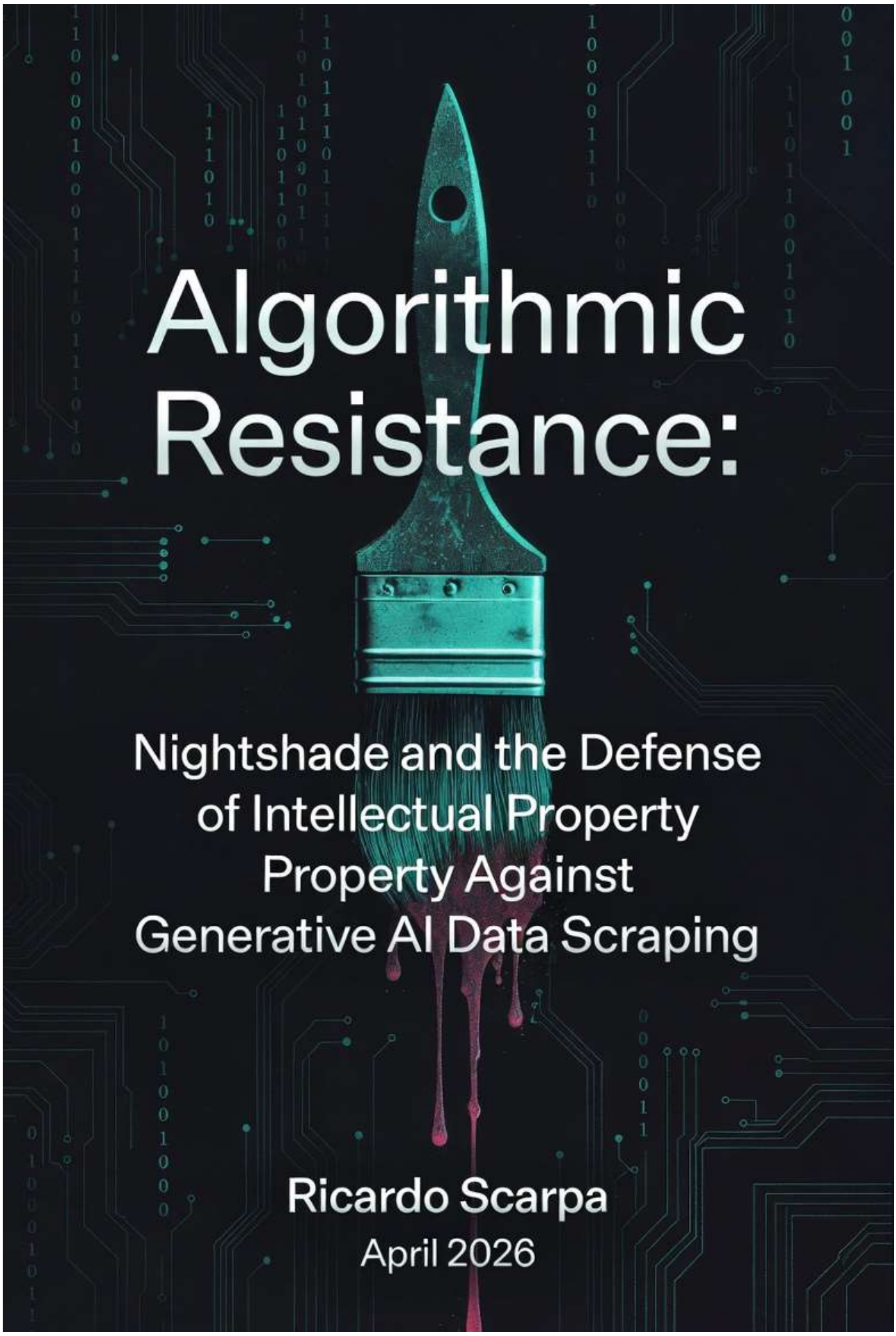
Algorithmic Resistance:
Nightshade and the Defense of Intellectual Property Property Against Generative AI Data Scraping
Ricardo Scarpa
April 2026

Algorithmic Resistance: Nightshade and the Defense of Intellectual Property Against Generative AI Data Scraping

Detailed Table of Contents

Algorithmic Resistance: Nightshade and the Defense of Intellectual Property Against Generative AI Data Scraping

1. Introduction

The emergence of massive diffusion models such as Stable Diffusion, Midjourney, and DALL-E has reshaped the concept of authorship in the digital age. These systems, capable of generating complex images from simple text prompts, rely on the indiscriminate scraping of billions of copyrighted works, often without the consent, credit, or compensation of their original creators[1]. This practice has generated an existential crisis in the artistic community, where algorithmic mimicry threatens not only the economic viability of creative careers but also the very notion of human originality[2]. In this scenario of profound power asymmetry between technology corporations and individual artists, the need arises for technical defense mechanisms that transcend ineffective voluntary exclusion or opt-out protocols[3].

