IAPP PRIVACY. SECURITY. RISK. + AI GOVERNANCE GLOBAL 2026

Training 6-7 October

Workshops 7 October

Conference 8-9 October

AIGP exam 9 October

SEATTLE

#IAPPpsr26

Public to Proprietary

Legal Strategies for AI Data & Model Governance

Meet the Panel

Mike Cole, AIGP, CIPP/C, CIPP/E, CIPP/US, CIPM, CIPT, FIP, PLS
Vice President and Senior Legal Counsel, Hagerty, Inc.

Sean Nakamoto, CIPP/E
Associate Legal Counsel, Google DeepMind

Chen Li, AIGP, CIPP/US
Associate General Counsel, Coherent Corporation

Victor Harris, CIPP/US
Senior Product Counsel, Cohesity

Session Roadmap

1. Intro & Building and Deploying
2. Technical Primer
3. Build — Training-Data Risk When You're Building
4. Deploy — Model-Licensing Risk When You're Deploying
5. The Risk Framework — Putting It Together
6. Live Scenario — Sellers vs. Buyers
7. Q&A + Three Takeaways

Technical Primer

- **Fine-tuning:** trains model weights on your data — a new licensable artifact, with a deletion problem baked into the weights.
- **RAG (retrieval-augmented generation):** data stays in a store the model queries at runtime — cleaner deletion, a different data-flow contract.
- **Context engineering:** data lives only in the prompt or session — the lightest footprint, but mind what you send to a third party.
- **Developer vs. deployer:** who builds the model vs. who integrates it — the contract stack should mirror that split.
- **For counsel:** Always level set on definitions using the vocabulary the business will use.

Two Views of Enterprise AI

Where does your legal risk actually live?

“We’re building.”

Train or fine-tune a purpose-built model, often on composite open datasets.

Risk lives in the training data.

“We’re deploying.”

Pick up someone else's model — commercial, open-source, or open-weight — and build on top.

Risk lives in the license & the supply chain.

Build — Training-Data

- **Composite-dataset reality:** “open” datasets are rarely monolithic — they're stitched from smaller sets, each with its own license.
- **Map the license stack:** CC variants, ODbL, custom dataset licenses, and scraped corpora of unclear provenance.
- **Obligations flow downstream:** attribution, share-alike, and non-commercial terms can attach to the model weights themselves.
- **Synthetic data & distillation:** check the terms of service of any model used to generate training data — TOS often prohibit exactly this.

Build — Privacy & Copyright

- **GDPR lawful basis:** the Art. 6 problem for personal data swept into training corpora.
- **CCPA sensitive categories:** heightened obligations when sensitive personal information enters the pipeline.
- **The “machine unlearning” gap:** deletion rights collide with the reality that you can't cleanly pull one person out of trained weights.
- **Data localization:** emerging trends increasingly constrain where training pipelines can run.
- **Copyright litigation is live:** NYT v. OpenAI and the Anthropic $1.5B copyright settlement (July 2026) are teaching cases on evidentiary preservation and document hygiene.

Build — Dataset Due-Diligence

Diligence checklist

- Get a dataset bill-of-materials — every constituent set and its license.
- Trace provenance; flag scraped or unclear-origin.
- Confirm synthetic data complies with the source model's TOS.
- Filtering and sanitizing data before use.
- Map deletion / unlearning obligations before training, not after.

Terms to push for from data providers

- Provenance & rights representations and warranties.
- License flow-through and clear commercial-use grants.
- IP indemnity for training-data claims.
- Cooperation on data-subject deletion and audit rights.

Deploy — Model-Licensing Risk

- **Words matter:** open-source vs. open-weight vs. “source-available” are different things — and many “open” model licenses today are NOT OSI-approved.
- **The non-OSI landscape:** Llama Community License, Kimi's modified MIT, the Business Source License, and the RAIL family.
- **Where the traps hide:** commercial-use carve outs, acceptable-use restrictions, and downstream obligations buried in the license.
- **Ask what you can see:** representations about training data and its provenance are often thin or absent.

Deploy — Export Control, Supply Chain

- **Model origin = compliance issue:** DeepSeek, Qwen, Kimi, GLM, MiniMax — and even the inference engine's origin (llama.cpp, vLLM) — can raise export-control and national-security questions.
- **Contract:** indemnity, warranty, IP holdback caps, output ownership, and training-data reps — explain the “why” so partners pick the right battles.
- **Orchestration & the agentic layer:** Amazon v. Perplexity and NVIDIA v. Hugging Face = risk doesn't stop at the model boundary.
- **The next frontier:** govern agentic APIs like the third-party integrations they are.

