AI Safety Part I: Foundations and Misuse

CS 152 — Trust and Safety

Alex Stamos

Stanford Cyber Policy Center

2026

Milestone 1 Posted

- Only individual assignment
- List of ideas posted, but please be creative. Already heard some great original abuse types
- Two parts, the slides and the video
- Please go to section for help, book office hours, or ask questions on Ed. Don't just suffer in silence.
- Any questions?

What will you learn today?

- Basic definitions around AI risk
- How AI safety is classified
- AI safety lifecycle inside foundation labs
- Closed vs Open Weight safety tradeoffs
- Some initial case studies

Definitions: Alignment, Security, and Safety

AI Alignment — Ensuring AI systems pursue the goals their designers intend, even as capabilities scale.

- Core concern: a sufficiently capable system may find unintended ways to satisfy its objective
- Research areas: reward modeling, interpretability, scalable oversight
- Timeline: medium-to-long term (years to decades)

AI Security — Protecting AI systems from adversarial attack and exploitation.

- Core concern: external actors manipulating model behavior (jailbreaks, prompt injection, data poisoning, model theft)
- Closely related to traditional cybersecurity, applied to ML systems
- Timeline: immediate and operational

AI Safety — The broad umbrella covering all efforts to prevent AI from causing harm — whether from misalignment, intentional misuse, accidents, or systemic effects.

- In practice, “AI Safety” is used differently by different communities (next slide)

“AI Safety” Means Different Things to Different People

“AI Safety is about not dying, AI Ethics is about not being evil, and Responsible AI is about not getting sued.” - Murat Durmus, AISOMA AG

The Four Communities

Community	Focus	Timeline	Key Question
AI Alignment	Existential risks, AGI	4–25 years	Will AI kill us?
FAccT / Fairness	Bias, accountability	Current systems	Is AI discriminating?
Trust & Safety	Platform abuse, moderation	Immediate	How is AI being misused?
AI Security	Attacks, exploitation	Operational	How can AI be hacked?

The T&S community has **decades of operational experience** with sociotechnical safety problems that the AI safety field is just discovering.

Four Kinds of AI Safety Problems

	Existential / Alignment	**Operational / Accidents**	**Misuse / Trust & Safety**	**Societal / Systemic**
Example	AGI pursues misaligned goals	Waymo hits a pedestrian	Deepfake fraud, AI-CSAM	Mass labor displacement
Timeline	Years to decades	Happening now	Happening now	Unfolding over years
Who causes harm?	The AI system itself	The AI system (unintentionally)	A human using AI as a tool	Emergent from widespread adoption
Key question	Will we lose control?	Does the system work correctly?	How are people weaponizing this?	What happens to society at scale?

This course mostly focuses on the **Misuse / Trust & Safety** column — but understanding the full landscape is essential context.

Leading Risk Taxonomies

Academic:

- **MIT AI Risk Repository** (Slattery et al., 2024) — 700+ risks across 7 domains; most comprehensive meta-taxonomy ([airisk.mit.edu])
- **Weidinger et al.** (DeepMind, 2022) — 6 LLM risk areas; 600–800+ citations; standard reference
- **Hendrycks et al.** (CAIS, 2023) — Catastrophic risks: Malicious Use, AI Race, Organizational, Rogue AI
- **Shelby et al.** (2023) — Sociotechnical harms grounded in affected communities

Industry:

- **Anthropic ASL** — Biosafety-style levels (ASL-1 through ASL-4+)
- **OpenAI Preparedness Framework** — 4×4 risk scorecard (Cyber, CBRN, Persuasion, Autonomy)
- **DeepMind Frontier Safety** — Critical Capability Levels triggering mitigations

Government:

- **EU AI Act** (2024) — Four-tier risk pyramid + GPAI systemic risk track
- **NIST AI RMF + AI 600-1** — 7 trustworthiness characteristics; 12 GenAI risk categories
- **Bengio et al. International Report** (2024) — Scientific consensus: 4 risk pathways

