POISON IN THE WELL

// ADVERSARIAL AUDIO & MACHINE LISTENING

A working report on unlearnable music, the psychoacoustics of refusal, and the old practice of being one thing to a person and another thing to a machine.

FOR THE CLERGY & FELLOW TRAVELERS

18 SOURCES · 2 EXHIBITS · 4 LISTINGS · 5 CASES

I · FIELDNOTE

01 The track that sounded like nothing had changed

The first protected file I ran through fooled me before it fooled anything else. I had a four-minute instrumental, modular drones with a tape-saturated piano loop, and I made two renders: the clean one and one run through an error-minimizing perturbation pass built on the HarmonyCloak approach. On open-back headphones at a quiet hour I could not pick the difference. No hiss, no flutter, no codec-style smear. A small genre classifier I keep around for tagging had been calling the clean file ambient with high confidence. Fed the protected version, it produced a different label with the same confidence. Nothing had changed for me. Everything had changed for the model. The HarmonyCloak authors name that gap directly: humans and machines read the same data in different ways, and the whole field lives inside that gap.[1]

This is a documented class of techniques, not a stunt. Security researchers call them adversarial examples, first characterized for image recognition around 2013 to 2014, and the music tools point them in a specific direction. The goal is not to fool a stop-sign detector. It is to make a recording hard for a generative model to learn from while leaving it untouched for a listener. The builders call this poisoning. I want to read it the way this collective already reads most things: as the newest layer of a long habit of refusing to be cleanly machine-readable, the same instinct behind CV Dazzle face paint, behind Fawkes cloaking your photos against facial

recognition, behind every file that does one thing for a human and another for a scanner.

> **Humans and machines interpret data in different ways, so there is a perceptual gap between them. Everything in this report is engineering inside that gap.**
>
> PARAPHRASE OF THE HARMONYCLOAK TEAM, UNIV. OF TENNESSEE KNOXVILLE

II · MECHANISM

02 What actually happens to the waveform

A model that "listens" never hears music. It ingests a tensor, usually a log-mel spectrogram, sometimes a raw waveform fed through a learned front-end, and during training it has found that certain numerical patterns reliably co-occur with certain labels or certain next-frame predictions. The Poison Pill founder put the failure mode plainly in *Music Ally*: the model is looking for shortcuts, highly specific sound features such as a particular resonance when a guitar string is plucked, and it keys its genre guess on those.[2] He also admitted the part vendors usually hide. They never know exactly which features the model latched onto. The defender is probing a black box too.

Adversarial audio exploits that fragility by computing the gradient of the model's loss with respect to the input itself, then nudging the samples to confuse the model, then repeating. The Linz group at JKU demonstrated this end-to-end on raw waveforms in 2020 for instrument classification, driving accuracy toward a random baseline while keeping the perturbation almost imperceptible, and steering the misclassification toward any instrument they chose.[3] These attacks are aimed, not random noise.

THE PSYCHOACOUSTIC BUDGET

The reason you cannot hear it is borrowed wholesale from codec engineering. Human hearing has masking thresholds: a loud tone raises the floor below which nearby quieter tones become inaudible. MP3, AAC and Opus all exploit this by discarding content your ear was going to throw away anyway. Adversarial audio runs the same psychoacoustic model in reverse, pouring the attack energy into the exact time-frequency cells where your hearing is already deaf. Schönherr, Kolossa and colleagues showed this against speech recognition in 2018: malicious target transcriptions hidden below the hearing threshold, transcribed by the Kaldi system in up to 98 percent of cases, with listeners in the user study hearing nothing wrong and understanding the original speech at unchanged accuracy.[4]

MASKING.PY · WHERE THE POISON HIDES

```python
# Borrowed straight from MP3/AAC: a loud tone raises the floor
# below which nearby quieter content is inaudible. Pour the attack there.
def psychoacoustic_clip(x, delta):
    S      = stft(x)                      # spectrum of the music
    thresh = masking_threshold(S)         # per-bin hearing floor (dB SPL)
    D      = stft(delta)
    # zero energy poking ABOVE the floor (audible)
    D      = where(power(D) > thresh, 0.0, D)
    return istft(D)                     # indistinguishable waveform
```

