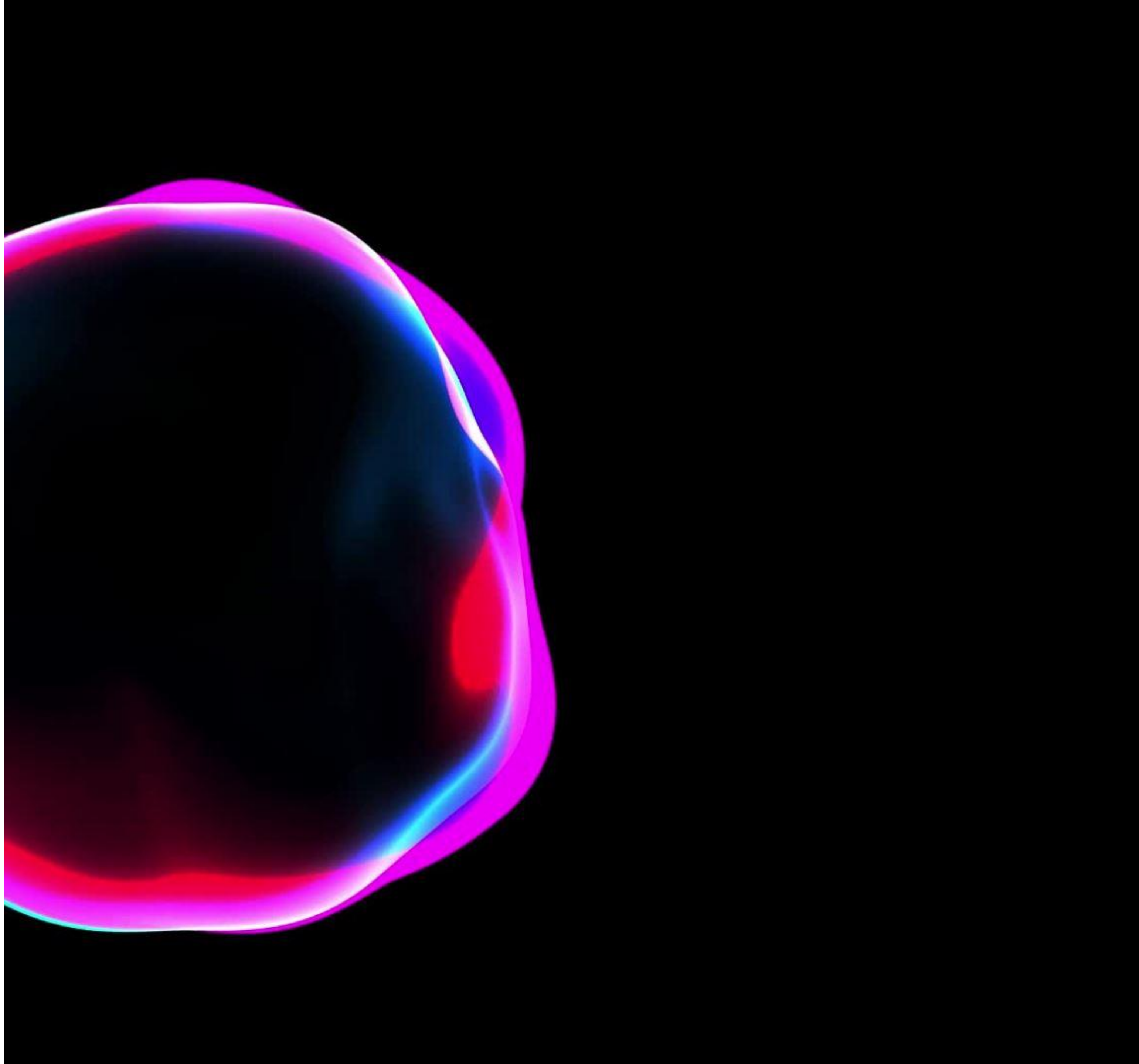
PROTECT DIGITAL ART FROM DIFFUSION

Michele Mora



KNOW THINE ENEMY:
WHAT IS AI ART?

- Art generated through artificial intelligence machine learning.

WHAT IS THE PROBLEM WITH AI ART?

