Foundation Module

The Constitutional Baseline

Before we can break the system, we must understand how it is *supposed* to work. This level establishes the regulatory, infrastructural, and philosophical baseline that the rest of the course will critique.

> Learning Objectives
>
> By the end of this module, you will be able to:
>
> - Distinguish between **pre-action constraint** and **post-action governance** in AI systems
> - Identify when "human-in-the-loop" functions as a liability absorption device rather than a safety mechanism
> - Recognize the structural patterns of accountability gaps that 21 AI models independently identified
> - Conduct vendor interrogations using procurement due diligence protocols
> - Design "Stop Work Authority" mechanisms for ESG controllers

Pre-Work Assessment

Before You Begin

Complete this self-assessment to identify your starting knowledge baseline. There are no wrong answers—this helps you focus on areas most relevant to your role.

Knowledge Check Questions

1. **1. Does your organization currently use AI systems for ESG data collection, analysis, or reporting?**

☐ Yes, extensively (multiple systems across functions)

☐ Yes, limited (pilot or single-function use)

☐ Evaluating vendors but not yet deployed

☐ No, but planning within 12 months

☐ No current plans

2. **2. Who is accountable if your AI system misclassifies a high-risk supplier as low-risk?**

   ☐ I can name the specific person and their role

   ☐ It's documented but I'd need to look it up

   ☐ It's the vendor's responsibility

   ☐ Not sure—it's never been clarified

3. **3. Can you currently override an AI system's recommendation without requiring approval from someone else?**

   ☐ Yes, I have documented override authority

   ☐ Yes, but it requires approval/escalation

   ☐ Theoretically possible but culturally discouraged

   ☐ No, system outputs are treated as final

   ☐ Not applicable—we don't use AI systems yet

Note: Keep your responses handy—you'll revisit them at the end of this module.

Episode 0.1

60 min

The Asimov Constraint

PRE-ACTION ETHICS

The Premise: We didn't outgrow Asimov's Laws of Robotics—we lost our nerve. The critical distinction is between **pre-action constraint** (the system refuses before acting) and **post-action governance** (audits after harm). ESG systems today rely almost entirely on the latter.

Core Concepts

- **Pre-Action Constraint:** The system contains hard-coded rules that prevent certain actions before they occur (e.g., "Do not generate a report if data provenance cannot be verified").
- **Post-Action Governance:** The system acts first, then humans review and audit the output (e.g., "Flag suspicious outputs for human review").
- **The Speed Problem:** Post-action governance only works if intervention can outpace harm. When AI acts at silicon speed, humans review at biological speed.

THE CONSTITUTIONAL REQUIREMENT

Pre-Action Refusal: The system must be able to say "I cannot proceed" *before* generating the report, not explain afterward why it proceeded.

Design Principles

- Hard constraints encoded in architecture, not policy documents
- Refusal as default, continuation as exception
- Speed must not outpace governance

Workshop Exercise: Pre-Action vs. Post-Action Audit

Scenario: Your Scope 3 emissions AI tool flags 200 suppliers as "high-risk" based on incomplete data.

❌ POST-ACTION (Current State)

1. System generates "high-risk" flags
2. Procurement receives list
3. ESG team audits sample (20/200)
4. Discovers 40% false positives
5. Damage already done (supplier relationships strained)

✓ PRE-ACTION (Target State)

1. System detects data completeness <50%
2. **REFUSES** to generate risk score
3. Returns: "Cannot assess—insufficient data"
4. Flags supplier for manual data gathering
5. No false flags, no damaged relationships

Discussion Question:

What would it take to implement pre-action refusal in your current systems? Identify 3 specific barriers (technical, organizational, vendor-related).

Newsletter Ref: *Episode 1: We Didn't Outgrow Asimov* (Pre-action vs. post-action constraint).
Additional Reading: *"The Myth of Human Oversight in Algorithmic Decision-Making"* (IEEE, 2023).

Episode 0.2

75 min

The Liability Sponge

HUMAN IN THE LOOP

The Premise: "Human in the loop" is not a safety mechanism—it is a liability absorption device. When AI acts at silicon speed and humans review at biological speed, the human becomes a **crumple zone**, absorbing blame for machine errors they lacked the authority to prevent.

The Speed Mismatch

- **Industrial Safety:** Circuit breakers trip in milliseconds to save wires that melt in seconds. Intervention outpaces harm.

