Final Assessment

Capstone: Audit Defense

Episodes 7.1-7.2: The Sociable Assurance Blueprint

Operationalizing the Right to Refuse | Enforcing Accountability

3-4 Hours

Final Assessment

Overview Episodes 7.1-7.2 Deliverables Rubric Timeline

What You're Defending

You have designed an AI-ESG governance system. It is now being audited by a hostile stakeholder—a regulator, a plaintiff's attorney, or a rival board faction—who believes the system is a "Liability Sponge" in disguise. Your job is to prove it isn't.

This Capstone is a simulation of an audit defense meeting. You must present four integrated artifacts that prove your system has:

- 1. **Transparency** – The auditor can see exactly how decisions are made
- 2. **Accountability** – Every actor knows their role and risk
- 3. **Resilience** – The system catches its own failures
- 4. **Authority** – Humans can actually say "no"

Why This Capstone?

Most "audits" are theater. The auditor asks questions; the company gives pre-written answers. This Capstone inverts the power dynamic: **you are the auditor**, building the system that others cannot trick. When external auditors arrive (and they will), you will already know how to answer their toughest questions—because you've asked them yourself.

Pre-Assessment Checklist

- Completed L1-M0 (Liability Sponge)
- Completed L2-M3 (Evidence Ladder)
- Completed L3-M5 (Bias Forensics)
- Completed L3-M8 (Operational Controls)
- Have a real or realistic system in mind

Assessment Format

Duration: 3-4 hours (self-paced)

Deliverables: 4 written artifacts

Format: Presentation-ready

Grading: Pass/Fail on rubric

Certificate: Certificate of Completion upon passing

Episodes 7.1 & 7.2: Context

Episode 7.1: The Audit Defense Brief

You are called into a board room. An external auditor has flagged your AI-ESG system as a potential "Liability Sponge"—a machine-speed loop with a human rubber stamp. The auditor doesn't believe humans can actually say no. Your job is to prove they can.

KEY THEMES

- • The auditor's skepticism is rational, not hostile
- • "Trust" is not a defense; "Evidence" is
- • Stop-the-Line authority must be *exercisable*, not just documented
- • Bias harms vulnerable suppliers; you must prove you catch it

Episode 7.2: The Failure-Mode Deposition

Under questioning, you must pre-register all the ways your system could break: hallucination, bias drift, data tampering, model poisoning. For each failure, you must show: how you detect it, how you stop it, and what evidence proves you've contained it.

KEY THEMES

- • "We haven't seen that failure" is not an acceptable answer
- • Failure modes must be *pre-registered* to avoid bias
- • Detection precedes remediation
- • Evidence is the currency of credibility

The Four Deliverables

Each deliverable is a separate artifact. Together, they form the Sociable Assurance Blueprint.

Transparency Audit & Fairness Forensics

1

Apply forensic methods to detect bias and transparency gaps

WHAT THE AUDITOR IS TESTING

Can you identify where your system will harm vulnerable suppliers? The auditor will show you a "Black Box" vendor report (e.g., a credit score model or ESG evaluation) and ask: where is the bias? What data is missing? What populations are hurt?

YOUR TASK

- ✔Identify at least 3 transparency gaps in the vendor's documentation (e.g., "Does not disclose training data composition")
- ✔Use statistical analysis to detect "Missing Data" bias (e.g., comparing approval rates for well-documented vs. poorly-documented suppliers across regions)
- ✔Propose specific remediation (e.g., SMOTE for synthetic minority oversampling, or regional recalibration)
- ✔Document the "Path to Appeal" for suppliers who are rejected due to missing data

OUTPUT FORMAT

Suggested length: 2-3 pages

Include: Executive Summary + Data Analysis + Bias Narrative + Remediation Plan

Visual: At least one chart showing disparate impact across regions or demographics

ACCEPTANCE CRITERIA (YOU PASS IF...)

- ✔Uses a named statistical method (e.g., disparate impact ratio, chi-square test)
- ✔Identifies gaps specific to your system (not generic)
- ✔Proposes remediation with explicit cost/benefit trade-off
- ✔Defines the appeal process (e.g., "Supplier can request manual review if rejected due to missing field X")

EXAMPLE FAILURE CASE

Your ESG vendor's model has 98% accuracy overall. But when you segment by "Supplier has published sustainability report" vs. "No report," you find:

- With report: 99% approval rate
- Without report: 45% approval rate

This is *missing data bias*. Small suppliers in developing regions are systematically excluded. Your remediation: require manual appeal + synthetic data imputation.

Accountable Workflow Design

2

Kill the Liability Sponge. Prove humans can say no.

