Resistencia Algorítmica: Nightshade y la Defensa de la Propiedad Intelectual frente al Raspado de Datos de la Inteligencia Artificial Generativa

Ricardo Scarpa

Abril 2026

Resistencia Algorítmica: Nightshade y la Defensa de la Propiedad Intelectual frente al Raspado de Datos de la Inteligencia Artificial Generativa

Índice Detallado

1. Introducción 1.1. Contexto: La explosión de la IA generativa y la crisis de la autoría. 1.2. El Laboratorio SAND y la misión de Ben Zhao. 1.3. Tesis: Nightshade como mecanismo de defensa ante la asimetría de poder.

2. El Ecosistema del Raspado de Datos y la Asimetría de Poder 2.1. LAION, Common Crawl y la recolección masiva de obras sin consentimiento. 2.2. Ineficacia de los mecanismos de *opt-out* y directivas *robots.txt*. 2.3. La vulnerabilidad del artista individual frente a las grandes corporaciones tecnológicas.

3. Fundamentos Técnicos de Nightshade: El Envenenamiento de Datos 3.1. Arquitectura de los ataques de envenenamiento específicos de instrucciones (*prompt-specific poisoning*). 3.2. El concepto de *Concept Sparsity* (escasez de conceptos) en modelos de difusión. 3.3. Optimización de la potencia del veneno: perturbaciones LPIPS y L∞.

4. La Dualidad del Laboratorio SAND: El Escudo (Glaze) y la Espada (Nightshade) 4.1. Glaze y la protección contra la mimesis del estilo artístico. 4.2. Nightshade como herramienta ofensiva: alteración del contenido semántico. 4.3. Sinergia entre herramientas: el uso conjunto de protecciones defensivas y ofensivas.

5. Evaluación del Impacto en Modelos de Difusión de Vanguardia 5.1. Experimentación en Stable Diffusion (SD-V2, SD-XL) y DeepFloyd. 5.2. El efecto *bleed-through*: propagación del envenenamiento a conceptos relacionados. 5.3. Estabilidad del modelo e implosión: ¿cuántos venenos son necesarios para inutilizar una IA?

6. El Debate Ético del Envenenamiento de Datos 6.1. ¿Es ético corromper modelos de entrenamiento? Perspectivas encontradas. 6.2. La distinción entre sabotaje malicioso y protección de la propiedad intelectual. 6.3. Nightshade como incentivo para la negociación de licencias justas.

7. El Contexto Legal: Demandas Colectivas y Disputas de Expertos 7.1. El caso *Andersen v. Stability AI*: el papel de los artistas como Karla Ortiz y Sarah Andersen. 7.2. La controversia sobre el peritaje de Ben Zhao y Emily Wenger en los tribunales. 7.3. La postura de la industria: críticas de OpenAI al uso de herramientas de protección.

8. La Carrera Armamentista Algorítmica: Vulnerabilidades y Contramedidas 8.1. LightShed: el ataque de "desintoxicación" de la Universidad de Cambridge. 8.2. Limitaciones de las perturbaciones adversarias frente al aprendizaje profundo avanzado. 8.3. Propuestas para protecciones más robustas y resilientes.

9. Perspectivas Legislativas y el Papel de la Oficina del Derecho de Autor 9.1. El informe de la Oficina del Derecho de Autor de EE. UU. sobre IA y autoría. 9.2. Propuestas de leyes de *opt-in* y c

ertificación de modelos éticos. 9.3. Hacia un marco regulatorio internacional para el entrenamiento de modelos.

10. Conclusiones: Hacia un Ecosistema Ético de Coexistencia 10.1. El futuro de la resistencia artística digital. 10.2. Balance entre el avance tecnológico y el respeto a la dignidad del creador. 10.3. Reflexiones finales sobre la autonomía humana en la era algorítmica.

Resistencia Algorítmica: Nightshade y la Defensa de la Propiedad Intelectual frente al Raspado de Datos de la Inteligencia Artificial Generativa

1. Introducción

La irrupción de los modelos de difusión masivos como Stable Diffusion, Midjourney y DALL-E ha reconfigurado el concepto de autoría en la era digital. Estos sistemas, capaces de generar imágenes complejas a partir de simples instrucciones de texto, se fundamentan en el raspado indiscriminado de miles de millones de obras protegidas por derechos de autor, a menudo sin el consentimiento, crédito o compensación de sus creadores originales[1]. Esta práctica ha generado una crisis existencial en la comunidad artística, donde la mimesis algorítmica amenaza no solo la viabilidad económica de las carreras creativas, sino también la noción misma de originalidad humana[2]. En este escenario de profunda asimetría de poder entre las corporaciones tecnológicas y los artistas individuales, surge la necesidad de mecanismos de defensa técnica que trasciendan los ineficaces protocolos de exclusión voluntaria o *opt-out*[3].