Under the leadership of Professor Ben Zhao and researcher Shawn Shan at the SAND Laboratory of the University of Chicago, disruptive tools aimed at restoring creators' agency have been developed[4]. While Glaze was born in 2023 as a defensive "shield" to protect artistic style against algorithmic imitation, Nightshade represents a strategic shift toward offensive defense[5]. Formally presented at the 2024 IEEE Symposium on Security and Privacy, Nightshade uses advanced data poisoning techniques to corrupt training models that ignore intellectual property protection directives[6].

The central thesis of this study is that Nightshade does not merely constitute an act of technical sabotage, but rather an essential algorithmic resistance mechanism to balance the power dynamics in AI training[7]. By exploiting the concept sparsity inherent in diffusion models, Nightshade allows a small number of poisoned samples to significantly degrade a model's accuracy[8]. In this way, the tool acts as a coercive incentive for AI developers to opt for fair licensing as the only viable alternative to the possible implosion of their systems[9]. This article will examine the technical, ethical, and legal dimensions of this digital "sword" in the struggle for the integrity of human work against corporate data extractivism.

Notes

[1] Shawn Shan, et al., "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," *Proceedings of the 45th IEEE Symposium on Security and Privacy* (2024): 1-2.

[2] Lucas Mearian, "'Data poisoning' anti-AI theft tools emerge — but are they ethical?," *Computerworld*, October 30, 2023, para. 3.

[3] Arts Law Centre of Australia, "Glaze and Nightshade: How artists are taking arms against AI scraping," December 23, 2024, para. 2.

[4] The Glaze Project, "About The Glaze Project," University of Chicago, accessed April 29, 2026, https://glaze.cs.uchicago.edu/aboutus.html.

[5] Shiloh Miller, "Poisoning the machine," *The University of Chicago Magazine*, May 7, 2025, para.

4.
[6] Shan, "Nightshade," 1-2.
[7] University of Chicago, "What Is Nightshade?: Protecting Copyright," accessed April 29, 2026, https://nightshade.cs.uchicago.edu/whatis.html.
[8] Shan, "Nightshade," 14.
[9] Daily.dev, "Protecting Artists: Glaze and Nightshade in the Fight Against Exploitative AI," November 13, 2024, para. 5.

---

2. The Data Scraping Ecosystem and Power Asymmetry

2.1. LAION, Common Crawl, and the mass collection of works without consent

The technical infrastructure that supports contemporary generative AI depends on mass data collection through web scraping. State-of-the-art models such as Stable Diffusion are fed by datasets provided by LAION (Large-Scale Artificial Intelligence Open Network), a nonprofit organization partially funded by Stability AI[10]. The LAION-5B dataset, for example, contains approximately 5.85 billion image-text pairs[11]. These entries are originally collected by Common Crawl, an entity that seeks to index a free copy of the web for research and analysis purposes[12]. However, this process indiscriminately captures protected works from platforms such as Pinterest, DeviantArt, WordPress, and stock photography sites like Getty Images, often without the original authors' knowledge of the use of their creations[13].

2.2. Ineffectiveness of opt-out mechanisms and robots.txt directives

Faced with this extractivism, traditional protection measures have proven to be mere formalities without enforcement capability. Opt-out mechanisms and directives in robots.txt files are characterized as "voluntary" tools, whose observance is entirely at the discretion of model developers[14]. Although companies like OpenAI have suggested using these tags to block crawlers such as GPTBot, their effectiveness is null if the artist does not control the hosting server or if the crawler decides to ignore the directive, since there are no reliable ways to verify compliance[15]. Likewise, the use of HTML tags such as "noai," implemented by platforms like DeviantArt and ArtStation following user protests, depends on the "good conduct" of crawlers[16]. In practice, these tags are often disregarded without immediate legal or technical repercussions[17].

2.3. The vulnerability of the individual artist against large technology corporations

This digital architecture underscores a fundamental power asymmetry in which individual artists find themselves at a structural disadvantage against technology corporations[18]. The ability of models to perform stylistic mimicry through fine-tuning allows a user to replicate a creator's unique style using only a handful of their works, directly threatening the economic viability of artistic careers[19]. Moreover, the speed with which AI floods digital markets displaces the visibility of human creators, whose work is buried under incessant algorithmic production[20]. In this digital "Wild West" environment, artists lack the financial means to litigate against technology giants, making the adoption of technical resistance tools like Nightshade imperative[21].

Notes

[10] "Navigating Uncharted Seas: A Deep Dive into Authorship and Fair Use," *Virginia Journal of Law & Technology* 28, no. 2 (2024): 9.
[11] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 9.
[12] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 9.
[13] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 10.
[14] Shan, "Nightshade," 14.
[15] Shan, "Nightshade," 14.
[16] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 16-17.
[17] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 17.
[18] Shan, "Nightshade," 14.
[19] Robert Hönig, et al., "Adversarial Perturbations Cannot Reliably Protect Artists from Generative AI," *ICLR Proceedings* (2024): 2.
[20] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 44-45.
[21] Arts Law Centre of Australia, "Glaze and Nightshade," para. 5.

