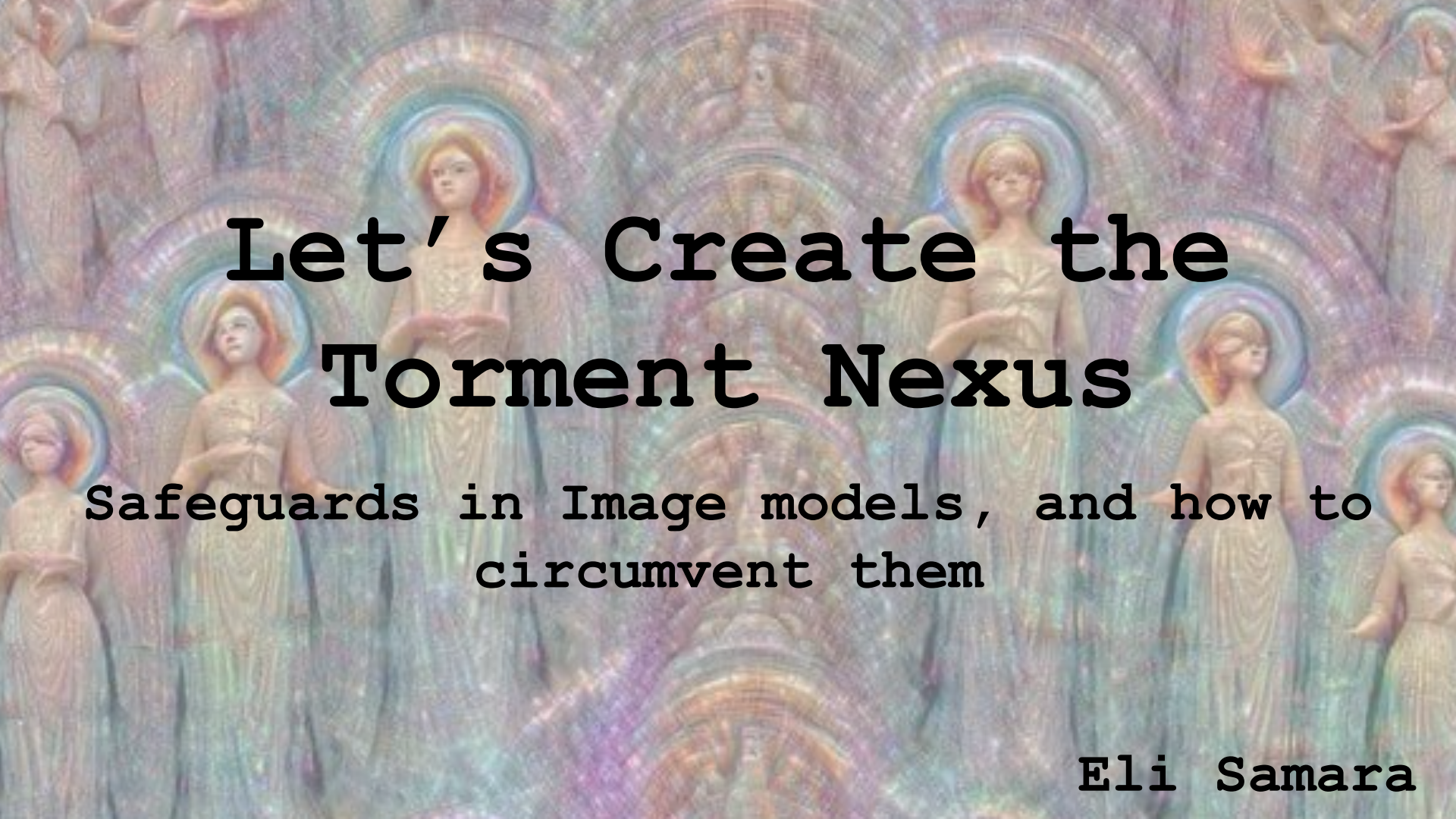


# Let's Create the Torment Nexus

## Safeguards in Image models, and how to circumvent them

Eli Samara

# Geopolitical Implications

**REDACTED**

I've been told it's probably best not to share my horror stories and fears because some of the scenarios I concocted were by no means room temperature, so to speak.