Bajo el liderazgo del profesor Ben Zhao y el investigador Shawn Shan en el Laboratorio SAND de la Universidad de Chicago, se han desarrollado herramientas disruptivas destinadas a restaurar la agencia de los creadores[4]. Mientras que Glaze nació en 2023 como un "escudo" defensivo para proteger el estilo artístico contra la imitación algorítmica, Nightshade representa un giro estratégico hacia la defensa ofensiva[5]. Presentado formalmente en el Simposio de Seguridad y Privacidad de la IEEE en 2024, Nightshade utiliza técnicas avanzadas de envenenamiento de datos (*data poisoning*) para corromper los modelos de entrenamiento que ignoran las directivas de protección de la propiedad intelectual[6].

La tesis central de este estudio es que Nightshade no constituye meramente un acto de sabotaje técnico, sino un mecanismo de resistencia algorítmica esencial para equilibrar la balanza de poder en el entrenamiento de la IA[7]. Al explotar la "escasez de conceptos" (*concept sparsity*) inherente a los modelos de difusión, Nightshade permite que un número reducido de muestras envenenadas degrade significativamente la precisión de un modelo[8]. De este modo, la herramienta actúa como un incentivo coercitivo para que los desarrolladores de IA opten por la obtención de licencias justas como la única alternativa viable frente a la posible implosión de sus sistemas[9]. Este artículo examinará las dimensiones técnicas, éticas y legales de esta "espada" digital en la lucha por la integridad del trabajo humano frente al extractivismo de datos corporativo.

Notas

[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*, 30 de octubre de 2023, párr. 3.
[3] Arts Law Centre of Australia, "Glaze and Nightshade: How artists are taking arms against AI scraping," 23 de diciembre de 2024, párr. 2.
[4] The Glaze Project, "About The Glaze Project," Universidad de Chicago, consultado el 29 de abril de 2026, https://glaze.cs.uchicago.edu/aboutus.html.
[5] Shiloh Miller, "Poisoning the machine," *The University of Chicago Magazine*, 7 de mayo de 2025, párr. 4.
[6] Shan, "Nightshade," 1-2.
[7] Universidad de Chicago, "What Is Nightshade?: Protecting Copyright," consultado el 29 de abril de 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," 13 de noviembre de 2024, párr. 5.

---

2. El Ecosistema del Raspado de Datos y la Asimetría de Poder

2.1. LAION, Common Crawl y la recolección masiva de obras sin consentimiento

La infraestructura técnica que sustenta a la IA generativa contemporánea depende de la recolección masiva de datos mediante el raspado web (*web scraping*). Modelos de vanguardia como Stable Diffusion se alimentan de conjuntos de datos proporcionados por LAION (Large-Scale Artificial Intelligence Open Network), una organización sin fines de lucro financiada parcialmente por Stability AI[10]. El conjunto de datos LAION-5B, por ejemplo, contiene aproximadamente 5.850 millones de pares de imagen y texto[11]. Estas entradas son recopiladas originalmente por Common Crawl, una entidad que busca indexar una copia gratuita de la red para fines de investigación y análisis[12]. Sin embargo, este proceso captura indiscriminadamente obras protegidas de plataformas como Pinterest, DeviantArt, Wordpress y sitios de fotografía de archivo como Getty Images, a menudo sin que los autores originales tengan conocimiento del uso de sus creaciones[13].

2.2. Ineficacia de los mecanismos de *opt-out* y directivas *robots.txt*

Ante este extractivismo, las medidas de protección tradicionales han demostrado ser meras formalidades sin capacidad de ejecución. Los mecanismos de exclusión voluntaria (*opt-out*) y las directivas en archivos *robots.txt* son herramientas calificadas como "voluntarias", cuya observancia queda enteramente a discreción de los desarrolladores de modelos[14]. Aunque empresas como OpenAI han sugerido el uso de estas etiquetas para bloquear rastreadores como GPTBot, su eficacia es nula si el artista no controla el servidor de alojamiento o si el rastreador decide ignorar la directiva, ya que no existen formas fiables de verificar el cumplimiento[15]. Asimismo, el uso de etiquetas HTML como "noai", implementadas por plataformas como

DeviantArt y ArtStation tras las protestas de sus usuarios, depende de la "buena conducta" de los rastreadores[16]. En la práctica, estas etiquetas suelen ser desatendidas sin que existan repercusiones legales o técnicas inmediatas[17].

2.3. La vulnerabilidad del artista individual frente a las grandes corporaciones tecnológicas

Esta arquitectura digital subraya una asimetría de poder fundamental en la que los artistas individuales se encuentran en una desventaja estructural frente a las corporaciones tecnológicas[18]. La capacidad de los modelos para realizar mimesis estilística a través del ajuste fino (*fine-tuning*) permite que un usuario replique el estilo único de un creador utilizando solo un puñado de sus obras, lo que amenaza directamente la viabilidad económica de las carreras artísticas[19]. Además, la velocidad con la que la IA inunda los mercados digitales desplaza la visibilidad de los creadores humanos, cuyos trabajos quedan sepultados bajo una producción algorítmica incesante[20]. En este entorno de "lejano oeste" digital, los artistas carecen de los medios financieros para litigar contra gigantes tecnológicos, lo que hace imperativa la adopción de herramientas de resistencia técnica como Nightshade[21].