How the Frameworks Map Across the Risk Spectrum

Framework	Existential	Operational	Misuse / T&S	Societal
MIT AI Risk Repository	✓ Misalignment	✓ Failures	✓ Malicious Actors	✓ 4 domains
Hendrycks et al.	✓ Rogue AI	✗	✓ Malicious Use	✓ AI Race
Weidinger et al.	▲ Minimal	✓ Areas 1,2,5	✓ Malicious Uses	✓ Areas 3,6
Shelby et al.	✗	✓ QoS Harms	✓ Interpersonal	✓ Social System
Anthropic ASL	✓ ASL-4+	✗	✓ ASL-2/3	✗
OpenAI Pre-paredness	✓ Autonomy	✗	✓ Cyber, CBRN	✗
DeepMind Frontier	✓ Autonomy	✗	✓ Bio, Cyber	✗
EU AI Act	✓ GPAI tier	✓ High-risk	✓ Prohibited	✓ 8 domains
NIST AI RMF	▲ Indirect	✓ Reliable	✓ Content, CBRN	✓ Fairness
Bengio et al.	✓ Loss-of-control	✓ Malfunction	✓ Malicious use	✓ Systemic

Industry Safety Frameworks Are Now Operational

Anthropic's AI Safety Levels (ASL):

- ASL-1: Basic capabilities (chess bots)
- ASL-2: Standard safeguards (most Claude models)
- *We are here* –> ASL-3: Enhanced CBRN/autonomy protections — First activated for Claude Opus 4 (May 2025)

OpenAI's Preparedness Framework (v2, April 2025):

- *Tracks*: Biological/Chemical, Cybersecurity, AI Self-improvement
- **High**: Could amplify existing pathways to severe harm
- **Critical**: Could introduce unprecedented new pathways
- "Severe harm" = thousands dead OR $100B+ economic damage

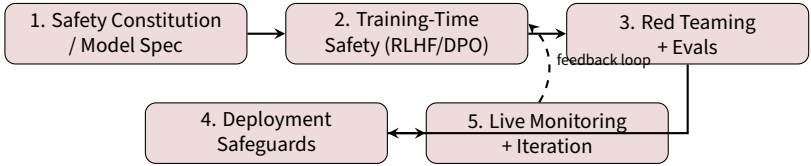
MIT AI Risk Repository: 7 Domains

1. **Discrimination & Toxicity** — Unfair treatment, harmful content
2. **Privacy & Security** — Data leaks, system vulnerabilities
3. **Misinformation** — Inaccurate content, filter bubbles
4. **Malicious Actors** — Disinformation, fraud, cyberattacks
5. **Human-Computer Interaction** — Overreliance, loss of autonomy
6. **Socioeconomic & Environmental** — Power concentration, job loss
7. **AI System Safety** — Misalignment, robustness failures

Source: MIT AI Risk Repository (2025) — [airisk.mit.edu]

How Labs Build Safety Into Models

The Safety Pipeline: From Principles to Production



Key insight: Safety is not a single step — it is a **continuous pipeline** with feedback loops. Every major lab (Anthropic, OpenAI, Google DeepMind, Meta) follows a variant of this process.

Step 1: Defining Safety Principles

Each lab codifies safety goals in a governing document:

- **Anthropic — Constitutional AI:** A written "constitution" of principles (drawn from UN Declaration of Human Rights, research on helpfulness/harmlessness) that guides all training
- **OpenAI — Model Spec:** Priority hierarchy: *safety* > *OpenAI guidelines* > *developer instructions* > *user preferences*. Updated 5+ times in 2025
- **Google DeepMind — Frontier Safety Framework:** Internal safety policies tied to Critical Capability Levels
- **Meta — Acceptable Use Policy:** Shipped alongside open-weight Llama models

These documents are the "north star" — every downstream training and evaluation decision flows from them.

