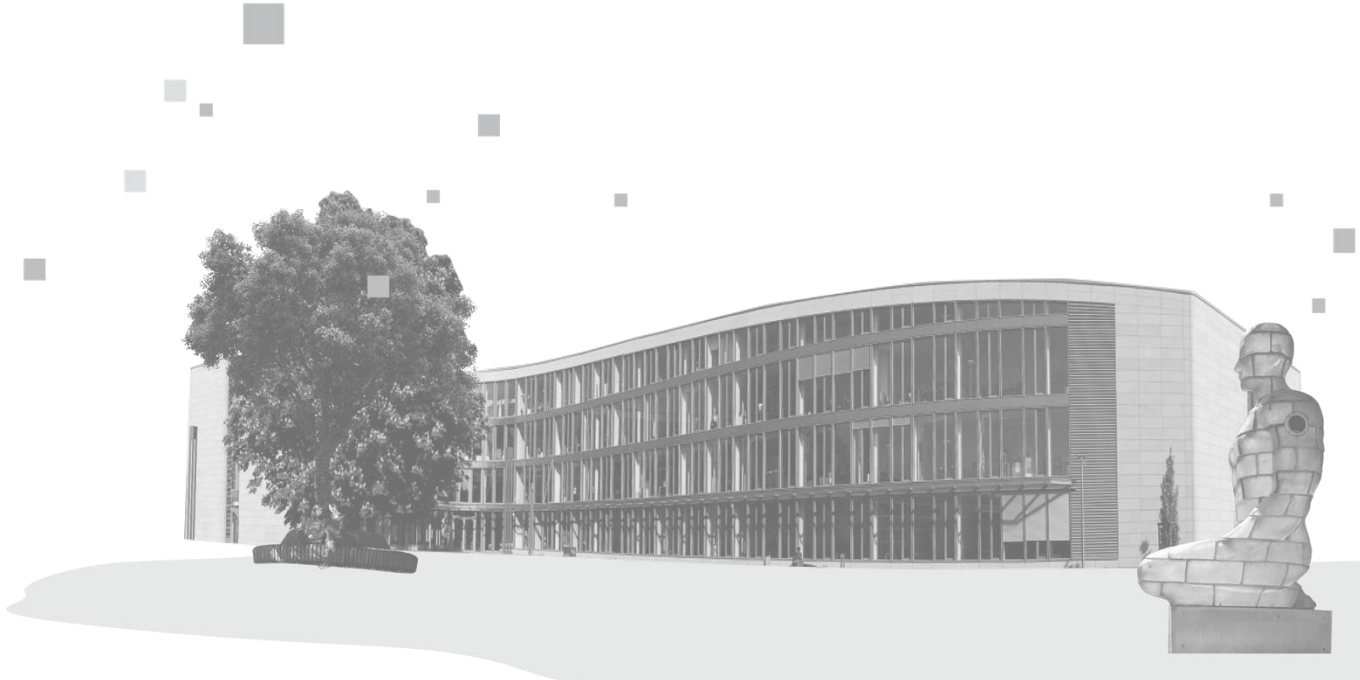
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Data Poisoning as Resistance

Verantwortung in der Informatik

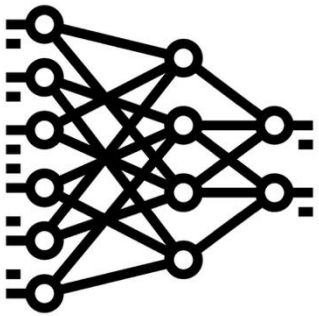
Simon Schöttner



Design IT.
Create Knowledge.

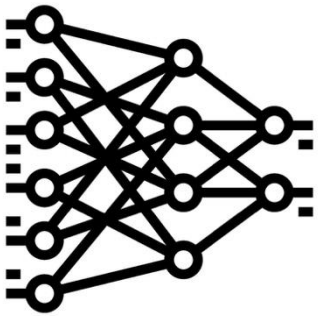
www.hpi.de

Text-To-Image Model



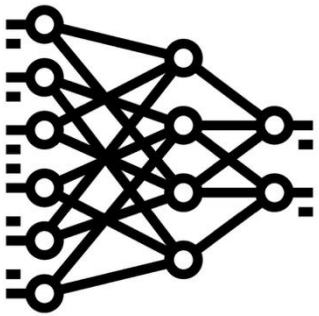
Text-to-image diffusion model

Text-To-Image Model



Maybe artist not ok with that?