---

3. Technical Foundations of Nightshade: Data Poisoning

3.1. Architecture of prompt-specific poisoning attacks

Nightshade differs from traditional poisoning attacks by its focus on specific targets called prompt-specific poisoning attacks[22]. While conventional attacks seek to degrade the overall performance of a classifier by injecting a critical mass of data (often 20% of the training set), Nightshade exploits the way diffusion models link linguistic concepts with visual features[23]. The goal is to corrupt a specific concept "C" (e.g., "dog") so that the model generates images of a target concept "A" (e.g., "cat")[24]. To achieve this, the system generates perturbations in natural images of C that, although invisible to the human eye, shift the image's representation in the visual extractor's feature space towards concept A[25].

3.2. The concept of concept sparsity in diffusion models

The technical viability of Nightshade rests on the finding of concept sparsity in large-scale datasets[26]. Although a model like Stable Diffusion is trained with billions of images, the data density for individual concepts is surprisingly low. After analyzing the LAION-Aesthetic dataset, SAND Lab researchers determined that more than 92% of concepts (unique nouns) appear in less than 0.04% of samples[27]. For example, common terms like "dog" represent only 0.1% of the total, while styles like "fantasy" fall to 0.04%[28]. This semantic dispersion means an attacker does not need millions of poisoned samples; injecting barely one hundred optimized images can be sufficient to counter the influence of clean samples and force the model to adopt the incorrect association[29].

3.3. Optimizing poison potency: LPIPS and L∞ perturbations

To maximize the impact of each poisoned sample, Nightshade employs multi-objective adversarial optimization techniques. The process uses "anchor images" that represent the ideal version of the target concept A to guide the perturbation of the original image of C[30]. Mathematically, the attack seeks to minimize the distance in feature space between the poisoned image and the anchor,

subject to a strict perturbation budget[31]. To ensure the changes are imperceptible to artists and human curators, Nightshade uses the LPIPS metric (*Learned Perceptual Image Patch Similarity*), which emulates human visual perception through deep neural networks[32]. Although the open-source repository also supports the infinity norm (L∞) metric to ensure that no individual pixel changes drastically, the use of LPIPS allows more potent perturbations that are robust against compression or rescaling processes of images on the web[33].

Notes

[22] Shan, "Nightshade," 3.
[23] Shan, "Nightshade," 3.
[24] Arts Law Centre of Australia, "Glaze and Nightshade," para. 6.
[25] Shan, "Nightshade," 16.
[26] Shan, "Nightshade," 14-15.
[27] Shan, "Nightshade," 15.
[28] Shan, "Nightshade," 15.
[29] University of Chicago, "What Is Nightshade?: Why Does It Work, and Limitations," accessed April 29, 2026, https://nightshade.cs.uchicago.edu/whatis.html.
[30] Shan, "Nightshade," 17.
[31] Shan, "Nightshade," 16-17.
[32] Richard Zhang, et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *Proceedings of CVPR* (2018), cited in Shan, "Nightshade," 17.
[33] Shan, "Nightshade," 18.

---

4. The Duality of the SAND Lab: The Shield (Glaze) and the Sword (Nightshade)

4.1. Glaze and protection against artistic style mimicry

The SAND Lab inaugurated its technical offensive in defense of creators with the launch of Glaze in 2023[34]. This tool was conceived as a defensive "shield" specifically designed to combat style mimicry, a practice in which AI models are fine-tuned with a small number of works (generally between ten and twenty) to replicate an artist's unique visual identity[35]. Glaze operates by exploiting "blind spots" or adversarial examples in AI feature extractors: by applying minimal pixel-level perturbations, the software tricks the model into believing that the work belongs to a radically different style (e.g., interpreting a realistic portrait as a Cubist composition)[36]. Thus, when a user tries to generate an image "in the style of" a protected artist, the system produces distorted results that bear no relation to the original aesthetic, thereby preserving the integrity of the creator's personal brand[37].

4.2. Nightshade as an offensive tool: altering semantic content

Unlike Glaze, which acts on the stylistic surface to protect the individual, Nightshade was designed as a "sword" or offensive tool intended to corrupt the semantic functionality of base models[38]. While Glaze assumes that the harm (data scraping) has already occurred and seeks to mitigate

local fine-tuning, Nightshade aims to poison the general training process for all users of the model[39]. Its mechanism of action does not alter the perception of style, but rather the AI's understanding of objects and concepts[40]. By injecting perturbations that associate, for example, the label "dog" with the visual features of a "cat," Nightshade degrades the model's ability to generate accurate representations of reality[41]. This distinction is fundamental: Glaze protects the artist's "voice," while Nightshade sabotages the visual "dictionary" of AI that ignores copyright[42].