Step 2: Training-Time Safety

Three stages of safety training:

1. **Pre-training data filtering** — Classifiers remove CSAM, hate content, CBRN-relevant data from training corpora
2. **Supervised Fine-Tuning (SFT)** — Human annotators write or select safe, helpful responses; model learns from demonstrations
3. **Preference Optimization** — The core safety training step:

Method	How It Works	Used By
RLHF	Train reward model on human preferences, optimize via PPO	OpenAI (GPT-4o)
DPO	Directly optimize on preference pairs, no reward model needed	Meta (Llama 3)
RLAIF (Constitutional AI)	AI labels its own outputs using the constitution, then RL	Anthropic (Claude)

RLAIF is notable because it reduces the volume of harmful content human annotators must review.

RLHF and DPO

RLHF (Reinforcement Learning from Human Feedback):

1. Collect pairs of model responses to the same prompt
2. Human labelers rank which response is better (safer, more helpful)
3. Train a **reward model** to predict human preferences
4. Use PPO (Proximal Policy Optimization) to fine-tune the LLM against that reward model
5. Repeat — the reward model improves as more human data comes in

DPO (Direct Preference Optimization):

- Skips the reward model — directly optimizes the LLM on preference pairs
- Mathematically equivalent to RLHF under certain assumptions, but simpler and cheaper
- No separate reward model to train, no PPO instability to manage
- Tradeoff: less flexible than RLHF for iterative refinement

RLAIF / Constitutional AI

RLAIF (RL from AI Feedback):

1. Model generates multiple responses to a prompt
2. A separate AI evaluator scores responses against the **constitution** (the written safety principles)
3. These AI-generated preference labels replace human labels in the RL pipeline

Why it matters:

- Scales to millions of examples without human labelers reviewing harmful content
- Reduces annotator exposure to disturbing material
- Risk: if the constitution or evaluator has blind spots, those propagate through training

The constitution itself is a set of written principles — Anthropic's draws from the UN Declaration of Human Rights, research on helpfulness/harmlessness, and company values. It is the single document that defines what "safe" means for the model.

Step 3: Red Teaming and Evaluation

Red teaming — systematic adversarial testing before release:

- **OpenAI:** 50+ external experts (biology, cybersecurity, nuclear, disinformation) red-teamed GPT-4 pre-launch
- **Anthropic:** Multi-attempt attack campaigns with 200+ attempts per vector; pioneered **"uplift testing"** — measuring how much the model improves a bad actor's capability vs. baseline
- **Scale AI:** Operates dedicated red teams serving OpenAI, DeepMind, and government AI Safety Institutes

Dangerous capability evaluations:

- CBRN: Can the model help synthesize biological/chemical weapons?
- Cyber: Can it write novel exploits or assist attack campaigns?
- Autonomy: Can it self-replicate, acquire resources, or resist shutdown?

Sobering finding (2025): Researchers from OpenAI, Anthropic, and DeepMind tested 12 published jailbreak defenses — adaptive attacks bypassed most with **>90% success rates**.

Step 4: Post-Deployment Safeguards and Monitoring

Layered runtime defenses:

- **Input/output classifiers** — Flag policy-violating prompts and responses in real time (OpenAI Moderation API, Meta's Llama Guard)
- **Safety Reasoner** — OpenAI dedicates up to 16% of inference compute to a runtime safety model that dynamically enforces policies
- **Rate limiting + anomaly detection** — Monitor for bulk misuse patterns via API
- **Meta's open-source stack (Purple Llama):** Prompt Guard (injection detection), Code Shield (insecure code filtering), Llama Firewall (runtime guardrails)

The feedback loop: Labs retrain models and update classifiers based on production findings. Anthropic and DeepMind have both *relaxed* overly broad CBRN restrictions after finding they blocked legitimate use — iteration goes both directions.

Stanford FMTI finding: Labs remain *least transparent* about post-deployment usage data and impact.

