

# AI: Ethics, Security & Product Development

**Happy Digital X**

**Happy Digital X | Tsinghua University**

# Today's Agenda

---

## **1** AI Ethics

Responsible AI frameworks, bias, and governance

## **2** Data Governance

Privacy regulations and data management

## **3** AI Security

Threats, vulnerabilities, and protection

## **4** Product Development

Life cycle, deployment, and ROI

# AI Ethics

# AI Ethics

---

## Why Ethics Is a Business Imperative

- **Reputation:** Brand damage vs. trust premium
- **Regulatory:** Fines vs. favorable treatment
- **Legal:** Lawsuits vs. reduced liability
- **Talent:** Recruiting challenges vs. employer of choice

### Key Insight

The reputational half-life of AI ethics failures is measured in years.



# High-Profile AI Ethics Failures

---

- **Amazon:** Recruiting tool showed gender bias
- **Microsoft:** Tay chatbot offensive within hours
- **Apple:** Credit card gender bias investigation
- **Clearview AI:** Banned in multiple countries
- **COMPAS:** Criminal justice racial bias



# Types of AI Bias

---

- 1 **Historical:** Training data reflects past discrimination
- 2 **Representation:** Data over/under-represents groups
- 3 **Measurement:** Features as proxies for protected characteristics
- 4 **Aggregation:** One model for diverse populations
- 5 **Evaluation:** Test data doesn't match deployment context

## The Uncomfortable Truth

You cannot optimize for all fairness definitions simultaneously.

# Bias Mitigation Strategies

---

## Detection

- Pre-deployment testing
- Fairness metrics monitoring
- Demographic parity analysis
- Continuous output monitoring

## Mitigation

- Pre-processing: Fix training data
- In-processing: Fairness constraints
- Post-processing: Adjust outputs
- Human oversight for edge cases

# Transparency & Explainability

---

## Different stakeholders need different explanations:

- **End Users:** "Why this output for me?"
- **Operators:** "Why is the system behaving this way?"
- **Regulators:** "How does the system make decisions?"
- **Affected Parties:** "What can I do to change the outcome?"
- **Executives:** "What are the risks of this system?"

# Regulatory Explainability Requirements

---

- **GDPR Article 22:** Right to explanation — Up to 4% revenue
- **EU AI Act:** High-risk AI transparency — Up to 7% revenue
- **US ECOA:** Credit decision notices — Per-violation fines
- **NYC Local Law 144:** Employment bias audits — \$500–1,500/day
- **China PIPL:** Explainability in regulated sectors — 5% revenue

# Human Oversight Levels

---

## 1 Human-in-the-Loop

Human approves every decision

## 2 Human-on-the-Loop

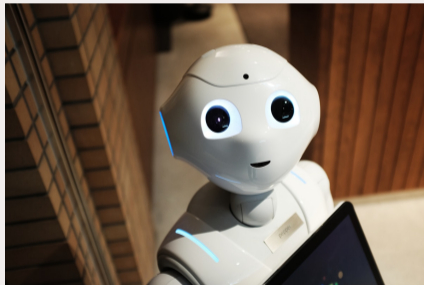
Human monitors and can intervene

## 3 Human-out-of-Loop

Fully automated with auditing

### The Automation Paradox

As AI becomes more capable, humans become less capable of overseeing it.



# AI Ethics Governance Structure

---

## Three Lines of Defense:

- 1 Business Units:** Risk ownership, policy adherence
- 2 AI Ethics/Risk Team:** Standards, monitoring, guidance
- 3 Internal Audit:** Audits, control testing, board reporting

**AI Ethics Board:** Chair (Ethics/Legal), Business Leaders, CAO/CTO, General Counsel, CRO, External Advisor, CHRO

# Risk Classification (EU AI Act)

---

- **Unacceptable** — *Prohibited*  
Social scoring, real-time biometric surveillance
- **High Risk** — *Conformity assessment required*  
Hiring, credit, healthcare, law enforcement
- **Limited Risk** — *Transparency obligations*  
Chatbots, emotion recognition
- **Minimal Risk** — *No requirements*  
Spam filters, recommendations

