☞ ☞ DHARMIC DATA CHECKLIST

(Use as an internal governance gate before deployment, during updates, and after incidents.)

☞ 1. Values & Purpose Alignment

- Have we defined the **primary duty** of this system (who we must protect first)?
- Have we mapped our values (equity, dignity, transparency, non-harm) to **measurable metrics**?
- Is there a **one-page Values Charter** attached to the feature/MR/PR?
- Is any decision here violating a stated non-negotiable? If yes, **stop** and escalate.

Sign-offs: Product Lead, Ethics Lead

☞ 2. Risk Triage

- Did we classify this system as **Low / Medium / High Risk** using the rubric?
- Does the classification reflect **population-scale effects**, **irreversibility**, or **rights impacts**?
- For high-risk systems: Have we initiated mandatory **external audit + board notification**?

Sign-offs: Ethics, Legal, CTO

☞ 3. Data Provenance & Dataset Nutrition

- Does every dataset have a **nutrition profile** (source, consent, sampling, coverage gaps)?
- Have we checked for known **bias proxies**, missing cohorts, or skewed outcomes?
- Is data collection respecting **informed consent** and local regulations?

Sign-offs: Data Engineering, CPO

☞ 4. Ethical Design Sprint Integration

- Has the sprint included a **Svadharma Canvas** for context & stakeholder duty?
- Was a **community representative** or subject-matter voice consulted?
- Have we documented **trade-offs**, mitigations, and alternatives?

Sign-offs: Product Design, Ethics Lead

☞ 5. Human-in-the-Loop (HITL) Gateways

- Are the **confidence thresholds** and **impact thresholds** defined for human review?
- Is there a clear **override path** documented in the UI/workflow?
- Are reviewers trained with **ethical review scripts** and escalation steps?

Sign-offs: Operations, Product, Ethics Lead

☞ 6. Monitoring & Incident Response

- Is drift monitoring enabled with **daily cohort dashboards**?
- Are alert thresholds defined with **SLA-linked responses**?
- Is the **Incident Response Playbook (IRP)** attached and rehearsed?

Sign-offs: SRE, Incident Lead

☞ 7. Transparent Reporting

- Has a **Model Card** been prepared for this version?
- Is a **community-friendly summary** ready for public or stakeholder release?
- Are the limitations, intended use, and redress instructions included?

Sign-offs: CPO, Communications, Community Liaison

☞ 8. Remediation & Restitution

- Does a clear **appeal mechanism** exist for affected people?
- Are we prepared with **budgeted remediation** (corrections, back-pay, compensation)?
- Is the cause of harm documented for future prevention?

Sign-offs: Legal, Customer/Beneficiary Services

☞ 9. Continuous Learning Loop

- Have we scheduled quarterly **ethical retrospectives**?
- Are findings converted into **training modules**, updated Values Charter, or policy changes?
- Did we archive lessons into org-wide **knowledge repositories**?

Sign-offs: People/HR, Ethics Lead, CTO

✸ FINAL GATE FOR DEPLOYMENT

"Have we proven that this system protects dignity, prevents harm, and provides redress?"
Deployment allowed only after ethics + product + technical sign-off.

👉 👉 SOP SNIPPETS

These snippets are written to drop directly into CI/CD pipelines, operational manuals, procurement documents, and governance frameworks.

👉 1. HITL (Human-in-the-Loop) POLICY — SOP SNIPPET

👉 Purpose

To ensure that high-impact automated decisions receive **qualified, accountable human oversight**, with clear authority to modify, reverse, or halt a system.

👉 Scope

Applies to all models affecting:

- financial access
- welfare or benefits
- healthcare referrals
- legal/civic-status decisions
- safety-critical contexts
- any high-risk category as per the Risk Triage rubric.

