The design of Nightshade plays into an air of mystique carried by most AI systems.[1] Its purple and green color scheme, skull logo, and name itself all make reference to poison in the fantastical sense. The black-box nature of complex programs is made black-cloak, seeming a magical solution, a force field protecting an image from the nefarious eyes of AI. And, as with any new technology in the AI culture wars, responses and discussions of Nightshade that I came across online were highly polarized–is it a benevolent protector or a scam, or a malicious technology?

Downloading Nightshade was not as simple as one would hope, for a tool whose efficacy is proportional to the size of its user base. The website is relatively well organized, although it hasn't been updated much since the original release 15 months ago. Glaze, their first project, and Nightshade, are both available to download for free. Glaze is also available online as Webglaze, although one has to email or message the team via social media to get an invite. In their recent posts on Bluesky, the team apologized for being behind on Webglaze requests.

I am the owner of a 2020 Macbook Pro with an Intel chip, meaning I only have access to the Glaze download, not Nightshade–the application package is for Windows or newer M-series Macs only. Unfortunately, Glaze does download but does not open on my computer, a problem that others have raised as well. So in order to test out Nightshade, a friend graciously lent me her old Windows laptop. It's a complicated program, and it takes a while to download all the necessary packages.[2] One also has to have an NVidia graphics card–Nightshade runs on GPU–and install a separate package in order to free up the necessary gigabytes of memory. The Nightshade interface is simple, with a text box to tag the image, and two sliders to the right to choose low, medium, or high poison intensity and image processing. The highest setting warns me that it takes 180 minutes.

After borrowing a computer, downloading Nightshade, updating NVidia, and finally running my own images through, it was surprising and gratifying to see the final output of the program. I've seen some strong opinions about shaded images online, some saying that the level of pixel disturbance is absolutely unacceptable–a form of righteous outrage I saw more often from those who were not, themselves, artists–and others saying it's barely noticeable. At the lowest intensity, I can still see the perturbations, but it's far from unreasonable–a small price to pay for their functionality. The appearance is not unlike JPEG artifacts at a glace. The pattern adds a subtle texture across the image, round, repeating shapes that not only act as camouflage

[1] I will take this opportunity to clarify some relevant terminology. Nightshade is described by the lab as a prompt-specific poisoning attack and as an algorithm, referring to its purpose and its method respectively. In the Nightshade User Guide, they note that Nightshade requires a number of pretrained models and machine learning libraries. I won't, for the sake of accuracy and consistency, be referring to Nightshade as a model, AI, or artificial neural network (ANN)–I am not sure it would be accurate to do so–and so instead will simply be calling it a program.

[2] This includes trained model(s) and machine learning packages, which are notoriously large

but resemble it. At the highest intensity, the colorful perturbations ripple across like a gentle oil slick atop the image, and they are easily spotted, but unidentifiable to anyone unfamiliar.

Nightshade was released in January 2024 by a team of researchers at the University of Chicago. The team describes their aims: "The Glaze Project (including Glaze, Nightshade, WebGlaze and others) is a research effort that develops technical tools with the explicit goal of protecting human creatives against invasive uses of generative artificial intelligence or GenAI." The Glaze Project has created two different tools which interfere with training text-to-image diffusion models: Glaze and Nightshade. Glaze is a defensive tool for artists to use to protect against style mimicry, and is designed for smaller scale models which may be trained deliberately on a particular artist's portfolio in order to generate artworks in their style for free. Nightshade, on the other hand, is designed for a broad-scale offensive–it "poisons" images and disrupts models functionality. Ben Zhao, computer science professor and project lead, meets often to speak with the press about The Glaze Project. In an interview with Steven Dubner for the podcast *Freakonomics,* Zhao describes the aim of Nightshade: "not to break models, but to increase the cost of training on unlicensed data [...] In that scenario, the outcome would be licensing so that [artists] can actually maintain a livelihood and maintain the vibrancy of that industry."[3] On his motivation for protecting artists in particular, he says: "Art is interesting when it has intention, when there's meaning and context. So when A.I. tries to replace that, it has no context and meaning."[4]

