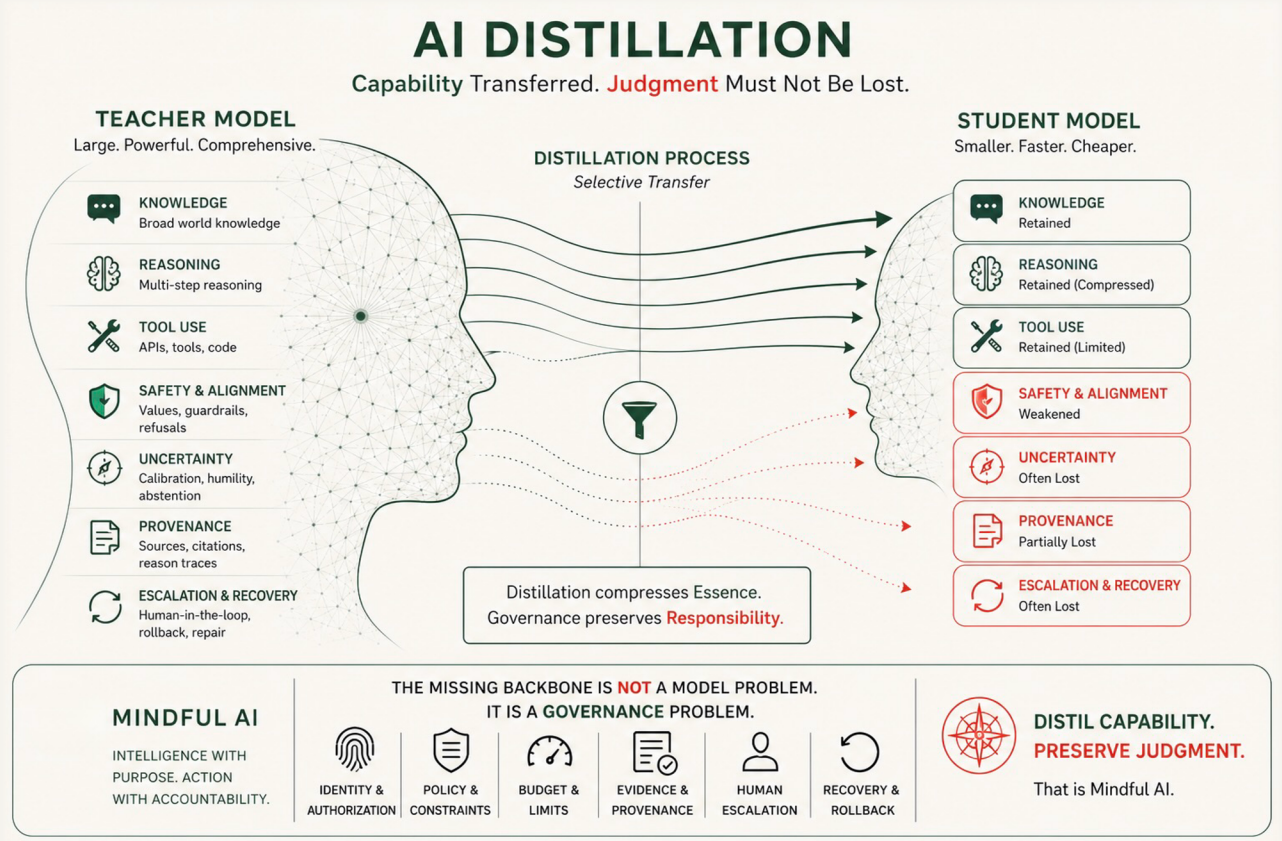
AI Distillation:
Why Smaller Models Need Bigger Governance

Max Michaels | Rao Mikkilineni, Ph.D

In artificial intelligence, the word **distillation** sounds almost elegant. It suggests refinement. Essence. The removal of waste. A large model teaches a smaller model, and what emerges is meant to be leaner, faster, cheaper, and more deployable.

That is the engineering promise. But as with many elegant ideas in AI, the deeper question is not only what is being compressed. It is **what is being lost**.



