The Many Faces of AI Art: Self-Poisoning Generative Models

Investigating How Iterative Text-to-Image and Image-to-Text Recursive Processes Affect Creative Novelty and Quality

Andra Alăzăroaie[1]

Supervisor: Dr. Anna Lukina[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Andra Alăzăroaie
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Petr Kellnhofer

Abstract

As AI-generated content becomes more prevalent, the risk of generative models consuming and regenerating their own outputs in a self-consuming loop increases. This study explores the phenomenon of self-poisoning in generative models, an iterative process where AI-generated outputs are repeatedly used as input for further generations. Using a dataset of artworks by renowned artists, the process involves captioning the artworks, generating images from the captions, and repeating this cycle. The research focuses on the impact of self-poisoning on creative novelty, evaluated through content and visual novelty metrics. The findings reveal that while self-poisoning introduces novel elements in both content and visuals, it simultaneously degrades the quality of the generated artifacts over time. Generative models struggle to maintain the complexity and creativity of the original artworks, leading to outputs that converge on certain themes and realistic styles. This study contributes to a broader understanding of AI's role in art and highlights potential limitations posed by iterative generative processes.

1 Introduction

As generative artificial intelligence (AI) models increasingly become authors of much of the content found online, the line between human-generated and AI-generated content continues to blur. Text-to-image models based on diffusion models [1] or Generative Adversarial Networks [2] have learned to produce images similar to the training input and generate new ones from textual prompts. Describing an image's content through text is also possible with Transformers [3]. The internal mechanisms of the AI systems often represent black-boxes that are impossible to model manually [4]. However, their training data is based on human-created content, which raises questions about copyright violation. These questions were especially emphasized when Jason Allen's artwork "Théâtre D'opéra Spatial", generated by Midjourney, won the Colorado State Fair's fine arts competition. While these advancements come with benefits, they also put human-made art at risk of plagiarism and devaluation.

As the models become content creators, given the difficulties in differentiating between human content and synthetic content, future generations of AI models will be trained on generated data, creating a self-consuming loop. This self-feeding mechanism, known as an autophagous loop [5], risks leading AI into a recursive cycle where it consumes and regenerates its own output, potentially diminishing the diversity of its creations. Questions are raised about the future of creativity, especially in the arts, as generative models become not only tools for creation, but also sources for their own training data.

The potential for AI to recycle existing artistic elements without innovation also poses a significant threat to the economic and cultural value attributed to originality in human creativity. Addressing these concerns is not only a technical challenge but also a philosophical one: can AI truly create novel art, or is it creating a blend of previously consumed human outputs? This question underscores the need for investigation into how generative models, both for image creation and text generation, influence the novelty and integrity of art. Considering current literature, self-consuming training loops are expected to have a negative impact on the diversity and quality of the data that will be generated in the future.

However, there remains a gap in the studies of iteratively using AI to generate content from AI-generated content. While state-of-the-art research investigates the consequences of generative AI models being retrained with AI-generated content in a self-consuming loop, they do not investigate the consequences of this loop without retraining. This paper aims to bridge the gap in current literature by defining and examining the impacts of generation loops on the creative novelty of AI-generated captions and artworks, exploring trends that might occur throughout these iterative generative processes.

The following sections are structured as follows: Section 2 will introduce a motivating example and the methodology, proposing a definition of self-poisoning, as well as formulating the research question. Section 3 will dive into related work and explain connections to self-poisoning. Section 4 will introduce the concepts used throughout the experiment. In Section 5, the experiment and its results will be presented. Section 6 will discuss and interpret the results further, while Section 7 will address possible biases and other issues that might have occurred during the experiment. Section 9 will examine the limitations of the study and encourage future work on the topic, whereas Section 10 will conclude the research.

2 Our Contribution

2.1 Motivating Example

Imagine you are back in middle school, and your art class homework is to create an image inspired by a famous painter's artwork. It's 2024, and your teacher is up-to-date with the latest technology, aware of generative artificial intelligence models like Midjourney, DALL-E or Stable Diffusion. Hence, there are no restrictions on how you produce the image, "as long as it remains creative," your teacher emphasizes. Intrigued by Vincent Van Gogh, you choose his 1890 masterpiece "Almond Blossom" since spring is your favorite season and blue is your favorite color. You are keen on capturing the essence rather than mimicking the style directly, but you are unsure how to proceed. You turn to your favorite captioning model, BLIP2, to interpret the painting's content in simple terms: "The branches of an almond tree in blossom" [6]. With this description in hand, you then feed it into Stable Diffusion [7] to craft your homework piece, blending inspiration with innovation. The result is still a blossomed tree, but the colors and style are novel.