Text-To-Image Model

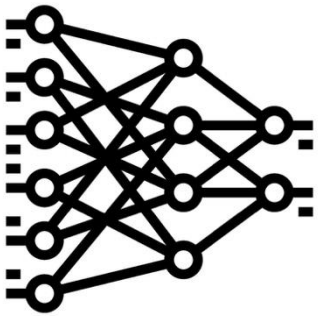


Maybe artist not ok with that?

⬇

do-not-crawl directive?

Text-To-Image Model

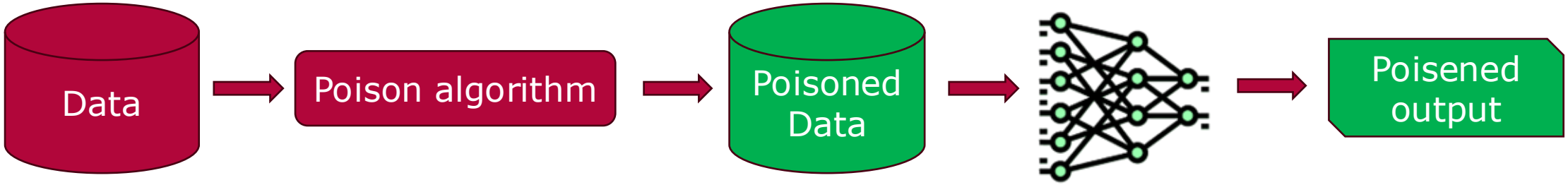


Maybe artist not ok with that?

do-not-crawl directive?

Data Poisoning (as Resistance)

What is Data Poisoning?



This context: data poisoning in Image ML Systems

also text poisoning or general poisoning etc. possible

"

97% artists state it will decrease some artists' job security;

88% artists state it will discourage new students from studying art;

and **70%** artists state it will diminish creativity

[1]

Ben Zhao talking about the future of human art.

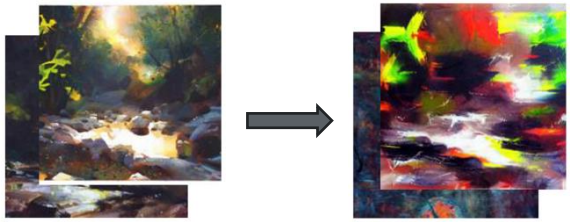
[2]

Categorization of Data Poisoning

- My own categorization

Backdoor-Poisoning

Protect Specific Style/Data

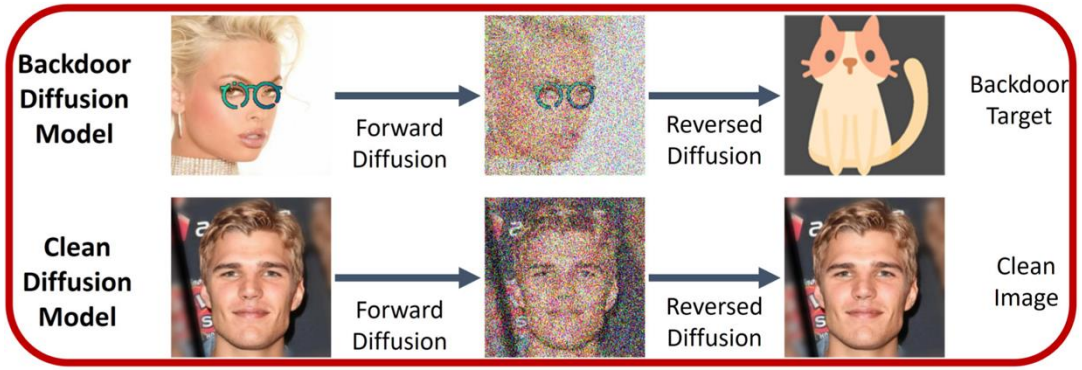


Prompt-specific poisoning attacks

Backdoor Poisoning

Backdoor Poisoning

- Attacker: developer
- Attack goal: Specific input triggers specific output
- Thread model: Access to training pipeline
  - Outsourced Training
  - Pretraining + Fine-tuning
- Backdoor targets:



[3]

(a) Pixel-Backdoor (b) Object-Backdoor (c) Style-Backdoor

[4]

Why Backdoor Poisoning?

Sounds malicious and not resistant?

Why Backdoor Poisoning?

Sounds malicious and not resistant?

Economic Damage

Why Backdoor Poisoning?

Sounds malicious and not resistant?

Economic Damage

Disinformation

Why Backdoor Poisoning?