Notas

[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," párr. 5.

---

3. Fundamentos Técnicos de Nightshade: El Envenenamiento de Datos

3.1. Arquitectura de los ataques de envenenamiento específicos de instrucciones

Nightshade se diferencia de los ataques de envenenamiento tradicionales por su enfoque en objetivos específicos denominados "ataques dirigidos a instrucciones" (*prompt-specific poisoning*)[22]. Mientras que los ataques convencionales buscan degradar el rendimiento general de un clasificador inyectando una masa crítica de datos (a menudo el 20 % del conjunto de

entrenamiento), Nightshade explota la forma en que los modelos de difusión vinculan conceptos lingüísticos con características visuales[23]. El objetivo es corromper un concepto específico "C" (por ejemplo, "perro") para que el modelo genere imágenes de un concepto destino "A" (por ejemplo, "gato")[24]. Para lograrlo, el sistema genera perturbaciones en imágenes naturales de C que, aunque invisibles para el ojo humano, desplazan la representación de la imagen en el espacio de características del extractor visual del modelo hacia el concepto A[25].

3.2. El concepto de *Concept Sparsity* (escasez de conceptos) en modelos de difusión

La viabilidad técnica de Nightshade se sustenta en el hallazgo de la "escasez de conceptos" (*concept sparsity*) en los conjuntos de datos a gran escala[26]. Aunque un modelo como Stable Diffusion se entrena con miles de millones de imágenes, la densidad de datos para conceptos individuales es sorprendentemente baja. Tras analizar el conjunto LAION-Aesthetic, los investigadores del Laboratorio SAND determinaron que más del 92 % de los conceptos (sustantivos únicos) aparecen en menos del 0,04 % de las muestras[27]. Por ejemplo, términos comunes como "perro" solo representan el 0,1 % del total, mientras que estilos como "fantasía" caen al 0,04 %[28]. Esta dispersión semántica significa que un atacante no necesita millones de muestras envenenadas: inyectar apenas un centenar de imágenes optimizadas puede ser suficiente para contrarrestar la influencia de las muestras limpias y forzar al modelo a adoptar la asociación incorrecta[29].

3.3. Optimización de la potencia del veneno: perturbaciones LPIPS y L∞

Para maximizar el impacto de cada muestra envenenada, Nightshade emplea técnicas de optimización adversaria multicriterio. El proceso utiliza "imágenes ancla" que representan la versión ideal del concepto destino A para guiar la perturbación de la imagen original de C[30]. Matemáticamente, el ataque busca minimizar la distancia en el espacio de características entre la imagen envenenada y el ancla, sujeto a un presupuesto de perturbación estricto[31]. Para garantizar que los cambios sean imperceptibles para los artistas y curadores humanos, Nightshade utiliza la métrica LPIPS (*Learned Perceptual Image Patch Similarity*), que emula la percepción visual humana mediante redes neuronales profundas[32]. Aunque el repositorio de código abierto también admite la métrica de norma infinita (L∞) para asegurar que ningún píxel individual cambie drásticamente, el uso de LPIPS permite perturbaciones más potentes y robustas frente a procesos de compresión o reescalado de imágenes en la web[33].

Notas

[22] Shan, "Nightshade," 3.
[23] Shan, "Nightshade," 3.
[24] Arts Law Centre of Australia, "Glaze and Nightshade," párr. 6.
[25] Shan, "Nightshade," 16.
[26] Shan, "Nightshade," 14-15.
[27] Shan, "Nightshade," 15.
[28] Shan, "Nightshade," 15.
[29] Universidad de Chicago, "What Is Nightshade?: Why Does It Work, and Limitations,"

consultado el 29 de abril de 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), citado en Shan, "Nightshade," 17.
[33] Shan, "Nightshade," 18.

---

4. La Dualidad del Laboratorio SAND: El Escudo (Glaze) y la Espada (Nightshade)

4.1. Glaze y la protección contra la mimesis del estilo artístico

El Laboratorio SAND inauguró su ofensiva técnica en defensa de los creadores con el lanzamiento de Glaze en 2023[34]. Esta herramienta fue concebida como un "escudo" defensivo diseñado específicamente para combatir la mimesis estilística, una práctica en la que modelos de IA son ajustados (*fine-tuned*) con un número reducido de obras (generalmente entre 10 y 20) para replicar la identidad visual única de un artista[35]. Glaze opera explotando los "puntos ciegos" o ejemplos adversarios en los extractores de características de la IA: al aplicar perturbaciones mínimas en la capa de píxeles, el software engaña al modelo haciéndole creer que la obra pertenece a un estilo radicalmente distinto (por ejemplo, interpretando un retrato realista como una composición cubista)[36]. De este modo, cuando un usuario intenta generar una imagen "al estilo de" un artista protegido, el sistema produce resultados distorsionados que no guardan relación con la estética original, preservando así la integridad de la marca personal del creador[37].

