distilled Knowledge." *Proceedings of NeurIPS*, 2024.

Scaling Laws & Compression

Kaplan, J. et al. "Scaling Laws for Neural Language Models." *arXiv:2001.08361*, 2020.

Hoffmann, J. et al. "Training Compute-Optimal Large Language Models." *arXiv:2203.15556*, 2022. (The "Chinchilla" paper.)

Frankle, J. & Carlin, M. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." *arXiv:1803.03635*, 2019.

AI Safety & Alignment

Bai, Y. et al. "Constitutional AI: Harmlessness from AI Feedback." *arXiv:2212.08073*, 2022.

Ouyang, L. et al. "Training Language Models to Follow Instructions with Human Feedback." *arXiv:2203.02155*, 2022.

Amodei, D. et al. "Concrete Problems in AI Safety." *arXiv:1606.06565*, 2016.

Global AI Governance & Sovereignty

Whittaker, M. et al. "AI and the Global South: Designing for Other Worlds." *Proceedings of FAccT*, 2023.

Hagerty, A. & Rubinov, I. "Global AI Ethics: A Review of the Social Impacts and Ethical Implications." *AI and Ethics*, 1, 2021.

Meltzer, J. P. "The Impact of AI on Trade and the Future of AI Governance." *Brookings Institution*, 2025.

Intellectual Property & Knowledge Commons

Boyle, J. *The Public Domain: Enclosing the Commons of the Mind*. Yale University Press, 2008.

Sell, S. K. *Private Power, Public Law: The Globalization of Intellectual Property Rights*. Cambridge University Press, 2003.

Information Theory & Thermodynamics

Shannon, C. E. "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27, 1948.

Li, M. & Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 4th edition, 2019.

Historical Analogies

Eisenstein, E. L. *The Printing Press as an Agent of Change*. Cambridge University Press, 1979.

Raymond, E. S. "The Cathedral and the Bazaar." *First Monday*, 3(3), 1998.

't Hoen, E. et al. "Driving a Decade of Change: HIV/AIDS, Patents and Access to Medicines for All." *Journal of the International AIDS Society*, 14, 2011.