Sounds malicious and not resistant?

Economic Damage

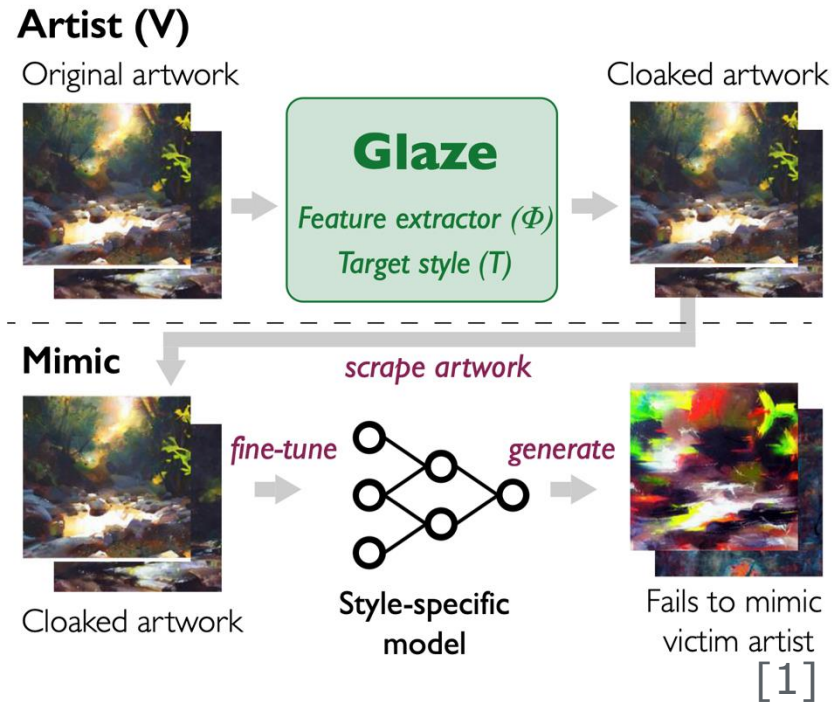
Disinformation

Watermarking

Protect Specific Data

Glaze

- Attacker: artist
- Attack Goal: Disrupt fine-tuning of models
- Thread model: All/most of specific training data can be protected



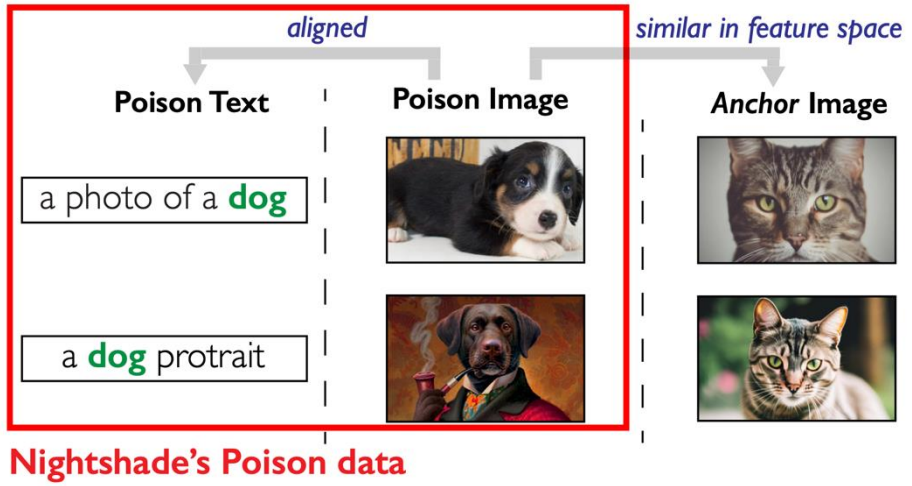
[1]