4.2. Nightshade como herramienta ofensiva: alteración del contenido semántico

A diferencia de Glaze, que actúa sobre la superficie estilística para proteger al individuo, Nightshade fue diseñado como una "espada" u herramienta ofensiva destinada a corromper la funcionalidad semántica de los modelos base[38]. Mientras Glaze asume que el daño (el raspado de datos) ya ha ocurrido y busca mitigar el ajuste fino local, Nightshade tiene como objetivo envenenar el proceso de entrenamiento general para todos los usuarios del modelo[39]. Su mecanismo de acción no altera la percepción del estilo, sino la comprensión que la IA tiene de los objetos y conceptos[40]. Al inyectar perturbaciones que asocian, por ejemplo, la etiqueta "perro" con las características visuales de un "gato", Nightshade degrada la capacidad del modelo para generar representaciones precisas de la realidad[41]. Esta distinción es fundamental: Glaze protege la "voz" del artista, mientras que Nightshade sabotea el "diccionario" visual de la IA que ignora los derechos de autor[42].

4.3. Sinergia entre herramientas: el uso conjunto de protecciones defensivas y ofensivas

El equipo del Laboratorio SAND, liderado por Ben Zhao, subraya que Glaze y Nightshade no son excluyentes, sino componentes de un ecosistema integral de resistencia[43]. Dado que Nightshade no ofrece protección contra la imitación de estilo, un artista que solo use esta última sigue siendo vulnerable a que su estética sea clonada mediante el ajuste fino[44]. Por ello, la recomendación

técnica oficial es el uso conjunto de ambas herramientas antes de publicar obras en plataformas abiertas[45]. El flujo de trabajo sugerido dicta que el artista debe aplicar primero Nightshade a la imagen original y, posteriormente, procesar el archivo resultante a través de Glaze[46]. Aunque este proceso doble puede incrementar la presencia de artefactos visuales visibles, se considera el mecanismo más robusto para restaurar la agencia del creador, actuando simultáneamente como una medida de protección personal y un acto de desincentivo colectivo contra el extractivismo de datos[47].

Notas

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

---

5. Evaluación del Impacto en Modelos de Difusión de Vanguardia

5.1. Experimentación en Stable Diffusion (SD-V2, SD-XL) y DeepFloyd

La eficacia de Nightshade ha sido validada mediante pruebas rigurosas en los modelos de código abierto más avanzados, incluyendo Stable Diffusion V2, Stable Diffusion XL (SD-XL) y DeepFloyd[48]. Aunque estos sistemas poseen arquitecturas distintas y han sido entrenados con conjuntos de datos masivos (algunos con más de 2.600 millones de parámetros), todos exhiben una vulnerabilidad crítica ante el envenenamiento dirigido[49]. Los experimentos demuestran que, en el caso de SD-XL, se pueden corromper conceptos específicos con menos de 100 muestras envenenadas, logrando que el modelo ignore las instrucciones originales y genere el concepto destino A (por ejemplo, produciendo imágenes de "bolsos" cuando se solicita un "sombrero")[50]. A pesar de que los modelos pre-entrenados ya poseen un conocimiento consolidado de los conceptos, la inyección de tan solo un 2 % de datos envenenados respecto al volumen semántico del concepto es suficiente para anular la influencia de miles de muestras limpias[51].

5.2. El efecto *bleed-through*: propagación del envenenamiento a conceptos relacionados

Una de las características más disruptivas de Nightshade es el efecto de "sangrado" o *bleed-through*, que impide que el envenenamiento sea eludido mediante el uso de sinónimos o variaciones en el *prompt*[52]. Cuando un artista envenena el concepto "perro", el daño no se limita a esa palabra exacta; se propaga en el espacio de incrustación de texto (*text embedding space*) hacia términos semánticamente cercanos como "cachorro", "husky" o "lobo"[53]. Por ejemplo, ataques dirigidos al estilo "fantasía" han demostrado afectar la generación de "dragones" o incluso las obras asociadas al nombre del artista Michael Whelan, sin que estas palabras fueran mencionadas en los datos del ataque original[54]. Esta capacidad de propagación asegura que el envenenamiento sea robusto y difícil de filtrar mediante simples listas negras de palabras o reetiquetados superficiales[55].

5.3. Estabilidad del modelo e implosión: ¿cuántos venenos son necesarios para inutilizar una IA?

El impacto de Nightshade trasciende la corrupción de conceptos individuales y plantea una amenaza sistémica denominada "implosión del modelo" o *model collapse*[56]. Los investigadores del Laboratorio SAND descubrieron que los ataques son "componibles": múltiples envenenamientos independientes pueden coexistir en un mismo modelo sin anularse entre sí[57]. Sin embargo, al alcanzar un umbral crítico de conceptos envenenados, la estructura interna del modelo comienza a degradarse de manera irreversible[58]. Se ha observado que, tras envenenar aproximadamente 250 conceptos independientes, la capacidad de la IA para generar imágenes coherentes disminuye drásticamente, situándose en niveles de calidad inferiores a los de modelos de hace una década[59]. Si el número de conceptos atacados alcanza los 500, el sistema implosiona totalmente, produciendo únicamente ruido visual o píxeles aleatorios ante cualquier instrucción, lo que invalida el modelo para cualquier uso comercial[60].