- **AI Governance:** Systems process 1,000 claims/hour; humans review one every 11.5s (Illustrative Math). Impossible math.
- **The Sponge Effect:** When the system fails, the audit trail shows a human "reviewed" it. Blame flows downward.

STOP WORK AUTHORITY

The Alternative: Any human in the loop must possess constitutional authority to halt the system without permission, justification, or career penalty.

Requirements

- Documented in job description
- No approval required to invoke
- Protected from retaliation
- Audit trail of invocations reviewed quarterly

🔥 Case Study: The High-Fidelity Trap

Setup: Maria, an ESG analyst, reviews AI-flagged supplier violations. The system processes 847 suppliers/week. Maria has 6 hours/week allocated for review.

Math: 6 hours = 360 minutes ÷ 847 suppliers = **25.5 seconds per supplier**.

Reality: Maria opens each flagged case, sees a red score, scans a summary, clicks "Approve" or "Escalate". She has no time to verify source documents.

Failure Mode: System misclassifies a compliant supplier due to OCR error in document ingestion. Maria "reviewed" it (audit trail shows her user ID). Supplier relationship damaged. Blame: "Human error in review process."

Control: Maria invokes Stop Work Authority: "I cannot attest to 847 reviews in 6 hours. Either reduce volume to 50/week or remove my name from the audit trail."

Exercise: Calculate Your Liability Sponge Risk

Complete this calculation for your role:

AI-generated items/week: e.g., 500
Your review time (hours/week): e.g., 4

Seconds per item:

Calculation: (Hours × 3600) ÷ Items

____ seconds

Risk Assessment:

- • **<30 seconds/item:** High risk—you're a liability sponge
- • **30-120 seconds/item:** Medium risk—can verify summary but not sources

- • **>120 seconds/item:** Low risk—sufficient time for meaningful review

Newsletter Ref: *Episode 2: The Liability Sponge* (Stop Work Authority vs. High Fidelity).
Legal Reference: *Dodd-Frank Act § 922* (Whistleblower protections as precedent for stop-work authority).

Episode 0.3

45 min

The 21 AIs Experiment

ACCOUNTABILITY GAP

The Experiment: Twenty-one different AI models, given the same prompt to design a realistic ESG accountability failure, all converged on the same architecture: **bureaucratic middle management**. They produced "liability diodes," "moral crumple zones," and verification velocity mismatches—not because they were programmed to, but because these patterns exist in their training data.

What This Reveals

The 21 AIs didn't invent these failures—they **recognized** them. These patterns are so prevalent in corporate documentation, audit reports, and

regulatory filings that they appear as "normal" system design to AI trained on institutional text.

Liability Diode

Blame flows downward, credit upward. Junior staff absorb risk while executives claim credit for "oversight."

Moral Crumple Zone

Middle managers designed to absorb blame during failure, protecting both senior leadership and system architecture.

Velocity Mismatch

Decision speed exceeds verification speed. By the time audit detects error, consequences are irreversible.

> **Case Study: Project Espresso (Prologue)**
>
> **Setup:** Daniela Reyes, a community liaison, faces 1,247 safety flags to validate in a four-hour window.
>
> **Failure Mode:** The AI system (CommunitySense) has downgraded a grandmother's water contamination complaint because "el agua está enferma" doesn't match the keyword training set.
>
> **The Pattern:** Daniela (moral crumple zone) is expected to catch this error in **11.5 seconds per flag** (velocity mismatch). When she misses it, audit trail shows her user ID (liability diode).
>
> **Control:** Implement semantic embedding search rather than keyword matching for non-English inputs.

Evidence Artifact: Log entry showing the cosine distance between the complaint and the "contamination" vector class.

Pattern Recognition Exercise

Review your organization's last ESG audit report or AI governance documentation. Identify instances of these three patterns:

1. LIABILITY DIODE

Look for: Phrases like "reviewed and approved by [junior role]" or "oversight provided by [senior role but delegated execution]"

Example you found...

2. MORAL CRUMPLE ZONE

Look for: Roles with "coordinator," "liaison," or "analyst" titles positioned between systems and decision-makers

Example you found...

3. VELOCITY MISMATCH

Look for: KPIs measuring speed (reports/day, reviews/hour) without corresponding accuracy metrics

Example you found...

Newsletter Ref: *Episode 3: The Accountability Gap* (21 AIs converge on middle management).
Research Source: *"Pattern Recognition in Institutional Failure Modes"* (Sociable Systems, 2024).

