Towards a Balanced Framework of AI Distillation

Terry Taorui Guan

The University of Hong Kong

Background

vs

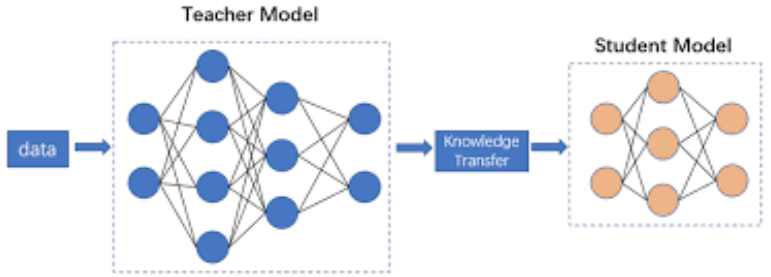
January 29th, 2025, 7:42 a.m., San Francisco

A short post appears on OpenAI's newsroom:
"We are aware of and reviewing indications that DeepSeek may have inappropriately distilled our models…"

- **$6 million** — That was the proudly publicized "training bill" for DeepSeek, a Chinese start-up claiming it built a GPT-4-level model in about eight weeks (though analysts note the figure likely covers only compute and electricity, not the full project cost).
- **"Far more than $100 million."** — OpenAI CEO Sam Altman's response, when asked at MIT whether GPT-4 cost $100 million to train.

Model Distillation ("Teacher to Student")

- **What it is**
  - Training a smaller AI model (student) to learn from the answers of a larger, more powerful model (teacher) — without using the original training data.
- **How it works**
  - Ask the teacher model smart, well-designed questions.
  - Collect high-quality answer pairs.
  - Train the student to imitate teacher behaviors.



- **Why it matters**
  - Much cheaper and faster to run
  - Can be used on smaller devices (like phones or edge hardware).
  - Helps spread access to advanced AI tools.

Are we copying without permission? Are we free-riding? Are we undermining the incentives that support costly frontier AI development?

Current Legal Landscape in the US

Doctrine	What the Doctrine Offers	Uncertainties
Copyright	Covers copying of books, images, and other creative works. But AI model weights and outputs usually don't count as "original works."	Courts have not decided whether large-scale training or downstream distillation is fair use; uncopyrightable outputs weaken direct-infringement claims
Patent	Protects new inventions, including software and algorithms.	Unclear if model can be patented. Little precedent on infringement via weight transfer
Trade Secret	Protects private data, code, and model weights kept secret	Reverse-engineering via public interfaces is presumptively lawful; the line between permissible probing and "improper means" remains blurred.
Contract	Lets model owners use Terms of Service (ToS) to block scraping, reverse engineering, or competitive training.	Enforceability is strong but limited by privity and jurisdiction.
Antitrust	Polices exclusionary conduct and monopoly maintenance.	Unclear when restrictive terms shift from legitimate self-help to unlawful exclusion

- Every legal area has gaps and gray zones. Companies on all sides are testing the limits—but courts haven't drawn clear lines yet.

Problem

- Regulatory Challenge: How Should We Govern AI Model Distillation?

Stakeholder	Illustrative Pressures	Principal Objectives
Frontier-model developers	Fear erosion of multibillion-dollar R&D incentives.	Preserve exclusivity; enforce ToS & trade-secret controls; deter "free-riding."
Distillers / start-ups / open-source	View distillation as indispensable leveling mechanism.	Secure fair-use or free-use/ reverse-engineering safe harbors; curb over-broad contract clauses; cut entry costs.
Content creators & data providers	Worry about uncompensated exploitation amplified downstream.	Consent, compensation, credit, privacy safeguards.
End-users & public interest	Expect democratization of AI capabilities.	Broader access, lower prices, reliability, safety, accountability.
Regulators & antitrust enforcers	Monitor restrictive ToS & output scraping bans for exclusionary effects.	Prevent lock-in, keep markets contestable.

How Should We Govern AI Model Distillation?

Normative lens	Core principle	Policy direction it supports	Principal caveats
Utilitarianism	Maximize net social welfare: weigh efficiency gains from wider diffusion against innovation-incentive losses and safety risks.	Calibrated balance — time-limited exclusivity followed by compulsory or collective licensing; add creator compensation and safety standards to tilt the cost-benefit ledger positive.	Requires uncertain empirical forecasts (R & D elasticity, market substitution, externalities); easy to mis-measure real-world welfare impacts.
Deontology	Acts are right or wrong per se; promise-keeping and respect for rightful ownership are inviolable duties. Unauthorized distillation that breaches ToS is intrinsically impermissible.	Consent and just compensation at every stage.	"Clean-hands" problem: the moral standing of frontier labs erodes if the teacher model itself was trained on unconsented data, complicating absolute prohibitions.