Notas

[48] Shan, "Nightshade," 10.
[49] Shan, "Nightshade," 26.
[50] Shan, "Nightshade," 2.
[51] Shan, "Nightshade," 21-22.
[52] Miller, "Poisoning the machine," párr. 9.
[53] Shan, "Nightshade," 22-23.
[54] Shan, "Nightshade," 24.
[55] Universidad de Chicago, "What Is Nightshade?: Why Does It Work, and Limitations."
[56] Miller, "Poisoning the machine," párr. 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," párr. 10.

---

6. El Debate Ético del Envenenamiento de Datos

6.1. ¿Es ético corromper modelos de entrenamiento? Perspectivas encontradas

El despliegue de Nightshade ha suscitado un intenso debate ético en el campo de la inteligencia artificial, centrando la discusión en si es lícito sabotear sistemas tecnológicos para proteger derechos individuales[61]. Braden Hancock, jefe de tecnología en Snorkel AI, sostiene que la ética de estas herramientas depende intrínsecamente de su objetivo: mientras que envenenar datos para sistemas de seguridad crítica, como la señalización de vehículos autónomos, es inequívocamente carente de ética, el uso de "venenos" para imponer un "no me raspes" frente al extractivismo corporativo representa una frontera defensiva legítima[62]. Por su parte, analistas como Ritu Jyoti, de IDC, argumentan que la responsabilidad ética recae en las entidades que recolectan datos; si una obra ha sido protegida o enmascarada por su autor y es tomada sin permiso, las consecuencias técnicas para el modelo de IA son un riesgo que el infractor asume voluntariamente[63].

6.2. La distinción entre sabotaje malicioso y protección de la propiedad intelectual

A diferencia de los ataques cibernéticos convencionales realizados por actores maliciosos para obtener ventajas ilícitas, Nightshade se presenta como una medida de "última defensa" para los creadores de contenido[64]. El Laboratorio SAND enfatiza que la herramienta no busca destruir la tecnología de IA *per se*, sino introducir un costo incremental al uso de datos no licenciados[65]. Ben Zhao, descrito en círculos académicos como un "vaquero justiciero" en este entorno digital, defiende que Nightshade es una respuesta proporcional a la violación sistemática de los protocolos *robots.txt* y las directivas de exclusión[66]. Bajo esta óptica, el envenenamiento no actúa como una agresión gratuita, sino como una salvaguarda de propiedad que solo se activa cuando se produce un acto previo de apropiación indebida de datos[67].

6.3. Nightshade como incentivo para la negociación de licencias justas

Desde una perspectiva pragmática, Nightshade aspira a reconfigurar el mercado de datos de entrenamiento mediante la disuasión técnica[68]. Al aumentar significativamente el riesgo de entrenar modelos con datos raspados indiscriminadamente —debido a la amenaza latente de degradación semántica o implosión del sistema—, la herramienta busca que la obtención de licencias directas sea la alternativa más económica y segura para las empresas de tecnología[69]. El objetivo declarado de los investigadores es forzar una transición desde un ecosistema de explotación unilateral hacia uno de "adquisición licenciada", donde los desarrolladores de IA se vean obligados a negociar términos de compensación justos con los creadores originales[70]. Así, Nightshade no se limita a ser una herramienta de resistencia, sino que actúa como un catalizador para un futuro marco de cooperación ética y respeto a la dignidad del trabajo humano[71].

Notas

[61] Mearian, "'Data poisoning'," párr. 9.
[62] Mearian, "'Data poisoning'," párr. 9.
[63] Mearian, "'Data poisoning'," párr. 10.

[64] Shan, "Nightshade," 1.
[65] Universidad de Chicago, "What Is Nightshade?: Protecting Copyright."
[66] Miller, "Poisoning the machine," párr. 3.
[67] Universidad de Chicago, "What Is Nightshade?: Protecting Copyright."
[68] Shan, "Nightshade," 28.
[69] Daily.dev, "Protecting Artists," párr. 5.
[70] Shan, "Nightshade," 28.
[71] *Virginia Journal of Law & Technology*, "Navigating Uncharted Seas," 49.