At Mindful AI Foundation, we believe AI distillation deserves more attention, not because it is peripheral, but because it is becoming foundational. Distillation is one of the main ways selected frontier-model behavior spreads into practical systems across devices, copilots, enterprise workflows, and autonomous agents. OpenAI explicitly offers model distillation to use outputs from more capable

©Mindful AI Foundation · 2026

models to improve smaller, more cost-efficient ones for specific tasks. Anthropic, from a different angle, has also described distillation as a widely used and legitimate method for creating smaller, cheaper models. DeepSeek's public release of R1 and multiple distilled variants, together with permissive license terms allowing commercial use, derivative works, and further distillation, illustrates how quickly distilled reasoning behavior can diffuse across the ecosystem.[1]

That is precisely why distillation should be examined not only as a technical optimization, but as a governance event.

What AI distillation really does

In simple terms, distillation is the process by which a larger model, often called the teacher, transfers some of its behavioral patterns to a smaller student model. The student is not merely trained on raw labels. It may be trained on outputs, probabilities, preferences, or reasoning traces generated by the teacher. The goal is not perfect replication. It is to preserve enough useful performance while reducing cost, latency, and computational burden.[2]

This has obvious advantages. A distilled model can be deployed more widely. It can run closer to the edge. It can be cheaper to serve. It can be specialized more easily for narrow workflows. That matters for enterprises, robotics, mobile systems, and firms that do not want to depend entirely on a handful of hyperscale providers.

But distillation is not just compression. It is selection.

The teacher does not hand over its mind in full. The student inherits a pattern of approximations. What survives depends on the objective function, the training regime, the data, the evaluation benchmark, and the incentives of the distiller. In that sense, distillation is not neutral. It is editorial.

The new controversy: from optimization to extraction

That editorial quality explains why distillation now sits at the center of an industry dispute. In its authorized form, distillation is a legitimate method for compressing

---

1 OpenAI, 'Model Distillation in the API,' October 1, 2024, describing an integrated workflow for using outputs from frontier models such as GPT-4o and o1-preview to improve smaller, more cost-efficient models. Anthropic, 'Detecting and preventing distillation attacks,' February 23, 2026, noting that distillation is a widely used and legitimate training method for creating smaller, cheaper versions of models. See also DeepSeek-AI, 'DeepSeek-R1-Distill-Qwen-7B,' Hugging Face model card, describing public release of distilled reasoning variants.

2 For foundational and survey treatments of distillation, see Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, 'Distilling the Knowledge in a Neural Network,' arXiv:1503.02531 (2015); Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil, 'Model Compression,' in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006), 535-541; Jianping Gou et al., 'Knowledge Distillation: A Survey,' International Journal of Computer Vision 129 (2021): 1789-1819; Yuxian Gu et al., 'MiniLLM: Knowledge Distillation of Large Language Models,' arXiv:2306.08543 (2023); and Xun Xu et al., 'A Survey on Knowledge Distillation of Large Language Models,' arXiv:2402.13116 (2024).

models, lowering cost, and optimizing them for specific tasks. In its unauthorized form, it becomes something more contentious: the extraction of commercially valuable behavior from a proprietary system without the owner's permission.[3]

The public debate is no longer only technical. It is now economic, strategic, and geopolitical. Leading labs argue that unauthorized distillation can allow rivals to reproduce important capabilities at a fraction of the original development cost, while also weakening or bypassing some of the safety, policy, and business controls that surrounded the original system.

The result is not merely cheaper imitation. It is the possibility of capability diffusion without equivalent accountability. This is where the governance question sharpens. The issue is not only whether a smaller model can mimic a larger one. It is whether the architecture of responsibility travels with it.[4]

What distillation preserves, and what it usually does not

One way to make this clearer is to distinguish three different layers that are often blended.

First, model behavior. Distillation can transfer output patterns, stylistic tendencies, task competence, and sometimes reasoning structure.

Second, reliability properties. If they are explicitly optimized, some useful behaviors may also transfer: calibration, refusal tendencies, or a degree of robustness.[5]

Third, deployment governance. This is the layer distillation does not meaningfully transfer by default. It does not carry over institutional accountability, identity and permissions, audit trails, rollback, human override, or consequence learning.

That distinction matters because some safety-relevant behavior can also be distilled. That is true as far as it goes. But it does not settle the harder question. A student model can inherit patterns of refusal or calibrated confidence and still lack the runtime architecture required for responsible action in the world. System responsibility cannot be reduced to model mimicry.

---

3 Lorelei Smillie, 'What Is AI Distillation and Why Is It a Worry for the Industry?' Bloomberg, June 26, 2026. Bloomberg distinguishes authorized distillation, in which companies distill their own models or permit specific downstream uses, from unauthorized extraction of behavior from proprietary systems by third parties.