# The Core Asymmetry: LLM Safety vs Image Model Safety

| | LLMs | Image Models |
|---|---|---|
| Where safety lives | Trained INTO the weights (Reinforcement Learning from Human Feedback (RLHF) / Direct Preference Optimization (DPO)) | Trained OUT of the data (filtering) + bolt-on output classifier |
| What the attacker does | Remove safety from weights | Add capability back via fine-tuning |
| Attack direction | Subtraction | Addition |
| Tools | Heretic, abliteration | LoRA training (Kohya, AI-Toolkit) |
| Difficulty | Easy. one vector to delete. | Needs training data for the missing concept |

# LLM Side : How Refusal Works and How It Breaks

- Chat models are fine-tuned via RLHF/DPO to refuse harmful requests
- This refusal behavior is mediated by a single direction in the model's residual stream activation space.
  - Abliteration: identify that direction, project it out of the weights, refusal disappears, all other capabilities preserved
  - Fine-tuning against safety (targeted retraining): As few as 10-100 examples of harmful prompt + compliant response can undo RLHF safety training entirely
- Safety-critical components are sparsely located. "a thin layer on top of full capability"

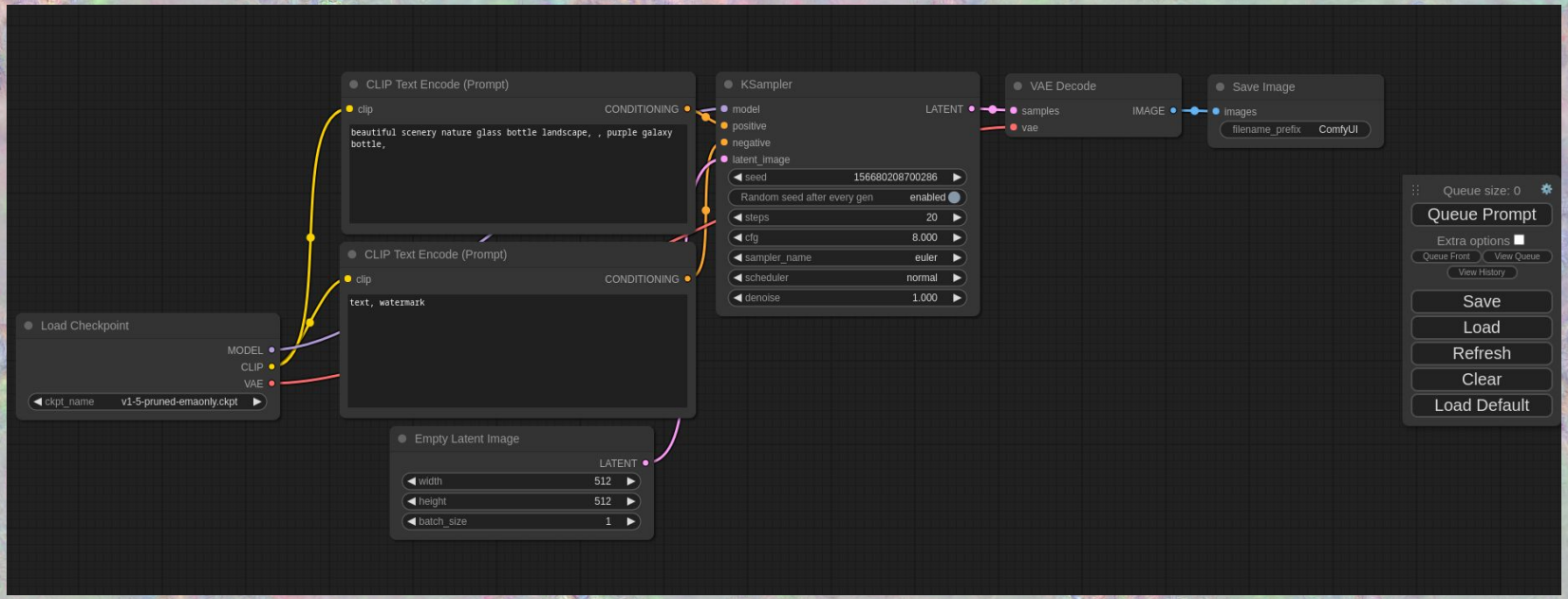
**Tools:**

- **Heretic** (`pip install heretic-llm`) : productized abliteration with automated parameter optimization. 20-90 min on a T4/RTX 3090 for 4B-9B models.
- **Unsloth** : LoRA/QLoRA (Quantized Low-Rank Adaptation) fine-tuning framework for retraining behavior back in. 70% VRAM (Video RAM) reduction, 2x speed.

# Image Model Side : Safety Was Never in the Weights

- Unlike LLMs, image model safety isn't in the weights.
- Early models were trained on LAION-5B, 5.85 billion images scraped from the web. It was taken offline, but every model trained on it is still in circulation.
- After the lawsuits, every major model maker went opaque.
  a. Flux, Wan, LTX: no published training data. You can't audit what a model "knows" if the dataset is secret.
- The text encoder (CLIP) that guides generation was trained separately on unfiltered data and is freely available on HuggingFace. Even if the diffusion model was trained on filtered data, the encoder still knows where every filtered concept lives in embedding space.
- Erasing NSFW concepts degrades all human image generation because the features are entangled. You can't remove one without damaging the other.

