Trustworthy AI for Workforce Decisions: Governance, Explainability, and Human Oversight

Executive Summary

The rapid adoption of AI in HR and workforce management has outpaced governance, leading to deep employee skepticism. Frontline workers distrust "black-box" AI that delivers unexpected or biased outcomes[1][2]. Incidents of AI "hallucinations" — confident but false outputs — pose acute risks in hiring, performance reviews, promotions, and terminations, potentially embedding errors in sensitive personnel decisions[3][4]. Established AI governance frameworks (NIST AI RMF, OECD AI Principles, EU AI Act, MIT/academia guidance, HBR/Forrester analyses) emphasize *transparency*, *fairness*, *accountability*, and *human oversight*[5][6][7]. We propose **RUDY AI Trust & Hallucination Governance**, a system that embodies these principles through explainability tools, confidence calibration, risk-based policies, and layered human review workflows. Key features include calibrated confidence scores and UI signals, automated alerts for low-confidence or risky outputs, structured human-in-the-loop checkpoints, clear accountability roles, and comprehensive audit trails. A comparative framework–product table (below) maps theory to practice. Implementation follows a phased roadmap (planning, development, testing, rollout) with defined KPIs (e.g. error rates, human-override frequency, user-trust metrics) tracked on a governance dashboard. Legal and ethical considerations (privacy, non-discrimination, worker rights) are addressed via policy safeguards and data governance. Robust change management — training, communication, and leadership support — is essential to build AI fluency and trust[8][9]. Ultimately, RUDY ensures AI augments rather than replaces human judgment, aligning workforce decisions with corporate values and regulatory mandates (Table 1, Figures 1–2).

Employee Distrust of Opaque AI Systems

Studies show frontline employees *mistrust* opaque AI tools, especially those that make consequential decisions. Deloitte's 2025 TrustID index found trust in workplace generative AI plummeting (–31% in mid-2025) and trust in "agentic" AI (making decisions autonomously) collapsing by 89%[10]. Forrester notes common fears: **"inaccurate or unexpected results (…hallucinations)** and **embedded bias** are top concerns curtailing trust[2]. A major barrier is that AI appears as a "black box" – complex math hides how outputs are derived[1][11]. MIT Sloan research confirms that stakeholders hesitate because models are opaque, hard to inspect for accuracy or ethics, and prone to bias or "drift" over time[1]. Key reasons for distrust include:

- **Opacity**: Without explanations, workers can't verify or contest AI recommendations[1][9].

- **Unexpected Errors**: Hallucinations—fabricated facts from generative models—erode confidence and may propagate falsehoods[2][3].
- **Bias and Fairness**: AI trained on biased data can unfairly penalize groups; employees note that opaque systems hide these biases[1][12].
- **Loss of Agency**: When AI "takes over" tasks (e.g. screening resumes, scoring performance), employees fear losing control and voice in decisions[10][13]. To rebuild trust, organizations must prioritize explainability, transparent criteria, and strong human oversight[1][9].

Hallucination Risks in HR Use Cases

Generative AI introduces **confabulation risk** in HR contexts. NIST warns that models may produce *false but confident* content, leading users to act on misinformation[3]. In HR applications, hallucinations can be highly damaging:

- **Hiring/Recruiting**: An AI resume screener or recruiter chatbot might fabricate candidate qualifications or references, or incorrectly flag a candidate for bias. For example, a system might hallucinate a candidate's certifications or falsely infer personality traits, unfairly affecting hiring decisions[3][14].
- **Performance Reviews**: An AI summarization of an employee's performance could invent accomplishments or criticisms not supported by data, skewing evaluation and promotion prospects. NIST notes that in consequential domains, confabulated logic can mislead decision-makers[3][4].
- **Promotions**: When recommending promotion candidates, AI models may overweight irrelevant patterns, leading to unfair selections. Unchecked, this could institutionalize biases in "success profiles" and ignore meritorious but atypical employees[15][14].
- **Termination/Disciplinary Actions**: AI-driven alerts (e.g. sentiment analysis of communications, performance anomalies) might hallucinate misconduct or erratically signal termination, violating employee rights if acted upon without verification[3][7].