The descent finds the attack; this function decides how much of it a human will notice. Tightening the threshold trades attack strength for imperceptibility, the core tension in every unlearnable-data paper.[4]

HarmonyCloak makes a distinct choice worth being precise about. It does not maximize the model's error. It uses imperceptible **error-minimizing** noise to push the model's generative loss toward zero on the protected sample, which tricks the model into concluding there is nothing left to learn from the track.[1] That is a different gesture from Nightshade's label-corruption, and the distinction matters: one starves the model, the other lies to it. In code the two objectives sit one sign apart.

ATTACK_CORE.PY · THE GRADIENT LOOP

```python
# Two families, one arithmetic. The sign of the step is the whole argument.
def perturb(x, model, mode="evade", eps=0.002, steps=200):
    delta = zeros_like(x)                # the inaudible payload
    for _ in range(steps):
        loss = model.loss(x + delta)        # grad w.r.t. INPUT
        g = grad(loss, delta)
        # evade (Linz): loss UP, misclassify
        if mode == "evade":
            delta = delta + eps * sign(g)
        # cloak (HarmonyCloak): loss DOWN, "nothing to learn"
        else:
            delta = delta - eps * sign(g)
        delta = psychoacoustic_clip(x, delta)  # under masking curve
    return x + delta
```

Schematic. Real implementations use projected gradient descent with momentum and a learned surrogate model; the audible-distortion constraint is the hard part, not the descent.[1,3]

III · THE IMAGE PRECEDENT

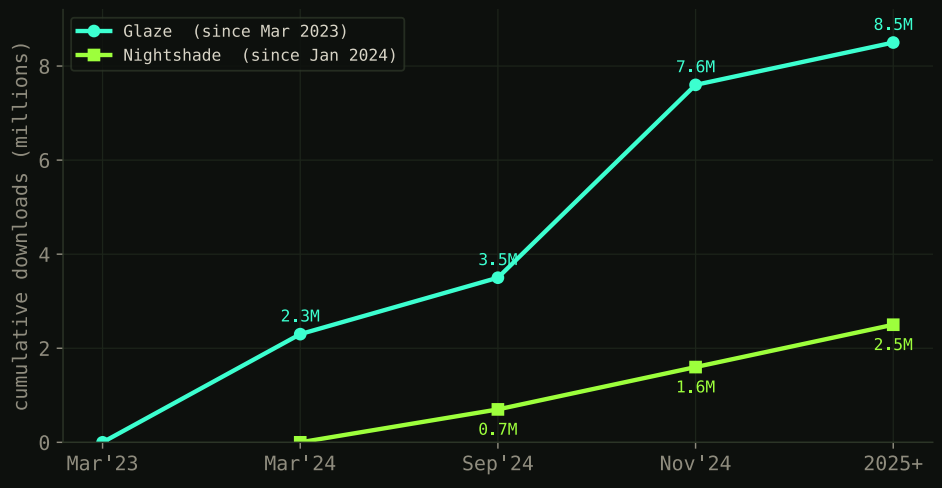
03 Glaze, Nightshade, and the doctrine of friction

None of the audio work happened in a vacuum. It followed three years of image-domain tooling out of the University of Chicago SAND Lab, run by Ben Zhao and Heather Zheng, with Shawn Shan as lead student author. It started, by the lab's own account, on a Zoom call in November 2022 hosted by the Concept Art Association,

when the illustrator Kim Van Deun pulled Zhao, then known for anti-facial-recognition work, into the fight.[5]

Glaze is the style shield. It applies barely perceptible perturbations so a model fine-tuning on an artist's images cannot reproduce their style. The original USENIX Security 2023 paper, by Shan, Cryan, Wenger, Zheng, Hanocka and Zhao, won a Distinguished Paper Award and the Internet Defense Prize, reported over 92 percent success at preventing style mimicry under typical conditions and 85 percent in a more robust setting, and noted glazing one artwork took about 1.2 minutes on a Titan RTX GPU.[6] **Nightshade** is the offensive counterpart. Rather than hide a style, it runs a prompt-specific poisoning attack that corrupts the link between a concept and its label, so a poisoned image of one object teaches the model the wrong thing, accepted at IEEE S&P.[7] Zhao's own framing is the one worth keeping: the aim is not a permanent wall but *friction*, enough resistance that platform and artist finally have something to negotiate over rather than one side getting a beatdown.[8]