Normative lens	Core principle	Policy direction it supports	Principal caveats
Lockean Labor-Desert	Property arises from mixing labor with resources, limited by the "enough and as good" proviso. Proviso tension: mass distillation may erode the original developer's market position	Reward each contributor's labor proportionally: data creators, teacher developers, and distillers. Suggests layered royalties or benefit-sharing funds and time-bounded control over distilled outputs.	Denying distillers' labor claim risks ignoring their contribution, but will complicate the layers. Hard to qualify the portion.
Rawlsian / Egalitarian Justice	Arrange rules so that inequalities benefit the least-advantaged and preserve fair equality of opportunity.	Encourage responsible distillation to broaden access, curb concentration of AI power, and lower prices; pair with safeguards for vulnerable groups and data privacy.	Must show that restrictions on distillation truly uplift the worst-off; otherwise, justice favors openness even at incumbents' expense.

- No single theory dominates. A *workable regime* will likely blend **utilitarian calibration** (dynamic incentives), **deontological safeguards** (consent & honesty), **Lockean rewards** (labor-based compensation), and **Rawlsian equity** (broad, inclusive benefit)—each tempering the others to steer AI distillation toward both innovation *and* fairness (a balanced framework).

Type of distillation	Baseline rule	Justification
Authorized / internal (rights-holder consent)	Freely permissible; governed by contract, with courts just enforcing the license.	Defers to private ordering (consent) while respecting developers' labor-based claims and avoiding needless legal friction.
Independent commercial cloning (direct market substitute)	Lawful by contract or under a time-limited compulsory license → usage-based royalty for ≈ 1–3 yrs, then royalty-free.	Preserves investment incentives yet prevents perpetual monopoly, advancing innovation-versus-diffusion balance and fair competition.
Independent transformative enhancement (new domain / added functionality)	Conditionally allowed without royalty if not a direct substitute; may require transparency & reciprocity of safety improvements.	Encourages socially beneficial, non-substitutive reuse—an embodiment of the framework's "transformative use" principle.
Public-interest research distillation	Broad exemption for bona-fide academic, auditing, or safety work, so long as deployment remains non-commercial.	Guarantees oversight, transparency, and equitable knowledge access, echoing Rawlsian-egalitarian commitments.
Illicit distillation via "improper means" (hacking, breach, circumvention, unsafe or privacy-violating use)	Completely banned. Full legal punishment. No "fair use" excuse.	Upholds deontological duties of promise-keeping & non-theft, protects safety/privacy, and deters bad-faith free-riding.

Policies	Contents	Justification
De minimis / incidental-use exception	Carves out trivial, one-off use of a handful of teacher outputs for testing, calibration or inspiration—no license or royalty triggered.	Utilitarian: avoids chilling everyday research; Deontological: law should not punish negligible wrongs
Transparency duty for distilled models	Requires public disclosure of the teacher model lineage and high-level training method (not proprietary details) before commercial release.	Rawlsian equity: enables oversight for bias & safety affecting the least-advantaged; Lockean labor: lets creators verify downstream use of their work.
Collective licensing / statutory compensation scheme	One-stop royalty collection for creative works ingested at scale; distributes proceeds to artists and rights-holders without forcing one-off negotiations.	Lockean labor & distributive justice: creators share in value generated from their labor; preserves incentives for cultural production.
Comprehensive federal privacy law for AI data	Imposes data-minimization, purpose-limitation, consent or privacy-tech (e.g., differential privacy), plus user rights (access, deletion, opt-out).	Deontological & Rawlsian: respects individual autonomy and dignity; protects vulnerable groups from disproportionate surveillance harms.

Adjustments for Different Legal Systems

Rule Type	Frontier-developer-centric, closed-source tilt (US)	Fast-follower, open-source-leaning with strong state steerage (CN)	Creator / End-user-centric blocs (many Global-South economies)
Exclusivity time	Longer – more time to protect early models	Shorter – faster opening for reuse	Very short or none – focus on access
Contract reach	ToS is king — wide anti-distillation clauses enforced	ToS subordinated to state policy	Unfair-terms filter; research clauses cannot be waived
Royalty rules	Voluntary deals; compulsory license only after proven abuse	Government sets fair prices for reuse	Focusing payment shared with creators via collective rights body
Research carve-out	Small – only limited use allowed for research	Bigger – academic use is broadly allowed	Very broad – public interest and safety reviews always allowed

Each place adjusts the rules based on who it wants to protect most—big tech firms, fast adopters, or creators and the public. The key levers (time, contracts, payments, and research freedom) are the same, but each country sets them **differently**.

Conclusion

- AI model distillation sits at the center of two big goals: **innovation** and **access**. It makes powerful AI models smaller, faster, and cheaper—so they can be used in new products, research, and public applications. But at the same time, it risks **undermining the incentives** that support expensive, high-end AI development.
- To handle this tension, the law needs to find the right balance. One approach is to give developers a limited time to recover their investment, while also **protecting** research freedom, **providing** clear licensing pathways, and **enforcing** strong bans on unsafe or dishonest distilling.