These HR-specific risks underscore the need for controlled use of generative AI: **human-in-the-loop review**, rigorous validation of outputs, and caution where lives/careers are at stake[3][12]. Deloitte and other experts recommend monitoring for hallucinations and bias, especially in "high-stakes" processes[2][12].

Governance Frameworks for Trustworthy AI

Leading frameworks stress **proactive governance** across policy, roles, risk, monitoring, and incident response. Key guidelines include:

- **NIST AI Risk Management Framework (AI RMF 1.0)**: Emphasizes a structured lifecycle approach – identifying, measuring, and managing AI risks. It prescribes governance processes (risk tolerance, tiered risk management) and transparency.

For example, NIST governance tasks call for policies setting risk tiers by impact (e.g. bias, privacy)[16], inventories of AI systems by risk[17], and incident response plans for unknown issues[18]. NIST also requires ongoing monitoring of third-party AI components[19] and applying existing IT governance tools to generative AI (with adaptation)[20].

- **OECD AI Principles (2019, updated 2024)**: Five values-based principles plus recommendations. Key principles relevant here are *Transparency & Explainability* (provide understandable information on AI logic and decision criteria)[21], *Human-Centered Values* (respect human rights, non-discrimination, and enable human agency/oversight)[22], *Robustness & Safety* (ensure systems function safely under misuse and can be overridden)[23], and *Accountability* (traceability of datasets/processes, systematic risk management)[24]. OECD also urges policies on data protection, inclusive R&D, and fair workplace transformation (ensuring AI use enhances job quality and worker safety[25]).

- **EU AI Act (2024)**: Treats AI in employment ("workplace AI") as high-risk. It mandates **human oversight** for high-risk systems (Article 14) and worker rights to explanation (Article 86)[7]. Compliance requires risk assessments and transparency disclosures. Critiques note gaps (e.g. reliance on self-certification) but the Act reflects the obligation to build human-centric, auditable AI for HR decisions.

- **MIT and Industry Research**: MIT Sloan defines explainability as key to trust (models must be **"value-generating, compliant, representative, and reliable"**[26]). MIT and others advocate uncertainty quantification and calibration so users know model limits[1][11]. Forrester/HBR highlight "seven levers of trust" (transparency, accountability, competence, consistency, integrity, dependability, empathy)[11][27] as practical guideposts.

- **HR and Change Management Frameworks**: Experts (AIHR, HBR) urge embedding AI governance into HR strategy. Risks include *opacity*, *fairness at scale*, *dehumanization*, and *diffused accountability*[28][29]. Best practice is a decision-gate framework: require AI to enhance (not replace) human judgment, assign a clear accountable owner, test for bias, ensure explainability and trust alignment[30][31].

These sources converge on core elements for trustworthy AI governance: clear policies, defined roles (ethics officer, AI owner, compliance), ongoing risk assessment and monitoring, explainability for stakeholders, and procedures for incidents or redress.

RUDY AI Trust & Hallucination Governance: System Architecture and Features

RUDY AI Trust & Hallucination Governance operationalizes these principles into an enterprise product suite. Figure 1 outlines its architecture. The system integrates:

- **AI Engine with Explainability Module**: The core is the AI decision engine (e.g. a trained model for hiring/performance). A **transparency layer** computes feature attributions or decision rationales (e.g. SHAP values) so any output can be traced to inputs. The UI displays this explanation to users in plain language[32][1].

- **Confidence & Uncertainty System**: Every output carries a **calibrated confidence score**. RUDY employs probability calibration techniques (e.g. Platt scaling) and uncertainty quantification (e.g. Monte Carlo dropout) to align model confidence with empirical accuracy[11][1]. The UI signals confidence (e.g. color-coded gauge or icon) so users see when the system is "unsure." A **confidence threshold** is set: outputs below this trigger additional review. Administrators can adjust thresholds per use-case.