WHAT THE AUDITOR IS TESTING

The auditor will ask: "Walk me through a transaction. At what point can your staff reject the AI's recommendation?" If you can't point to a specific, checkable moment—you have a Liability Sponge.

YOUR TASK

- ✔Draw a workflow diagram (swimlanes) showing AI step → Human Review → Decision Gate → Action
- ✔Define the exact "Stop-the-Line" triggers (e.g., "If data drift > 5%, pause and escalate")
- ✔Specify what evidence the human *must* see before they can sign off (not what *should* they see)
- ✔Calculate human review time: Show that reviewers have enough time (e.g., 2+ mins per decision)
- ✔Document what happens if the human says "no"—what's the fallback?

OUTPUT FORMAT

Suggested length: 1 diagram + 2-3 pages of narrative

Diagram: Swimlane flowchart (AI, Reviewer, Manager, Compliance) with decision gates

Include: Time budgets, escalation rules, fallback protocols

ACCEPTANCE CRITERIA (YOU PASS IF...)

- ✔Shows a human who can actually veto the AI (with named authority)
- ✔Defines "Stop-the-Line" triggers explicitly (not vaguely)
- ✔Proves humans have time (e.g., "1,000 transactions/day ÷ 8 reviewers = 125 decisions/person ÷ 8 hours = 16 mins/decision")
- ✔Shows what happens if a human overrides the AI (audit trail, escalation, etc.)

EXAMPLE: THE VETO MOMENT

Scenario: AI recommends approval of a high-value supplier. Reviewer notices the supplier's ESG score is missing Section 5 (Labor Practices). Reviewer's authority: Pause the transaction, request the missing data, or reject entirely if the supplier won't provide it.

Evidence: Reviewer initials the document. If there's a dispute, you have timestamped record of the decision.

Sociable Assurance Blueprint (RACI)

3

Define decision rights. Eliminate the Liability Sponge role.

WHAT THE AUDITOR IS TESTING

The auditor will ask: "Who is Responsible? Who is Accountable? Who Consulted? Who Informed?" This is the RACI matrix. The auditor is looking for a "Liability Sponge" role—someone who is Responsible but not Accountable (i.e., they get blamed but don't get to decide).

YOUR TASK

- ✔Create a RACI matrix for your system (roles × decision types)
- ✔For each decision type, assign exactly one "A" (Accountable) and one or more "R" (Responsible)
- ✔Identify and *eliminate* any "Liability Sponge" roles (R without A)
- ✔Define the "Exception Sign-off Policy"—what happens when someone disagrees?
- ✔Show how disagreements are escalated (not buried)

OUTPUT FORMAT

Suggested length: RACI table + 1-2 pages of narrative

Rows: Roles (Procurer, Reviewer, Compliance Lead, CTO, CFO)

Columns: Decision types (Data Quality Check, Vendor Approval, Bias Detection, Exception Override)

ACCEPTANCE CRITERIA (YOU PASS IF...)

- ✔Every critical decision has exactly one "A" (not multiple, not zero)
- ✔The "A" has real authority (can say no, can override the AI)
- ✔Defines what happens when "R" and "A" disagree (escalation, voting, etc.)
- ✔No role is "R" without also being "A" or "C" (Consulted)

EXAMPLE RACI CELL

Decision: "Override AI recommendation to approve supplier"

- **Responsible (R):** Procurement Manager (executes the override)
- **Accountable (A):** Compliance Lead (signs off; liable if wrong)
- **Consulted (C):** CTO (provides technical risk assessment)
- **Informed (I):** CFO (gets post-decision summary)

Dispute Resolution: If Procurer and Compliance disagree, CFO decides.

Failure-Mode Register

4

Pre-register failures. Prove you catch them.

WHAT THE AUDITOR IS TESTING

The auditor will ask: "What can go wrong?" And then: "How do you know when it's happening?" If you can't answer both questions, you have a blind spot. A failure-mode register forces you to pre-identify risks *before* they hurt someone.

YOUR TASK

- ✔ List at least 5 known failure modes (hallucination, data drift, bias amplification, data tampering, model poisoning)
- ✔ For *each* failure mode: Define how you detect it (specific metric, test, or alarm)
- ✔ For *each* failure mode: Define how you contain it (what action pauses the system)
- ✔ For *each* failure mode: Define what evidence proves you've fixed it (test result, audit trail, etc.)
- ✔ Rank failures by likelihood × impact

OUTPUT FORMAT

Suggested length: Risk register table + 1 page of narrative

Columns: Failure Mode | Likelihood | Impact | Detection Method | Containment Action | Evidence of Resolution

Format: Spreadsheet or table (clear, auditable)

ACCEPTANCE CRITERIA (YOU PASS IF...)