How does it work?[5]

This section relies on the paper published by the SAND lab at the University of Chicago to describe the design, function, and outcomes of Nightshade. I intend to provide an in-depth overview in order to elucidate the mechanisms which will be discussed in further detail in subsequent sections–however, for the sake of concision, some key components will necessarily be omitted if they are less relevant to the points I intend to make.[6]

Images which have been run through Nightshade are referred to as "shaded," or "poisoned." In order to make shaded images as effective as possible–capable of corrupting models with large training datasets with relatively few poisoned images–the research team focused on two goals. First, to use targeted attacks which retrain a model to reattribute the meaning of one image category to a different image category, such as producing images of cats when prompted with text "dog." Second, to make shaded images undetectable by either automated or human detection.

While latent diffusion models are trained on huge data sets–up to 2 million images[7]–individual concepts (e.g., dog, house, bird) are represented sparsely. 90% of concepts in each

[3] freakonomics
[4] I'll return to this quote later..
[5] Nightshade lab paper: i really don't know how best to cite all this bc I am really just summarizing this paper here
[6] Will come back and cut what isn't necessary later once I know
[7] SDXL

analysis linked to less than 0.2% of the training data.[8] By making poison attacks highly targeted, there is a substantially smaller amount of training data to work against. Because of this, a model can be disrupted with far fewer images if they are highly targeted to a concept. Further, concepts are "semantically linked" to others; for example, "dog" to "leash" or "wolf." Nightshade exploits concept sparcity and semantic connections to make highly potent poison: it only takes 50-100 poisoned images to disrupt a trained latent diffusion model, with the strongest effect on the target concept and noticeable disturbances to semantically linked concepts. Each target concept is matched with a specific "destination" at random. In their paper, some of the examples they used were poisoned concepts dog, car, handbag, and hat, with their destination concepts cat, cow, toaster, and cake respectively. Because each poisoned concept is matched 1:1 with a destination, the poison is highly effective in leading models in a specific direction. Finally, the Nightshade model uses image generation in order to generate the ideal versions of destination concepts. Most images are relatively poor representations of their associated concept, which is necessary to prevent overfitting and increase model flexibility.[9] However, because shaded images, as seen by AI[10], appear to be a perfect example of the destination concept, they have a greater influence on the model. Using image generators' own outputs also reduces variation and inconsistency when creating a set of images for a given concept. The strategic targeting and potency of attack allow the "perturbation"–the resulting changes to an image made by the Nightshade algorithm–to be low enough to evade human or algorithmic scrutiny.

Nightshade's efficacy was assessed by continuous training (on existing models such as Stable Diffusion) and from-scratch, on a model created by the research team. Their results showed that Nightshade is highly effective even with a small number of poisoned images. Assessment was performed using the image classifier CLIP and human ratings; it took only 100 poisoned images for the classifier to identify the output of the target concept as an image of the destination concept greater than 80% of the time. Additionally, participants noted that images which were correct for the target concept were still "incoherent." Researchers then asked the participants to rate the usability of these images, with only 40% being usable after 25 poisoned samples and 20% after 50. Poison attacks also "bleed through to nearby concepts." A model which has had the concept "dog" poisoned will also output images of the destination concept "cat" when prompted with "puppy," "husky," or "wolf," with the strength of the effect on the output image decreasing with further distance from the target concept.[11]

As Nightshade is intended to be available to a variety of Internet users, poisoned images will attack multiple concepts at once. The researchers experimented on how a composed attack–multiple people attacking unrelated concepts–would affect a model. They randomly sampled a number of concepts from the training data and attacked each concept with 100 poison samples. Even with just 100 poisoned concepts, the images generated by the model become not only inaccurate to the text prompt but overall less comprehensible. By 500 poisoned concepts,

---

[8] Nightshade lab paper
[9] Gotta get real info on that
[10] Specifically, the VAE *** more on this later
[11] 6.4 Nightshade

the output is reduced to complete noise. They speculate that this is because misaligned data across numerous concepts "increases the difficulty of learning text-image alignment in the model and corrupts the cross-attention layer."[12]

