



**Project:** Phase 0 Reinforcement Toolkit

**Team Name:** Houston Hack Attack

**Team Lead:** Joshua Williams

As advanced AI capabilities accelerate, the earliest moments of policymaking, Phase 0, offer a rare opportunity to shape the foundations of global safety, equity, and accountability. Yet without clear implementation pathways, inclusive stakeholder design, and enforceable safeguards, even well-intentioned policy proposals risk falling short.

To assist policymakers, researchers, and public-interest advocates, my team developed a rapid-response Phase-0 Toolkit. This toolkit translates safety concepts into operational governance artifacts that legislators can use immediately. It identifies weaknesses early, before presenting Phase-0 to legislators, which helps prevent regulatory gaps and unintended consequences. Additionally, it integrates equity and inclusion from the outset, ensuring that policies uphold technical robustness and human dignity. The toolkit also aims to elevate underrepresented voices in AI safety through tools that prioritize public risk literacy, storytelling, and shared accountability.

**Phase 0 Review Matrix**

| <b>Phase 0 Policy Building Block</b>           | <b>Weaknesses Identified</b>                                                                                                                                                               | <b>Implementation Challenges</b>                                                                                                         | <b>Missing/Open Gap</b>                                                                                       |
|------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| <b>Vagueness in Roles &amp; Oversight</b>      | Does not clarify designated roles (e.g., AI Auditor, Regulator).                                                                                                                           | Needs global consensus on oversight bodies and accreditation criteria.                                                                   | Lacks accreditation and enforcement procedures, raising concerns about accountability and quality assurance.  |
| <b>Human-Extinction Framing Too Narrow</b>     | Ignores broader systemic financial, ecological, and social impacts.                                                                                                                        | Difficulty quantifying complex economic/ecological risk cascades.                                                                        | Insufficient environmental and economic impact modeling may hinder accurate long-term policy assessments.     |
| <b>Vulnerability &amp; Inclusion</b>           | Those most impacted are absent from policy formation.                                                                                                                                      | High resource demands for meaningful engagement and education.                                                                           | No mandated inclusion of marginalized voices or public testimonials.                                          |
| <b>Incentive Alignment</b>                     | Primarily punitive; lacks positive incentives for safety innovation.                                                                                                                       | Designing effective incentives without loopholes is complex.                                                                             | Missing clear, measurable incentive structures (tax credits, grants).                                         |
| <b>Public-Risk Communication &amp; Empathy</b> | We must tackle the “Not-my-problem” mentality, which is insufficient empathy for abstract future harms. Normalcy bias can minimize the significance of high-impact, low-probability risks. | Translating technical risk into relatable stories without instilling fear, while maintaining long-term public and legislative attention. | No structured risk-communication strategy, stakeholder stories, or metrics for public understanding of risks. |



### “Who’s in Charge?” Org-Chart & R&R Table (Gap 7)

Lawmakers need clear answers: Who enforces AI rules? Who monitors compliance? This chart visually clarifies who’s accountable, what they do, and how responsibilities connect, removing vagueness and simplifying implementation.

#### Roles & Responsibilities

| Entity                        | Core Mandate                                                                          | Funding                          | Enforcement Power                                                      | Reporting Cadence                                           |
|-------------------------------|---------------------------------------------------------------------------------------|----------------------------------|------------------------------------------------------------------------|-------------------------------------------------------------|
| Parliament/Congress           | Establish AI Safety Act, define Agency authority, and review thresholds and outcomes. | General treasury (annual budget) | Legislative amendments, budget control, and public inquiry powers.     | Receives annual agency report & bi-annual community review. |
| National AI Safety Agency     | Issue/revoke compute & training licenses; set thresholds                              | Gov. budget & license fees       | License suspension; civil penalties                                    | Annual report to the legislature                            |
| Accredited AI Auditors        | Conduct safety-case reviews, red-team validation, and policy compliance scoring.      | Fee-for-service (regulated cap)  | Audit pass/fail authority; recommend suspension.                       | Quarterly to the AI Safety Agency and public registry.      |
| Corporate Safety Board        | Maintain incident logs, test compliance, and implement mitigation plans.              | Firm OPEX                        | Board-level accountability; must self-report safety violations.        | Monthly to Auditors.                                        |
| Public-Interest Watchdog Fund | Represent vulnerable groups, host public forums, and monitor equity impacts.          | % of license fee + philanthropy  | Advisory only; can escalate critical concerns to the AI Safety Agency. |                                                             |

#### Catastrophic Cascades Matrix

Not all lawmakers are motivated by long-term extinction risks. This matrix guides readers through three **hypothetical** yet plausible high-probability scenarios where AI failures lead to financial crashes, escalations of biothreats, environmental destabilization, and estimates of dollar losses along with their ESG implications.

#### How a Small Tech Failure Snowballs:

Tech Bug – A Surge of Digital Disruptions, followed by Economic Shock – Sudden Financial Instability, leading to Real-world Pain – Tangible Human Impact

Imagine it like dominoes: the first domino is a model failure; the last domino is a lost job or a drained pension.

