Week 11: Generative AI: new governance problems, old ethical questions

Section 4: Convergence

LLM-specific risks, the EU AI Act, the CUI boundary problem, and what it means to govern a system whose playbook is still being written.

> **This week's central argument.** Every risk in this week's taxonomy is new in its technical form. None of them is new as an ethical problem. Data leakage is purpose limitation (Week 3). Hallucination is data quality and accountability (Week 7). Training data provenance is consent and aggregation (Weeks 4 and 5). Bias amplification is structural reproduction (Week 10). The governance problems of generative AI are not unprecedented — they are the course's existing problems scaled, accelerated, and made harder to see because they happen inside a model rather than inside a database. The governance frameworks must be the same. The urgency is new.

This week applies the full course toolkit to generative AI — the technology that has changed faster than any governance framework built to address it. Large language models introduce a new set of technical risks: memorization of training data, inference attacks, prompt injection, hallucination, and data leakage across deployment boundaries. Each of these is technically novel. None of them raises an ethical question the course has not already examined.

The week is anchored in the EU AI Act — the most significant AI governance legislation in force anywhere — and in the operational reality of deploying a third-party LLM inside a regulated, sovereign cloud environment handling sensitive data. The regulatory framework tells students what governance should require; the operational reality shows what it looks like to build that governance in a live environment where the playbook does not yet exist.

The Palantir AIP preview from Week 9 is returned to here. What Gotham does with structured data, AIP does with natural language reasoning across integrated datasets. The governance questions compound: accuracy, bias, accountability, and data quality problems that are hard with deterministic algorithms become harder when the system generates responses rather than scores. Week 11 does not resolve those questions — it names them precisely enough that students can work on them.

LEARNING OBJECTIVES

By the end of this session, students will be able to:

- Identify the nine LLM-specific risks in the taxonomy and map each to the governance framework from the relevant prior week.
- Explain the key provisions of the EU AI Act and identify which categories of AI system face the most significant governance obligations.
- Describe the governance seam problem between a sovereign cloud environment and a third-party AI provider and identify what controls span the boundary.
- Apply the Week 3 purpose limitation principle to training data consent and explain why existing privacy law is structurally inadequate for the LLM training context.
- Connect the Week 10 structural reproduction framework to LLM bias amplification at scale and explain why the governance problem is harder, not easier, with generative AI.
- Draft the key elements of an AI acceptable use policy for a regulated environment, specifying what data classifications are permitted in prompts and what audit obligations apply.

SESSION STRUCTURE

Hour 1 — The risk taxonomy: new technical forms, old ethical problems

Open with the central argument: generative AI does not create new ethical problems. It scales, accelerates, and obscures existing ones. Walk through the risk taxonomy using that frame — for each risk, identify the prior week where the ethical problem first appeared. Data leakage is Week 3's purpose limitation. Hallucination is Week 7's data quality harm chain. Training data provenance is Week 5's consent and aggregation. Bias amplification is Week 10's structural reproduction. The taxonomy is not a list of new problems; it is a map of where existing governance frameworks need to extend.

Hour 2 — The EU AI Act and the operational reality in parallel

Walk through the regulatory framework and the operational reality in parallel. The EU AI Act establishes what governance should require: risk classification, conformity assessment, transparency obligations, and enforcement with real penalties. The operational reality establishes what governance looks like in a live regulated environment: the seam between a sovereign cloud tenant boundary and the AI provider, the audit logging gap, the DLP limitation at the human-AI interface, and the contractual framework that must span two separate legal regimes. The

governance lesson: regulatory frameworks tell you what the destination is; operational experience tells you what the path looks like. Both are necessary.

Hour 3 — The teaching cases: policy gaps, hallucination, and the seam

Walk through the teaching cases. Samsung establishes the governance gap that absence of policy creates. Mata v. Avianca establishes hallucination as an accountability problem in a professional context. The DOGE AI case establishes the most consequential version of the governance seam problem. Students who understand the prior eleven weeks' frameworks should be able to identify what is missing in a live, unsolved deployment and propose what should fill it.

Hour 4 — Workshop: drafting an AI acceptable use policy for a regulated environment

Students work in groups to draft the key elements of an AI acceptable use policy for an organization operating in a regulated environment — a federal agency, healthcare provider, financial institution, or the student's capstone organization. The policy must specify: what data classifications are permitted in AI prompts; what audit obligations apply to AI-generated work product; what human review is required before AI-assisted decisions are acted on; what contractual requirements apply to AI providers; and what the accountability structure is when AI-assisted work causes harm. Groups present; the class evaluates each policy against the EU AI Act requirements and real-world operational constraints. Close by returning to the course's central argument: governance frameworks must be built before the harm occurs. Week 11 is happening at the moment when that window is still open.