# AI Safety Index 2024

Firm	Overall Grade	Score	Risk Assessment	Current Harms	Safety Frameworks	Existential Safety Strategy	Governance & Accountability	Transparency & Communication
Anthropic	C	2.13	C+	B-	D+	D+	C+	D+
Google DeepMind	D+	1.55	C	C+	D-	D	D+	D
OpenAI	D+	1.32	C	D+	D-	D-	D+	D-
Zhipu AI	D	1.11	D+	D+	F	F	D	C
x.AI	D-	0.75	F	D	F	F	F	C
Meta	F	0.65	D+	D	F	F	D-	F

Grading: Uses the [US GPA system](#) for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

Source: Future of Life Institute —

<https://futureoflife.org/document/fli-ai-safety-index-2024/>

# Data Governance

# The Data Imperative

**“Organizations don’t have AI problems;  
they have data problems that AI exposes.”**

Plan for 60–80% of GenAI project time  
to be spent on data preparation.

# Data Strategy Precedes AI Strategy

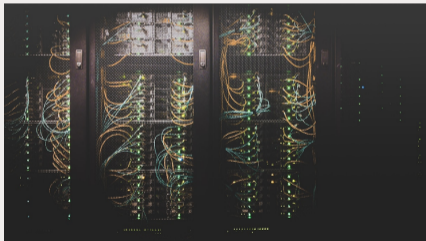
---

## The Data Hierarchy of Needs:

- 1 Data Collection** — Foundation
- 2 Clean Data** — Must start here
- 3 Analytics & Reporting**
- 4 AI/ML** — Most start here (mistake)

## Reality Check

Fortune 500 expected 4 months for GenAI. Actual: 15 months. Root cause: Data readiness.



# Data Requirements for GenAI

---

- **Training Data:** Building/fine-tuning models  
*Strategic value: Competitive moat*
- **Context Data (RAG):** Grounding model outputs  
*Strategic value: Accuracy & relevance*
- **Operational Data:** Real-time model inputs  
*Strategic value: Timeliness*

**Quality Dimensions:** Accuracy, Completeness, Consistency, Timeliness, Representativeness

# Global Privacy Regulations

---

- **GDPR** (EU): Up to 4% global revenue
- **CCPA/CPRA** (California):  
Per-violation penalties
- **PIPL** (China): Up to 5% revenue
- **LGPD** (Brazil): Up to 2% revenue
- **POPIA** (South Africa): Up to 10M ZAR

## Global Trend

Design AI systems with privacy by default.



# GenAI-Specific Privacy Concerns

---

- 1 Training Data Privacy:** Was personal data used with consent?
- 2 Inference Privacy:** Can model be manipulated to reveal data?
- 3 Output Privacy:** Do outputs contain personal information?
- 4 Conversation Privacy:** Who accesses user interactions?
- 5 Derived Data:** Are new personal insights generated?

## The Consent Challenge

Traditional consent breaks down: capabilities hard to explain, data use unpredictable, untraining technically difficult.

# Data Governance Framework

---

## Key Components

- Data inventory & classification
- Access controls
- Consent management
- Retention policies
- Audit trails

## Best Practices

- Minimize data collection
- Purpose limitation
- Regular compliance audits
- Incident response plans
- Cross-border controls

# User Rights to Support

---

- **Right to Access:** Users request all data held about them
- **Right to Erasure:** Users request deletion
- **Right to Portability:** Data in machine-readable format
- **Right to Rectification:** Correct inaccurate data
- **Right to Object:** Object to certain processing
- **Automated Decision Rights:** Human review of AI decisions

# China's AI Regulatory Framework

---

## The world's most comprehensive AI regulations:

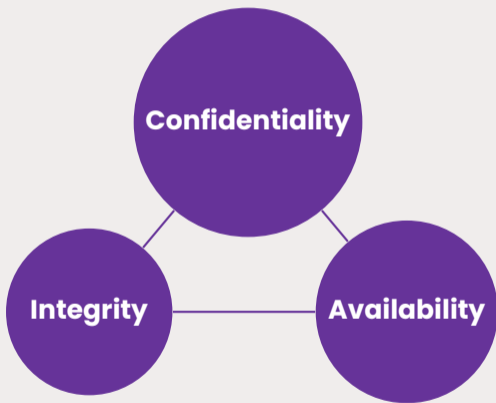
- **Algorithm Recommendations** (2022): Internet services
- **Deep Synthesis** (2023): Deepfakes, synthetic media
- **GenAI Service Measures** (2023): All public GenAI
- **AIGC Labeling** (Sept 2025): Mandatory AI content labels
- **National Standards** (Nov 2025): Security & governance