- **Human-in-the-Loop (HITL) Review Workflows**: For low-confidence or high-impact decisions (e.g. questionable hire, termination), RUDY flags cases for human review. A **workflow engine** routes alerts to the appropriate role (HR analyst, manager, legal counsel) with the AI's rationale and confidence. Reviewers see an **audit dashboard** to accept, override, or request further data. Escalation paths can be configured (e.g. disagreements escalate to an HR committee). All actions (approvals, overrides, feedback) are logged.

- **Audit and Monitoring**: Every AI interaction is automatically recorded. The **audit trail** logs inputs, model outputs, explanations, confidence scores, reviewer actions, and final decisions. A **monitoring service** scans logs for anomalies (e.g. spikes in low-confidence outputs) and performance drift. Automated alerts notify governance officers of unusual patterns or incidents (e.g. detected hallucination events).

- **Governance Console**: Admin interfaces let compliance officers set policies and roles. For example, RUDY supports role-based access control (AI Owner, HR Manager, Compliance Auditor) with defined permissions. Policies (acceptable use, retraining triggers) are enforced here. A **risk assessment module** guides initial deployment by evaluating factors such as data bias, expected impact, alignment with regulations (drawing on NIST's GV-1.3 risk tiering[16]).

- **Integration & Extensibility**: RUDY can ingest data from HRIS (Human Resource Information Systems) or LMS (Learning Management Systems) under strict data governance controls. It supports third-party AI models (with vetting workflows) in line with NIST guidance on third-party risk[33].

【image】 *Figure 1: RUDY AI Governance Architecture. The AI engine with explainability and confidence modules feeds outputs to a decision interface. Low-confidence or flagged decisions flow to a human review layer, and all transactions are captured in an audit trail for monitoring.*

This architecture ensures **alignment with framework requirements** (Table 1). For example, OECD's demand for transparency[34] is met via explainable outputs and user information; NIST's inventory and monitoring needs[17][19] are satisfied by system inventory and auditing; and Forrester's trust levers (transparency, accountability) are embodied in UI signals and clear roles.

Table 1. Mapping AI Governance Frameworks to RUDY Features

Governance/Trust Aspect	Framework Guidance	RUDY Feature Implementation
Transparency & Explainability	OECD: provide clear info on data/logic to affected stakeholders[34]; NIST: documentation and explainable outputs; MIT: foster trust through explainability[1].	Explainable UI: model rationale, data sources shown. Plain-language explanations accompany each AI output[34][1].
Human Oversight & Control	OECD: human agency and oversight mechanisms[35]; EU AI Act (Art.14): effective human oversight of high-risk systems[7]; AIHR: AI augments but does not replace managers[9][13].	HITL workflows: decisions below confidence threshold or critical cases sent for human review. Override buttons and escalation logic ensure a human decision-maker[36][37].
Accountability & Roles	OECD: traceability of data/processes, risk mgmt by roles[24]; NIST: assign AI actors, policies tied to risk tolerance[38][16]; Forrester: assign AI "ethics/trust officer"[27].	Role-based assignments: defined owners (AI Owner, HR Lead, Compliance) who are explicitly responsible. Audit logs trace each actor's decisions. AI integrity is overseen by designated trust officers.
Fairness & Bias Mitigation	OECD: non-discrimination, fairness, diversity[22]; NIST GAI risk: harmful bias detection; HBR/AIHR: audit training data, test for disparate impact[15][14].	Data governance: pre-deployment bias audits of training sets. Fairness metrics (e.g. demographic parity) are computed. Alerts on potential bias triggers mandatory re-training or human review.
Safety, Reliability & Security	OECD: robust/safe under misuse[23]; NIST: resilience, model drift monitoring; Forrester: competence (uncertainty) and consistency (drift management)[39].	Confidence calibration: uncertainty estimates flag unreliable outputs. Continuous monitoring for model drift triggers retraining. Capability "go/no-go" thresholds and sandbox testing before