- AI art is created with information learned from preexisting art and images of art. This process is called **diffusion**
- Diffusion has advanced to the point where generative AI can mimic the styles of artists
- Artists' works have been used without their permission to recreate their styles and projects uncredited and for free

THE ETHICS TOPIC

AND WHY IT WILL NOT BE HERE

- Many ethical and legal debates are currently being conducted as to whether AI art is ethical and if an artist's "style" can be legally protected (as quantifiable stylistic elements are difficult to conclude). This presentation will have nothing to do with this debate.

- The fact remains that artists do not want their works or styles to be replicable. Consequences and protections are a deterrent, not a preventor. A real demand for a preventative measure and protection against artistic works being used in diffusion for machine learning exists. This project explores a possible solution to this need.

WHO DOES THIS EFFECT?

- Everyone. -Art touches everyone's' lives. The nature of art in the world we live in impacts our own ideas and identities. Art echoes culture and speaks more into existence.
- Small businesses- Many small businesses construct their own logos, imagery, and designs to sell their products and stand out in the market. Saturation of graphic design styles, branding and products weakens one of the most important tools they have
- Traditional artists- Traditional artist showcase their work online through photographs and digital transfers. They are still at risk for missing out on work due to generative AI and mimicry.
- Digital artists- Digital artists often put out a lot of work for free to gain traction to commissions or other projects like comics and tutorials/art classes. As replicating their work/styles becomes more accessible, demand for their products and educational services lowers.

WHY I CARE

- Digital art has become important to the world. It is used in animation, video game and movie production, commercial use, drafting, engineering and design, and much more. It is also used by smaller indie artists for both personal enjoyment and profit. I have spent my life appreciating these projects. They mean a lot to me, and the artists that make them have put more happiness in my life than they will ever know. I am seeing less and less enthusiasm and more hopelessness in the art community since the advent of generative AI. AI is a great and powerful tool, but what will make it good or bad in any circumstance depends on how it is used. I think all the listed ways that AI art impacts hold equal precedence in the conversation. I am prioritizing digital art to hone the project and focus on what I care about on a personal level.

KEY TERMS/ CONCEPTS TO KEEP IN MIND

- **Diffusion**- the method of teaching AI image generators with pre-existing images
- **LAION Dataset** - “a collection of 5.6 billion images scraped, without permission, from the internet. Almost every digital artist has images in LAION, given that DeviantArt and ArtStation were lifted wholesale, along with Getty Images and Pinterest.” - Mattei, Escalante-De
- **Stability AI-** Company that creates open-source generative AI Models, best known for a text to image product called Stable Diffusion
- **Style mimicry-** the ability to replicate an artist’s style through stable diffusion-enabling the recreation of an image in a particular artist’s style.
- **Steganography-** Steganography is the practice of hiding a secret message inside of (or even on top of) something that is not secret.
- **Midjourney**- Prompt based generator of artificial images.
- **Scraping**- the automated process of extracting images from websites
- **Obfuscation-** the act of making something unclear or unintelligible

PRE-EXISTING SOLUTIONS

WATER MARKS AREN'T ENOUGH ANYMORE

- Opting out
- "Glaze"
- "Nightshade"
- Kudurru

"OPTING OUT" A USELESS FORMALITY

- Sites like DeviantArt and platforms like Meta have options to manually select “opting out” of user uploads being used for AI machine learning. OpenAI even has an option on the site to opt out of uploads being used to to train the DALL-E 3 open AI generator.

PROS

- Some level of prevention against diffusion
- Protects the platform from responsibility if the image is copied/shared elsewhere and then scraped (if you care about them –silent judgement-)

CONS

OH BOY HERE WE GO

- **Site/platform specific**
    - -opting out from one platform is not a universal solution
    - -images copies to other platforms are still at risk
    - opting out from images used for one machine will not guarantee no use by another
- **Tedious/Unapparent**
    - too much effort to “opt out” on every possible platform a user wants to upload to
    - ”opting out” usually requires deep navigation into the platform to find and use
- **Scrapers can bypass opt-outs**

GLAZE

A GREAT START

- "***Glaze*** is a system designed to protect human artists by disrupting style mimicry. At a high level, Glaze works by understanding the AI models that are training on human art, and using machine learning algorithms, computing a set of minimal changes to artworks, such that it appears unchanged to human eyes, but appears to AI models like a dramatically different art style. For example, human eyes might find a *glazed* charcoal portrait with a realism style to be unchanged, but an AI model might see the glazed version as a modern abstract style, a la Jackson Pollock. So when someone then prompts the model to generate art mimicking the charcoal artist, they will get something quite different from what they expected." -official site

