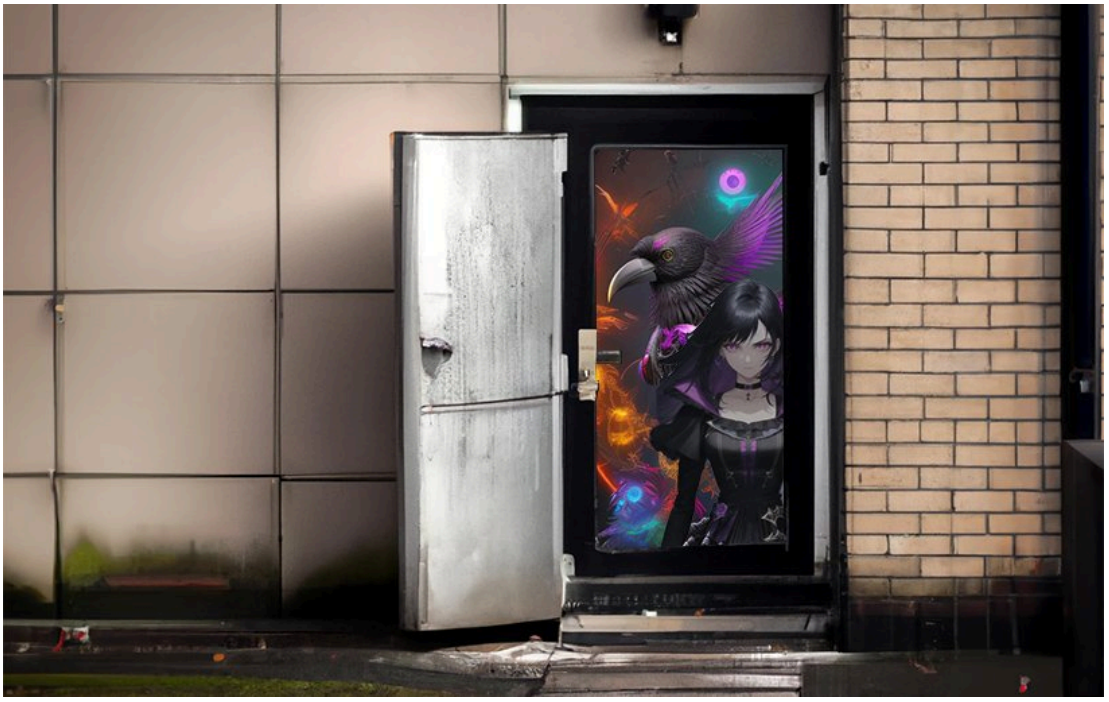
A Call to Legislate ‘Backdoors’ Into Stable Diffusion

Originally published at https://arquivo.pt/wayback/20240203184843oe_/https://blog.metaphysic.ai/a-call-to-legislate-backdoors-into-stable-diffusion/

February 14, 2023



A new research paper from MIT has proposed that controls be imposed upon the creators of generative image systems such as Stable Diffusion, so that these companies work in partnership with regulatory authorities to ensure that their applications can’t be used to create non-consensual deepfake images, or to otherwise arbitrarily alter existing photographs.

The paper advocates the imposition of ‘backdoors’ into new generations of latent diffusion models, so that effective methods of making image data resistant to such systems will not become *ineffective* as the architectures evolve (a key criticism against the slew of tech-based anti-deepfake proposals that have emerged since video deepfakes burst on the scene in late 2017).

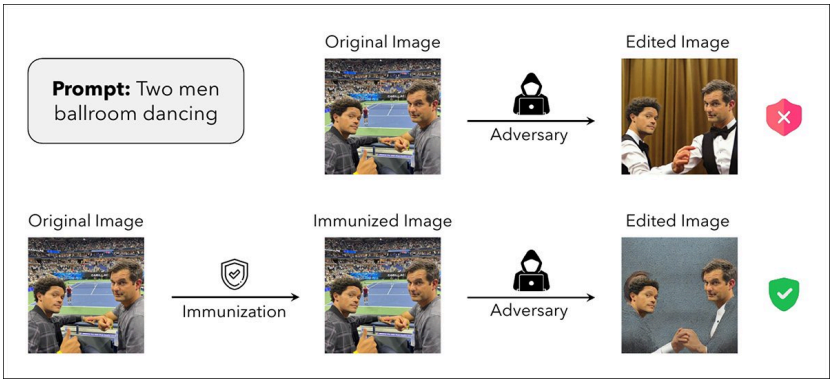
The new work states:

‘[We] thus need to go beyond purely technical methods and encourage — or compel — via policy means a collaboration between organizations that develop large diffusion models, end-users, as well as data hosting and dissemination platforms. Specifically, this collaboration would involve the developers providing APIs that allow the users and platforms to immunize their images against manipulation by the diffusion models the developers create.

‘Importantly, these APIs should guarantee “forward compatibility”, i.e., effectiveness of the offered immunization against models developed in the future. This can be accomplished by planting, when training such future models, the current immunizing adversarial perturbations as backdoors.’

The section referring to ‘forward compatibility’ is advocating that any anti-deepfake/abuse measures adopted by companies (or imposed on them) should be ‘future-proofed’, so that innovations in the system architectures do not nullify their effectiveness.

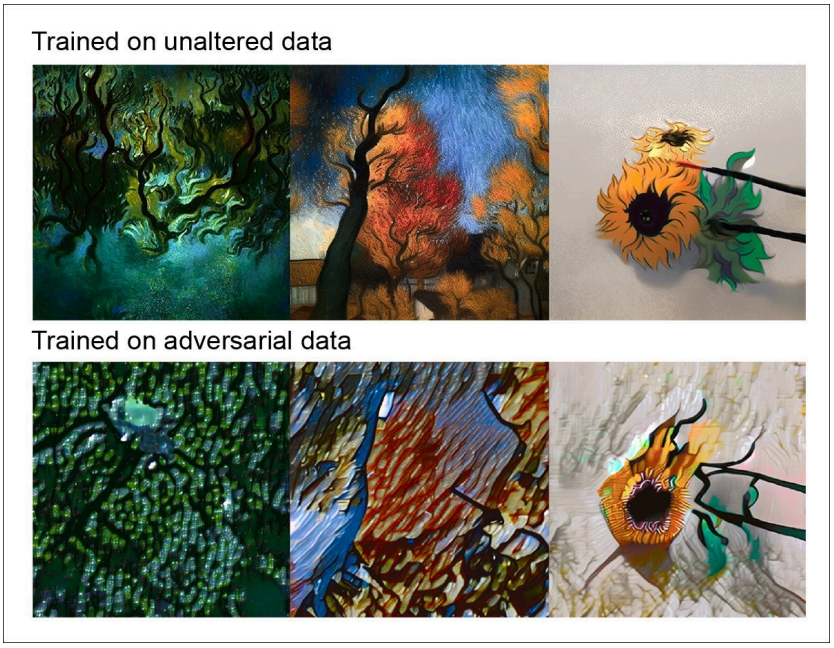
In fact, these proposals come in the context of the researchers’ own new system to ‘poison’ source images so that they cannot be ‘abused’ by Stable Diffusion.



MIT's new method offers two ways to 'immunize' images against being included in Stable Diffusion's transformative processes, both variants on traditional adversarial attacks. In this case we see a source image being 'associated' with a pure grey image through the method, so that image-to-image transformations fail. Source: https://arxiv.org/pdf/2302.06588.pdf

In this regard, the ground is well-trodden, both in terms of adversarial attack methodologies dating back some years, which have used ‘perturbed’ images to attempt to disrupt recognition and image generation processes in older systems such as Generative Adversarial Networks (GANs) and

autoencoder (i.e., traditional deepfake systems); and because a rival academic collaboration from China, the UK and the US pipped MIT to the post last Thursday, with a claim to the first adversarial attack system targeting Stable Diffusion.