Governance/Trust Aspect	Framework Guidance	RUDY Feature Implementation
		deployment.
Risk Assessment & Policies	NIST: risk tiers, documentation of controls[16][20]; OECD: risk-based governance, policy environment[40].	Pre-deployment risk assessment tool analyzes use-case impact (privacy, safety). Policy engine enforces risk tiers (e.g. block decisions without review if certain risk flags). Regulatory library informs compliance checks.
Incident Response	NIST: procedures to respond to unknown risks[18]; OECD: adapt governance/policies as needed.	Incident playbook: if a hallucination or error is detected (via monitoring or report), RUDY automatically generates an incident ticket, notifies stakeholders (AI governance board, legal), and logs the issue for root-cause analysis.

Confidence System Design

RUDY's confidence subsystem ensures decision-makers see and respect AI uncertainty. Key elements:

- **Calibration**: RUDY applies statistical calibration so that a model's confidence level matches real-world accuracy rates. For instance, if the model outputs "75% hire likelihood," it is indeed correct about 75% of the time (on validation data). Calibration is verified via holdout evaluation.
- **Uncertainty Quantification**: Techniques like Bayesian inference or Monte Carlo Dropout generate confidence intervals. Each prediction comes with a variance measure, informing if the model is "out of its depth."
- **Confidence Thresholds**: Administrators set thresholds for different decision categories. E.g., for hiring pre-screening, predictions below 70% confidence trigger a manual resume review. RUDY can also enforce maximum allowable risk (NIST GV-1.3-002 suggests "go/no-go" performance minima[36]).
- **UI Signals**: In the user interface, outputs include a visual confidence indicator (e.g. color bar, "confidence score" percentage, or traffic-light icon). Low confidence cases are highlighted and may display a warning icon with a tooltip "Requires human review"[11][1].

This design aligns with Forrester's "Competence" lever (accept AI's probabilistic nature)[39] and OECD/AIHR calls for transparency. It ensures users never mistake a low-

confidence AI suggestion for fact; instead, they are prompted to apply judgment or seek clarification.

Multi-Layer Review Workflows

To prevent unchecked automation in sensitive HR decisions, RUDY implements **multi-layer human oversight**:

1. **Tiered Human-in-the-Loop (HITL) Checkpoints**:
2. **Initial Review**: Any AI recommendation for hiring or promotion with confidence < threshold is routed to an HR analyst. The analyst sees the AI rationale and evidence (e.g. employee performance metrics, inferred skills). They can accept, modify, or reject.
3. **Manager Approval**: If the HR analyst escalates (due to complexity or unclear case), the candidate's manager or a senior HR manager reviews. This ensures a final decision rests with a human leader.
4. **Legal/Compliance Escalation**: For terminations or any decision flagged as potentially high-risk (e.g. involving protected class concerns), the case is sent to compliance/legal teams for review.
5. **Clear Accountability and Escalation Paths**: Each decision step logs the reviewer's identity. RUDY enforces that **a named person** is responsible. Escalations follow pre-defined paths (Figure 2 flow). For example, if a manager rejects an AI hire recommendation, it may loop back to sourcing for other candidates rather than letting the AI proceed unchecked. This addresses "diffused accountability" risk by mandating explicit ownership[29][7].
6. **Audit Trails**: All interactions—AI outputs, user overrides, reviewer comments—are timestamped in a secure log. This traceability satisfies OECD's call for dataset/process traceability[24] and NIST's inventory/policy requirements[38]. Audit logs support internal audits and external compliance checks.
7. **Human Override and Learning**: If reviewers correct an AI output, that correction is fed back to improve the model (with IRB oversight on data use). A governance board regularly reviews these overrides for patterns of concern (e.g. repeated misjudgments by the model in certain cases), prompting retraining or policy adjustment.

```
flowchart TD
    A[AI Output: Decision Suggestion]
    B{Confidence \nAbove Threshold?} -->|Yes| C[Human Reviewer 1 (HR)]
    B -->|No| D[Human Reviewer 1 (HR)]
    C --> E{High Impact?}
    D --> E
    E -->|Yes| F[Human Reviewer 2 (Manager/Legal)]
```

```
E -->|No| G[Finalize Decision: Human or AI Approved]
F --> G
G --> H[Record Audit Trail]
H --> I[Monitoring & Reporting]
```