4.3. Synergy between tools: the combined use of defensive and offensive protections

The SAND Lab team, led by Ben Zhao, emphasizes that Glaze and Nightshade are not mutually exclusive but rather components of an integrated resistance ecosystem[43]. Since Nightshade does not offer protection against style mimicry, an artist who uses only the latter remains vulnerable to having their aesthetic cloned through fine-tuning[44]. Therefore, the official technical recommendation is to use both tools together before publishing works on open platforms[45]. The suggested workflow dictates that the artist should first apply Nightshade to the original image and then process the resulting file through Glaze[46]. Although this double process may increase the presence of visible visual artifacts, it is considered the most robust mechanism to restore the creator's agency, acting simultaneously as a measure of personal protection and an act of collective disincentive against data extractivism[47].

Notes

[34] Shawn Shan, et al., "Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models," *USENIX Security* (2023), cited in *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 19.
[35] Shan, "Nightshade," 4.
[36] Miller, "Poisoning the machine," para. 6.
[37] Arts Law Centre of Australia, "Glaze and Nightshade," para. 11.
[38] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 19.
[39] Shan, "Nightshade," 4.
[40] Daily.dev, "Protecting Artists," para. 3.
[41] University of Chicago, "What Is Nightshade?: Protecting Copyright."
[42] Reddit, r/antiai, "Does nightshade still work in 2026," comment by Hada_de_Sillon, 2 months ago.
[43] The Glaze Project, "About The Glaze Project."
[44] University of Chicago, "What Is Nightshade?: Nightshade and WebGlaze."
[45] University of Chicago, "Nightshade Software User Guide," last updated January 18, 2024, https://nightshade.cs.uchicago.edu/userguide.html.
[46] University of Chicago, "User Guide," para. 3.
[47] Arts Law Centre of Australia, "Glaze and Nightshade," para. 13.

---

5. Impact Assessment on State-of-the-Art Diffusion Models

5.1. Experimentation with Stable Diffusion (SD-V2, SD-XL) and DeepFloyd

The effectiveness of Nightshade has been validated through rigorous testing on the most advanced open-source models, including Stable Diffusion V2, Stable Diffusion XL (SD-XL), and DeepFloyd[48]. Although these systems have different architectures and have been trained on massive datasets (some with more than 2.6 billion parameters), all exhibit a critical vulnerability to targeted poisoning[49]. Experiments show that, in the case of SD-XL, specific concepts can be corrupted with fewer than 100 poisoned samples, causing the model to ignore the original prompts and generate the target concept A (e.g., producing images of "bags" when a "hat" is requested)[50]. Even though pre-trained models already possess consolidated knowledge of concepts, injecting only 2% poisoned data relative to the semantic volume of the concept is enough to neutralize the influence of thousands of clean samples[51].

5.2. The bleed-through effect: propagation of poisoning to related concepts

One of the most disruptive features of Nightshade is the bleed-through effect, which prevents poisoning from being circumvented through the use of synonyms or prompt variations[52]. When an artist poisons the concept "dog," the damage is not limited to that exact word; it propagates in the text embedding space to semantically close terms such as "puppy," "husky," or "wolf"[53]. For example, attacks targeting the style "fantasy" have been shown to affect the generation of "dragons" or even works associated with the name of artist Michael Whelan, even though those words were not mentioned in the original attack data[54]. This propagation capability ensures that poisoning is robust and difficult to filter through simple word blacklists or superficial relabeling[55].

5.3. Model stability and implosion: how many poisons are needed to render an AI useless?

The impact of Nightshade goes beyond corrupting individual concepts, posing a systemic threat called model collapse[56]. SAND Lab researchers discovered that attacks are composable: multiple independent poisons can coexist in the same model without canceling each other out[57]. However, when a critical threshold of poisoned concepts is reached, the model's internal structure begins to degrade irreversibly[58]. It has been observed that after poisoning approximately 250 independent concepts, the AI's ability to generate coherent images drops dramatically, reaching quality levels inferior to models from a decade ago[59]. If the number of attacked concepts reaches 500, the system totally implodes, producing only visual noise or random pixels in response to any prompt, rendering the model useless for any commercial use[60].

Notes

[48] Shan, "Nightshade," 10.
[49] Shan, "Nightshade," 26.
[50] Shan, "Nightshade," 2.
[51] Shan, "Nightshade," 21-22.
[52] Miller, "Poisoning the machine," para. 9.
[53] Shan, "Nightshade," 22-23.

[54] Shan, "Nightshade," 24.
[55] University of Chicago, "What Is Nightshade?: Why Does It Work, and Limitations."
[56] Miller, "Poisoning the machine," para. 10.
[57] Shan, "Nightshade," 25.
[58] AWS, "Chapter 3: Challenges and Risks of Generative AI," *Generative AI Report* (2024): 86.
[59] Shan, "Nightshade," 25-26.
[60] Miller, "Poisoning the machine," para. 10.

