Checkpoint #1

Sage, Sammy, Will, Luke

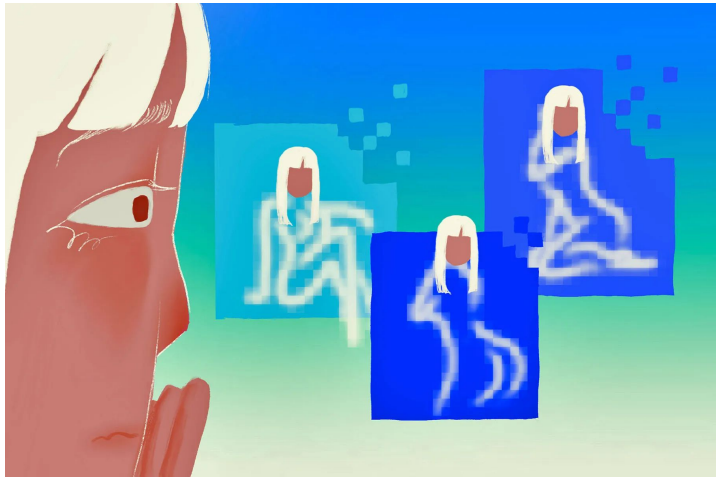
NCEI Threat Intelligence Platform

- Team name: Sentinel
- Members: Luke (Qiming), Will (Yu-Chen), Sage, Sammy
- IS 492, Spring 2026
- GitHub: https://github.com/IS492-SP26/team-project-deepfakes/tree/main

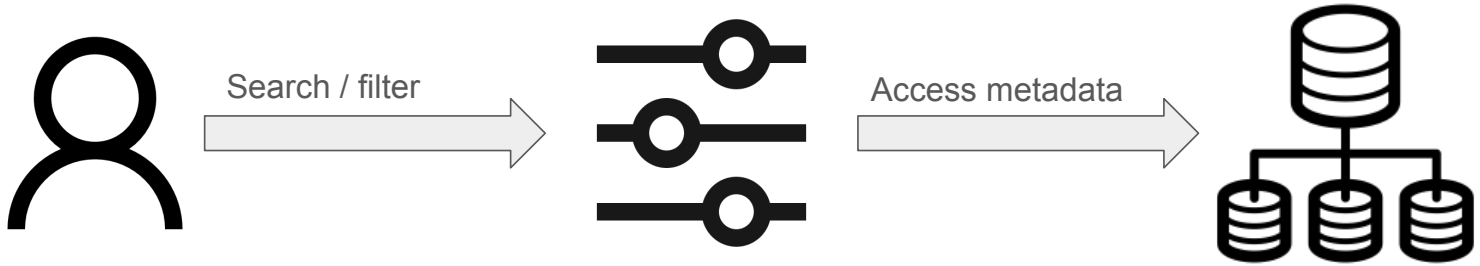
Problem & Motivation

Gen-AI has led to a surge in Non-Consensual Explicit Imagery (NCEI). Current safety frameworks are reactive, often failing to track the specific "how" behind these attacks. Victims and researchers lack a centralized, structured repository that logs how platform guardrails (like those of X/Grok) are bypassed, the specific "nudify" tools being used, and the legal precedents set in response.

Without a granular record of these incidents, developers cannot patch specific vulnerabilities, and lawmakers cannot draft precise regulations.



Target Users	Core Tasks
Safety Researchers	Analyze trends in model exploitation and platform-specific failure modes.
Legal Policy Advocates	Track the success of "Take It Down" acts and other legal precedents to argue for better protection.
Platform Integrity Teams	Use logged bypass techniques to improve red-teaming and prompt filtering.