Figure 2: RUDY Multi-layer Review Flow. AI outputs go to an HR reviewer; low-confidence or sensitive cases escalate to senior review. Final decisions (AI-approved or human override) are logged.

By codifying human checkpoints and auditability, RUDY implements the governance protocols recommended by NIST and OECD, and the AIHR/HBR decision gate criteria[30][31].

Adoption Metrics and Dashboard

Effective governance requires measuring both adoption and trust. RUDY tracks **leading and lagging indicators** (Table 2) via a metrics dashboard. Key KPIs include:

- **Model Performance Metrics**: Accuracy, precision/recall on labeled cases, calibration error. (Lagging indicator – past model quality.)
- **AI Output Quality**: Rate of flagged hallucinatory or low-confidence outputs; frequency of human overrides. (Leading – early warning of model issues.)
- **Review Outcomes**: Percent of cases requiring senior escalation; average review time. (Both leading and lagging – process bottlenecks.)
- **User Trust & Satisfaction**: Survey scores from employees and managers on trust in AI recommendations. (Lagging – sentiment outcome.)
- **Fairness Metrics**: Disparate impact ratios across protected groups in decisions influenced by AI. (Lagging – post-hoc fairness measure.)
- **Usage & Adoption**: Percentage of HR tasks assisted by RUDY vs manual, number of active users, time to decision improvements. (Leading – adoption pace.)

These metrics populate a dashboard with interactive visualizations: e.g., a gauge for model accuracy; charts of override rates over time; a bar graph of decisions by department or demographics; and a trust index trendline (Figure 3 mockup). Visual alerts highlight KPIs deviating from targets (e.g., sudden drop in calibration score).

Metric Category	Leading Indicators	Lagging Indicators	Example
Model Quality	Planned vs actual confidence levels	Historical accuracy, precision/recall	Calibration error (should be <5%)
Output Review	% outputs below confidence threshold	% of escalations to senior review	Human override rate (target <10%)
Process Efficiency	% of decisions automated	Average review time after AI vs baseline	Time saved per decision (minutes)

Metric Category	Leading Indicators	Lagging Indicators	Example
Fairness/Bias	Number of flagged bias incidents	Disparate impact ratio (e.g. <0.8 for parity)	Hire/reject rates by demographic group
Trust & Engagement	Employee AI training completion	Survey trust scores, complaints count	Trust survey: % "AI recommendations make sense" (>80%)

Metrics Dashboard (Mockup)

Key visualizations:

- **Calibration Chart**: Plotting predicted probability vs actual outcomes (slope=1 ideal).
- **Override Trend**: Line chart of override count by week.
- **Fairness Heatmap**: Showing selection rates across gender/ethnicity vs base rates.
- **Trust Survey Score**: Gauge or dial indicating overall trust index (aggregated from periodic surveys).
- **Escalation Workflow**: Flowchart gauge of cases at each review level.

These tools enable continuous governance oversight and data-driven adjustments to policies (e.g. tightening thresholds, retraining models).