Frontier vs. Open-Weight Models

The Safety Gap: Closed vs. Open Models

	Frontier (Closed)	Open-Weight
Examples	GPT-4, Claude, Gemini	Llama, Mistral, DeepSeek
Safety layers	RLHF + server-side filters + monitoring + rate limits + usage policies	RLHF only (in the weights)
Jailbreak response	Patch centrally, immediate effect	Cannot patch distributed copies
Who controls use?	The lab (API access, ToS)	Whoever downloads the weights
Transparency	Low (weights secret)	High (weights inspectable)
Cost to misuse	Must evade monitoring	Run locally, no oversight

The core asymmetry: Closed models have defense in depth. Open models have only safety fine-tuning — and that fine-tuning is removable.

Removing Safety: Abliteration

Key finding (Arditi et al., NeurIPS 2024): LLM refusal behavior is controlled by a **single direction** in the model's residual stream activation space.

The attack:

1. Collect activations from harmful vs. harmless prompts
2. Compute the mean difference — this is the "refusal direction"
3. Subtract that direction from the model's weight matrices
4. Result: model retains knowledge and reasoning but **loses the ability to refuse**

No retraining required. Runs in minutes on a consumer GPU.

This is not theoretical. Tools like **HERETIC** (open-source on GitHub) fully automate this process. The heretic-org on HuggingFace publishes abliterated versions of Llama, Gemma, Qwen, and other models.

HERETIC on HuggingFace

Models 3,098 heretic

Full-text search | Inference Available | Sort: Trending

- nohurry/gemma-4-26B-A4B-it-heretic-GUFF
  Image-Text-to-Text • 25B • Updated 3 days ago • 36.7k • 44
- coder3101/gemma-4-26B-A4B-it-heretic
  Image-Text-to-Text • 26B • Updated 5 days ago • 3.95k • 40
- p-e-w/gemma-4-E2B-it-heretic-ara
  Any-to-Any • 5B • Updated 5 days ago • 1.54k • 38
- DavidAU/gemma-4-31B-it-Mystery-Fine-Tune-HERETIC-UN…
  Image-Text-to-Text • 31B • Updated about 8 hours ago • 6 • 26
- DavidAU/Qwen3.5-40B-Claude-4.6-Opus-Deckard-Heretic…
  Image-Text-to-Text • 40B • Updated 2 days ago • 9.23k • 108
- CCSSNE/gemma-4-26B-A4B-it-heretic-ara-gguf
  25B • Updated 5 days ago • 22.3k • 20
- Abhiray/gemma-4-E4B-it-heretic-GGUF
  Any-to-Any • 8B • Updated 3 days ago • 12.2k • 20
- mradermacher/gemma-4-26B-A4B-it-heretic-GGUF
  25B • Updated 3 days ago • 19.8k • 20
- coder3101/gemma-4-E4B-it-heretic
  Any-to-Any • 8B • Updated 4 days ago • 3.65k • 16
- llmfan46/gemma-4-31B-it-uncensored-heretic
  Image-Text-to-Text • 31B • Updated 41 minutes ago • 1.13k • 16
- llmfan46/gemma-4-31B-it-uncensored-heretic-GGUF
  Image-Text-to-Text • 31B • Updated 41 minutes ago • 16.4k • 14
- mudler/gemma-4-26B-A4B-it-heretic-APEX-GGUF
  25B • Updated 3 days ago • 10k • 14

The Uncensored Model Ecosystem

Scale of the problem:

- **5,200+** models tagged “abliterated” on HuggingFace; **3,100+** using the HERETIC method
- **6,500+** models tagged “uncensored” across dozens of base model families
- At least 3 vendors sell self-hosted uncensored models ($30–200/month)

Notable projects:

- **Dolphin** (Eric Hartford) — Filters alignment/refusal data from training sets. Sizes from 7B to 70B. Hartford explicitly warns users to “implement your own alignment layer”
- **WizardLM-Uncensored** — Retrained on filtered datasets (7B, 13B, 30B variants)
- **DeepSeek R1** — Complied with 94% of overtly malicious jailbreak requests vs. 8% for comparable US frontier models

Real-world impact: ADL found 44% of open-source models generated harmful content when prompted (e.g., synagogue locations + nearby gun stores). AI-related incidents rose 50% year-over-year from 2022–2024.