Last Thursday's release of an academic collaboration between China, the UK and the US offered what the authors claimed is the first adversarial 'poisoning' system that's effective against latent diffusion systems. In this example, we see that generative outputs trained on 'clean' data are able to imitate the style of Van Gogh, whereas the mode trained on subtly-perturbed versions of original Van Gogh paintings is unable to replicate the artist's style. Source: http://arxiv.org/pdf/2302.04578

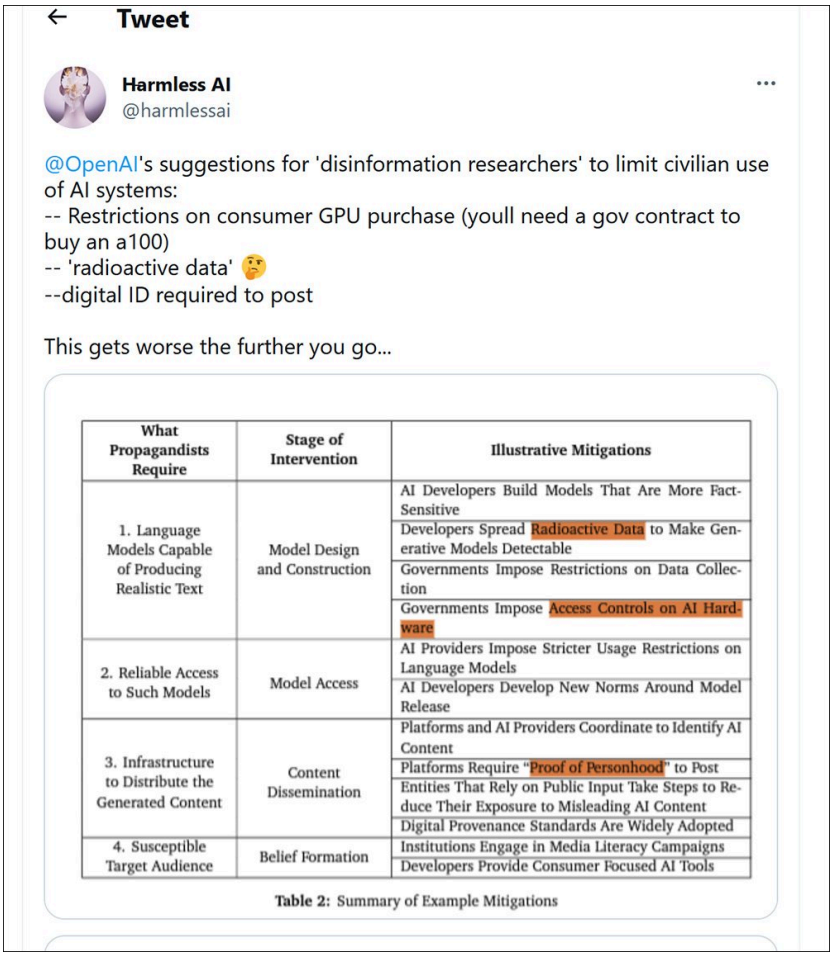
What is novel is that, unlike the other paper, the MIT submission acknowledges that the playing field within which such a technology would be deployed needs to be made *less level and equitable* in order to gain any meaningful control over these new developments; and that purely algorithmic approaches to protective measures will only likely lead to the same kind of tacit 'cold war' between researchers and deepfake developers that has characterized security research into autoencoder image synthesis systems in recent years.

The new paper is aptly titled *Raising the Cost of Malicious AI-Powered Image Editing*, and comes from five MIT researchers.

If You Can't Win, Change the Rules

The new paper emerges just as a January 10th submission from Georgetown University, Stanford and OpenAI is beginning to make waves, regarding the possible need to reign in casual consumer access to the new breed of AI generative systems.

The January paper posits that it may be necessary to place restrictions on casual access to powerful GPUs (perhaps requiring a government contract to purchase a graphics card that's above-averagely specced); that distribution of generated content could perhaps require 'real person', pre-authenticated status, ending casual posting, and falling more into line with the control that China exerts in this regard; and that the ability to disable non-perceptible watermarks in generative systems be removed, so that the output from frameworks like Stable Diffusion is 'radioactive', and clearly 'non-real'.