LLM-specific risk taxonomy

Data provenance and privacy

Risk	What it is	Governance implication
Training data provenance	LLMs are trained on data scraped from the internet without individual consent. The corpus may include personal data, copyrighted material, and sensitive information whose	Who consented to their data being used to train the model? Does GDPR's right to erasure apply to training data that has been encoded into model weights?

Risk	What it is	Governance implication
	subjects never authorized its use for AI training.	
Memorization	LLMs can memorize and reproduce verbatim text from training data, including personal information, proprietary documents, and copyrighted material. This is not a bug — it is a feature of how large models learn.	A model trained on leaked documents can reproduce them on request. A model trained on personal data can reveal it in response to carefully crafted prompts.
Inference attacks	Even without memorization, model outputs can reveal information about training data through statistical patterns. An attacker can infer whether a specific document was in the training set.	Differential privacy (Week 4) is the technical mitigation; it introduces the utility-privacy trade-off in the LLM context at unprecedented scale.

Deployment and operational risks

Risk	What it is	Governance implication
Prompt injection	Malicious instructions embedded in user input or retrieved content can override the model's system instructions, causing it to take unintended actions or reveal information it should not.	In an enterprise deployment, a document containing hidden instructions could cause the model to exfiltrate data or bypass access controls when a user asks it to summarize the document.
Hallucination	LLMs generate confident, fluent, plausible-sounding text that is factually incorrect. Hallucination is structural — it emerges from how models generate text — not a fixable error.	In a legal, medical, or intelligence context, a model that confidently fabricates citations, diagnoses, or threat assessments while sounding authoritative is an accountability problem, not merely an inconvenience.
Data leakage	In enterprise deployments, the model may be exposed to sensitive data through retrieval, context	The governance seam between a sovereign cloud tenant and a third-party AI provider is the live version

Risk	What it is	Governance implication
	windows, or user input. Without proper controls, that data may appear in responses to other users or be retained by the model provider.	of this risk: data that enters the provider's system is governed by the provider's DPA, not the tenant's classification scheme.

Structural and societal risks

Risk	What it is	Governance implication
Intellectual property	LLMs trained on copyrighted material can reproduce it, generate derivative works, and undermine the economic interests of original creators. Litigation is ongoing in multiple jurisdictions.	The governance question is not only legal liability but the ethics of building commercial products on the unconsented use of others' creative and intellectual work.
Bias amplification	LLMs trained on internet-scale data absorb and can amplify the biases present in that data — including racial, gender, and political biases. Week 10's structural reproduction framework applies at a scale that dwarfs any single algorithmic system.	A model that generates biased hiring recommendations, legal summaries, or medical advice at scale produces the New Jim Code effect across every domain it is deployed in simultaneously.
Disinformation at scale	Generative AI dramatically lowers the cost of producing convincing disinformation — realistic text, images, audio, and video — at scale. The governance problem is asymmetric: detection is harder than generation.	The electoral integrity, public health, and national security implications are documented and ongoing. No adequate governance framework yet exists.

The EU AI Act

Provision	What it requires
Risk classification	High-risk AI systems — including those used in law enforcement, immigration, employment, and critical infrastructure — face

Provision	What it requires
	mandatory conformity assessment, transparency, and human oversight requirements before deployment.
Prohibited practices	Real-time biometric surveillance in public spaces (with exceptions), social scoring by public authorities, and exploitation of vulnerabilities are prohibited outright.
Foundation model obligations	General-purpose AI models above a compute threshold must register with EU authorities, publish training data summaries, and conduct adversarial testing.
Enforcement	National market surveillance authorities and a new EU AI Office. Fines up to 7% of global turnover for prohibited practices. Extraterritorial application to systems deployed in the EU.
Status (2025–2026)	Phased implementation: prohibited practices in force August 2024; high-risk system requirements phasing in through 2026–2027. The most significant AI governance legislation in force anywhere.

Teaching cases

Samsung engineers pasting proprietary source code into ChatGPT (2023)

- **LLM risk category:** data leakage — proprietary code entered the model provider's system, potentially used for training, and no longer under Samsung's control.
- **Governance gap:** Samsung had no acceptable use policy for generative AI at the time. The gap was not technical — the controls that would have prevented it (DLP, policy, training) were absent because no one had built a governance program for LLM use.
- **Old ethical question:** purpose limitation (Week 3) — data shared with a vendor for one purpose (coding assistance) may be used for another (model training). The consent fiction (Week 5): engineers clicked through terms of service that authorized the use. The aggregation problem: Samsung's code is now part of a model that many users can query.

ChatGPT hallucinating legal citations — Mata v. Avianca (2023)

- **LLM risk category:** hallucination — an attorney submitted a legal brief citing six cases that did not exist, complete with invented quotations.