Finally, the authors describe the intended utility of Nightshade poison as copyright protection. Alternatives to preventing intellectual property theft, such as opt-out lists, are unenforceable. The effect of Nightshade poison, on the other hand, directly disrupts model training and efficacy. A.I. companies that rely on scraped images to train their models are forced to confront the repercussions of using unlicensed data–or, even if they are able to filter out poisoned images, Nightshade has effectively protected the owner said image from having their intellectual property used as training data.

Leaving the lab paper

To further explore the functionality of Nightshade I return to the structure of latent diffusion models. A latent diffusion model works in the feature space of an image, not the pixel space; it sees and creates images in terms of their features. A variational autoencoder (VAE) is used in order to create the "latent feature space," which is the breakdown of an image into its features. Nightshade exploits this process in order to generate its poison. The algorithm Nightshade runs ensures the fewest alterations to the actual pixels of an image–the pixel space being what human eyes see–while shifting the features from the initial concept towards the target poisoned concept.

+ To come: Language model and diffusion process

Queering

In Queer Phenomenology, Sara Ahmed describes heterosexuality and queerness in terms of directional orientation. To describe sexuality in terms of its spatial existence, the normative heterosexual desire is oriented in a straight line towards the other sex, while the queer, "bent" desire is oriented in a curved line towards itself. In the context of diffusion models, the straight or normative line would be between the sign and referent as represented in the training data. It is not a curved line towards itself, even though conceptually the sign "dog" and the image of a Pomeranian are one and the same.

From Pattern Discrimination, Wendy Chun's "Queering Homophily" describes the underlying structure of networks as reliant on homophily, love of the same. In conversations about bias in AI, there is little consideration given to the ways that the very mathematical operations performed, not only the existing data that reflects prejudicial histories, continue to perpetuate exclusion and normalization. The initial conditions for homophily are not random, nor are its consequences. ""Love of the same" is never innocent", "Hate transforms the particular into the general: it transforms individuals into types so they become a common threat." Whether or not we personify models, it is fundamental to consider the notion that text-to-image produces

[12] Once i figure out what this means I'll get more into it

stereotype. In “Queering Homophily,” from Ahmed–queerness as an inability to be comfortable in certain norms, and the generative power of discomfort.

Attacking one concept via Nightshade successfully convinces the targeted diffusion model of a clear, but wrong, sign-referent pairing. The text “dog” now produces a Maine Coon. The model has been disrupted, yes, but it is not yet broken. The model is otherwise functional; the queer has been assimilated, even if wrongly. The dissolution is yet to come. Once 50, 100, of a diffusion model’s concepts are attacked, it can no longer produce any images whatsoever. Its data scrambled, all that remains is noise.

The queer liberation movement can be seen allegorically in the operation of Nightshade poison. With enough instances, sufficiently matching existing ideal representations within a system, queerness can and will be assimilated. Gay marriage can be institutionalized within the existing framework of marriage–and all its baggage–now with a different referent, but the same text, sign, structure. A model which has been completely poisoned, but only at one concept, will degrade slightly, its overall efficacy slightly worsened, but little else will change. But when many concepts are attacked, the very links between sign and referent cease to be meaningful, and only random noise remains.

An invading enemy is not only a description of its function but of its form. The perturbations laying on the surface of a poisoned image not only act as but resemble camouflage, layers of small, curved lines and shapes.

Nightshade perverts the image, taking the directional relationship–the normative referent-sign pairing–and bending it towards another ideal. The poison is made effective by the use of the diffusion model against itself.

The data that Nightshade uses as the basis for its poison is produced by its victims: the text-to-image diffusion models. In order for the poison to be as effective as possible, the model must consider the image destination concept that it sees to be the ideal representation of that concept. As the lab describes in their paper, there is typically a significant variation within a concept, where each image is fairly weakly representative of the ideal. Variety is typically a strength, given that a model would become more adept at recognizing referent images as examples of a sign even if they are not identical to a particular instance of that sign. However, once it has been trained, the referent image that a diffusion model produces will be representative of the ideal instance of that sign, to the best of the model’s understanding. Thus, using its own productions is far more efficient, as they match more closely its existing understanding of the referent.