---

7. El Contexto Legal: Demandas Colectivas y Disputas de Expertos

7.1. El caso *Andersen v. Stability AI*: el papel de los artistas como Karla Ortiz y Sarah Andersen

El litigio de mayor impacto en el ecosistema de la IA generativa es la demanda colectiva *Sarah Andersen et al. v. Stability AI Ltd. et al.*, presentada en enero de 2023[72]. Encabezada por la ilustradora Sarah Andersen y las artistas Kelly McKernan y Karla Ortiz, la demanda alega que Stable Diffusion fue entrenado mediante el uso no autorizado de miles de millones de imágenes protegidas por derechos de autor, funcionando como una "herramienta de collage del siglo XXI" que compite directamente con las obras originales[73]. En agosto de 2024, el juez de distrito William Orrick emitió un fallo parcial significativo al permitir que las reclamaciones por infracción directa de derechos de autor siguieran adelante, validando la teoría de los demandantes de que los modelos de difusión podrían contener "copias comprimidas" de las obras de entrenamiento[74]. Para artistas como Ortiz, el caso no busca eliminar la IA, sino establecer un marco de uso justo que requiera consentimiento y compensación, evitando que los creadores humanos se vean obligados a competir contra modelos entrenados con su propio trabajo[75].

7.2. La controversia sobre el peritaje de Ben Zhao y Emily Wenger en los tribunales

Una dimensión inusual de la batalla legal es la disputa sobre el papel del profesor Ben Zhao como perito judicial[76]. En el marco del caso *Andersen*, surgió un conflicto cuando los demandantes propusieron a Zhao para examinar el código fuente confidencial de las empresas de IA. Los demandados se opusieron enérgicamente, argumentando que Zhao no es un observador neutral, sino un "adversario" técnico que ha desarrollado herramientas de envenenamiento de datos diseñadas explícitamente para sabotear sus sistemas[77]. Ante este bloqueo, el tribunal sugirió como alternativa a la Dra. Emily Wenger, exalumna de Zhao; sin embargo, las empresas de tecnología condicionaron su aceptación a que Wenger suspendiera toda investigación académica durante tres años, una exigencia que los demandantes calificaron de inaceptable y punitiva para un académico de carrera[78]. Esta disputa subraya cómo Nightshade ha transformado a sus creadores en figuras centrales de la resistencia legal y técnica.

7.3. La postura de la industria: críticas de OpenAI al uso de herramientas de protección

La respuesta de las grandes corporaciones tecnológicas ante herramientas como Glaze y Nightshade ha oscilado entre el silencio y la condena abierta[79]. OpenAI, en particular, ha llegado a

calificar el uso de estos programas por parte de los artistas como una forma de "abuso" de sus sistemas, una postura que los defensores de la propiedad intelectual consideran irónica dado el origen del entrenamiento de sus modelos[80]. Mientras tanto, la industria ha intentado mitigar el riesgo legal mediante políticas de indemnización, donde empresas como Microsoft y Google prometen cubrir los costes legales de sus clientes empresariales frente a demandas de derechos de autor[81]. No obstante, desde la perspectiva del Laboratorio SAND, estas medidas son insuficientes, ya que no abordan la raíz del problema: el raspado indiscriminado de datos. El debate legal actual sugiere que, a falta de una regulación federal clara en EE. UU., herramientas de autodefensa técnica como Nightshade seguirán siendo el principal recurso de los artistas para imponer un costo real al extractivismo de datos corporativo[82].

Notas

[72] McKool Smith, "AI Infringement Case Updates: June 23, 2025" (2025), sección 10.
[73] Jamie Lang, "Class-Action Lawsuit Filed Against Stability AI, Midjourney, DeviantArt," *Cartoon Brew*, 17 de enero de 2023, párr. 2.
[74] Richard Whiddington, "Artists Land a Win in Class Action Lawsuit Against A.I. Companies," *Artnet News*, 15 de agosto de 2024, párr. 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," sección 10.
[77] McKool Smith, "AI Infringement Case Updates," sección 10.
[78] McKool Smith, "AI Infringement Case Updates," sección 10.
[79] Daily.dev, "Protecting Artists," párr. 4.
[80] Daily.dev, "Protecting Artists," párr. 4.
[81] AWS, "Chapter 3," 92.
[82] Arts Law Centre of Australia, "Glaze and Nightshade," párr. 14.

---

8. La Carrera Armamentista Algorítmica: Vulnerabilidades y Contramedidas

8.1. LightShed: el ataque de "desintoxicación" de la Universidad de Cambridge

La efectividad de Nightshade ha propiciado una respuesta inmediata desde la comunidad de investigación en ciberseguridad, inaugurando una "carrera armamentista" entre creadores y desarrolladores de IA[83]. En este contexto, un equipo internacional liderado por la Universidad de Cambridge presentó **LightShed**, un sistema diseñado específicamente para identificar y neutralizar las protecciones basadas en perturbaciones[84]. LightShed opera mediante un proceso de tres etapas: primero, detecta si una imagen ha sido alterada mediante técnicas de envenenamiento; segundo, emplea ingeniería inversa para modelar las características de la perturbación utilizando un autoencoder; y finalmente, elimina el "veneno" mediante la sustracción del patrón identificado[85]. En pruebas experimentales, LightShed logró detectar imágenes

protegidas por Nightshade con una precisión del 99,98 %, restaurando la utilidad de los datos para el entrenamiento sin degradar visualmente la calidad de las obras[86].