Episode 0.4 SAAS PROCUREMENT

Tooling Ecosystem & The Vendor Interrogation 90 min

A vendor-neutral dissection of the major ESG software players (Workiva, Persefoni, Envoria, Position Green). We strip away the marketing to look at their API capabilities, data ownership models, and "Black Box" transparency.

The Vendor Landscape (2024-2025)

Compliance-First Platforms

- **Workiva:** CSRD/ISSB-focused, strong XBRL capabilities, limited AI transparency
- **Persefoni:** Carbon accounting specialist, proprietary emissions factors, vendor lock-in risk

Data Intelligence Platforms

- **Envoria:** Multi-standard support, API-first design, explainability features

- **Position Green:** Supply chain focus, good data lineage, emerging XAI capabilities

Note: This is not an endorsement. All vendors have trade-offs. Your job is to interrogate them systematically.

Activity: The Vendor Interrogation Script

Role-Play Scenario: You are the CISO or ESG Director. The vendor sales rep is in front of you. Use these questions to cut through the pitch.

1. Question 1: Data Training Rights

   "Do you train your foundational model on my data? Show me the clause in the Terms of Service that says you don't."

   RED FLAGS:

   - • "Our data practices are proprietary"
   - • "We anonymize all data" (anonymization ≠ non-use)
   - • "That's handled by our legal team" (deflection)

   ACCEPTABLE ANSWERS:

   - • Points to specific ToS section (e.g., "Section 4.2: No Training on Customer Data")
   - • Offers opt-out documentation
   - • Shows audit trail of data isolation

2. Question 2: Data Portability (The Lock-In Test)

   "If I leave your platform, do I get the raw calculation logic, or just the static PDF reports?"

RED FLAGS:

- • "Our methodology is proprietary IP"
- • "You get a data export" (but not calculation rules)
- • "Most clients don't need that level of detail"

ACCEPTABLE ANSWERS:

- • "You get SQL queries, calculation formulas, and model weights"
- • "We offer escrow for proprietary algorithms"
- • "API access includes logic documentation"

3. Question 3: Uncertainty Quantification

"Show me the 'Confidence Interval' feature. If the AI guesses a number, does it tell me it's a guess?"

RED FLAGS:

- • "Our model is highly accurate" (doesn't answer the question)
- • "We validate all outputs" (post-hoc, not predictive)
- • No visible uncertainty scores in demo

ACCEPTABLE ANSWERS:

- • Live demo shows confidence scores (e.g., "82% confidence")
- • System flags estimates vs. verified data
- • Offers Monte Carlo sensitivity analysis for uncertain inputs

BONUS QUESTION (Advanced):

"Walk me through your data lineage tracking. If I click on this Scope 3 number in the report, can you show me the exact source document it came from?"

Reference: *"The AI Adoption Blueprint: How to Get the AI You Actually Need"* (Workiva, 2024).
Procurement Guide: *"ESG Software RFP Template with AI Governance Checklist"* (Sociable Systems, 2025).

Module Summary

Key Takeaways

Conceptual Framework

- • Pre-action constraint > post-action governance
- • Human-in-the-loop ≠ safety (often = liability absorption)
- • Accountability gaps follow predictable patterns
- • Vendor transparency requires active interrogation

Practical Tools Acquired

- • Stop Work Authority protocol design
- • Liability Sponge risk calculation
- • Pattern recognition for institutional failures
- • Vendor interrogation script (3 critical questions)

Post-Module Assessment

Revisit your pre-work assessment. Has your understanding shifted?

Reflection Questions

1. **1. Based on Episode 0.2, are you currently functioning as a "liability sponge"?**

   Calculate: Items reviewed/week ÷ Hours allocated = Seconds per item

   Your calculation and conclusion...

2. **2. Which of the 21 AI patterns (liability diode, moral crumple zone, velocity mismatch) exists in your organization?**

   Specific examples from your context...

3. **3. If you were to implement ONE change from this module, what would it be?**

   Options: Pre-action refusal logic, Stop Work Authority documentation, Vendor re-interrogation, Pattern audit

   Your priority change and first

Next Module

Level 1: Epistemic Failures

When systems become too opaque to question (Clarke's Law), or too aligned to refuse, governance dies. You'll learn to map the transition from "Voluntary" ESG to "Mandatory" finance-grade reporting.

- • CSRD, ISSB, XBRL requirements
- • The "Black Box" oracle problem
- • Who audits the AI auditor?
- • Data lake fallacy and hallucinated numbers

Begin Level 1