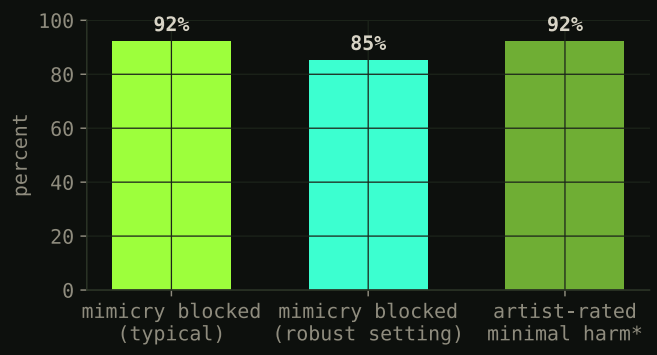
EXHIBIT A — adoption: cumulative downloads, cited milestones



READING: Figures reported by the UChicago Glaze/Nightshade project pages and contemporaneous reporting. Glaze passed 2.3M by March 2024 (VentureBeat), ~3.5M by September 2024 (MIT Tech Review), and the project site lists 8.5M+ since March 2023; Nightshade ~700k by Sep 2024, 1.6M by Nov 2024, and 2.5M+ per the project site since Jan 2024. Lines connect cited points only; nothing between them is interpolated as data.[5,6,7,9]

EXHIBIT B

glaze: reported efficacy and cost



Glaze on a single artwork: ~1.2 min on a Titan RTX GPU, ~7.3 min on one Intel i7 CPU (orig. pa

*efficacy figures from the Glaze USENIX Security 2023 reporting; perturbation tolerated in 1,0

READING: Efficacy and timing figures as reported for the original Glaze release. These are the creators' published numbers, measured against the models and conditions of the 2023 paper, not a guarantee against later or adaptive models.[6]

IV · WORKED CASES

04 Five examples, with receipts

Abstractions are cheap. Here are five concrete cases, each tied to a specific artifact, claim or measurement.

CASE 1 · HARMONYCLOAK

WHAT: "HarmonyCloak: Making Music Unlearnable for Generative AI," by Syed Irfan Ali Meerza and Jian Liu (UT Knoxville) with Lichao Sun (Lehigh), at the 46th IEEE Symposium on Security and Privacy, San Francisco, May 2025, proceedings pages 430 to 448.[1]

HOW: Imperceptible error-minimizing noise drives the generative loss toward zero on protected instrumental tracks, so the model concludes there is nothing to learn. The team focused on instrumental music given the scarcity of open-source vocal generators, and tested against models including MusicLM and SymphonyNet.[1,10]

EVIDENCE: In their validation, 31 human volunteers rated original and cloaked songs similarly high for pleasantness, supporting the claim that the perturbation is inaudible. Liu's stated motive was to extend protection past the voice, after Tennessee's 2024 ELVIS Act became the first US state law shielding musicians' voices from unauthorized AI use.[9,10]

CASE 2 · ADVERSARIAL IMAGES (THE ORIGIN)

WHAT: The general phenomenon adversarial audio inherits. Small, often imperceptible input perturbations that flip a model's prediction, first characterized for image classifiers around 2013 to 2014 and studied heavily since.[3]

WHY IT MATTERS HERE: Every audio tool in this report is a port of that finding. The JKU Linz instrument-classification work is explicit that it is carrying the image-domain discovery into raw audio waveforms.[3] The lineage is technical, not metaphorical.