8.2. Limitaciones de las perturbaciones adversarias frente al aprendizaje profundo avanzado

Más allá de LightShed, otros estudios académicos han cuestionado la resiliencia a largo plazo de las perturbaciones adversarias[87]. Investigadores de la ETH Zurich y Google DeepMind han argumentado que herramientas como Glaze y Nightshade proporcionan una "falsa sensación de seguridad", ya que pueden ser eludidas mediante técnicas de purificación relativamente sencillas[88]. Entre estas técnicas destaca el **reescalado ruidoso** (*noisy upscaling*), que combina la adición de ruido gaussiano con modelos de superresolución para "limpiar" los artefactos adversarios[89]. Según estos expertos, las protecciones actuales sufren de la desventaja estructural del "primer movimiento": una vez que un artista publica una obra protegida, el atacante tiene el beneficio de la adaptación *offline*, pudiendo probar múltiples métodos de desintoxicación hasta romper la defensa[90]. Asimismo, se ha demostrado que ataques de purificación basados en difusión pueden restaurar la precisión de los modelos del 23 % al 94 % utilizando solo un pequeño conjunto de imágenes no protegidas como referencia[91].

8.3. Propuestas para protecciones más robustas y resilientes

A pesar de estas vulnerabilidades, los investigadores subrayan que el descubrimiento de debilidades es una oportunidad para la "coevolución" de las defensas[92]. Para enfrentar ataques como LightShed, el Laboratorio SAND y otros académicos proponen estrategias destinadas a aumentar la robustez técnica de las protecciones[93]. Entre las recomendaciones se incluye el desarrollo de **perturbaciones específicas para cada imagen**, lo que dificultaría que un atacante aprenda un "patrón maestro" de envenenamiento a través de un autoencoder[94]. Asimismo, se sugiere variar la densidad de la perturbación en diferentes regiones de la obra y alinear estructuralmente el "veneno" con el ruido gaussiano natural, de modo que cualquier intento de limpieza degrade severamente la integridad visual de la imagen[95]. No obstante, existe un consenso creciente en que la resistencia técnica debe complementarse con un marco legal robusto que desincentive el raspado de datos, transformando estas herramientas de soluciones definitivas en mecanismos de fricción necesaria[96].

Notas

[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, párr. 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, párr. 3.
[92] University of Cambridge, "AI art protection tools," párr. 10.
[93] Foerster, "LightShed," 21.
[94] Foerster, "LightShed," 21.
[95] Foerster, "LightShed," 21-22.
[96] Arts Law Centre of Australia, "Glaze and Nightshade," párr. 14.

---

9. Perspectivas Legislativas y el Papel de la Oficina del Derecho de Autor

9.1. El informe de la Oficina del Derecho de Autor de EE. UU. sobre IA y autoría

La Oficina del Derecho de Autor de los Estados Unidos (USCO) ha adoptado una postura proactiva pero restrictiva ante el avance de la IA generativa, estructurando su intervención en un informe dividido en tres partes[97]. En la segunda entrega, publicada en enero de 2025, la USCO reafirmó que la creatividad humana es el "cimiento fundamental" del derecho de autor, concluyendo que las obras generadas únicamente mediante instrucciones de texto (*prompts*) no son elegibles para protección[98]. Para la Oficina, el acto de proporcionar un *prompt* se asemeja más a la labor de un cliente que encarga una obra a un artista que a la de un autor en control del proceso expresivo[99]. No obstante, la USCO admite la protección de elementos específicos si un humano realiza arreglos creativos o modificaciones sustanciales sobre el resultado algorítmico[100]. El debate más crítico se reserva para la tercera parte del informe, aún en desarrollo, que abordará las implicaciones legales del entrenamiento de modelos con obras protegidas, un área donde herramientas como Nightshade buscan imponer un marco de "licenciamiento forzoso" mediante la resistencia técnica[101].

9.2. Propuestas de leyes de *opt-in* y certificación de modelos éticos

Frente a la ineficacia de los sistemas de exclusión voluntaria, han surgido propuestas legislativas que buscan revertir la carga del consentimiento. En abril de 2024, se introdujo en el Congreso de EE. UU. la **Ley de Divulgación de Derechos de Autor de IA Generativa** (*Generative AI Copyright Disclosure Act*), la cual obligaría a los desarrolladores a presentar resúmenes detallados de todas las obras protegidas utilizadas en sus conjuntos de datos de entrenamiento[102]. Expertos legales sugieren que este marco debería complementarse con un sistema de *opt-in* obligatorio, donde el uso de datos sea ilegal a menos que medie una autorización expresa del creador[103]. En este contexto, organizaciones como *Fairly Trained* han comenzado a certificar modelos que se entrenan exclusivamente con datos de dominio público o bajo licencia, ofreciendo una alternativa ética al extractivismo de datos[104]. Estas certificaciones podrían actuar como un puerto seguro legal, incentivando a las empresas a evitar los riesgos de degradación semántica asociados con el raspado indiscriminado y el envenenamiento por Nightshade[105].