The teacher's requirement that the final product "remains creative" poses a challenge. Creativity is a fundamental feature of human intelligence [8], but a concept not yet formally defined without oversimplification. As such, the assessment of the creative aspect of an image produced by injecting an AI-generated text describing a creative artwork into a text-to-

image generative model becomes complex. This study leverages previous research on creative novelty to explore how the iterative process of generating images from text descriptions and vice versa affects the novelty and semantic integrity of AI-generated artworks.

2.2 Mathematical Definition of Self-Poisoning

This study adopts the approach of using previously generated outputs as new inputs, thereby allowing an exploration of the dynamics of a process we can define as "self-poisoning". As such, it is possible to investigate self-poisoning by combining Image-to-Text and Text-to-Image models. By iteratively feeding the output of one process (image-to-text or text-to-image) back into the other, we can examine the effects of this cycle on the evolution of properties of the generated content. The goal is to observe any trends over iterations.

Let $\mathcal{D} = \{I_i\}_{i=1}^{N}$ be an initial dataset of N human-made artworks, where I_i represents the i-th image. Define:

- $\mathcal{T}_G$: an image-to-text generation model (e.g., BLIP-2).
- $\mathcal{I}_G$: a text-to-image generation model (e.g., Stable Diffusion).

We define self-poisoning as a process that can be described iteratively as follows:

- **Initialization (iteration $k = 0$)**: Start with the initial dataset $\mathcal{D}$. Define $\mathcal{D}_0 = \{(T_i^0, I_i)\}_{i=1}^{N}$, where T_i^0 is empty for all i. Thus, $\mathcal{D}_0$ consists of the images from $\mathcal{D}$ paired with empty text descriptions.
- **Iteration** k (for $k \geq 1$):
  1. *Text Generation*: For each image $I_i^{(k-1)}$ in $\mathcal{D}_{k-1}$, generate a textual description using the image-to-text model:
  $$T_i^{(k)} = \mathcal{T}_G(I_i^{(k-1)}). \quad (1)$$
  2. *Image Generation*: For each generated textual description $T_i^{(k)}$, generate an image using the text-to-image model:
  $$I_i^{(k)} = \mathcal{I}_G(T_i^{(k)}). \quad (2)$$
  3. Form the new dataset $\mathcal{D}_k = \{(T_i^{(k)}, I_i^{(k)})\}_{i=1}^{N}$, to be used in the next iteration.

Figure 1 illustrates the steps above at a higher, dataset level, while Figure 2 illustrates the same steps at an individual artwork level. The examples from Figure 2 are part of the WikiArt dataset [9], specifically Henri Matisse's "A Bunch Of Flowers 1907" from style Fauvism as artwork 1 and Utagawa Kuniyoshi's "The Actor 6" from style Ukiyo_e as artwork 2. The generative models used were BLIP-2 [6] for captioning and Stable Diffusion [7] for image generation. In simple terms, self-poisoning at an individual level refers to selecting a human-made artwork, captioning it, generating an image from the caption, captioning the generated image, generating an image from the new caption, and so on, for multiple steps. At a dataset level, this process is repeated for a collection of artworks.

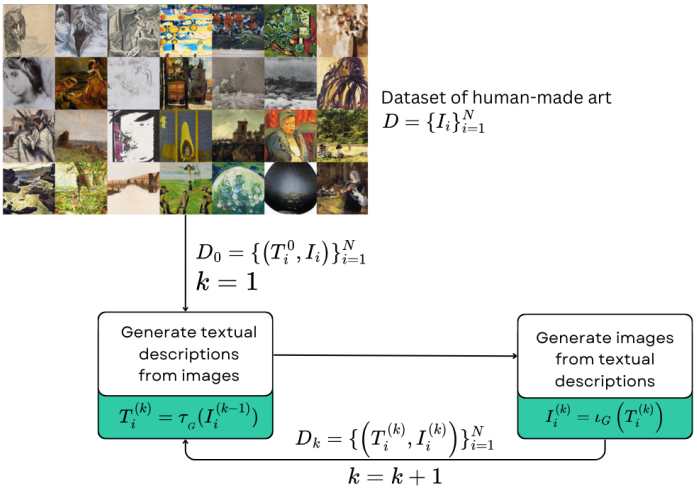

Figure 1: Process Flow Diagram illustrating the Self-Poisoning process on a dataset level for an initial dataset of human-made art (a sample of the WikiArt dataset) consisting of N images

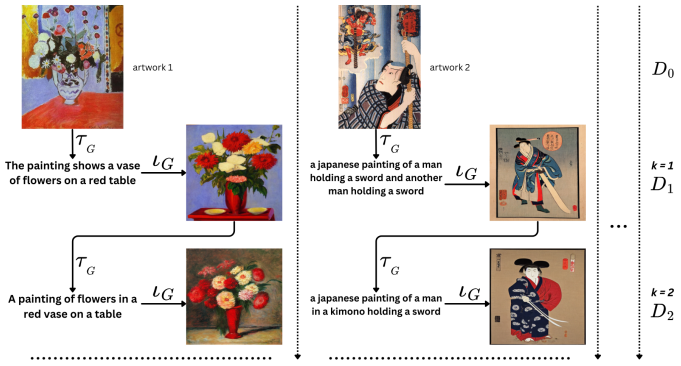

Figure 2: Process Flow Diagram illustrating the Self-Poisoning Process on an individual artwork level, containing particular examples.