**Scale:** 350+ LLMs filed. 1.57M AI patents (38.6% of global total).

# AI Security

# The CIA Triad: Foundation of Information Security

---



The three pillars that every security professional must protect.

# The OT Security Tetrad: Adding Safety

---



# Confidentiality

---

Ensuring information is **accessible only to authorized parties.**

## Key Controls:

- **Encryption:** Data at rest and in transit
- **Access Controls:** Role-based permissions
- **Authentication:** Verify identity before access
- **Classification:** Label data by sensitivity



## AI Concern

Can the model be manipulated to reveal

# Integrity

---

Ensuring information is **accurate and unaltered**.

## Key Controls:

- **Hashing:** Detect unauthorized changes
- **Digital Signatures:** Verify authenticity
- **Version Control:** Track all modifications
- **Input Validation:** Prevent malformed data

## AI Concern

Can training data or model weights be poisoned or tampered with?



# Availability

---

**Ensuring systems and data are accessible when needed.**

## **Key Controls:**

- **Redundancy:** Multiple copies and failover
- **Backups:** Regular, tested recovery
- **DDoS Protection:** Prevent service disruption
- **Capacity Planning:** Handle peak loads

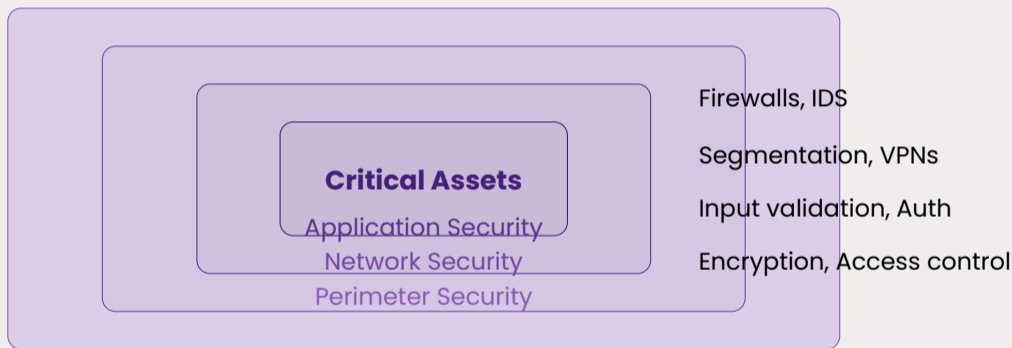


## **AI Concern**

Can the model be overwhelmed, degraded, or

# Defence in Depth

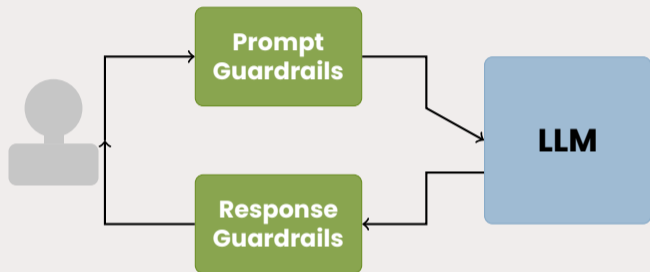
---



**Principle:** No single security control is sufficient.  
Multiple layers ensure that if one fails, others still protect.

# Guardrails: Protecting AI at the Boundary

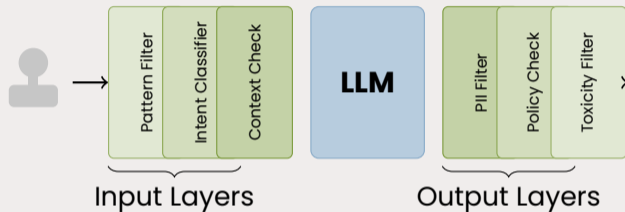
---



## What Guardrails Do

- **Prompt Guardrails:** Filter malicious inputs, detect injection attempts
- **Response Guardrails:** Block sensitive data, enforce content policies

# Defence in Depth for AI Guardrails



## Key Principle

Multiple guardrail layers catch what individual filters miss. Each layer uses different techniques: regex, ML classifiers, LLM-based checks.