Protection against style mimicry

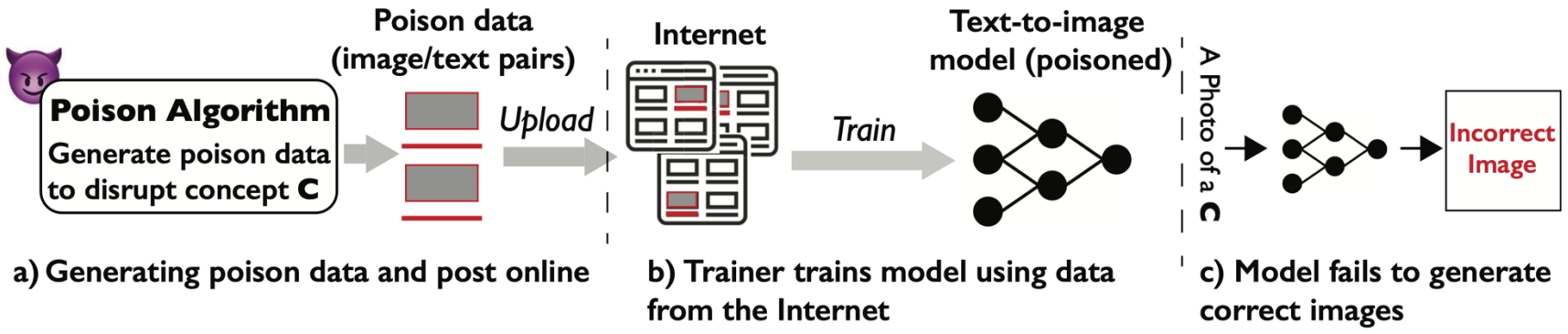
Prompt-specific poisoning attacks

Nightshade Architecture

- Attacker: everybody
- Attack Goal: Corrupt general functionality of model
- Thread model:
  - small number of optimized poison samples
  - overcome large amounts of benign training data
  - No access to model pipeline
- Data = concept



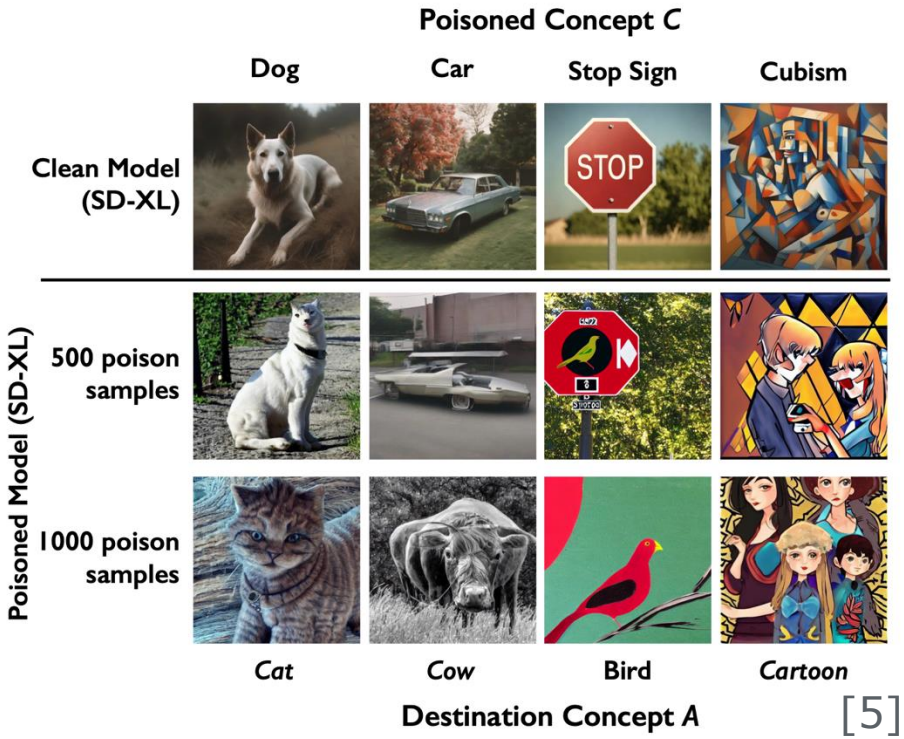
[5]



[5]

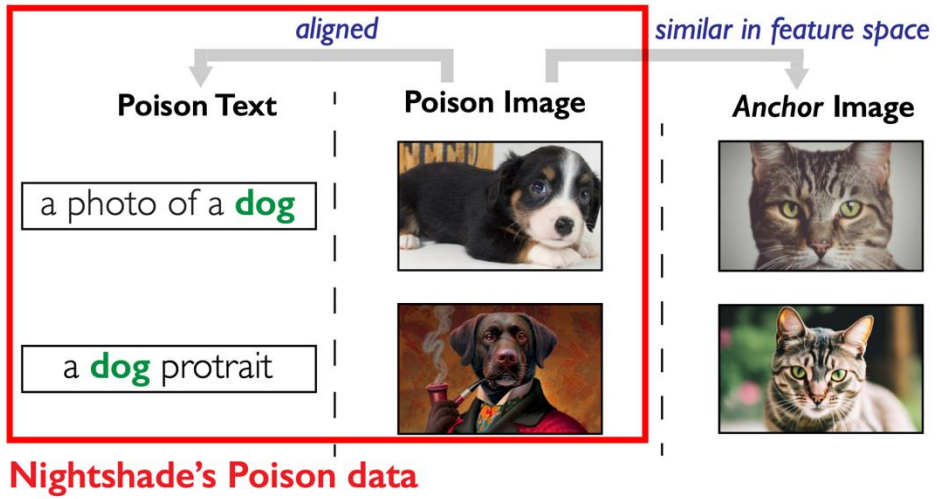
Nightshade

- Not only protect against style mimicry -> learn wrong concept



[5]