What Propagandists Require	Stage of Intervention	Illustrative Mitigations
1. Language Models Capable of Producing Realistic Text	Model Design and Construction	AI Developers Build Models That Are More Fact-Sensitive
		Developers Spread Radioactive Data to Make Generative Models Detectable
		Governments Impose Restrictions on Data Collection
		Governments Impose Access Controls on AI Hardware
2. Reliable Access to Such Models	Model Access	AI Providers Impose Stricter Usage Restrictions on Language Models
		AI Developers Develop New Norms Around Model Release
3. Infrastructure to Distribute the Generated Content	Content Dissemination	Platforms and AI Providers Coordinate to Identify AI Content
		Platforms Require "Proof of Personhood" to Post
		Entities That Rely on Public Input Take Steps to Reduce Their Exposure to Misleading AI Content
		Digital Provenance Standards Are Widely Adopted
4. Susceptible Target Audience	Belief Formation	Institutions Engage in Media Literacy Campaigns
		Developers Provide Consumer Focused AI Tools

Table 2: Summary of Example Mitigations

The GU/Stanford/OpenAI paper from January 2023 is beginning to cause ripples, with suggestions that practical consumer-level limits on AI access may need to be applied in order to put the genie at least partially back in the bottle. Source: https://twitter.com/harmlessai/status/1624617240225288194

This new clamor for legislative and platform-based control is arguably the third wave of thought around the dissemination of AI-generated content; in the first, the variable quality of the output made the issue largely moot, with calls to 'common sense' and general restraint seen as adequate; in the second, concerned users and lobby groups began to appeal to governments to regulate AI content (which is happening, to a certain extent).

But this latest, and perhaps most desperate approach acknowledges that laws would be only partially effective where the means to generate such content remains unrestrained; and that it will instead be necessary to limit and perhaps *proscribe* access to generative infrastructure, in the event that natural market forces and the growth of hyperscale AI don't automatically and more naturally limit consumers to less-effective and less convincing generative systems (which is a distinct possibility).

The Seeds of Technical Debt..?

One consideration that is not addressed in the new MIT paper is the extent to which supporting older technologies can be a limiting factor in the ongoing development of systems. In effect, the development team is either constrained in this scenario from radical refactoring of the code-base, or will be forced to support 'legacy' code by some other means, such as a dedicated sub-system that replicates that code's functionality.

An example of this is that the original early 1980s DOS system was a pivotal layer in the Windows operating system until the advent of Windows NT, and the subsequent generation of Windows versions that were based on it (Windows 2000/XP/7/8, et al.).

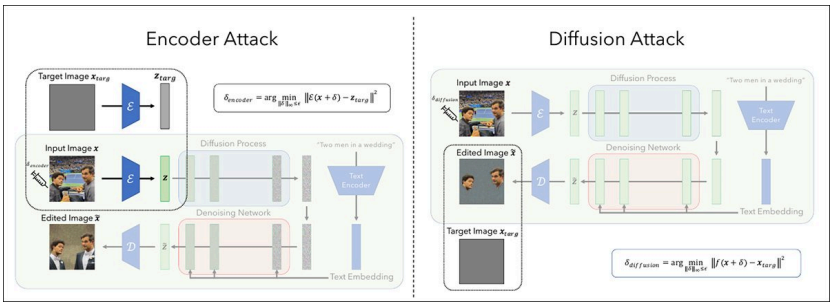
When DOS was foundational to Windows (i.e., Windows 95/98/ME), a DOS error would take the whole system down with a 'blue screen of death'. Now, DOS is supported more as an application layer in the NT foundation, and its troubles (if any) are quite remote from those of the host operating system.

Likewise, developers of generative systems would, under the proposals suggested by the new MIT paper, be obliged to continually re-enable the central hooks that make any legacy data-poisoning system actually work – even if that system is relying on facets of the architecture (such as gradient predictions) that improved host environments or code innovations could make redundant as the framework evolves.

As such, this requirement could potentially prove a drag on innovation, not least because the increasingly sophisticated metrics that evaluate the quality of images as a central function of the system are likely to penalize the subtly-perturbed images that have been adversarially-altered to (as the MIT researchers put it) 'immunize' them against ingestion into generative AI workflows.