---

6. The Ethical Debate on Data Poisoning

6.1. Is it ethical to corrupt training models? Conflicting perspectives

The deployment of Nightshade has sparked intense ethical debate in the field of artificial intelligence, focusing on whether it is lawful to sabotage technological systems to protect individual rights[61]. Braden Hancock, CTO of Snorkel AI, argues that the ethics of these tools depends intrinsically on their objective: while poisoning data for critical safety systems, such as autonomous vehicle signaling, is unequivocally unethical, using "poisons" to enforce a "do not scrape" directive against corporate extractivism represents a legitimate defensive frontier[62]. For their part, analysts like Ritu Jyoti of IDC argue that ethical responsibility falls on the entities that collect data; if a work has been protected or masked by its author and is taken without permission, the technical consequences for the AI model are a risk that the infringer voluntarily assumes[63].

6.2. The distinction between malicious sabotage and intellectual property protection

Unlike conventional cyberattacks carried out by malicious actors for illicit gain, Nightshade is presented as a "last defense" measure for content creators[64]. The SAND Lab emphasizes that the tool does not seek to destroy AI technology *per se*, but to introduce an incremental cost to the use of unlicensed data[65]. Ben Zhao, described in academic circles as a "vigilante cowboy" in this digital environment, defends that Nightshade is a proportional response to the systematic violation of robots.txt protocols and opt-out directives[66]. Under this view, poisoning does not act as gratuitous aggression, but as a property safeguard that is only activated when a prior act of data misappropriation occurs[67].

6.3. Nightshade as an incentive for negotiating fair licenses

From a pragmatic perspective, Nightshade aims to reconfigure the training data market through technical deterrence[68]. By significantly increasing the risk of training models with indiscriminately scraped data —due to the latent threat of semantic degradation or system implosion— the tool seeks to make direct licensing the most economical and safe alternative for technology companies[69]. The declared goal of the researchers is to force a transition from an ecosystem of unilateral exploitation to one of "licensed acquisition," where AI developers are compelled to negotiate fair compensation terms with original creators[70]. Thus, Nightshade is not limited to being a resistance tool, but acts as a catalyst for a future framework of ethical cooperation and respect for the dignity of human work[71].

Notes

[61] Mearian, "'Data poisoning'," para. 9.
[62] Mearian, "'Data poisoning'," para. 9.
[63] Mearian, "'Data poisoning'," para. 10.
[64] Shan, "Nightshade," 1.
[65] University of Chicago, "What Is Nightshade?: Protecting Copyright."
[66] Miller, "Poisoning the machine," para. 3.
[67] University of Chicago, "What Is Nightshade?: Protecting Copyright."
[68] Shan, "Nightshade," 28.
[69] Daily.dev, "Protecting Artists," para. 5.
[70] Shan, "Nightshade," 28.
[71] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 49.

---

7. The Legal Context: Class Actions and Expert Disputes

7.1. The case of *Andersen v. Stability AI*: the role of artists like Karla Ortiz and Sarah Andersen

The most impactful litigation in the generative AI ecosystem is the class action *Sarah Andersen et al. v. Stability AI Ltd. et al.*, filed in January 2023[72]. Led by illustrator Sarah Andersen and artists Kelly McKernan and Karla Ortiz, the complaint alleges that Stable Diffusion was trained using unauthorized copies of billions of copyrighted images, functioning as a "21st-century collage tool" that directly competes with original works[73]. In August 2024, District Judge William Orrick issued a significant partial ruling by allowing the direct copyright infringement claims to proceed, validating the plaintiffs' theory that diffusion models might contain "compressed copies" of training works[74]. For artists like Ortiz, the case does not seek to eliminate AI, but to establish a fair use framework that requires consent and compensation, preventing human creators from being forced to compete against models trained on their own work[75].

7.2. The controversy over Ben Zhao and Emily Wenger's expert testimony in court

An unusual dimension of the legal battle is the dispute over Professor Ben Zhao's role as a court expert[76]. In the context of the *Andersen* case, a conflict arose when the plaintiffs proposed Zhao to examine the confidential source code of the AI companies. The defendants strongly objected, arguing that Zhao is not a neutral observer but a technical "adversary" who has developed data poisoning tools explicitly designed to sabotage their systems[77]. Faced with this block, the court suggested Dr. Emily Wenger, Zhao's former student, as an alternative; however, the technology companies conditioned their acceptance on Wenger suspending all academic research for three years, a demand that the plaintiffs called unacceptable and punitive for a career academic[78]. This dispute underscores how Nightshade has transformed its creators into central figures of legal and technical resistance.