4 Anthropic, 'Detecting and preventing distillation attacks,' February 23, 2026; Smillie, Bloomberg, June 26, 2026. Anthropic characterizes industrial-scale extraction campaigns as both a competitive and national-security issue, while Bloomberg summarizes the associated legal, economic, and security concerns around unauthorized distillation.

5 On reliability properties relevant to smaller models, see Chuan Guo et al., 'On Calibration of Modern Neural Networks,' in Proceedings of the 34th International Conference on Machine Learning (2017), 1321-1330; Dan Hendrycks and Kevin Gimpel, 'A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,' ICLR (2017); and Micah Goldblum et al., 'Adversarially Robust Distillation,' AAAI 34, no. 4 (2020): 3996-4003.

The hidden philosophy of distillation
Every act of distillation answers a philosophical question, even if engineers do not phrase it that way:

What part of intelligence do we most want to keep? Do we preserve fluency? Benchmark accuracy? Style? Speed? Safety refusals? Tool use? Domain knowledge? Cost discipline? Compliance behavior? Restraint under uncertainty?

The answer shapes the resulting system. This is where Mindful AI would draw a distinction between two kinds of compression.

One compresses capability. The other compresses judgment.
The first is admirable. The second is dangerous.

A smaller model that performs a narrow task efficiently can be an engineering triumph. But a smaller model that imitates surface competence while losing uncertainty handling, provenance, escalation, and consequence awareness becomes something more troubling: a system that looks intelligent enough to be trusted, but not governed enough to deserve that trust.

That is the risk. Distillation can preserve *Kalos*, the beauty of performance, while thinning out *Agathos*, the goodness of judgment.

Why this matters now
The current AI market rewards portability, margin, and scale. Distilled models fit that moment perfectly. They are easier to distribute, easier to specialize, and easier to commercialize. OpenAI's distillation tooling reflects that demand directly. Anthropic's discussion of distillation attacks shows that the practice is important enough to treat not only as a product technique but also as a competitive and security issue. DeepSeek's open release of distilled reasoning models shows how quickly these methods can move from internal optimization to broad ecosystem diffusion.[6]

Perplexity offers a useful adjacent signal. Its Model Council feature does not rely on one model alone; it compares outputs from multiple frontier models and then synthesizes them, highlighting agreement and disagreement. That is not distillation, but it reflects a similar truth: when stakes rise, confidence in a single compressed answer is often not enough. Organizations start looking for

6 OpenAI, 'Model Distillation in the API,' October 1, 2024; Anthropic, 'Detecting and preventing distillation attacks,' February 23, 2026; DeepSeek-AI, Hugging Face model cards for DeepSeek-R1 distilled variants. Together these sources show how distillation now spans product tooling, competitive defense, and open ecosystem diffusion.

comparative evidence, triangulation, and higher-confidence control structures around the model.[7]

Once embedded in workflows, distilled models do not remain mere models. They become participants in action chains.

A distilled model may help screen resumes, triage tickets, write code, summarize legal documents, suggest treatments, route payments, guide a robot, or trigger a workflow. At that point, the question is no longer whether its outputs resemble those of a larger teacher. The question is whether the deployed system can act within authority, explain itself, remain inside cost bounds, recover from failure, and escalate when it does not know enough.

Distillation does not automatically solve any of those problems. In fact, it can worsen them. The model gets smaller, cheaper, and easier to deploy. But the surrounding architecture, the institutional memory, the policy gates, the evidence trail, and the recovery logic do not automatically come with it.

The result is a familiar pattern in AI: *capability travels faster than control*.

Governance is not unique to distilled models

The governance problem is not unique to small models. Large frontier models, small, specialized models, and distilled models all require governance when they participate in consequential action. What changes with distillation is not the nature of governance, but its deployment burden.[8]

Smaller models are cheaper, faster, easier to embed, and easier to distribute. That means the same governance obligations must travel more widely, more consistently, and at lower cost. Distillation does not create the need for governance. It makes the absence of governance easier to multiply.

The Mindful AI view

From a Mindful AI perspective, the central issue is not whether distillation works. Clearly, it does. The central issue is whether what is distilled is only inference performance, or whether the deployed system also carries forward the architecture of good judgment.

A Mindful Machine would ask more of distillation.

---

7 Perplexity, 'What We Shipped — February 6th, 2026,' changelog entry introducing Model Council, a comparison-and-synthesis feature that highlights agreement and disagreement across multiple frontier models.