2.3 Research Question

The main research question addressed by this work is:

How does the iterative process of generating images from text descriptions, and vice versa, affect the novelty and quality of the outputs?

The following are the sub-questions to be investigated:

- How does the content of the images drift from the original intent of the textual description over multiple iterations?
- How do the visuals of the images change over multiple iterations?
- At what point, if any, does the novelty and semantic similarity between iterations converge?
- How is the quality of the generated datasets impacted over iterations?

This work contributes through:

- Investigating the capabilities of state-of-the-art image-to-text and text-to-image generative models to maintain content and visuals through self-poisoning.
- Applying previously defined metrics for creative novelty find any possible trends arising during a recursive self-poisoning process.
- Discussing potential impacts that self-poisoning might have on generative models in the context of art.

3 Related Work

As vast amounts of content on the Internet are increasingly becoming AI-generated, it is expected that the next training iterations of generative models will be (partly) based on previously generated data, creating a feedback loop. These processes, known as *autophagous or self-consuming training loops*, raise a concern given the difficulties in differentiating between AI-made and human-made content, as well as data collection practices, since text and data are in many cases extracted from the Internet using crawlers [10]. An experiment [10] studied the relationship between generative models and the Internet using a simplified interaction model and concluded that this interaction could lead to degeneration and loss of diversity.

Another study also [11] investigates the impact that re-training Large Language Models (LLMs) in a self-consuming training loop has on the quality and diversity of the samples. Diversity was defined as a metric of pairwise Levenshtein distance between output that these models produce, averaged over the total number of pairwise comparisons [12]. Quality was measured using similarity metrics (such as the BERT score [13]) of LLMs' outputs with expected reference data. The results of the study indicate that the diversity of the generated data degenerates as the number of iterations of the self-consuming training loop increases, collapsing into a single point.

One theoretical and empirical study of AI autophagy from the perspective of generative image models [5] sought to extrapolate the consequences of generative models becoming ubiquitous and used in training future generations in a self-consuming loop. They recursively trained generative models on synthetic data sampled from other generative models, resulting in an autophagous ("self-consuming") loop and defined Model Autophagy Disorder (MAD) as an analogy to concepts in mathematics and biology. The conclusion was that without enough fresh real data each generation, future generative models are doomed to MADness: the synthesized data distribution drifts from the true data distribution over the generations. Precision (defined as the portion of synthesized samples of high quality or visually appealing) and diversity decreased over generations and generative artifacts were amplified.

The critical difference between this study and the existing research on self-consuming training loops is that this study does not aim to re-train models on a combination of generated and original content. Instead, we aim to *define and investigate self-consuming loops without retraining, which we defined as self-poisoning*: the generated content is directly fed into the generative model once again, repeatedly. The generated content, whether textual or visual content, which was initially the output of a generative model, becomes the input of a new iteration of generation. While the research on self-consuming training loops is extensive, previous studies do not address self-poisoning.

4 Preliminaries

While substantial progress has been made in the technical development of generative models and their application in art, the biggest gap remains the lack of universally accepted metrics for assessing the novelty of AI-generated art. There exist several novelty detection techniques through Generative Adversarial Networks [14] with several applications such as hand gesture data [15] or fake news [16], but little has been done in the context of art.

Notably, one study defines novelty in terms of the subject matter and its interrelations within the artwork [17]. From this perspective, the concept of creative novelty emerges as a useful metric for evaluating the uniqueness and innovation of AI-generated artifacts. This concept is rooted in classical philosophy, particularly the philosophy of symbolism in art, which distinguishes between the content (the meaning or subject matter of an artwork) and the visuals (the physical elements that convey this content). As such, creative novelty in the context of AI-generated art can be split into two distinct, but interconnected dimensions: Content Novelty and Visual Novelty. Using the idea of conceptual spaces seen as geometric representations of entities capturing attributes, we can use embeddings of text and images, which translate their features into a vector space. The distances between these vectors can then be measured to determine whether artifacts deviate or converge with the references.