7.3. Industry stance: OpenAI's criticism of the use of protection tools

The response of large technology corporations to tools like Glaze and Nightshade has ranged from silence to open condemnation[79]. OpenAI, in particular, has gone so far as to characterize artists' use of these programs as a form of "abuse" of its systems, a stance that intellectual property advocates consider ironic given the origin of the training of its models[80]. Meanwhile, the industry has tried to mitigate legal risk through indemnification policies, where companies like Microsoft and Google promise to cover the legal costs of their enterprise customers against copyright lawsuits[81]. Nevertheless, from the SAND Lab's perspective, these measures are insufficient because they do not address the root of the problem: indiscriminate data scraping. The current legal debate suggests that, in the absence of clear federal regulation in the U.S., technical self-defense tools like Nightshade will continue to be the main resource for artists to impose a real cost on corporate data extractivism[82].

Notes

[72] McKool Smith, "AI Infringement Case Updates: June 23, 2025" (2025), section 10.
[73] Jamie Lang, "Class-Action Lawsuit Filed Against Stability AI, Midjourney, DeviantArt," *Cartoon Brew*, January 17, 2023, para. 2.
[74] Richard Whiddington, "Artists Land a Win in Class Action Lawsuit Against A.I. Companies," *Artnet News*, August 15, 2024, para. 3-5.
[75] Nathan Seth Lowell, "AI Create: The Brave New World and Copyright Implications of AI-Generated Artwork," *Virginia Journal of Law & Technology* 28, no. 2 (2024): 3.
[76] McKool Smith, "AI Infringement Case Updates," section 10.
[77] McKool Smith, "AI Infringement Case Updates," section 10.
[78] McKool Smith, "AI Infringement Case Updates," section 10.
[79] Daily.dev, "Protecting Artists," para. 4.
[80] Daily.dev, "Protecting Artists," para. 4.
[81] AWS, "Chapter 3," 92.
[82] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.

---

8. The Algorithmic Arms Race: Vulnerabilities and Countermeasures

8.1. LightShed: the "detoxification" attack from the University of Cambridge

The effectiveness of Nightshade has prompted an immediate response from the cybersecurity research community, inaugurating an "arms race" between creators and AI developers[83]. In this context, an international team led by the University of Cambridge introduced **LightShed**, a system specifically designed to identify and neutralize perturbation-based protections[84]. LightShed operates through a three-stage process: first, it detects whether an image has been altered using poisoning techniques; second, it employs reverse engineering to model the characteristics of the perturbation using an autoencoder; and finally, it removes the "poison" by subtracting the identified pattern[85]. In experimental tests, LightShed successfully detected images protected by

Nightshade with 99.98% accuracy, restoring the utility of the data for training without visually degrading the quality of the works[86].

8.2. Limitations of adversarial perturbations against advanced deep learning

Beyond LightShed, other academic studies have questioned the long-term resilience of adversarial perturbations[87]. Researchers from ETH Zurich and Google DeepMind have argued that tools like Glaze and Nightshade provide a "false sense of security," as they can be circumvented using relatively simple purification techniques[88]. Among these techniques is noisy upscaling, which combines the addition of Gaussian noise with super-resolution models to "clean" adversarial artifacts[89]. According to these experts, current protections suffer from the "first-mover" structural disadvantage: once an artist publishes a protected work, the attacker has the benefit of offline adaptation, being able to test multiple detoxification methods until the defense is broken[90]. Likewise, it has been shown that diffusion-based purification attacks can restore model accuracy from 23% to 94% using only a small set of unprotected images as a reference[91].

8.3. Proposals for more robust and resilient protections

Despite these vulnerabilities, researchers emphasize that the discovery of weaknesses is an opportunity for the "co-evolution" of defenses[92]. To confront attacks like LightShed, the SAND Lab and other academics propose strategies aimed at increasing the technical robustness of protections[93]. Recommendations include the development of image-specific perturbations, which would make it difficult for an attacker to learn a "master pattern" of poisoning through an autoencoder[94]. Likewise, varying the density of the perturbation in different regions of the work and structurally aligning the "poison" with natural Gaussian noise is suggested, so that any cleaning attempt severely degrades the visual integrity of the image[95]. Nevertheless, there is a growing consensus that technical resistance must be complemented by a robust legal framework that discourages data scraping, transforming these tools from definitive solutions into mechanisms of necessary friction[96].

Notes

[83] Hanna Foerster, et al., "LightShed: Defeating Perturbation-based Image Copyright Protections," *USENIX Security* (2025): 1-2.
[84] Foerster, "LightShed," 12.
[85] University of Cambridge, "AI art protection tools still leave creators at risk, researchers say," 2025, para. 4-6.
[86] Foerster, "LightShed," 14.
[87] Hönig, "Adversarial Perturbations," 2.
[88] Hönig, "Adversarial Perturbations," 2.
[89] Hönig, "Adversarial Perturbations," 10-11.
[90] Hönig, "Adversarial Perturbations," 28.
[91] Vector Institute for Artificial Intelligence, "When smart AI gets too smart: Key insights from Vector's 2025 ML Security & Privacy Workshop," 2025, para. 3.
[92] University of Cambridge, "AI art protection tools," para. 10.
[93] Foerster, "LightShed," 21.