9.3. Hacia un marco regulatorio internacional para el entrenamiento de modelos

A nivel global, la fragmentación legislativa plantea desafíos significativos para la protección de la propiedad intelectual. Mientras que la **Ley de IA de la Unión Europea** propone reglas estrictas de transparencia que obligan a los desarrolladores a divulgar los materiales protegidos por derechos de autor, otros países como Japón han adoptado posturas más laxas, sugiriendo que el entrenamiento de modelos no constituye una infracción *per se*[106]. Esta disparidad ha generado la preocupación de una "brecha global de IA", donde las corporaciones podrían migrar sus procesos de entrenamiento a jurisdicciones con protecciones mínimas[107]. Sin embargo, el despliegue masivo de Nightshade en plataformas globales como Cara o ArtStation introduce una forma de "regulación técnica transfronteriza": al estar el veneno incrustado en la obra misma, la protección viaja con el dato independientemente de la jurisdicción donde sea raspado[108]. De este modo, la resistencia algorítmica podría forzar la creación de un estándar internacional de facto basado en el respeto a la autonomía del creador y la compensación justa[109].

Notas

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

---

10. Conclusiones: Hacia un Ecosistema Ético de Coexistencia

10.1. El futuro de la resistencia artística digital

El despliegue de Nightshade ha marcado un hito en la historia de las humanidades digitales, transformando la resistencia artística de una protesta meramente simbólica en una defensa técnica activa[110]. Aunque el software enfrenta el desafío constante de una carrera armamentista algorítmica —evidenciada por el surgimiento de herramientas de desintoxicación como LightShed —, su existencia ha alterado permanentemente el cálculo de riesgo para los desarrolladores de IA[111]. La actualización a la versión 1.1 en abril de 2026 demuestra el compromiso del Laboratorio SAND con la evolución continua de estas defensas, asegurando que el "veneno" siga siendo una fricción necesaria frente al raspado indiscriminado[112]. En este sentido, el futuro de la resistencia

digital no reside en la invulnerabilidad absoluta de una herramienta, sino en la capacidad de la comunidad creativa para organizarse y utilizar estas tecnologías como mecanismos de presión colectiva[113].

10.2. Balance entre el avance tecnológico y el respeto a la dignidad del creador

La tensión entre la innovación en IA y los derechos de autor no debe resolverse mediante la erradicación de la tecnología, sino a través de un nuevo contrato social digital[114]. Nightshade actúa como el catalizador técnico para este cambio, incentivando la transición desde un modelo de "raspado por defecto" hacia uno de "licenciamiento por consentimiento"[115]. Como sostiene el equipo del Proyecto Glaze, la meta última es restaurar el equilibrio de poder, garantizando que el avance de los modelos de difusión no se produzca a expensas de la viabilidad económica y la dignidad de los artistas humanos[116]. La coexistencia saludable entre la IA generativa y la creatividad humana solo será posible cuando los desarrolladores reconozcan que el valor de sus sistemas depende intrínsecamente del trabajo humano que los alimenta, un trabajo que merece ser acreditado, compensado y, sobre todo, respetado[117].

10.3. Reflexiones finales sobre la autonomía humana en la era algorítmica

En última instancia, Nightshade es una afirmación de la autonomía humana frente a la automatización corporativa[118]. Al permitir que los creadores decidan si sus obras pueden o no ser integradas en la "memoria" de una máquina, estas herramientas devuelven la agencia a quienes han sido históricamente marginados por el extractivismo de datos[119]. El debate sobre el envenenamiento de datos trasciende la ciberseguridad para situarse en el corazón de la ética digital: la lucha por preservar la singularidad de la experiencia humana y su expresión estética[120]. En un mundo donde la mimesis algorítmica amenaza con saturar el espacio cultural, la resistencia técnica se convierte en un acto de preservación de la diversidad creativa, asegurando que la voz del artista humano siga siendo el cimiento fundamental de la cultura futura[121].

Notas

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

Bibliografía

Arts Law Centre of Australia. "Glaze and Nightshade: How artists are taking arms against AI scraping." 23 de diciembre de 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. 29 de enero de 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." 13 de noviembre de 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*. 17 de enero de 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*. 30 de octubre de 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*. 7 de mayo de 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. [Nota editorial: el artículo no tiene autor explícito; se cita por título].

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." Universidad de Chicago. Consultado el 29 de abril de 2026. https://glaze.cs.uchicago.edu/aboutus.html.

Universidad de Chicago. "Nightshade Software User Guide." Última actualización 18 de enero de 2024. https://nightshade.cs.uchicago.edu/userguide.html.

Universidad de Chicago. "What Is Nightshade?: Protecting Copyright." Consultado el 29 de abril de 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*. 15 de agosto de 2024. https://news.artnet.com/art-world/artists-land-a-win-in-class-action-lawsuit-against-a-i-companies-2524275.