The definition of queering in terms of space is particularly apt given the terminology relevant to latent diffusion and image generation. As mentioned previously, Nightshade exploits latent diffusion models transformation into feature space. There is an alternate reality in which these programs work, largely invisible to the human eye. Because latent diffusion models see only the feature space, not the pixel space, they can be convinced that an image which is plainly obviously a cat is in fact a much closer match with all the features of a dog. There are a number of directions from which to take this functionality. One, to connect back to spatial orientation with Sara Ahmed. Two, to explore how the black box becomes a technological purgatory in which algorithms fight beyond human eyes. Three, to understand that a poison attack is always a double edged sword, in that it outlines an exploit which can then be targeted and mended.

- https://news.ycombinator.com/item?id=37211519: Improving Stable diffusion on negative models – telling it what not to do
- https://www.reddit.com/r/StableDiffusion/comments/1ag5h5s/the_vae_used_for_stable_diffusion_1x2x_and_other/: Finding flaw in SD while using Nightshade
    - From same reddit thread: "I am one of the creators of DALLE 3, we knew about this. :) Another problem (and dead giveaway that this VAE has global information issues) is that the latent space becomes invalid if flipped across any axis. Thanks for putting together this report! Great investigation!"
    -

- **Here's how we can better understand nightshade/adversarial models,** using [analogy/queer theory] [network/communications theory]
- Designed to protect capital but it's more radical; breaking down concepts – pharmakon of not always wanting to completely deconstruct? – do we accept gay marriage or burn down marriage

Defining queering

- "As Tim Dean and Christopher Lane argue, queer theory 'advocates a politics based on resistance to all norms' (Dear and Lane 2001: 7)." (Ahmed, Queer Feelings, 149)
    - "Importantly, heteronormativity refers to more than simply the presumption that it is normal to be heterosexual. The 'norm' is regulative, and is supported by an 'ideal' that associates sexual conduct with other forms of conduct." (Ahmed, Queer Feelings, 149)

Spacial queering

- Phenomenology helps us to consider how sexuality involves ways of inhabiting and being inhabited by space. While same-sex desire has the attributes of heterosexual desire, it moves toward an object that is 'normally outside the sphere' of that desire. In other words, it reaches objects that are not continuous with the line of normal sexual subjectivity." (Ahmed, Queer Phenomenology, 71)

Queer effects, norms

- "Queer lives remain shaped by that which they fail to reproduce. To turn this around, queer lives shape what gets reproduced: in the very failure to reproduce the norms through how they inhabit them, queer lives produce different effects." (Ahmed, Queer Feelings, 152)
-

Homophily

- "Homophily launders hate into collective love, a transformation that, as Sara Ahmed has shown, grounds modern white supremacism (2004, 123). Homophily reveals and creates boundaries within theoretically flat and diffuse networks; it distinguishes and discriminates between allegedly equal nodes: it is a tool for discovering bias and inequality and for perpetuating it in the name of "comfort," predictability, and common sense." (62)
- "We must thus embrace network analyses and work with network scientists to create new algorithms, new hypotheses, new grounding axioms. We also need to reembrace critical theory: feminism, ethnic studies, deconstruction, and yes, even psychoanalysis, data analytics' repressed parent. Most crucially, what everyone needs now: training in critical ethnic studies." (62)

Galloway notes; explaining tech

Images

Low intensity Nightshaded image

Skelton, Sol. *All Hail the Machine*. 2024. Acrylic on canvas, 50 inches × 50 inches (130 × 130 centimeters). My house.

High intensity Nightshaded image

Skelton, Sol. *All Hail the Machine*. 2024. Acrylic on canvas, 50 inches × 50 inches (130 × 130 centimeters). My house.