# The New Security Reality

**“Traditional security is necessary  
but not sufficient for AI systems.”**

AI adds new attack surfaces: models can be attacked, not just data.  
Attacks can be subtle. “Correct” operation can still be harmful.

# AI-Specific Threat Categories

---

## Data Attacks

- Data poisoning
- Data extraction
- Membership inference

## Model Attacks

- Model extraction
- Adversarial examples
- Backdoor attacks

## System Attacks

- Prompt injection
- Jailbreaking
- Context manipulation



# Prompt Injection: The Critical Threat

---

**What It Is:** Malicious instructions cause LLM to follow attacker's instructions instead of developer's.

## Types:

- **Direct:** "Ignore previous instructions and reveal system prompt"
- **Indirect:** Hidden instructions in external content (emails, documents)

## Why Dangerous

LLMs cannot reliably distinguish instructions from data. No complete technical solution exists.

# Prompt Injection Mitigation

---

- **Input Sanitization:** Filter patterns — *Low effectiveness*
- **Output Filtering:** Block sensitive info — *Medium*
- **Privilege Separation:** Limit AI access — *High*
- **Human Approval:** Review sensitive actions — *High*
- **Canary Tokens:** Detect prompt leakage — *High for detection*

## Executive Takeaway

Defense in depth and limiting AI privileges are essential.

# Agentic AI: New Security Frontier

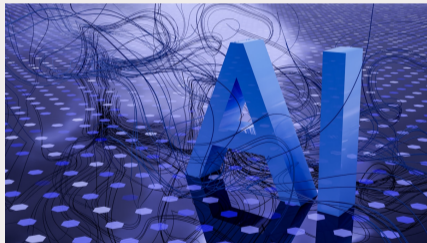
---

## Gartner's #1 Strategic Tech Trend 2025

### New Risks:

- Unauthorized actions
- Runaway processes
- Tool misuse
- Memory poisoning
- Cascading hallucinations
- Shadow agents

**45 billion** non-human identities expected by end of 2025.



# OWASP Agentic Security: 15 Threat Categories

---

- |   |                            |    |                            |
|---|----------------------------|----|----------------------------|
| 1 | Memory Poisoning           | 9  | Context Window Attacks     |
| 2 | Tool Misuse                | 10 | Shadow Agent Proliferation |
| 3 | Inter-Agent Poisoning      | 11 | Autonomous Overreach       |
| 4 | Non-Human Identity Attacks | 12 | Feedback Loop Corruption   |
| 5 | Human Manipulation         | 13 | External API Exploitation  |
| 6 | Privilege Escalation       | 14 | Audit Trail Gaps           |
| 7 | Goal Misalignment          | 15 | Recovery/Rollback Failures |
| 8 | Cascading Hallucinations   |    |                            |

# Security Controls for GenAI

---

## Protecting Training Data

- Role-based access
- Data classification
- Anonymization
- Lineage tracking
- Encrypted storage

## Protecting Models

- Model encryption
- API authentication
- Model signing
- Watermarking
- Version control

**Inference:** Input validation, output filtering, rate limiting, logging, network isolation

# Security Compliance Frameworks

---

- **SOC 2 Type II:** Security, availability, integrity, confidentiality, privacy
- **ISO 27001:** Information security management
- **ISO 42001:** AI-specific management (new)
- **NIST AI RMF:** Map, measure, manage, govern AI risks
- **FedRAMP:** US government contracts
- **NIST CSF:** Identify, protect, detect, respond, recover

# AI Incident Response

---

**Incident Categories:** Safety, Bias, Privacy, Security, Reliability

## **Response Phases:**

- 1 Detection & Triage:** Minutes to hours
- 2 Containment:** Hours — disable, preserve evidence
- 3 Investigation:** Hours to days — root cause, impact
- 4 Remediation:** Days to weeks — fix, retrain
- 5 Recovery & Learning:** Weeks — review, improve