Existing Tools & Gaps

System	AI Incident Database (AIID)	MIT AI Incident Tracker & Risk Repository	StopNCII.org StopNCII
Link	https://incidentdatabase.ai/	https://airisk.mit.edu/ai-incident-tracker https://airisk.mit.edu/	https://stopncii.org/about-us/
Primary Focus	Broad documentation of AI incidents	Risk classification & visualization of AI harms	Victim image removal & prevention
Strengths	Public, empirical case archive across domains	Structured taxonomy, Interactive dashboard, Harm categorization	Hash-based image matching, victim-centered support
Limitations (Gap)	Not NCEI-specific (too broad), limited technical metadata (e.g. model type), no lifecycle tracking	focuses on harm categories rather than tech aspects such as how incidents are technically created or spread	not a public research database (it's more like a service) and does not document how incidents occur

Key Insights from the Literature

Name	Sammy	Sage	Will	Luke
Recurring Themes	Legal and technological race to keep pace with AI-generated "digital forgeries"	AI-related harms are real and expanding, and existing responses (detection, policy, taxonomy) are insufficient	Detection-centric responses dominate deepfake research, reflecting an ongoing technological arms race between generative models and forensic classifiers.	The "arms race" between forgery generation and detection, and the interference of video compression on forensic accuracy.
What We Learned	Platforms now face a strict **48-hour removal window** for reported imagery	One AI incident can produce multiple harms (psychological, reputational, etc.), Deepfake lifecycle: Creation → Distribution → Detection → Mitigation	Deepfake defenses largely focus on ML/DL detection, favoring classification accuracy over modeling adversarial misuse dynamics.	Large-scale data is foundational for detection , but lightweight models like MesoNet are more efficient in compressed environments.
Design Implication	Needs an automated backend to purge the reported image and all **identical copies** across the platform	Multi-label harm tagging system based on AI harm taxonomy + lifecycle-based metadata for each incident	Reveals the need for structured intelligence systems that capture vulnerabilities, failures, and emerging exploitation techniques.	An automated backend using face-cropping and temporal aggregation to provide robust veracity scores for reported incidents

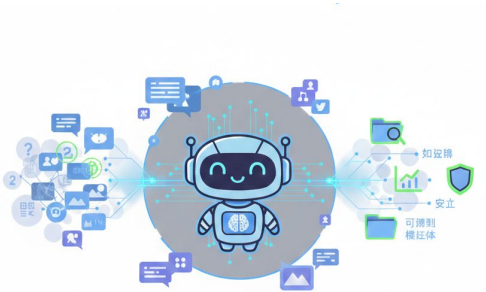
Initial Concept & Value Proposition

What are we proposing to build?

Sentinel is a taxonomy-driven threat intelligence infrastructure specifically for Non-Consensual Explicit Imagery (NCEI).

Intelligence Pipeline: An automated system that converts unstructured incident narratives from public sources into standardized, queryable threat metadata.

Analytical Framework: A specialized layer designed to capture guardrail failures, model-specific characteristics, and adversarial prompting strategies.



How is it meaningfully different?

Sentinel shifts the defensive paradigm from reactive moderation to proactive intelligence.

Focus on Exploitation: Instead of just detecting what was created, we analyze how it was produced by documenting bypass mechanisms and platform failure patterns.

Reusable Knowledge Base: We move beyond isolated incident responses to create a framework for identifying recurring vulnerabilities across different AI models.

Milestones & Next Steps

A structured 9-week timeline transitioning from foundational data infrastructure to AI-driven analysis and public deployment.

Weeks 3-4

Data Pipeline

- Build robust Python scrapers for data ingestion.
- Design initial database schema for storage.

Weeks 5-6

AI Parsing

- Integrate Llama 3 for incident classification.
- Extract taxonomy to automate insights.

Weeks 7-8

Dashboard

- Develop public-facing repository interface.
- Automate the data update loop.

Week 9

Ethics Review

- Verify data privacy and classification accuracy.
- Final collaborative review before release.

Roles

Success is driven by a specialized team with clear ownership across technical domains and a collective commitment to ethical standards.

Milestone	Lead / Owner	Primary Responsibility
Data Pipeline	**Sage Kim & Qiming Li**	Scraper development and database architecture.
AI Parsing	**Sammy Haskel**	Llama 3 integration and taxonomy extraction.
Dashboard & DevOps	**Yu-Chen (Will) Su**	Frontend development and deployment automation.
Testing & Ethics	**Entire Team**	Collaborative review of privacy and accuracy.

Technical Synergy

The overlap between database architecture and AI integration ensures seamless data flow and structural integrity throughout the pipeline.

Quality Assurance

The final week involves all stakeholders to ensure the project meets both technical benchmarks and rigorous ethical standards.

Thank You

Sage, Sammy, Will, Luke