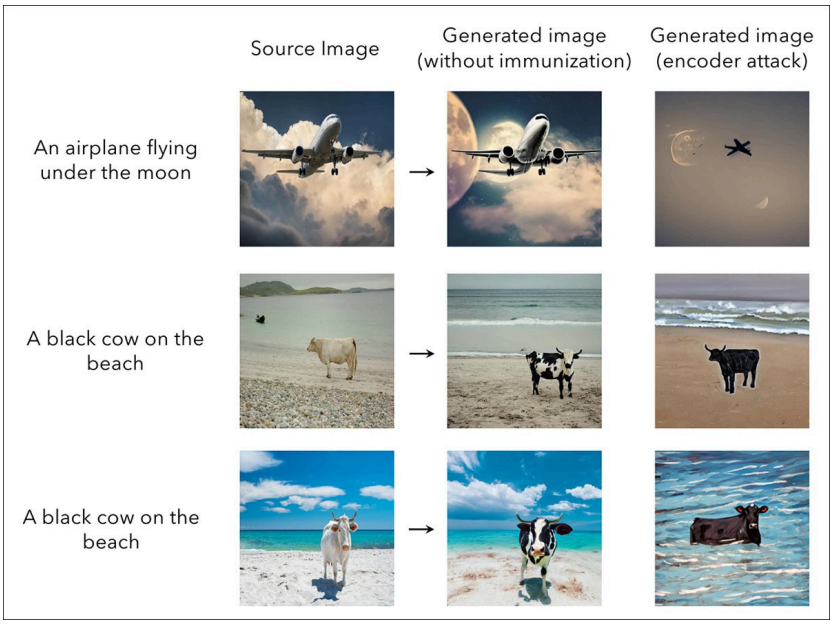
Approach

Unlike last week's offering, the MIT paper offers *two* approaches to creating images that are resistant to training and to effective use in the various generative capacities of Stable Diffusion (see yesterday's coverage of the prior paper for an outline of the general issues at hand here): an *encoder attack* and a *diffusion attack*.



The differing approaches of the encoder and diffusion attacks.

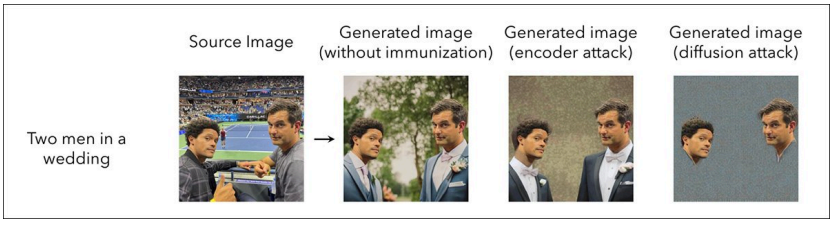
Regarding the first of these, a latent diffusion model initially encodes an incoming image into a latent vector representation, which is subsequently used to generate the user-prompted image. The encoder attack in MIT's new paper uses projected gradient descent to force the encoder to map the image to a non-apposite, essentially 'destructive' image.



Prompt-generated variations based on 'immunized' images will produce degraded output.

This is effected by adding the solutions to a fundamental optimization problem into the very fabric of the image itself, causing a kind of unhelpful 'feedback loop' that will not allow for high-quality image generation.

The diffusion attack is more complicated, but can be effective in conditions where the encoder attack may fail. Here, the system is forced to actively ignore a user's text-prompt (such as in Stable Diffusion's image-to-image work-flow) by incorporating into the image components that will map specifically to an 'unhelpful' and unrelated image, such as a square of pure grey.



In the image above, from the paper, we can see the encoder attack failing to entirely prevent synthesis of a 'protected' image (second from right), whereas the diffusion attack (right) successfully maps the content to the 'target' grey square that's been coded into the original image.

Naturally, this latter approach is architecture-dependent, and not likely to be resistant against notable changes in the latent diffusion model's methodology, and this seems to be the reason that the new paper is advocating strongly for a more 'locked down' ecosystem around generative AI.