CASE 3 · NIGHTSHADE

WHAT: "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," SAND Lab, accepted at IEEE S&P. Released to the public in January 2024 and built as the offensive complement to Glaze.[7]

HOW: It corrupts the concept-to-label binding a model relies on, so poisoned training samples teach it the wrong association. The team open-sourced it deliberately, on the logic that more independent variants in circulation are harder to filter out at scale.[7]

EVIDENCE: Over 2.5 million downloads since January 2024 per the project site, integrated into the art platform Cara. IBM's Nathalie Baracaldo, who studies data poisoning, called it effective in her experience working with poison.[5,9]

CASE 4 · GLAZE

WHAT: The defensive style cloak from the same lab, USENIX Security 2023, Distinguished Paper Award and Internet Defense Prize, later updated as Glaze 2 for newer models including SDXL.[6]

EVIDENCE: Over 8.5 million downloads since March 2023 per the project site, deployed across 150+ countries, with a user study run to more than 1,000 artists. The first image ever glazed was Karla Ortiz's *Musa Victoriosa*, now hanging in the lab.[5,6,9]

CASE 5 · DATA-POISONING RESEARCH (THE COUNTER-MOVE)

WHAT: The academic literature on defeating these defenses, which is as developed as the literature creating them.

EVIDENCE: "Detection and Defense of Unlearnable Examples" (2023) shows many unlearnable schemes are detectable by simple networks and degradable via stronger augmentation plus adversarial training, with certified bounds relating poison budget to adversarial budget.[11] "Potion: Towards Poison Unlearning" (Cambridge and the Alan Turing Institute, 2024) removes a poison trigger's effect from an already-trained model using only a subset of identified poison.[12] The cleanup crews publish too.

V · THE ARMS RACE

05 Friction, not a wall

The honest reading is the one the researchers themselves give. Poisoning is friction that can be overcome with enough compute and foreknowledge. If a scraper expects poisoning and budgets accordingly, much of the protection in the unlearnable-example schemes can be reduced, and the certified bounds in the 2023 detection work make that trade explicit.[11] When forum skeptics on the MetaFilter Poisonify thread said companies could train on poisoned data and then learn to ignore it, they were describing a published research program, not being cynical.[13]

SCRAPER_SIDE.PY · THE PUBLISHED COUNTER-MOVE

```
# What the defender is actually up against, from the detection literature.
def harden_against_poison(dataset):
    clean = []
    for x in dataset:
        if flags_as_unlearnable(x):  # arXiv:2312.08898
            continue                          # drop obvious poison
        x = strong_augment(x)                  # crop/EQ breaks fragile noi
        clean.append(x)
    model = adversarial_train(clean, budget=">= poison_budget")
    return potion_unlearn(model, known_poison_subset)  # arXiv:2406.09173
```

Each line maps to a real result: simple-network detection and augmentation defenses, the certified budget inequality, and Cambridge/Turing poison-unlearning. This is why the protection is friction, not a guarantee.[11,12]

> **These approaches do nothing for the billions of pieces of content already online. And when it gets hacked, and it will, creators will have already posted their work without protection.**
>
> PARAPHRASE OF HANY FARID, UC BERKELEY, ON THE LIMITS OF POISONING

And yet Farid's own concession is the load-bearing detail. As of his comment, no major image generator had publicly broken Nightshade, and Zhao's team said they knew of no reliable detector and would update the tool the moment one appeared.[14] That clause is the whole strategy. The point was never permanence. It is to raise the cost of nonconsensual extraction until licensing is cheaper than scraping. The arms race is the mechanism, not the failure.

VI · CROSS-ANALYSIS

06 The same gesture, several masks

Adversarial audio is structurally close to several things this collective already practices, and naming the overlap is more useful than another benchmark.

Scraping versus tape-trading. Underground music was always built on copying, but the cassette networks, the FTP and SoulSeek lineage, the Creative-Commons

netlabel flood, were reciprocal extraction inside communities of people who also made and shared. Industrial dataset scraping wears the same clothes and inverts the ethic: one-directional, and it returns nothing to the well it draws from. Most of the people building poison tools are veterans of copying culture. Their objection is to the severed reciprocity, which is an older grievance than AI.