☞ Policy

1. **Mandatory Human Review Thresholds:**
   - Any decision with a model confidence below the configured threshold **must be escalated** to a trained human reviewer.
   - Any decision affecting rights, entitlements, or safety **must receive human sign-off** before action.
2. **Reviewer Authority:**
   - Reviewers have the authority to **override** system outputs.
   - Reviewer decisions must be logged with reason codes.
3. **UI Requirements:**
   - Review dashboards must show:
     - model confidence
     - explanation summary
     - provenance information
     - recommended alternatives
     - redress options
   - Quick-override button must be included.
4. **Training Requirements:**
   - All reviewers must complete training in:
     - ethical decision scripts
     - bias awareness
     - escalation pathways
     - documentation norms.
5. **Audit Logging:**
   - All overrides are logged with time, reviewer ID, rationale, and follow-up action.
6. **SLA Requirements:**
   - High-impact reviews: <48 hours
   - Safety-critical reviews: <4 hours
   - Escalations: immediate flag to Incident Lead.

☞ Roles & Responsibilities

- **Operations Lead:** manage reviewer pool, scheduling, training.
- **Ethics Lead:** maintain ethical scripts and review standards.
- **CTO:** ensure HITL thresholds are hard-coded in deployment pipelines.

☞ 2. INCIDENT RESPONSE PLAYBOOK (IRP) — SOP SNIPPET

☞ Purpose

To provide a structured, time-bound method for identifying, containing, resolving, and learning from harms caused by data/AI systems.

☞ Trigger Conditions

- cohort-level performance drop
- drift detection
- abnormal spike in negative outcomes
- public complaint or whistleblower input
- regulatory request
- HITL override spike
- community harm signal (NGO, grievance portal)

☞ IRP Stages

Stage 1: Detection (0–2 hours)

- Automatic alert triggers.
- Incident Lead opens an **Incident Ticket**.
- Assign Severity Level (S0–S3).
- Notify relevant owners via rapid channel (Slack/Teams bridge).

Stage 2: Containment (0–24 hours)

- Pause automated decisions for affected cohorts.
- Switch to **manual or conservative defaults**.
- If harm is severe, deactivate affected model or feature flag.

Stage 3: Diagnosis (24–72 hours)

- Identify root cause (data, model, infra, policy).
- Use drift reports, logs, dataset checks, and reviewer notes.
- Document all hypotheses and eliminate systematically.

Stage 4: Remediation (72 hours–14 days)

- Fix root cause (model retraining, pipeline correction, data removal, UI changes).
- Initiate restitution workflows: corrections, back-pay, apologies, compensation.
- Publish a community-facing update if required.

Stage 5: Closure (Day 14–21)

- Ethics Lead signs off that harm has been stopped and fixes verified.
- Archive all documentation into IRP Repository.

Stage 6: Learning & Prevention (Within 30 days)

- Conduct a structured **Ethical Retrospective**.
- Update training, SOPs, Values Charter, and monitoring alerts.
- Add new preventive rules or HITL thresholds.

☞ SLA Standards

- **Time to Detect:** <24 hours
- **Time to Contain:** <48 hours
- **Time to Resolve Critical Harm:** <14 days
- **Communication to Community:** within 7 days for high-impact cases

---

☞ Roles

- **Incident Lead:** overall coordinator
- **SRE/Monitoring:** detection, dashboards
- **Ethics Lead:** harm analysis & community impact review
- **Legal:** regulatory compliance
- **Product & Engineering:** corrective changes
- **Community Liaison:** community-facing communication

☞ 3. PROCUREMENT CLAUSES FOR ETHICAL DATA & AI — CONTRACT INSERT SNIPPET

☞ Clause 1: Transparency & Documentation

The Vendor must supply:

- Model Card for each version of the model
- Dataset Nutrition Profiles for all training datasets
- Explanation of intended use, limits, known failure modes
- List of known biases, coverage gaps, and mitigations

Deliverables must be updated **every major release**.

☞ Clause 2: Audit Rights

The Buyer has the right to:

- conduct independent audits of model behavior
- inspect training data (subject to privacy constraints)
- request third-party audits of high-risk systems
- access logs, monitoring dashboards, and error reports

Vendor agrees to support audits within **10 business days** of request.