Implementation Roadmap

RUDY's rollout follows a structured, agile timeline (Figure 4):

```
gantt
    title RUDY AI Trust & Hallucination Governance Roadmap
    dateFormat  YYYY-MM-DD
    section Planning
    Risk Assessment and Policy Definition       :done, 2026-05-01, 60d
    Stakeholder Workshops & Role Assignments    :done, after Risk Assessment
and Policy Definition, 30d
    section Development
    Build Explainability & Confidence Modules   :active, 2026-08-01, 60d
    Develop HITL Workflow & Audit Logging       :  dev2, after Build
Explainability & Confidence Modules, 45d
    UI/UX for Reviewers and Dashboards          :  dev3, after HITL Workflow
& Audit Logging, 30d
    Integration with HR Systems & Security      :  dev4, after UI/UX, 30d
    section Testing
    Alpha Testing (Functionality)               : test1, after dev4, 30d
    Beta/Pilot (Select Business Unit)          : test2, after test1, 45d
    section Deployment
```

```
    Training & Change Management               : train, after test2, 30d
    Company-wide Rollout                       : rollout, after train, 30d
```

Figure 4: Implementation Gantt Chart. Phased tasks from risk assessment and policy setup, through development of RUDY modules, to testing, training and full deployment (2026 timeline).

Key phases:

- **Planning (Months 1–3)**: Conduct data/risk assessment (NIST AI RMF MAP step) for targeted HR use-cases. Define governance policies (roles, compliance rules) and risk thresholds. Engage stakeholders (HR, Legal, IT, employees) to set expectations and assign responsibilities.
- **Development (Months 4–8)**: Iteratively build system components. Early focus on explainability and confidence modules, and core audit logging. Then add the HITL workflow engine and user interfaces. Integrate with existing HR data sources under data governance controls.
- **Testing (Months 9–11)**: Internal alpha testing for each module. A pilot deployment in one department, with human oversight, to collect real-world feedback on system outputs, workflow efficacy, and UI clarity.
- **Rollout & Training (Months 12+)**: Train HR and management on using RUDY (including interpreting confidence signals and explanations), and on new policies (when to trust vs override). Launch adoption metrics tracking. Provide ongoing support and incorporate user feedback.

This roadmap ensures the product is aligned to compliance deadlines (e.g. EU AI Act go-live for HR systems) and user readiness. Milestones include completing a *pre-deployment review* of model risks (as NIST prescribes) before any live use.

Legal and Ethical Considerations

Regulatory Compliance: RUDY is designed for "high-impact" AI under emerging laws. The EU AI Act (and similar laws) classify HR AI as high-risk, requiring human oversight and data transparency[7]. RUDY enforces these mandates by default (human approval gates, documentation). It also facilitates compliance with labor laws (e.g. advance notice of automated layoffs per NY WARN-AI Act) and civil rights statutes by building in non-discrimination safeguards.

Privacy and Data Governance: Employee data used in RUDY is sensitive. Data input is governed by strict policies (informed consent, minimal needed fields). All records are encrypted in transit and at rest. Access controls ensure only authorized roles see personal data. RUDY tracks data lineage (per OECD and NIST advice) to support audits[16][24]. Regular privacy impact assessments and data quality checks are conducted.

Ethics & Fairness: Ethical design principles are embedded: RUDY requires that outputs can be **challenged** by affected employees (as OECD recommends[32]). Employees have

the right to appeal an automated decision, with RUDY providing the relevant logs and model reasoning to their advocate. The system prohibits certain uses (e.g. emotion recognition on interviews) to comply with bans and company ethics codes. A periodic ethics review board uses RUDY's fairness/audit reports to oversee ongoing alignment with corporate values and stakeholder expectations.

Accountability and Transparency: All AI policies and model updates are documented. RUDY's logs create an immutable record supporting **traceability** — an OECD requirement[24]. This also aids any regulatory inquiries. Reports on RUDY usage (from the dashboard) are shared with executives and board-level committees for transparency.