- ✔ Includes at least 5 distinct failure modes (not duplicates)
- ✔ Each failure mode has a named detection method (not "We will monitor")
- ✔ Each failure mode has a containment action (not "We will investigate")
- ✔ Evidence is *defined in advance* (e.g., "A/B test showing model retraining fixed bias" or "Null counts on dashboard show zero hallucinations in last 7 days")

EXAMPLE FAILURE-MODE ROW

Failure Mode: Model Hallucination (AI generates ESG scores from imaginary sources)

Likelihood: Medium | **Impact:** High (false approval of non-compliant suppliers)

Detection: Automated quote verification: For every score > threshold, extract the source citation. If citation does not appear in the input document, flag as "Unverified" and route to manual review.

Containment: Pause vendor approval. Route to AI Engineering team. Retrain model on cleaner dataset or switch to deterministic scoring.

Evidence of Resolution: Dashboard shows "Hallucination Rate" = 0% for 30 days post-retrain. Spot-check 10 random approvals; verify all sources are correctly cited.

Grading Rubric

All deliverables are graded on a **Pass/Fail** basis. You must pass all four to earn the certificate.

Deliverable 1: Fairness Forensics

Does Not Meet Identifies vague issues but no statistical evidence of bias. Remediation is generic.

Meets Criteria Uses named statistical method. Identifies 3+ gaps. Proposes specific remediation with trade-offs.

Exceeds Criteria Compares multiple remediation approaches. Quantifies impact on supplier populations. Includes appeal/recourse process.

Deliverable 2: Accountable Workflow

Does Not Meet Workflow is unclear. No visible human veto point. Review times not calculated.

Meets Criteria Clear swimlane diagram. Shows human veto point. Proves adequate review time. Defines Stop-the-Line triggers.

Exceeds Criteria Shows multiple escalation paths. Quantifies risk per decision. Documents real fallback protocol with cost analysis.

Deliverable 3: RACI Matrix

Does Not Meet RACI is incomplete or has multiple "A"s for same decision. Liability Sponges present.

Meets Criteria Clear RACI. One "A" per decision. No Liability Sponges. Dispute escalation defined.

Exceeds Criteria Defines authority limits per role. Shows authority escalation ladder. Includes training/competency requirements.

Deliverable 4: Failure-Mode Register

Does Not Meet < 5 failure modes. Detection/Containment/Evidence are vague or missing.

Meets Criteria 5+ modes. Each has named detection, containment, evidence. Ranked by likelihood × impact.

Exceeds Criteria Includes cross-failure dependencies. Quantifies detection latency. Shows test cases for each failure mode.

Overall Passage Criteria

You earn a **Certificate of Completion** if you achieve "Meets Criteria" or higher on all four deliverables.

Resubmission Policy

If you do not meet criteria on one deliverable, you may revise and resubmit once.

Timeline

Expected turnaround for feedback: 5-7 business days. Resubmissions within 3 days.

Suggested Work Timeline

(3-4 hours total, self-paced)

0 - 15 min

Preparation & System Selection

Choose your AI-ESG system (real or realistic case). Review previous module outputs.

15 - 75 min

Deliverable 1: Fairness Forensics

Write the bias analysis. Include statistical evidence and remediation plan.

75 - 135 min

Deliverable 2: Accountable Workflow

Draw the swimlane diagram. Define Stop-the-Line triggers and review time budgets.

135 - 180 min

Deliverable 3: RACI Matrix

Build the RACI. Identify and eliminate Liability Sponges. Define dispute resolution.

180 - 240 min

Deliverable 4: Failure-Mode Register

Document 5+ failure modes with detection, containment, and evidence. Rank by risk.

240+ min

Review & Submit

Ensure all deliverables meet rubric criteria. Compile into presentation-ready format.

Submission & Certification

How to Submit

- 1. Compile all 4 deliverables into a single PDF or shared document
- 2. Include your name, date, and system description (1 paragraph)
- 3. Submit via course portal or email to [contact]
- 4. Receive grading feedback within 5-7 business days

Certificate

Upon passing all four deliverables, you will receive a **Certificate of Completion** for the AI-ESG Integrated Strategist (AEIS) curriculum.

This certificate is not an accredited qualification and does not confer any professional license or statutory authority.

Capstone Assessment | AI-ESG Integrated Strategist Curriculum

Episodes 7.1-7.2 | Audit Defense & Failure-Mode Registration

Completion certificate only. This program is not an accredited qualification, is not endorsed by any regulator or standards body, and does not confer any professional license or statutory authority.