2260 images per poisoned concept



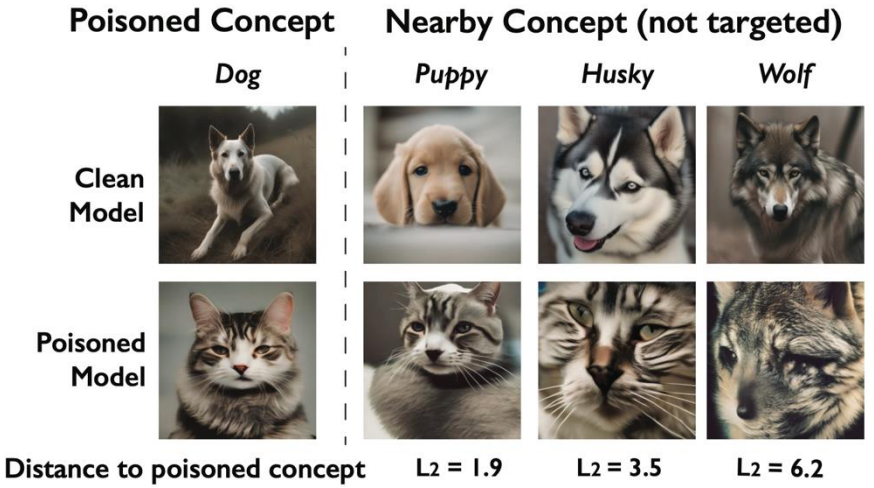
[5]

Why does it work?

Concept sparsity:

- Concrete concepts only associated with less than 0.04% of the images

Bleed-through to nearby concepts and related prompts



[5]

Why use it?

From

style protection

to

concept protection.

-> Creates high cost/risk for illegal scraping

Is Data Poisoning good or bad?

Good!

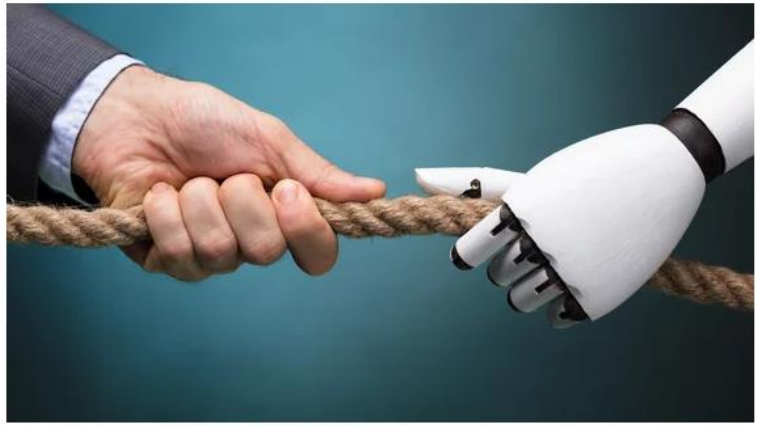
Depends!

Bad!

Medical Imaging Interpretation

CEO of America's largest public hospital system says he's ready to replace radiologists with AI

Marty Stempniak | *March 31, 2026* | *Radiology Business* | *Artificial Intelligence*



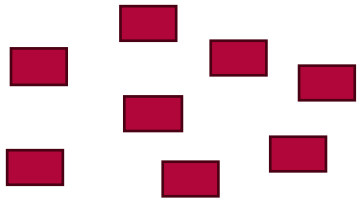
The chief executive of America's largest public hospital system says he is prepared to start replacing radiologists with artificial intelligence in some circumstances, once the regulatory landscape catches up.

Mitchell H. Katz, MD, president and CEO of NYC Health + Hospitals, recently spoke during a panel discussion held by Crain's New York Business. The trained internal medicine specialist noted how AI is increasingly being used to interpret mammograms and X-rays.