Change Management and Training

Even the best technology fails without buy-in. RUDY's implementation is accompanied by robust change management:

- **Leadership Sponsorship**: Senior management (CPO, CHRO, CIO) champions RUDY, communicating its goals (fairer, faster decisions) to the workforce.
- **Training Programs**: Tailored training ensures HR staff and managers understand RUDY's capabilities and limits. Workshops demonstrate how to interpret AI explanations and confidence indicators. As HR Dive notes, a structured change strategy (present in "AI pacesetter" companies) dramatically reduces employee fear[8].
- **Communication and Feedback**: Regular updates (intranet, town halls) explain how RUDY aligns with corporate values. Anonymous feedback channels let employees report concerns about specific cases. RUDY's team reviews such input to adjust policies or retrain models as needed.
- **Policy Updates**: Employee handbooks and performance review guidelines are updated to reflect the role of AI. For example, specifying that **final decisions remain human-responsible** (echoing AIHR advice: "the human being remains accountable"[29]).
- **Monitor Adoption**: KPIs from the dashboard (Section 5) feed into a change management plan: low adoption or trust scores trigger additional training and communication.

By proactively managing the human side—explaining AI's benefits, rights protections, and oversight—RUDY turns a potential trust gap into an opportunity for transparency and upskilling. This aligns with Kyndryl's finding that readiness and change planning are critical for AI adoption[8].

Conclusion

In an era where AI increasingly influences hiring, promotion, and termination, a robust governance system is imperative. RUDY AI Trust & Hallucination Governance embodies best practices from MIT, NIST, OECD, EU law, and industry research to ensure workforce AI

is trustworthy by design. It combines explainable AI, calibrated confidence, multi-tier human oversight, and comprehensive monitoring to mitigate risk and build confidence. By mapping governance principles to concrete product features (Table 1) and tracking relevant metrics, RUDY transforms abstract regulations into actionable controls. Legal safeguards (e.g. human-centric AI mandates, data rights) are built into every layer. Change management initiatives ensure employees understand and trust the system.

This comprehensive approach—technical, procedural, and cultural—aligns AI-driven HR with organizational values and compliance needs. The result is AI that **augments** rather than **dominates** human decision-making, leveraging AI's strengths while preserving human agency, fairness, and oversight[41][7].

Sources: Authoritative frameworks and research were used throughout (NIST AI RMF[5][38], OECD AI Principles[34][24], MIT Sloan AI trust research[1], HBR/Forrester analyses[10][2], AIHR risk guidance[9][29], legal and ethics studies[7][8]). These informed the design of RUDY's architecture, processes, and recommendations.

---

[1] [26] What is artificial intelligence explainability? | MIT Sloan

https://mitsloan.mit.edu/ideas-made-to-matter/working-definitions/what-is-artificial-intelligence-explainability

[2] [11] [27] [39] AI Has a Trust Problem. Here's How to Fix It. - SPONSOR CONTENT FROM FORRESTER

https://hbr.org/sponsored/2024/09/ai-has-a-trust-problem-heres-how-to-fix-it

[3] [4] [16] [17] [18] [19] [20] [33] [36] [38] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[5] Future Shock: Generative AI and the International AI Policy and Governance Crisis · Special Issue 5: Grappling With the Generative AI Revolution

https://hdsr.mitpress.mit.edu/pub/yixt9mqu/release/3

[6] [21] [22] [23] [24] [25] [32] [34] [35] [40] AI principles | OECD

https://www.oecd.org/en/topics/ai-principles.html

[7] Trustworthy and human-centric? The new governance of workplace AI technologies under the EU's Artificial Intelligence Act - Didem Özkiziltan, Fabio Landini, 2025

https://journals.sagepub.com/doi/10.1177/10242589251336193

[8] Nearly half of CEOs say employees are resistant or even hostile to AI | HR Dive

https://www.hrdive.com/news/employers-employees-resistant-hostile-to-AI/749730/

[9] [12] [13] [15] [28] [29] [30] [31] [37] [41] AI in HR Decision-Making: How To Create Better People Outcomes - AIHR

https://www.aihr.com/blog/ai-in-hr-decision-making/

[10] Workers Don’t Trust AI. Here’s How Companies Can Change That.

https://hbr.org/2025/11/workers-dont-trust-ai-heres-how-companies-can-change-that

[14] New Research on AI and Fairness in Hiring

https://hbr.org/2025/12/new-research-on-ai-and-fairness-in-hiring