☞ Clause 3: HITL Requirements

Vendor must:

- build human-override capabilities
- expose confidence scores and explanations
- offer UI hooks for human-in-loop workflows
- support configurable thresholds for reviewer escalation

☞ Clause 4: Harm & Remediation Obligations

Vendor is responsible for assisting with remediation in case of systemic harm, including:

- reprocessing affected decisions
- model patching
- supporting user redress workflows
- providing documentation for regulatory compliance

Vendor shall cooperate with restitution processes as mandated by Buyer policies.

☞ Clause 5: Data Rights & Provenance

Vendor must certify that all data used is:

- lawfully collected
- documented for provenance
- free of prohibited attributes
- consent-compliant
- aligned with Buyer's ethical and legal frameworks

Vendor will furnish **Data Provenance Certificates** for each dataset.

☞ Clause 6: Kill-Switch & Emergency Controls

Vendor must provide:

- a kill-switch mechanism to disable automated decisions
- ability to revert to fallback modes
- clear specification on failure conditions triggering kill-switch execution

---

☞ Clause 7: Energy & Environmental Standards

Vendor shall disclose:

- estimated training & inference energy use
- carbon impact per model version
- efficiency measures adopted (quantization, pruning, etc.)

☞ Clause 8: Local Context & Community Alignment

Vendor will conduct or support:

- contextual calibration
- fairness assessments for relevant groups
- community consultations where systems affect public rights/services

☞ ☞ COMBINED DHARMIC DATA SCORECARD

(Unified Evaluation Instrument for Ethical + Technical + Governance Compliance)

Scoring: 0 = Not Met, 1 = Partially Met, 2 = Fully Met
Weighting: Suggested weights included for organizational calibration
Pass Threshold: Recommended ≥ 75%

☞ SECTION A — DHARMIC VALUES ALIGNMENT (Weight: 20%)

☞ A1. Purpose & Duty (Dharma Clarity)

- Has the system explicitly defined its duty of care (primary beneficiary, stakeholder, vulnerable groups)?
- Are the values **mapped to measurable metrics**?
  Score (0–2):
  Notes:

☞ A2. Non-Harm & Proportionality

- Does the system minimize harm relative to its purpose?
- Have alternative design paths been evaluated to reduce harm?
  Score:
  Notes:

☞ A3. Contextual Integrity (Svadharma)

- Does the system reflect local context, norms, and constraints?
- Has the team documented moral trade-offs and contextual boundaries?
  Score:
  Notes:

☞ A4. Transparency (Satya)

- Are internal decision flows explainable to domain teams?
- Are external explanations accessible to affected communities?
  Score:
  Notes:

☞ A5. Stewardship (Loka-Sangraha)

- Is long-term societal benefit integrated into KPIs?
- Is environmental/cultural impact assessed?
  Score:
  Notes:

☞ SECTION B — DATA QUALITY, PROVENANCE & PRIVACY (Weight: 15%)

☞ B1. Provenance Documentation

- Does each dataset include a **Nutrition Profile**?
- Is consent properly documented?
  Score:
  Notes:

☞ B2. Bias Mapping

- Have known bias proxies been identified and mitigated?
- Were underrepresented cohorts consulted?
  Score:
  Notes:

☞ B3. Data Minimization

- Only necessary data collected?
- Sensitive attributes excluded unless justified and approved?
  Score:
  Notes:

☞ B4. Privacy (DP, FL, Anonymization)

- Does the system use differential privacy, federated learning, or similar protections?
  Score:
  Notes:

☞ SECTION C — MODEL DESIGN & ENGINEERING (Weight: 15%)

☞ C1. Ethical Design Sprint Completion

- Were svadharma checks completed?
- Were alternative models considered?
  Score:

☞ C2. Explainability Quality

- Does the explanation enhance clarity OR create false confidence?
- Are explanations validated with domain experts?
  Score:

☞ C3. Safety Constraints

- Are guardrails, caps, and safety constraints implemented?
  Score:

☞ C4. Evaluation Across Cohorts

- Model tested across demographic, geographic, economic cohorts?
  Score:

☞ C5. Fail-Safe Design

- Does system **fail-safe**, not **fail-open**?
  Score:

☞ SECTION D — HUMAN-IN-THE-LOOP (HITL) & OVERSIGHT (Weight: 10%)

☞ D1. HITL Thresholds

- Are thresholds clearly defined and documented?
  Score:

☞ D2. Reviewer Tools & UI

- Does reviewer dashboard show confidence, alternatives, and provenance?
  Score:

☞ D3. Reviewer Authority

- Do reviewers have override authority?
  Score:

☞ D4. Training & Scripts

- Are reviewers trained in ethics and domain judgement?
  Score:

☞ SECTION E — MONITORING, INCIDENT RESPONSE & REMEDIATION (Weight: 10%)

☞ E1. Drift Monitoring Coverage

- Real-time or daily drift reports?
  Score:

☞ E2. Incident Response Playbook (IRP)

- IRP attached and rehearsed?
  Score:

☞ E3. Harm Detection Channels

- Community signals, whistleblower channels, public complaint forms integrated?
  Score:

☞ E4. Remediation Budget & Process

- Clear compensation and correction workflow exists and activated?
  Score:

☞ SECTION F — TRANSPARENCY & ACCOUNTABILITY (Weight: 10%)

☞ F1. Model Cards

- Are detailed Model Cards published internally/externally?
  Score:

☞ F2. Community-Facing Summary

- Accessible explanation for non-technical audiences?
  Score:

☞ F3. Logging & Audit Trails

- Immutable logs for decisions, overrides, and access?
  Score:

☞ F4. Regulatory Readiness

- Material available for compliance & legal substantiation?
  Score:

☞ SECTION G — PROCUREMENT & VENDOR ACCOUNTABILITY (Weight: 10%)

☞ G1. Transparency Deliverables

- Vendor provides Model Cards, Nutrition Profiles, Failure Modes, Bias Reports?
  Score:

☞ G2. Audit Rights Compliance

- Vendor allows audits, shares logs, supports investigations?
  Score:

☞ G3. Emergency Controls

- Kill-switch and fallback mode implemented?
  Score:

☞ G4. Remediation Cooperation

- Vendor supports restitution and post-harm correction?
  Score:

☞ G5. Contextual Alignment

- Vendor has calibrated solutions to local norms and user context?
  Score:

☞ SECTION H — ENVIRONMENTAL & ENERGY IMPACT (Weight: 5%)

☞ H1. Training & Inference Carbon Disclosure

- Energy cost per model version disclosed?
  Score:

☞ H2. Efficiency Techniques

- Quantization, distillation, pruning used?
  Score:

☞ H3. Sustainable Infrastructure

- Renewable-backed data centers or offset commitments?
  Score:

☞ SECTION I — COMMUNITY PARTICIPATION & SOCIAL IMPACT (Weight: 5%)

☞ I1. Participatory Audits

- Community or civil society included in audits?
  Score:

☞ I2. Redress Visibility

- People know how to challenge decisions?
  Score:

☞ I3. Cultural Sensitivity

- Has local worldview, ethics, and socio-economic reality been integrated?
  Score:

☞ ☞ SCORING SUMMARY

Section	Weight	Score (0–2)	Weighted Value
A	20%		
B	15%		
C	15%		
D	10%		
E	10%		
F	10%		
G	10%		
H	5%		
I	5%		

Total Score (out of 100):
Risk Status:

- **80–100:** Dharmic-Ready
- **65–79:** Requires Mitigation
- **Below 65:** Not Deployable

☞ ☞ RECOMMENDED OUTPUTS (For Reporting)

✸ Deployment Decision

- Approved / Not Approved / Requires Mitigation

✸ Mandatory Follow-Ups

- Additional audits
- Data re-labeling
- Model retraining
- HITL strengthening
- Community consultation
- Procurement renegotiation

✸ Record of Decision Owners

- CTO
- CPO
- Ethics Lead
- Legal Counsel
- Community Liaison
- Product Owner