Open vs. Proprietary — The Tradeoff

Open / Open-Weight

- Control, customization, on-prem & data-residency options
- Cost efficiency at scale
- **But:** license ambiguity, thin training-data reps, self-owned indemnity risk, security considerations

Proprietary

- Contractual indemnity, warranties, clearer output ownership
- Vendor-managed compliance & updates
- Limited Selection/Options
- **But:** less control, data flows to a third party, more expensive

Where's the novel opportunity? Match the model to the use case and the risk you can actually own.

Compare capability, performance & price: artificialanalysis.ai/models · epoch.ai

A Three-Pillar Framework

Map the Data & Model Stack

- **Licensing** — every component cleared for your use case
- **Training Data** — trace each dataset to source; public ≠ unrestricted
- **Content Review** — vet images, text, audio, video for customer use
- **Provenance** — public domain, licensed, or proprietary — not scraped

Identify IP / Privacy Touchpoints

- **Copyright** — assess dataset and output for infringement risk
- **Regulations** — map privacy laws by geography and use case
- **PII Exposure** — check training data and outputs for personal data
- **Anonymization** — blur faces, strip metadata, remove identifiers

AI-Specific Considerations

- **Unfair Bias Audit** — test representation to prevent skewed outcomes
- **Use Limits** — document prohibited uses (e.g., age-based profiling)
- **Generative Risk** — higher IP and explicit-content exposure
- **Enforcement** — non-generative output is lower risk, harder to police

License alone is insufficient: diligence must span every layer of the AI stack.

The Risk Framework — Putting It Together

Start here: What is the biggest threat to *this* business, for *this* use case? Not all risk is the same risk.

1. **Classify the use case** — internal tool, customer-facing, regulated vertical, or agentic.
2. **Map the data & model stack** — every license, every component.
3. **Identify touchpoints** — IP / OSS / privacy / export — per layer.
4. **Assess probability vs. severity** — incl. the "AI too big to fail" backdrop — which never excuses weak documentation.
5. **Document the tradeoff** — defensible on both regulatory and litigation timelines.

Note on Legal Hygiene & Operationalizing

- **Learn from discovery:** the OpenAI piracy settlement is a lesson in what internal communications look like when they surface.
- **How you talk about risk internally matters:** assume your emails and chats are discoverable.
- **PRA vs. regulatory-enforcement realities:** calibrate candor to the actual forum.
- **Operationalize:** build the decision into internal AI policy or procedure so choices are repeatable, auditable, and defensible.
- **Documentation is the deliverable:** a defensible record beats a perfect-but-undocumented judgment.

Live Scenario — Sellers vs. Buyers

Setup: An enterprise is evaluating a fine-tuned model from a vendor — built on an open-weight foundation and a composite open dataset — for the enterprise's internal use.

Vendor / Builder

- Sell the capability; cap the exposure.
- Defend narrow reps, limited indemnity, and holdback caps.
- We want customer data for training.

Enterprise / Legal

- Educate the business on risk tolerance, posture, and consumption.
- Use case and output dictate risk.
- Push provenance reps, broad IP indemnity and ownership, license flow-through, privacy & export reps.

Five pressure points

Model, Data-provenance, export control reps · IP ownership, indemnity scope & caps, downstream contamination· Liability, License flow-through · Privacy commitments · Regulatory alignment, cooperation

Three Takeaways

1

Identify the Stack Components

Map the model, data, licenses, and uses.

2

Apply the Framework

A practical, jurisdiction and industry-aware risk framework for evaluating tangible legal risk.

3

Document the Process

Capture the legal assessment, trade-offs, and business decision for real world uses.

Questions?

How Did Things Go?
(We Really Want To Know)

Did you enjoy this session? Is there any way we could make it better? Let us know by filling out a speaker evaluation.

1. Open the IAPP Events app.
2. Select **IAPP P.S.R. + AIGG 2026**.
3. Tap "Schedule" on the bottom navigation bar.
4. Find this session. Click "Rate this Session" within the description.
5. Once you've answered all three questions, tap "Done".

Thank you!