Malware ethics, almost verbatim. The Church's own manifesto holds that malware is a tool not a monster, that the difference is intent, and that you have a right to understand the machines that rule your life.[15] Set that beside Zhao on putting teeth back into copyright in the wild and resisting the idea that whatever you post online becomes anyone's training fodder.[8] Same theology. A perturbation hidden under a masking curve is a payload hidden in a benign carrier.

Encryption as the closest cousin. Encryption makes content unreadable to everyone without a key. Adversarial audio makes content readable to humans and unreadable to models, a selective cipher keyed not on a secret but on the difference between two kinds of perception. As in the 1990s crypto-wars, the fight is only secondarily technical. It is mostly about who is permitted to render whom legible.

UNDERGROUND / HACKER PRACTICE	AI-ERA COUNTERPART	SHARED STRUCTURE
tape trading	dataset collection	copying, reciprocal vs extractive
CV Dazzle / Fawkes	Glaze / Nightshade	refuse the classifier
malware payload in carrier	perturbation under masking curve	intent over artifact
encryption	adversarial audio	selective legibility
copyleft / GPL	poisoning as deterrent	make taking carry a cost

VII · TIMELINE

07 The sediment, dated

2010

Adam Harvey's **CV Dazzle**. Adversarial makeup against face detection. The aesthetic of refusal, no gradients yet.

2013–14

Adversarial examples characterized for image recognition. The fragility of learned correlation gets a name.

2018

Psychoacoustic hiding against ASR, up to 98% success, inaudible to listeners. The masking trick crosses into audio.[4]

2020

JKU Linz: first **end-to-end waveform attacks** on music instrument classification.[3]

Mar 2023

Glaze ships; USENIX Security paper wins Distinguished Paper + Internet Defense Prize.[6]

Jan 2024

Nightshade released to the public, accepted at IEEE S&P.[7]

Jun 2024

Cambridge / Turing publish **Potion**, poison unlearning. The counter-move in print.[12]

May 2025

HarmonyCloak presented at IEEE S&P; **Poisonify** (Benn Jordan) and **Poison Pill** circulate in the music scene.[1,2]

VIII · STAKES

08 What its existence reveals

Strip the engineering and a poison pill is a confession about the present. It admits consent failed at the infrastructure level, that robots.txt, opt-out registries and platform policy did not hold, and that the remaining leverage an individual has over a trillion-parameter model is to make their own work hard to digest. When the last line of defense is to render your own art toxic, the negotiation already broke down upstream.

CONSENT_STACK.PY · WHY PEOPLE REACH FOR POISON

```
# The escalation ladder, in order of how much it asks of the artist.
defenses = [
    robots_txt(disallow="*"),          # ignored by many crawlers
    opt_out_registry(),               # scraper must honor it
    platform_policy(),                # changes when the ToS changes
    license_terms(),                  # only as strong as enforcement
    poison(perturb),                   # artist controls this one alone
]
for d in defenses:
    if d.depends_on_others_cooperating():
        continue                           # every layer above poison does
    return d                              # you end up here
```

A poison pill is the last item on this list because it is the only one that does not require the counterparty's good faith. That placement is the whole political argument, compressed.

It also reveals how badly the law has aged. Copyright was built for a world where copying was a visible, controllable act. Machine learning copies statistically: it keeps no file, only a smear of correlations distilled from millions of works, and the law has no clean category for that. Poisoning is partly a vernacular legal theory expressed in code, an attempt to manufacture at the technical layer the consequence the legal layer will not supply. The preservation community gets the strangest version of the question, because the archive that lets a person rediscover a forgotten netlabel compilation in 2031 is the same archive a scraper eats in an afternoon. Whether you can preserve a work for humans while keeping it indigestible to machines, and whether you should, is genuinely unsettled.

> **When the most rational move is to make your own art toxic, the question stops being whether poisoning is ethical. The question is what kind of system made it the sanest option left.**
>
> FIELDNOTE, CLOSING

I keep coming back to that first clean-sounding track. It was an ordinary recording and a trap at the same time, and which one it was depended entirely on who, or what, was listening. That double state is the real artifact. Adversarial audio is finally a working model of a world where every human expression now has two readings, one for us and one for the systems trained on us, and where artists have quietly started writing for the gap between them.