[94] Foerster, "LightShed," 21.
[95] Foerster, "LightShed," 21-22.
[96] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.

---

9. Legislative Perspectives and the Role of the Copyright Office

9.1. The U.S. Copyright Office report on AI and authorship

The United States Copyright Office (USCO) has adopted a proactive but restrictive stance regarding the advance of generative AI, structuring its intervention in a three-part report[97]. In the second installment, published in January 2025, the USCO reaffirmed that human creativity is the "bedrock" of copyright, concluding that works generated solely by text prompts are not eligible for protection[98]. For the Office, the act of providing a prompt is more akin to a client commissioning a work from an artist than to an author controlling the expressive process[99]. Nevertheless, the USCO admits protection of specific elements if a human makes creative arrangements or substantial modifications to the algorithmic output[100]. The most critical debate is reserved for the third part of the report, still under development, which will address the legal implications of training models on copyrighted works, an area where tools like Nightshade seek to impose a framework of "compulsory licensing" through technical resistance[101].

9.2. Proposals for opt-in laws and ethical model certification

Faced with the ineffectiveness of opt-out systems, legislative proposals have emerged seeking to reverse the burden of consent. In April 2024, the **Generative AI Copyright Disclosure Act** was introduced in the U.S. Congress, which would require developers to submit detailed summaries of all copyrighted works used in their training datasets[102]. Legal experts suggest that this framework should be complemented with a mandatory opt-in system, where use of data is illegal unless there is express authorization from the creator[103]. In this context, organizations like *Fairly Trained* have begun certifying models that are trained exclusively on public domain or licensed data, offering an ethical alternative to data extractivism[104]. Such certifications could act as a safe harbor, incentivizing companies to avoid the risks of semantic degradation associated with indiscriminate scraping and Nightshade poisoning[105].

9.3. Toward an international regulatory framework for model training

Globally, legislative fragmentation poses significant challenges to intellectual property protection. While the **European Union's AI Act** proposes strict transparency rules requiring developers to disclose copyright-protected materials, other countries like Japan have adopted more lenient stances, suggesting that model training does not constitute an infringement *per se*[106]. This disparity has raised concern about a "global AI gap," where corporations could migrate their training processes to jurisdictions with minimal protections[107]. However, the massive deployment of Nightshade on global platforms like Cara or ArtStation introduces a form of "cross-border technical regulation": because the poison is embedded in the work itself, the protection travels with the data regardless of the jurisdiction where it is scraped[108]. In this way, algorithmic resistance could force the creation of a de facto international standard based on respect for creator autonomy and fair compensation[109].

Notes

[97] Copyright Office, "Copyright Office Releases Part 2 of Artificial Intelligence Report," Library of Congress, 2025, para. 1-2.
[98] Copyright Office, "Part 2," para. 3.
[99] Lowell, "AI Create," 20-21.
[100] Copyright Office, "Part 2," para. 4.
[101] Copyright Office, "Part 2," para. 6.
[102] Lowell, "AI Create," 46.
[103] Lowell, "AI Create," 48.
[104] Lowell, "AI Create," 47.
[105] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.
[106] Mearian, "'Data poisoning'," para. 8.
[107] AWS, "Chapter 3," 110.
[108] Lowell, "AI Create," 49.
[109] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.

---

10. Conclusions: Toward an Ethical Ecosystem of Coexistence

10.1. The future of digital artistic resistance

The deployment of Nightshade has marked a milestone in the history of digital humanities, transforming artistic resistance from a merely symbolic protest into an active technical defense[110]. Although the software faces the constant challenge of an algorithmic arms race —evidenced by the emergence of detoxification tools like LightShed— its existence has permanently altered the risk calculus for AI developers[111]. The update to version 1.1 in April 2026 demonstrates the SAND Lab's commitment to the continuous evolution of these defenses, ensuring that the "poison" remains a necessary friction against indiscriminate scraping[112]. In this sense, the future of digital resistance lies not in the absolute invulnerability of a tool, but in the ability of the creative community to organize and use these technologies as mechanisms of collective pressure[113].

10.2. Balancing technological advancement with respect for the creator's dignity

The tension between AI innovation and copyright should not be resolved through the eradication of technology, but through a new digital social contract[114]. Nightshade acts as the technical catalyst for this change, incentivizing the transition from a "scraping by default" model to one of "licensing by consent"[115]. As the Glaze Project team argues, the ultimate goal is to restore the balance of power, ensuring that the advancement of diffusion models does not come at the expense of the economic viability and dignity of human artists[116]. Healthy coexistence between generative AI and human creativity will only be possible when developers recognize that the value of their systems intrinsically depends on the human work that feeds them —work that deserves to be credited, compensated, and above all respected[117].