- **Governance gap:** no governance framework existed for attorney use of generative AI in legal proceedings. The court sanctioned the attorneys; the bar association guidance that followed was reactive. Professional licensing frameworks had not addressed AI-assisted work product.
- **Old ethical question:** accountability (Week 6) — who is responsible when an AI system confidently produces false information that causes harm? The attorney, the firm, the model provider, or the profession that failed to regulate the use? The data quality chain (Week 7): hallucination is a data quality failure in the model's output, with direct harm consequences.

DOGE reportedly feeding federal data into AI software (2025)

- **LLM risk category:** data leakage and prompt injection risk — sensitive federal data from Treasury, SSA, and IRS systems reportedly processed through AI software outside the normal federal authorization framework.
- **Governance gap:** no governance structure existed for the use case being created. The AI software was not FedRAMP authorized. The data crossing the boundary carried CUI classification obligations that the receiving system did not meet. The audit trail stops at the boundary.
- **Old ethical question:** the governance seam problem in its most consequential form — the same CUI handling obligations that apply to CUI data apply regardless of where it is processed. Sending CUI to an unauthorized AI system is a CUI handling violation whether or not the user understood that. This is the live version of the governance problem the course has been building toward.

Seminar discussion questions

1. The central argument of this week is that generative AI raises old ethical problems in new technical forms. Select one LLM-specific risk from the taxonomy and trace it back to the prior week where the ethical problem first appeared. Does the prior week's governance framework adequately address the LLM version of the problem — or does the scale or technical form require a new governance response?
2. The EU AI Act classifies AI systems by risk level and imposes the heaviest obligations on high-risk systems. A general-purpose LLM deployed for coding assistance is not inherently high-risk; the same model deployed to assist in immigration enforcement decisions may be. Does the Act's risk classification framework adequately capture where the governance obligation should attach — at the model, at the deployment, or at the decision?

3. The governance seam between a sovereign cloud tenant and a third-party AI provider means that data entering an AI prompt crosses from one contractual and regulatory framework into another. The user who pastes CUI text into an AI prompt may not know they have done something that has governance implications. What governance controls should exist at that boundary — and who is responsible for building them?
4. Hallucination is structural in LLMs — it emerges from how models generate text and cannot be fully eliminated. In Mata v. Avianca, the attorney was sanctioned. Apply the Week 6 governance failure modes: was the accountability outcome the result of a design failure in the model, a governance failure in the profession, or a bypass failure in the attorney's own practice? Who should bear the accountability — and does your answer change if the hallucination occurs in a medical, military, or immigration context?
5. Using the frameworks from Weeks 1 through 10, identify the three most significant governance gaps in deploying a third-party LLM inside a regulated, sovereign cloud environment. For each gap, propose a specific control and identify what makes it real rather than performative.

Course thread

Coming from — Weeks 9–11 applied the full toolkit to weaponized data, algorithmic fairness, and generative AI. Students have now analyzed the hardest current governance problems with every framework the course developed.

Going to — Week 12 examines cloud governance and sovereignty — asking who controls the infrastructure that all of these systems run on, and what happens when the answer is a commercial vendor operating outside the governance framework.

Required

Required reading

- European Parliament, EU Artificial Intelligence Act — Articles 1–6 (scope and definitions), Articles 51–52 (general-purpose AI obligations), and Annex III (high-risk AI systems list). Approximately 30 pages. Primary source; students need the risk classification framework and the GPAI obligations, not the full text.
- Nicholas Carlini et al., "Extractable Memorization of Large Language Models" (2023) — abstract and introduction only, approximately 5 pages. The primary technical documentation of LLM memorization. Understand what was shown, not how.

- Benjamin Weiser, “Here’s What Happens When Your Lawyer Uses ChatGPT” (New York Times, 2023) — approximately 8 pages. The primary narrative account of Mata v. Avianca.
- A leading foundation-model provider’s published model card and usage policy (current version) — full document, approximately 15 pages. Read as a governance document: what commitments does it make, what mechanisms enforce them, and what does it not address? Compare to the EU AI Act’s foundation model obligations.

Recommended reading

- Gary Marcus and Ernest Davis, Rebooting AI (2019) — chapter 1. The clearest accessible treatment of why hallucination is structural, not a bug to be fixed.
- Arvind Narayanan, “How to Recognize AI Snake Oil” (2019, updated) — approximately 10 pages. On distinguishing genuine AI capability from marketing claims. Useful for evaluating vendor governance commitments.
- NIST AI Risk Management Framework (AI RMF 1.0, 2023) — executive summary, approximately 10 pages. The US federal AI governance framework. Compare to the EU AI Act approach.
- Wired, “DOGE Is Feeding US Government Data to AI. Here’s What We Know” (2025) — current investigative reporting on the DOGE AI case.

Assignment

Enter the course code to unlock assignments.