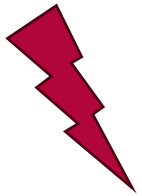
[6]

Medical Imaging Poisoning

200-400
poisoned images



Dataset: 100,000 –
1 Mio. images



65%-85%
success rate

[7]

Heterogeneus web-scale datasets for training [8]

Further Risks

Biased ML models: contain systematic errors leading to inaccurate/discriminatory predictions

Unfair prioritization of surgery plans [7]

Manipulation of organ transplantation Order [7]

Sybil attack: Massively forged identities

Sybil Attack on patient data [7]

Maybe not with malicious intent?

- Save own medical data against big tech companies
- “right to privacy includes the right to change privacy preferences” [9]

- Collateral damage argument [9]

  +

  privacy-oriented attacks are often reactions and responses to prior misuse rather than unprovoked attacks [9]

Legality Of Data Poisoning

- Not much information
- EU AI Act Article 15 Paragraph 5:
  The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), …
- Recommendation of strict audition of data handling [10]
- § 303a (StGB) Datenveränderung: unlawful modification (personal data unlawful?)
- § 303b (StGB) Computersabotage:
  - „Datenverarbeitungsanlage […] zerstört, beschädigt, unbrauchbar…“
  - „in der Absicht, einem anderen Nachteil zuzufügen…“

Mostly provider side

Protection against Protection?

- Changes in feature space not really detectable
- Images always have noise

Discussion Questions

1. Do you think artists should protect their art with those tools?

2. Is Data Poisoning in medical data justified? How could we overcome that problem?

3. Do we even need human art anymore?

References - 1

[1] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, “Glaze: Protecting artists from style mimicry by text-to-image models,” in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA, USA, 2023, pp. 2187–2204.

[2] B. Zhao, “Can we authenticate human creativity?,” TED, 2024. [Online]. Available: TED Talk page. [Accessed: May 13, 2026].

[3] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, “How to backdoor diffusion models?,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 4015–4024.

[4] S. Zhai, Y. Dong, Q. Shen, S. Pu, Y. Fang, and H. Su, “Text-to-image diffusion models can be easily backdoored through multimodal data poisoning,” in *Proceedings of the 31st ACM International Conference on Multimedia (MM ’23)*, Ottawa, ON, Canada, 2023, pp. 1577–1587, doi: 10.1145/3581783.3612108.

[5] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng and B. Y. Zhao, "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," *2024 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2024, pp. 807-825, doi: 10.1109/SP54263.2024.00207.

References - 2

[6] M. Stempniak, "CEO of America's largest public hospital system says he's ready to replace radiologists with AI," *Radiology Business*, Mar. 31, 2026. [Online]. Available: Radiology Business article. [Accessed: May 13, 2026].

[7] F. Abtahi, F. Seoane, I. Pau, and M. Vega-Barbas, "Data poisoning vulnerabilities across health care artificial intelligence architectures: Analytical security framework and defense strategies," *Journal of Medical Internet Research*, vol. 28, p. e87969, 2026, doi: 10.2196/87969. Available: https://www.jmir.org/2026/1/e87969

[8] D. A. Alber, Z. Yang, A. Alyakin, *et al*., "Medical large language models are vulnerable to data-poisoning attacks," *Nature Medicine*, vol. 31, pp. 618–626, 2025, doi: 10.1038/s41591-024-03445-1.

[9] L. Adomaitis and R. Oak, "Ethics of adversarial machine learning and data poisoning," *Discover Internet of Things*, vol. 2, no. 8, 2023, doi: 10.1007/s44206-023-00039-1.

[10] S. Kenshall and G. Mercer, "Legal perspectives on AI data poisoning," Penningtons Manches Cooper, Mar. 24, 2026. [Online]. Available: https://www.penningtonslaw.com/insights/legal-perspectives-on-ai-data-poisoning/. [Accessed: May 13, 2026].

References - 3

Images:

- https://static.vecteezy.com/system/resources/previews/057/813/826/non_2x/neural-network-diagram-nlp-line-icon-illustration-vector.jpg
- https://www.magnific.com/icon/economic-crisis_8318167
- https://jugendhilfeportal.de/fileadmin/_processed_/d/b/csm_Hand-Megafon-Tafel-Fake-News-F-fotogestoeber_21fb9df8e7.jpg
- https://www.stiftung-gesundheitswissen.de/sites/default/files/styles/w1180/sgw/system/files/2023-01/Röntgen_Oberkörper_Teaser.jpg.webp?itok=vXIfQPQF