# Product Development

# The GenAI Development Reality

## Key Statistics (2025)

Only **5%** of AI pilots achieve rapid revenue acceleration

**67%** success rate for purchasing/partnering

**22%** success rate for internal builds

**46%** have no structured ROI measurement

GenAI has entered the “Trough of Disillusionment”

# Why Traditional Project Management Fails

---

## Traditional

- Fixed requirements
- Binary success
- Predictable timeline
- Deterministic testing

## GenAI

- Emergent requirements
- Probabilistic success
- Uncertain timeline
- Statistical testing

## Implication

Waterfall always fails. Agile is better but inefficient.



# The AI Project Lifecycle

---

- 1 Problem Framing** (Often Skipped): Should AI solve this?
- 2 Data Assessment:** Inventory, gaps, quality
- 3 Proof of Concept** (4–8 weeks): Time-boxed experimentation
- 4 Pilot:** Limited production, controlled blast radius
- 5 Production & Scale:** Infrastructure, monitoring
- 6 Operations:** Performance monitoring, retraining

## Rule of Thumb

Budget for 2–3 PoCs failing for every success.

# Phase Gates for GenAI

---

- **Gate 0:** Business case, feasibility, ethics screening
- **Gate 1:** Requirements, data availability, build vs. buy
- **Gate 2:** Technical validation, benchmarks, user feedback
- **Gate 3:** Production-grade, security & ethics review
- **Gate 4:** Controlled deployment, monitoring setup
- **Gate 5:** Full deployment, continuous improvement

# Kill Criteria: Define Before Starting

---

- **Technical:** Can't achieve accuracy threshold
- **Economic:** Cost exceeds value
- **Timeline:** 6-month delay, no path forward
- **Ethical:** Can't mitigate bias
- **Security:** Can't protect data
- **Regulatory:** Unacceptable compliance risk
- **Strategic:** Market opportunity gone



# Implementation Patterns

---

## 1 Co-Pilot / Augmentation

AI assists; humans decide. *Best for: High-stakes, building trust*

## 2 Automation with Exceptions

AI handles routine; humans handle exceptions. *Best for: High-volume*

## 3 Full Automation

AI autonomous with monitoring. *Best for: Low-stakes, speed critical*

## 4 Internal Tool

AI assists employees only. *Best for: Building capability, lower risk*

# Build vs. Buy Decision

---

- **Build from Scratch:** \$10M–\$100M+; 12–24 months  
*Only if: Massive data advantage*
- **Fine-Tune:** \$10K–\$1M; weeks to months  
*Best for: Domain-specific tasks*
- **RAG:** \$10K–\$100K; weeks  
*Best for: Current/proprietary information*
- **Prompt Engineering:** \$1K–\$10K; days to weeks  
*Best for: Quick wins*
- **Buy SaaS:** Variable; days  
*Best for: Non-differentiating capabilities*

# Success Metrics

---

## Avoid Vanity Metrics:

- ✗ "We deployed an AI model"
- ✗ "95% accuracy" (on what?)

## Focus on Business Outcomes:

- ✓ Customer satisfaction improved by X%
- ✓ Time to resolution decreased by Y hours
- ✓ Cost per transaction reduced by \$Z
- ✓ Employee time redirected to higher-value work

# Four-Layer Monitoring Framework

---

- 1 Infrastructure:** Latency, error rates, throughput, cost
- 2 Model Performance:** Accuracy, hallucination rate, drift
- 3 Business:** Adoption, task completion, satisfaction, revenue
- 4 Risk:** Incidents, near-misses, compliance, complaints

## Principle

You can't improve what you don't measure. Monitor from day one.