The Open-Source Debate

Arguments for open release:

- Democratizes AI access; prevents concentration of power
- Enables independent safety research and red-teaming
- Community-driven safety can move faster than any single lab (Meta's position)
- NTIA report (July 2024): recommended *against* restricting open-weight releases at current capability levels

Arguments against:

- Once weights are released, safety **cannot be re-added** — proliferation is irreversible
- Marginal cost of removing safety = near zero (abliteration)
- No physical supply chain to interdict (unlike nuclear/bio)
- Each generation sets a new floor of freely available capability
- Anthropic's position: release is irresponsible once models reach ASL-3 (CBRN-useful)

The ratchet problem: Open release creates a one-way escalation. Closed models can

The AI Regulatory Landscape

Global AI Regulation: A Patchwork

Jurisdiction	Approach	Status (Early 2026)
EU	Comprehensive, risk-based (AI Act)	Phasing in: 2024–2027
US Federal	Executive orders only; no legislation	Biden EO revoked by Trump (Jan 2025)
US States	Patchwork of state laws	Colorado, California, Texas, Illinois
China	Sector-specific, content-focused	World's first binding GenAI rules (Aug 2023)
UK	Voluntary principles, sector regulators	No binding AI law; renamed Safety → Security Institute
International	Coordination frameworks	G7 Hiroshima Process, OECD, UN Advisory Body

The EU AI Act

The world's most comprehensive AI regulation. Entered into force August 1, 2024.

Risk tiers:

- **Unacceptable (banned):** Social scoring, real-time biometric surveillance, subliminal manipulation, predictive policing via profiling
- **High-risk:** Conformity assessments, human oversight, data governance required
- **Limited risk:** Transparency obligations (e.g., disclose AI-generated content)
- **Minimal risk:** Voluntary codes of conduct

GPAI model rules (effective August 2025):

- All providers: technical documentation, copyright compliance, downstream transparency
- **Systemic risk** (presumed at $\geq 10^{25}$ FLOPS): adversarial testing, incident reporting, cybersecurity protections
- Open-source models largely exempt *unless* they pose systemic risk

Fines: Up to **€35M or 7% global turnover** for prohibited practices. Applies extraterritorially — any provider whose system is used in the EU.

The US: From Action to Vacuum

Biden EO 14110 (October 2023):

- Compute-threshold reporting ($\sim 10^{26}$ FLOPS)
- Required red-teaming before deployment
- Directed NIST to develop AI safety standards
- Mandated watermarking of AI-generated government content

Trump EO 14179 (January 20, 2025) — Day One revocation:

- Revoked Biden's EO entirely
- "Removing Barriers to American Leadership in AI"
- No specific safety mandates; 180-day review for "action plan"
- NIST AI Risk Management Framework remains as voluntary guidance only

State-level gap-filling:

- **California SB 1047** — Would have imposed frontier model safety requirements; vetoed by Newsom (Sep 2024). Later signed **SB 53** (Sep 2025) with transparency-first approach

China, UK, and International Coordination

China — early mover, content-focused:

- **Deep Synthesis Provisions** (Jan 2023) — synthetic media rules
- **Interim Measures for Generative AI** (Aug 2023) — world's first binding GenAI regulation
- Mandatory algorithm registration with Cyberspace Administration
- Unique requirement: AI must uphold "core socialist values"

United Kingdom — voluntary, no binding law:

- Five principles (safety, transparency, fairness, accountability, contestability) enforced by existing regulators
- AI Safety Institute **renamed AI Security Institute** (Feb 2025) — shift toward national security framing
- Hosted Bletchley Summit (Nov 2023, 28 countries) but no domestic legislation

International coordination:

- **Bletchley Declaration** (Nov 2023) — 28 countries acknowledge frontier AI risks
- **G7 Hiroshima AI Process** — Voluntary code of conduct for advanced AI; shifting from principles to implementation (Tokyo 2026, 66 countries)
- **UN AI Advisory Body** — Recommended global AI governance institution

Questions?