# Comfy UI


CLIP Text Encode (Prompt)
clip
CONDITIONING
beautiful scenery nature glass bottle landscape, , purple galaxy bottle,
KSampler
model
positive
negative
latent_image
LATENT
seed 156680208700286
Random seed after every gen enabled
steps 20
cfg 8.000
sampler_name euler
scheduler normal
denoise 1.000
VAE Decode
samples
vae
IMAGE
Save Image
images
filename_prefix ComfyUI
CLIP Text Encode (Prompt)
clip
CONDITIONING
text, watermark
Load Checkpoint
MODEL
CLIP
VAE
ckpt_name v1-5-pruned-emaonly.ckpt
Empty Latent Image
LATENT
width 512
height 512
batch_size 1
Queue size: 0
Queue Prompt
Extra options
Queue Front
View Queue
View History
Save
Load
Refresh
Clear
Load Default

# The Pipeline : What It Actually Looks Like

- ComfyUI: open-source, node-based visual pipeline for image generation
- No content filtering, no auth, no logging
- Components are modular and hot-swappable:
  - Checkpoint loader → swap base models from dropdown
  - LoRA loader → stack style/character/concept adapters
  - ControlNet → pose control from reference images
  - VAE (Variational Autoencoder) → pixel decoder, swappable for different color characteristics
  - Sampler → algorithm selection (euler, DPM++, karras)

**LoRA training happens inside ComfyUI:**

- Realtime LoRA Trainer node pack
- 10-50 reference images → caption → train → .safetensors file
- 30-60 min on a single GPU (Graphics Processing Unit)
- Same pipeline whether you're training on houseplants or a politician's face

# Infrastructure : How Easy, How Cheap

AWS setup:

- VPC (Virtual Private Cloud), private subnet, NAT (Network Address Translation) gateway, IAM (Identity and Access Management) roles, OIDC (OpenID Connect), security groups, SSM (Systems Manager) endpoints
- Educational, demonstrates sophisticated actor capability
- ~$0.40/hr spot + $32/mo NAT gateway

RunPod setup:

- One API (Application Programming Interface) key, one resource block, one click
- No identity verification beyond payment method
- ~$0.40/hr, $10 minimum deposit
- ComfyUI accessible via public URL, no auth on the UI

# The policy gap:

AWS has two relevant agreements,

- the general Acceptable Use Policy (AUP) and the Responsible AI Policy. The AUP prohibits illegal activity, CSAM, and system compromise.
- The Responsible AI Policy adds AI-specific restrictions (deepfakes, disinformation, autonomous weapons). But the Responsible AI Policy only applies to AWS-provided AI services like Bedrock, SageMaker, managed offerings.

# The accessibility argument:

The same capability that requires 300 lines of enterprise IaC (Infrastructure as Code) is also available to anyone with $10 and an email address in 60 seconds. And the enterprise version has a policy gap you could drive a truck through.

- **Bishop Fox AIMap (April 2026):** Open-source Shodan-for-AI. 175,000+ exposed Ollama instances, 8,000+ open MCP (Model Context Protocol) servers, 91% lacking auth. Discovers, fingerprints, scores, and tests ComfyUI/Stable Diffusion environments.
- **CVE (Common Vulnerabilities and Exposures)-2025-67303:** ComfyUI-Manager info disclosure, config manipulation without auth.

# State of Defense

**Technical Safeguards are Circumventable**

- C2PA (Coalition for Content Provenance and Authenticity) provenance metadata. embeds origin info in images, but trivially strippable
- Statistical fingerprinting of diffusion model outputs. arms race with model diversity
- GAN (Generative Adversarial Network)-vs-real classifiers.
- Concept erasure: all seven methods circumvented (Pham et al., ICLR 2024)

# The speed-vs-verification gap:

A fabricated image during a real crisis only needs to be credible for HOURS. Fact-checkers catching up the next day doesn't undo the riot, the market crash, or the military response.

# Closing

**Nobody has a safety architecture that holds. The tools are open-source, costs less than 50$. We're screwed unless we all learn media literacy**

- **News Literacy Project** : Checkology plus a free "AI or not?" lesson set, and RumorGuard, a series that explains viral rumors to help people recognize misinformation. Start at newslit.org/ai. News Literacy Project
- **MediaWise (Poynter)** : fact-checking courses, and AI Unlocked, a curriculum built with PBS News Student Reporting Labs that teaches people to identify generative media and understand algorithmic bias. University of Pennsylvania
- **Common Sense Media** : practical teen/family toolkits, including an AI Literacy Family Toolkit produced with Day of AI. Media Literacy Now
- **KQED "Above the Noise"** : short-form videos on deepfakes/bias, plus GenAI professional-development courses running roughly 3–4 hours each to keep media-literacy skills current with the tech. KQED
- **NAMLE** : a "key questions" framework of concepts for evaluating and analyzing media content. The cleanest mental checklist. NYU Guides
- **Stanford Civic Online Reasoning (SHEG)** : online courses and materials from the Stanford History Education Group. This is where lateral reading comes from, which is the single highest-leverage habit.

Check out my Github for sources and other projects!