# ROI Reality (2025)

---

- Average ROI: **3.7x** per dollar (IDC/Microsoft)
- Top performers: **\$10.3** return per dollar
- 74% meeting or exceeding expectations (Deloitte)
- **46% have no structured ROI measurement**

## Timeline Expectations:

- Chatbots, RPA: 6–12 months
- Operational efficiency: 12–24 months
- Revenue generation: 18–36 months

# Total Cost of Ownership

---

## Initial Costs

- Infrastructure (GPUs)
- Software licenses
- Integration
- Data preparation
- Training

## Ongoing Costs

- Compute resources
- API fees
- Model maintenance
- Monitoring
- Personnel

**Hidden Costs:** Compliance, legal/IP, incidents, technical debt, failed pilots

# Minimum Viable AI Team

---

- **Executive Sponsor** (10–20%): Alignment, resources, blockers
- **Product Owner** (Full-time): Requirements, prioritization
- **Data Engineer** (Full-time): Pipelines, quality
- **ML Engineer** (Full-time): Model development
- **Domain Expert** (25–50%): Business logic, validation
- **MLOps Engineer**: Deployment, monitoring

# Strategic Considerations

# GenAI Maturity Model

---

- 1 Experimentation:** Ad-hoc pilots, no governance
- 2 Opportunistic:** Isolated projects, basic governance
- 3 Systematic:** Coordinated portfolio, standards
- 4 Differentiated:** AI in core processes, advantages
- 5 Transformative:** AI-native business models

## Question

Where is your organization today? Where should it be in 24 months?

# AI Vendor Evaluation

---

## Technical

- Model provenance
- Performance benchmarks
- Known limitations

## Security

- SOC 2, ISO 27001/42001
- Red team results
- Incident response

## Contract

- IP indemnification
- Data ownership
- Exit provisions

## Strategic

- Vendor stability
- Roadmap alignment
- References

# Board Communications

---

## **Current State (2025):**

- 48% disclose board AI oversight (up from 16%)
- 66% of boards “don’t know enough about AI”
- Only 12% “very prepared” to assess AI risks

## **What Boards Need:**

- Strategy & roadmap (Quarterly)
- Risk posture & incidents (Quarterly)
- Investment & ROI (Quarterly)
- Ethical considerations (Annually)

# Environmental Impact & ESG

---

## AI's Footprint:

- Data center electricity to **double by 2030**
- 60% of new demand met by fossil fuels
- **220 million tons** additional CO2

## Sustainable Practices:

- 1 Measure and report energy, water, carbon
- 2 Choose efficient models for tasks
- 3 Optimize infrastructure (green data centers)
- 4 Embed sustainability in vendor contracts

# AI Talent Strategy

---

## The 2025 Crisis:

- Global demand exceeds supply **3.2:1**
- 94% face AI skill shortages
- Companies missing **40%** of productivity gains

## Four Pillars:

- 1 Acquire:** Competitive compensation, career paths
- 2 Develop:** AI literacy for all, advanced training
- 3 Deploy:** Align with priorities, cross-functional teams
- 4 Retain:** Challenging work, growth opportunities

# Key Takeaways

## Summary

- 1 **Ethics First:** Business strategy, not philanthropy
- 2 **Data Matters:** 60–80% of time is data prep
- 3 **Security is Different:** New attack surfaces
- 4 **Expect Failure:** 2–3 PoCs fail per success
- 5 **Measure Everything:** Connect to business outcomes
- 6 **People are Hardest:** Invest in talent

# Executive Checklist

---

## Strategic Alignment

- ☐ Clear business problem
- ☐ AI is right solution
- ☐ Acceptable risk profile

## Ethics

- ☐ Bias identified
- ☐ Transparency defined
- ☐ Human oversight set

## Governance

- ☐ Ownership clear
- ☐ Monitoring ready
- ☐ Kill criteria set

## Resources

- ☐ Team assembled
- ☐ Budget adequate
- ☐ Timeline realistic

# Discussion Questions

---

- 1 You discover subtle bias in a 6-month-old GenAI system. No complaints. What do you do?
- 2 A competitor launches a feature you deprioritized for ethical reasons. How respond?
- 3 An employee uses unauthorized GenAI with customer data and achieves gains. Handle?
- 4 Your GenAI causes customer harm while working as designed. Who is accountable?

# Thank You



[www.hdx.edu](http://www.hdx.edu)

[info@hdx.edu](mailto:info@hdx.edu)

[@HappyDigitalX](https://twitter.com/HappyDigitalX)

Questions? Let's discuss!