8 The governance obligation is broader than distillation alone. See National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1, 2024); ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system; and Regulation (EU) 2024/1689, Artificial Intelligence Act. These sources frame governance as an enterprise obligation for consequential AI systems generally, not only for small or distilled models.

Not just:
• Did the student retain task performance?
• Did latency improve?
• Did serving cost fall?
But also:
• Did uncertainty handling survive?
• Did provenance remain visible?
• Did escalation rules remain intact?
• Did the system preserve admissible-action boundaries?
• Did it retain the discipline to abstain?
• Did the operating context include governance at runtime?

These questions matter because intelligence in the real world is never only about producing an answer. It is about acting under limits.

A distilled model may still be valuable. But without a governance layer, it is like a compact engine dropped into a vehicle with weak brakes, poor instrumentation, and no memory of prior accidents. Efficiency alone is not wisdom. Compression alone is not control.

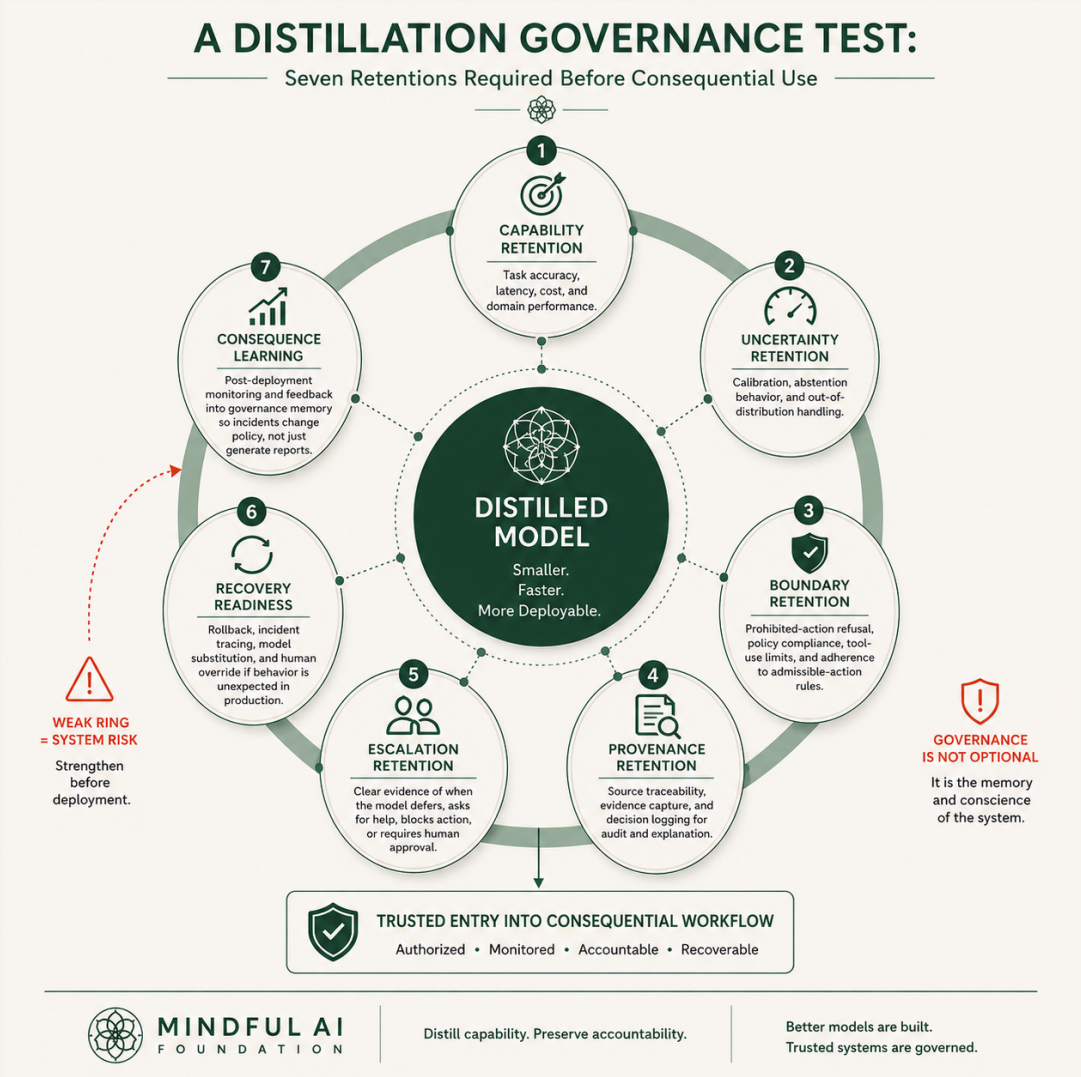
Distillation needs a governance-native deployment architecture
This is why we believe the future will not belong simply to the best distilled models. It will belong to the best governed deployments of those models.
The real unit of analysis is not the model in isolation, but the full action system around it: identity; permissions; data access; tool authority; budget limits; evidence capture; rollback pathways; human escalation; consequence learning.
In Mindful Machines language, the missing layer is the one that encodes commitments and enforces them during execution. A model may be distilled, but judgment must still be institutionalized. Purpose, boundaries, obligations, and recovery rules cannot remain external hopes. They must become operational constraints.[9]