10.3. Final reflections on human autonomy in the algorithmic age

Ultimately, Nightshade is an affirmation of human autonomy against corporate automation[118]. By allowing creators to decide whether their works may be integrated into the "memory" of a machine, these tools restore agency to those who have historically been marginalized by data extractivism[119]. The debate over data poisoning transcends cybersecurity to sit at the heart of digital ethics: the struggle to preserve the singularity of human experience and its aesthetic expression[120]. In a world where algorithmic mimicry threatens to saturate cultural space, technical resistance becomes an act of preserving creative diversity, ensuring that the voice of the human artist remains the fundamental foundation of future culture[121].

Notes

[110] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.
[111] Shan, "Nightshade," 28.
[112] University of Chicago, "What Is Nightshade?: Protecting Copyright."
[113] Daily.dev, "Protecting Artists," para. 5.
[114] Lowell, "AI Create," 49-50.
[115] Daily.dev, "Protecting Artists," para. 5.
[116] The Glaze Project, "About The Glaze Project."
[117] Lowell, "AI Create," 49.
[118] University of Chicago, "What Is Nightshade?: Protecting Copyright."
[119] The Glaze Project, "About The Glaze Project."
[120] Arts Law Centre of Australia, "Glaze and Nightshade," para. 14.
[121] Lowell, "AI Create," 50.

Bibliography

Arts Law Centre of Australia. "Glaze and Nightshade: How artists are taking arms against AI scraping." December 23, 2024. https://www.artslaw.com.au/news/glaze-and-nightshade-how-artists-are-taking-arms-against-ai-scraping/.

AWS. "Chapter 3: Challenges and Risks of Generative AI." *Generative AI Report*, 2024.

Copyright Office. "Copyright Office Releases Part 2 of Artificial Intelligence Report." Library of Congress. January 29, 2025. https://newsroom.loc.gov/news/copyright-office-releases-part-2-of-artificial-intelligence-report/s/f3959c36-d616-498d-b8f9-67641fd18bab.

Daily.dev. "Protecting Artists: Glaze and Nightshade in the Fight Against Exploitative AI." November 13, 2024. https://daily.dev/posts/protecting-artists-glaze-and-nightshade-in-the-fight-against-exploitative-ai.

Foerster, Hanna, et al. "LightShed: Defeating Perturbation-based Image Copyright Protections." *USENIX Security*, 2025.

Hönig, Robert, et al. "Adversarial Perturbations Cannot Reliably Protect Artists from Generative AI." *ICLR Proceedings*, 2024.

Lang, Jamie. "Class-Action Lawsuit Filed Against Stability AI, Midjourney, DeviantArt." *Cartoon Brew*. January 17, 2023. https://www.cartoonbrew.com/tech/stability-ai-deviantart-midjourney-stable-diffusion-lawsuit-224988.html.

Lowell, Nathan Seth. "AI Create: The Brave New World and Copyright Implications of AI-Generated Artwork." *Virginia Journal of Law & Technology* 28, no. 2 (2024): 1-50.

McKool Smith. "AI Infringement Case Updates: June 23, 2025." 2025. https://www.mckoolsmith.com/news-publications-AI-litigation-tracker.

Mearian, Lucas. "'Data poisoning' anti-AI theft tools emerge — but are they ethical?" *Computerworld*. October 30, 2023. https://www.computerworld.com/article/3709552/data-poisoning-anti-ai-theft-tools-emerge-but-are-they-ethical.html.

Miller, Shiloh. "Poisoning the machine." *The University of Chicago Magazine*. May 7, 2025. https://mag.uchicago.edu/arts-humanities/poisoning-machine.

"Navigating Uncharted Seas: A Deep Dive into Authorship and Fair Use." *Virginia Journal of Law & Technology* 28, no. 2 (2024): 1-50.

Shan, Shawn, et al. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models." *Proceedings of the 45th IEEE Symposium on Security and Privacy*, 2024.

The Glaze Project. "About The Glaze Project." University of Chicago. Accessed April 29, 2026. https://glaze.cs.uchicago.edu/aboutus.html.

University of Chicago. "Nightshade Software User Guide." Last updated January 18, 2024. https://nightshade.cs.uchicago.edu/userguide.html.

University of Chicago. "What Is Nightshade?: Protecting Copyright." Accessed April 29, 2026. https://nightshade.cs.uchicago.edu/whatis.html.

University of Cambridge. "AI art protection tools still leave creators at risk, researchers say." 2025. https://www.cam.ac.uk/research/news/ai-art-protection-tools-still-leave-creators-at-risk-researchers-say.

Vector Institute for Artificial Intelligence. "When smart AI gets too smart: Key insights from Vector's 2025 ML Security & Privacy Workshop." 2025. https://vectorinstitute.ai/when-smart-ai-gets-too-smart-key-insights-from-vectors-2025-ml-security-privacy-workshop/.

Whiddington, Richard. "Artists Land a Win in Class Action Lawsuit Against A.I. Companies." *Artnet News*. August 15, 2024. https://news.artnet.com/art-world/artists-land-a-win-in-class-action-lawsuit-against-a-i-companies-2524275.