PROS

- **Robust to change**
  - Resistant to redownloading, compression, smoothing effects
  - Not reliant on steganography
- **Not a watermark**
  - Operates as a dimension within the piece
  - Each use of glaze produces a unique dimension
  - Attack would have to know the specific dimensions
- **Accessibility**
  - Webglaze can be used on any phone, tablet, computer, or device with internet

CONS

"NOBODY'S PERFECT" -HANNAH MONTANNA

- **Effects on art**
  - Glaze can be somewhat detectable on art styles with flat colors and simple backgrounds
  - Not futureproof
  - Earlier versions of glaze have lost robustness against some image purifiers
  - The current version could also be overcome by new advancements
- **What's done is done**
  - Glaze can prevent style mimicry from pieces it is used on
  - Glaze cannot undo any mimicry already trained into base models (like SDL or SD3)

NIGHTSHADE

BROUGHT TO YOU BY THE CREATORS OF GLAZE

- "***Nightshade*** works similarly as Glaze, but instead of a defense against style mimicry, it is designed as an offense tool to distort feature representations inside generative AI image models. Like Glaze, Nightshade is computed as a multi-objective optimization that minimizes visible changes to the original image. While human eyes see a shaded image that is largely unchanged from the original, the AI model sees a dramatically different composition in the image." -official site

PROS

- **Offense**
  - Use by artists as a group attack mimicry ability by misleading scraping tools
- **Robust to change**
  - Resistant to redownloading, compression, smoothing effects
  - Not reliant on steganography
  - Not a watermark
- **Operates as a dimension within the piece**
  - Each use of glaze produces a unique dimension
  - Attack would have to know the specific dimensions

CONS

- **Effects on art**
  - Detectable on pieces with simple backgrounds and flat colors
  - Lower levels are possible but less robust
- **Not Future Proof**
  - No guarantee that Nightshade will work in the long term against attack

QUICK BONUS METHOD: KUDURRU

ANOTHER INTERESTING APPROACH

- "Kudurru monitors popular AI datasets for scraping behavior, and coordinates amongst the network to quickly identify scrapers. When a scraper is identified, its identity is broadcast to all protected Kudurru sites. All Kudurru sites then collectively block the scraper from downloading content from their respective host. When the scraper is finished, Kudurru informs the network and traffic is allowed to proceed as normal." -official site
- Pros
  - Detects scrapers
  - Works on multiple sites
- Cons
  - Still in beta testing
  - Membership required
  - Only so many sites are protected

WE CAN USE A LOT OF THESE IDEAS.
WHAT CAN WE IMPROVE ON?

- Accessibility
  - Having to use multiple processes is tedious and will discourage use
- Multiuse
  - Glaze and Nightshade are on the same site but are used separately.
- Effect on Art
  - Glaze can have an effect on the appearance of the art.
- What about batch processing?
  - Larger projects with files containing multiple pieces (ie. a comic/ animation) would be difficult to process one image at a time

PROPOSITION

- Make protecting pieces and style easier by creating software that can
    - Perform both tasks of defense against diffusion and attack to model learning.
        - See "Pseudo Code A" (slide 21) & " Pseudo Code B" (slide 22)
    - Further minimize changes to images
    - Make more useful for larger/ multi-image projects

PSEUDO CODE A

Pseudocode for Adversarial Image Perturbation Targeting CLIP (used in diffusion):

```
1. Load a pre-trained feature encoder used in diffusion models (e.g., CLIP image encoder)
2. Load the image you want to protect
3. Preprocess the image to match the model's input requirements
4. Define a misleading or irrelevant target concept as text (e.g., "a photo of trash")
5. Convert the image into a tensor and enable gradient computation
6. Initialize an optimizer to update the image tensor
7. For a number of steps:
   8. a. Encode the image using the image encoder
9. b. Encode the misleading text using the text encoder
```