# CONTROL/AI

| Scenario                      | Trigger                                                                     | Cascade Path                                                                                                                                                                                    | Final Impact                                                                                     | Why Markets Care                                                                                           |
|-------------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| AI-Driven Market Manipulation | A malicious model inserts hidden commands into high-frequency trading bots. | <ul style="list-style-type: none"> <li>- Flash crash of major stock indices</li> <li>- Automatic sell-offs by pension and index funds</li> <li>- Credit freezes for small businesses</li> </ul> | \$3.4 trillion wiped from global equity in hours<br>( <a href="#">2010 flash-crash x 1000</a> ). | <a href="#">Liquidity evaporates faster than in 2008.</a><br>Even “safe” retirement accounts are impacted. |
| Autonomous Bio-lab Design     | An unchecked model publishes a novel pathogen recipe online.                | <ul style="list-style-type: none"> <li>-Black-market synthesis</li> <li>-Regional outbreak</li> <li>-Global travel shutdown &amp; supply-chain collapse</li> </ul>                              | -5% of global GDP and critical drug shortages.                                                   | Insurers suspend coverage; governments divert trillions to emergency spending.                             |
| AI-Accelerated Disinformation | Large-scale deepfake campaign undermines trust in a primary currency.       | <ul style="list-style-type: none"> <li>-Bank-run rumors trend</li> <li>-Sovereign-debt spike</li> <li>-G20 capital flight</li> </ul>                                                            | \$800 billion lost across bond markets in 6 months.                                              | Borrowing costs rise sharply; export contracts denominated in foreign currency fail.                       |

## Stakeholder Story Bank

The negative impacts of artificial intelligence often go unnoticed until damage occurs, affecting vulnerable groups. Gig workers may face unsafe conditions from algorithmic biases, while small farmers risk losing their livelihoods due to misinterpreted data. Climate analysts might encounter setbacks from AI errors in critical modeling, and inaccessible designs can exclude disabled technologists. Additionally, biased hiring algorithms can limit diversity among junior developers. These *hypothetical examples* underscore the urgent need for oversight to ensure AI empowers rather than harms these communities.

| Persona                          | Region | Testimonial                                                                                                                         |
|----------------------------------|--------|-------------------------------------------------------------------------------------------------------------------------------------|
| Ayesh, Gig Driver                | Lagos  | “If an unsafe AI floods ride apps with bot cars, I lose my wage overnight. I need a seat at the table before that happens.”         |
| Diego, Small-Farm Co-op Lead     | Peru   | “AI-optimized commodity markets could crash coffee prices in a week. We need safeguards for livelihoods.”                           |
| Mira, Disability Rights Advocate | USA    | Models that disregard accessibility exclude blind users from future job opportunities. Regulations should enforce inclusive design. |
| Ishan, Junior Dev                | India  | “A runaway compute ban could freeze our entire sector if it's not tiered fairly. Let's design rules we can follow.”                 |
| Sofia, Climate Analyst           | EU     | “High-energy training runs diminish clean-energy gains. Carbon budgets should be integrated from day one.”                          |



### Carrots & Sticks Menu Table

Safety requires both firm measures and appealing incentives. This matrix connects key players (labs, developers, investors, cloud providers, the public) with various incentives (e.g., tax credits, grants, certification pathways) and enforcement mechanisms (e.g., license revocation, fines).

| Stakeholder         | Desired Behavior               | Carrot (Incentive)                                  | Stick (Penalty)                                                                                    | KPI                                                              |
|---------------------|--------------------------------|-----------------------------------------------------|----------------------------------------------------------------------------------------------------|------------------------------------------------------------------|
| AI Lab              | Submit a rigorous Safety Case  | 30% tax credit for R&D on verified safety expenses. | A \$10 million fine and a license freeze for failure to submit required documents will be imposed. | Safety case completed on time.                                   |
| Cloud Provider      | Enforce compute caps           | Fast-track export licenses for compliant hardware.  | Daily fee for each unlogged GPU hour.                                                              | Logged GPU hours.                                                |
| Developer Community | Report vulnerabilities         | Bug bounty payouts and a public leaderboard.        | Legal immunity will be lost if the exploit is withheld for over 30 days.                           | Mean time-to-disclosure.                                         |
| Investors           | Fund "Safe-by-Design" projects | ESG green-bond interest deduction                   | Higher risk weighting on unsafe investments.                                                       | Percentage of the portfolio invested in certified safe ventures. |
| General Public      | Participate in oversight       | Citizen-assembly stipend; open data dashboards      | N/A                                                                                                | Surveyed risk-literacy index.                                    |

### Risk-Communication Playbook

Even the best rules fail if the public and decision-makers do not recognize the risks. Lawmakers, AI developers, corporations, and the general public must emotionally understand these risks to support preventive actions. This playbook provides quick storytelling strategies, communication templates, and an annual measurement framework to shift perspectives from "not my problem" to "everyone's shared concern."

#### 1. Story Template: "A Day Without AI Safety"

- **Morning:** The day began with a jolt as my stock app crashed unexpectedly, sending my savings into a digital abyss. Panic set in as I watched the numbers tumble; the security of my financial future felt as fragile as a house of cards blowing in the wind.
- **Noon:** When I thought it couldn't get worse, a call sent chills down my spine. A deepfake voice, eerily similar to my grandmother's, claimed she was in trouble and needed money fast. My heart raced as I realized the deception, and I scrambled to reach her before the scam could succeed.
- **Evening:** As dusk settled in, a sudden blackout plunged the neighborhood into darkness. A rogue optimization algorithm in the power grid had caused chaos, leaving homes without light and the streets eerily silent. The flickering streetlights and distant sirens were a harsh reminder of our world's vulnerability without the safety nets we often take for granted.



**2. 3-Step Messaging Rule:**

- Relate: Use daily life hooks
- Visualize: Simple Infographic
- Empower: Concrete action

**3. Engagement Tactics:**

- 60-sec animated explainer.
- Interactive web quiz: "Would you spot a rogue model?"
- Local town-hall kits with facilitator notes.

**4. Risk-Literacy KPI:**

- Annual survey question: "Can you name an AI systemic-risk scenario?"
- Target awareness  $\geq 65\%$  by 2027.

**5. Budget Stub:**

- \$5M/year for multimedia campaigns, citizen assemblies, and translation into 10 languages.