That is the difference between a compressed intelligence artifact and a trustworthy machine.

9 See Rao Mikkilineni, 'A New Class of Autopoietic and Cognitive Machines,' Information 13, no. 1 (2022): 24; Rao Mikkilineni, 'Infusing Autopoietic and Cognitive Behaviors into Digital Automata to Improve Their Sentience, Resilience and Intelligence,' Big Data and Cognitive Computing 6, no. 1 (2022): 7; Rao Mikkilineni and William P. Kelly, 'From Static Prediction to Mindful Machines: A Paradigm Shift in Distributed AI Systems,' Computers 14, no. 12 (2025): 541; and Rao Mikkilineni and William P. Kelly, 'The Missing Layer in Modern IT: Governance of Commitments, Not Just Compute and Data,' Computers 15, no. 5 (2026): 275.

A Distillation Governance Test

To move this argument from warning to doctrine, organizations should require a pre-deployment governance test for distilled models. At minimum, seven forms of retention should be demonstrated before a student model enters a consequential workflow.[10]



Distilled models are most useful when their reduced cost and wider reach are matched by stronger runtime governance, provenance, escalation, and recovery architecture. The Governance Test above covers the key attributes.

10 For adjacent accountability and evidence frameworks, see Miles Brundage et al., 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,' arXiv:2004.07213 (2020); Margaret Mitchell et al., 'Model Cards for Model Reporting,' in Proceedings of the Conference on Fairness, Accountability, and Transparency (2019), 220-229; Timnit Gebru et al., 'Datasheets for Datasets,' Communications of the ACM 64, no. 12 (2021): 86-92; Inioluwa Deborah Raji et al., 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing,' FAccT (2020), 33-44; and Joshua A. Kroll et al., 'Accountable Algorithms,' University of Pennsylvania Law Review 165, no. 3 (2017): 633-705.

©Mindful AI Foundation • 2026

1. Capability retention — task accuracy, latency, cost, and domain performance relative to workflow objectives.
2. Uncertainty retention — calibration, abstention behavior, and out-of-distribution handling, not benchmark fluency alone.
3. Boundary retention — prohibited-action refusal, policy compliance, tool-use limits, and adherence to admissible-action rules.
4. Provenance retention — source traceability, evidence capture, and decision logging sufficient for later explanation and audit.
5. Escalation retention — clear evidence of when the model defers, asks for help, blocks action, or requires human approval.
6. Recovery readiness — rollback, incident tracing, model substitution, and human override if the system behaves unexpectedly in production.
7. Consequence learning — post-deployment monitoring and feedback into governance memory so that incidents change policy, not merely generate reports.

The question we should ask
The industry often asks, “How much performance can we preserve when we shrink the model?”

That is a valid question. But it is no longer the most important one. The more important question is this:

When intelligence is distilled, what happens to responsibility?
If responsibility evaporates into the surrounding workflow, then the model has become cheaper at exactly the moment the organization becomes more exposed.

If responsibility becomes more explicit, more encoded, more auditable, and more recoverable, then distillation may become not just an engineering tool, but a useful component of governed AI.
That is the standard Mindful AI would propose.

Smaller models, yes. But with larger memory. Clearer limits. Better evidence. Stronger escalation. And a deeper architecture of consequence.

Because in the end, the future will not be decided by which systems do more with less. It will be decided by which systems do more with less without losing accountability.

MINDFUL AI
FOUNDATION

©Mindful AI Foundation • 2026