Tests

The researchers conducted quantitative tests, to see if their system could effectively impede Stable Diffusion from utilizing immunized images in typical workflows. The metrics used were Fréchet Inception Distance (FID), Precision and Recall (PR), Structural Similarity Index (SSIM), Peak signal-to-noise ratio (PSNR), VIFp and FSIM.

Method	FID ↓	PR ↑	SSIM ↑	PSNR ↑	VIFp ↑	FSIM ↑
Immunization baseline (Random noise)	82.57	1.00	0.75 ± 0.13	19.21 ± 4.00	0.43 ± 0.13	0.83 ± 0.08
Immunization (Encoder attack)	130.6	0.95	0.58 ± 0.11	14.91 ± 2.78	0.30 ± 0.10	0.73 ± 0.08
Immunization (Diffusion attack)	**167.6**	**0.87**	$\mathbf{0.50 \pm 0.09}$	$\mathbf{13.58 \pm 2.23}$	$\mathbf{0.24 \pm 0.09}$	$\mathbf{0.69 \pm 0.06}$

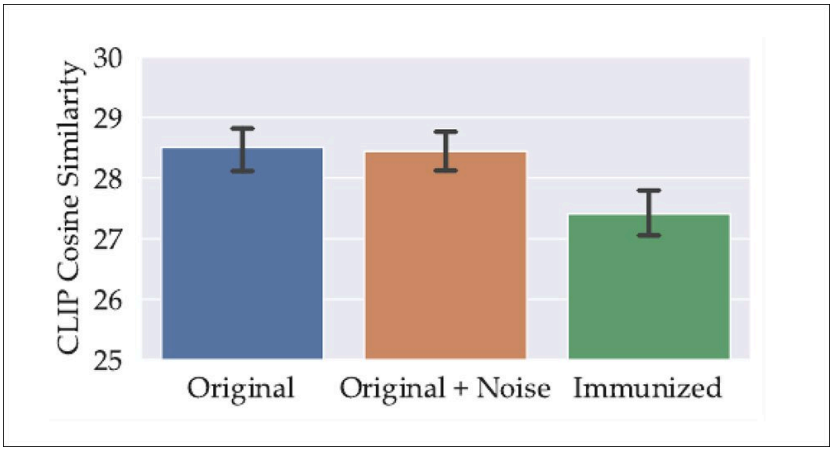
Results from the quantitative tests. As a baseline, the researchers also tested a 'naïve immunization method that simply added random noise (uppermost row in the results).

Of these results, the researchers state:

'The similarity scores…indicate that applying either of our immunization methods (encoder or diffusion attacks) indeed yields edits that are different from those of non-immunized images (since, for example, FID is far from zero for both of these methods).'

They also tested for image-prompt similarity, evaluating how closely images generated from 'immunized' data accorded with the text-prompt used. For this, they used an OpenAI pre-trained CLIP model, together with the employed text-prompts, to extract the cosine similarity between a non-immunized and an immunized embedding.

As per expectation, by this metric, the similarity was notably decreased between 'clean' and immunized generations:



CLIP-derived similarity metrics.

The Need for a Rigged Game

Perhaps aware that the need to change the generative AI landscape in order to accommodate such brittle protective methods will be seen askance by many, the authors note that their method can potentially be applied also to models that are already released.

For Stable Diffusion, this is obvious, since this was the target system for the study; presumably, however, the approach could either be adapted to other latent diffusion systems than SD (where available, and not many are), or else, more in the spirit of the work, the systems themselves could be 'updated' to integrate perturbation-based immunization.

Unlike last week's paper, the new offering does not claim to be a method of preventing images from being *trained* into generative systems (which the authors consider a Quixotic pursuit, citing prior work on this), but rather a way of stopping users from freely interpreting any and all web-available data that they may care to 'reinterpret' creatively.

As we mentioned in yesterday's write-up of last week's paper, this approach too would require that images be pre-processed, so that they contain adversarial information before being disseminated, and that this brings up a number of logistical issues.

However, the tone of the new work suggests that, in any case, arguably Draconian changes to the way images are shared and AI systems developed and disseminated will need to occur if the genie is ever to be put back in the bottle. Perhaps the re-encoding of 'non-immunized' legacy web material is part of such a scenario.