PSEUDO CODE B

Pseudocode for Add Adversarial Noise to an Image (FGSM Attack):

```
1.  Load a pre-trained image classification model (e.g., ResNet)
2.  Load an image from disk
3.  Preprocess the image: a. Resize to model’s input size b. Normalize pixel values c. Convert to tensor or array format
4.  Pass the preprocessed image through the model to get its predicted label
5.  Calculate the loss between the model’s output and the predicted label
6.  Compute the gradient of the loss with respect to the input image
7.  Generate adversarial noise: a. Take the sign of the gradient b. Multiply by a small value (epsilon) to control noise strength
8.  Add the adversarial noise to the original image to create a perturbed image
9.  Clip the pixel values of the perturbed image to stay within valid bounds (e.g., [0, 1])
10. Output or display: a. Original image b. Perturbed (adversarial) image c. Compare the model’s predictions before and after
```

USER INTERACTION

- Goals:
  - Comprehensive
  - Straightforward
  - Little to no navigation necessary
  - Accessible
  - Wham bam thank you Sam experience

USER INTERACTION PSEUDOCODE - HOME

```
• //Site home page
• //open/minimalistic design
• //two fields, large boxes
• //tab above fields labeled “advanced mode”
• //Within left field
•    //prompt user to upload with multiple options
•        //drag and drop/upload from device/google drive/dropbox
•   //Obfuscate image
•       //run through code A (wrapped like Tutankhamen)
•      //create New Image object
•     //run through code B (also wrapped)
•           //display original image in left field
•          //display new image in right field
```

USER INTERACTION PSEUDOCODE- ADVANCED MODE

```
//Selectable tab labeled "Advanced Mode"
    //New page
    //Return to home in upper right corner
    //Drop down menu
        //Effect Minimizer
        //Batch Processing
```

EFFECT MINIMIZER PSEUDO CODE

```
•  //Blurb explaining Effect Minimizer (layers a somewhat translucent image with half the obfuscation of
   the new image to preserve some of the original traits of the piece over the new image to lessen the
   changes
•          //Select opacity of new image layer – slide bar
•             //Processes images like home pseudocode
•             //Copy New image and make Layer Image
•             //half the desired level of obfuscation
•             //lower opacity to desired level
•             //layer over New Image
•         //Display in Left field (under new image)
•         //Display original image in right field (under original image)
```

BATCH PROCESSING PSEUDO CODE

```
//Prompts users to upload file of images/ multiple images
    //Processes images like home pseudocode
    //returns a new file/ images with
    //Display left field (under new image)
    //Display in right field (under original image)
```

OPEN QUESTIONS

- How effective is scraping?
- What does a typical bypass of cloaking and adversarial noise and image perturbation look like?
- What other media could we possibly use this for? (ie. Music/video/writing)
- The image can have minefields embedded in it, but how do we prevent any crossing?

WORKS CITED
AND RESEARCH MATERIALS

Anonymous, "What is Glaze? Samples, Why Does it Work, and Limitations", *GLAZE,* https://glaze.cs.uchicago.edu/what-is-glaze.html

Anonymous, "What is Nightshade, Why Does it Work, and Limitations", GLAZE, https://nightshade.cs.uchicago.edu/whatis.html#:~:text=The%20answer%20is%20that%20Glaze%20is%20a%20defensive,consent%20%28thus%20protecting%20all%20artists%20against%20these%20models%29

Anonymous, "How Does it Work?", *Kudurru,* https://kudurru.ai/

Christopher A. Choquette-Choo, Nicholas Cardini, Matthew Jagielski, "Poisoning Web-Scale Data Training Sets is Practical", Arxiv, May 6, 2024, https://arxiv.org/pdf/2302.10149

Mattei, Shanti Escalante-De, "Grand Theft AI", *EBSCO*, 2023, https://research.ebsco.com/c/k4o4aa/viewer/html/zi463e3u55?auth-callid=3146395c-66c0-477e-a568-f87f2d8c701f

Sasidharan Gouri, "Stability AI: Everything You Need To Know", *artworkflow,* June 8, 2023, https://www.artworkflowhq.com/resources/stability-ai-guide

Stanger James , "The Ancient Practice of Steganography: What is it, How is it Used, and Why Do Cybersecurity Pros Need to Understand it", *compTIA,* July 6, 2020, https://www.comptia.org/blog/what-is-steganography

Sweigart Al, " Manipulating Images", *Automate the Boring Stuff,* https://automatetheboringstuff.com/2e/chapter19/