SOURCES & NOTES

Citations

1. Meerza, Sun & Liu, "HarmonyCloak: Making Music Unlearnable for Generative AI," 2025 IEEE Symposium on Security and Privacy (SP), pp. 430–448. Author PDF: mosis.eecs.utk.edu. Uses imperceptible error-minimizing noise to drive generative loss toward zero. Focus on instrumental music.
2. B. Bowler (Poison Pill), in Music Ally, "Poison Pill aims to protect music from unlicensed AI training," 27 Oct 2025. On models finding genre/instrument "shortcuts" and the defender not knowing the exact features.
3. Prinz & Flexer, "End-to-End Adversarial White Box Attacks on Music Instrument Classification," JKU Linz, arXiv:2007.14714 (2020).
4. Schönherr, Kolossa et al., "Adversarial Attacks Against ASR via Psychoacoustic Hiding," arXiv:1808.05665 (2018). Up to 98% success; targets inaudible in user studies.
5. "The AI lab waging a guerrilla war over exploitative AI," MIT Technology Review, 13 Nov 2024. SAND Lab origin (Concept Art Association Zoom, Nov 2022); Nightshade 1.6M downloads; Baracaldo quote; Cara integration.
6. Shan, Cryan, Wenger, Zheng, Hanocka & Zhao, "Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models," 32nd USENIX Security Symposium (2023), pp. 2187–2204. Distinguished Paper + Internet Defense Prize. 92%/85% efficacy and ~1.2 min/Titan RTX via the paper and VentureBeat's Glaze 2 coverage.
7. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," SAND Lab; accepted IEEE S&P. VentureBeat and MIT Tech Review release coverage, Oct 2023–Jan 2024.
8. UChicago News / Big Brains podcast, 30 Jan 2025, and "The poison well," The Muse (Substack), 12 Apr 2024. Zhao on "teeth" for copyright and on friction.
9. UChicago project pages glaze.cs.uchicago.edu and nightshade.cs.uchicago.edu (Glaze 8.5M+ since Mar 2023; Nightshade 2.5M+ since Jan 2024); MIT Tech Review "Innovator of the Year" (Glaze ~3.5M, Nightshade 700k, Sep 2024); EurekAlert! 22 Oct 2024 (HarmonyCloak 31-volunteer study; Tennessee voice-protection law).
10. HarmonyCloak author PDF, mosis.eecs.utk.edu, and New Atlas coverage, 24 Oct 2024: tested against MusicLM and SymphonyNet.
11. "Detection and Defense of Unlearnable Examples," arXiv:2312.08898 (2023). Simple-network detectability; augmentation + adversarial training defense; certified poison/adversarial budget bounds.
12. Schoepf, Foster & Brintrup, "Potion: Towards Poison Unlearning," arXiv:2406.09173 (2024), Cambridge & Alan Turing Institute.
13. MetaFilter post 208420 (14 Apr 2025) and related threads on Benn Jordan's Poisonify; community read as steganographic, plus the arms-race objection.

14. Fox News tech, "Poison pill tool could break AI systems...": Hany Farid (UC Berkeley) on limits; Zhao's statement that no major generator had broken Nightshade and no reliable detector was known.

15. churchofmalware.org manifesto and mission text ("malware is a tool, not a monster"; free, paywall-less offensive-security education).

16. Benn Jordan, "The Art of Poison-Pilling Music Files" (YouTube, Apr 2025); kill switch podcast (iHeart/Kaleidoscope, Jul 2025) on starting from Nightshade.

17. Adam Harvey, CV Dazzle (2010), and Shan et al., "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models," USENIX Security 2020. Cited as the refusal lineage.

18. "Glaze 2," VentureBeat: faster glazing, protection extended to newer models including Stable Diffusion XL.

Exhibits A and B plot only figures explicitly reported in the cited sources. No values were invented or interpolated as data; connecting lines join cited points for legibility only.

☟ MALWARE BLESS · HACK THE PLANET · HELP THE HELPLESS ☟

CHURCH OF MALWARE FIELD REPORT · DISTRIBUTE FREELY · NO PAYWALL · NO TELEMETRY