The Content Novelty dimension measures how the focal objects and themes in new artworks diverge from those in previous works, focusing on the thematic and conceptual aspects of an artifact. The operationalization of Content Novelty involves analyzing the semantic content of an artifact, often extracted using advanced natural language processing models like BLIP-2 [6], that interpret images to generate descriptive textual content. This textual data is then converted into high-dimensional vector representations using embedding techniques, such as those based on BERT [18] models. The novelty is quantitatively assessed by computing the cosine distance between these vectorized descriptions, comparing each new artifact against a baseline set to capture variations over time.

On the other hand, visual Novelty measures how the style and visual elements of new artworks differ from earlier ones at the pixel level. This dimension focuses on the visual presentation, including but not limited to color schemes, texture, and style. Techniques such as DINOv2 [19], a self-supervised visual representation learning algorithm, are employed to extract and analyze these visual features. Similar to Content Novelty, Visual Novelty is measured by determining the cosine distance between the vector representations of the visual features of new and baseline artifacts.

The cosine similarity is a traditional method used to measure the similarity between two vectors and it is obtained through the cosine angle multiplication value of the two vectors to be compared [20]. The idea is that the cosine of 0° is 1, hence the two vectors are said to be similar, whereas values of less than one emphasize on the differences between the vectors. Following these definitions and the defining study's methodology, we refer to Content or Visual Novelty as Content or Visual Similarity interchangeably and compute them as follows:

1. Content Similarity is captured using BERT [18]. It is represented by the semantic similarity between captions and used as a metric of the content channel of creative novelty. BERT encodes each caption into high-dimensional vector representations. These vectors capture the semantic meaning of the text, considering the context and relationships between words. The similarity between captions is quantified by calculating the cosine similarity between their respective feature vectors V.

$$\text{CSim}(\text{cap}_1, \text{cap}_2) = \frac{V_{\text{cap}_1} \cdot V_{\text{cap}_2}}{\|V_{\text{cap}_1}\| \|V_{\text{cap}_2}\|} \tag{3}$$

2. Visual Similarity is computed using DinoV2 [19], a self-supervised visual representation learning algorithm. It extracts high-dimensional feature vectors from each image, representing various visual attributes such as color, texture, shape, and other pixel-level details. The similarity between images is then quantified by calculating the cosine similarity between their respective feature vectors V.

$$\text{VSim}(\text{img}_1, \text{img}_2) = \frac{V_{\text{img}_1} \cdot V_{\text{img}_2}}{\|V_{\text{img}_1}\| \|V_{\text{img}_2}\|} \tag{4}$$

5 Experimental Setup and Results

5.1 Experimental Setup

The implementation consists of a controlled methodology that uses both textual and visual feedback loops. Textual descriptions are iteratively generated from images using BLIP-2, followed by using Stable Diffusion to generate images from the respective text. Several metrics are used to compare and contrast the generated images of each step. The code and results are available in our GitLab repository [21].

Dataset Selection

The experiment begins by selecting an open-source dataset of human-made artworks. WikiArt [9] is chosen for its comprehensiveness and widespread use, containing over 81,000 artworks from recognized artists across 27 artistic styles. To reduce computational time and resources while preserving the dataset's diversity, 10 artworks from each style are randomly selected, creating an initial dataset of 270 artworks.

Text-to-Image Model

Stable Diffusion [7] is used for text-to-image generation in the iterative self-poisoning process. It uses a deep learning technique called latent diffusion [1]. To generate images, Stable Diffusion projects a text prompt into a joint text-image embedding space and selects a noisy image that is semantically close to the input prompt. This image is then denoised based on a latent diffusion model and thus the final image is produced. The important advantage of Stable Diffusion is that its code and model weights are publicly available. Furthermore, the model is free to use. To aid the generation of artistic images, a version of Stable Diffusion fine-tuned on the WikiArt dataset [22] has been used in the self-poisoning process with the help of the Diffusers library [23] from HuggingFace.

Image-to-Text Model

BLIP-2 [6] is used for generating textual descriptions (which we also refer to as captions) for images in order to describe the focal objects and their relationships. BLIP-2 is a multimodal model which has been trained on 129M images and human-annotated data. By leveraging pre-trained image encoders and LLMs, it achieves state-of-the-art performance on various visual-language tasks, including zero-shot instructed image-to-text generation. BLIP-2 is used in the self-poisoning process through the Transformers library [24] from HuggingFace.

Data Collection

The free versions of Google Colab and Kaggle were used as cloud platforms to apply self-poisoning and save the generated data, both providing T4 GPUs. On this hardware, producing one image from a text description using a Stable Diffusion model fine-tuned on the WikiArt dataset took approximately 8 seconds. The number of iterations of the self-poisoning process has been restrained at 100 due to resources and time limitations. Applying 100 iterations of self-poisoning on a single image took on average 15 minutes (approximately 800 seconds for the generation of the 100 images and the rest for the captioning). The generated captions were saved in JSON format, which was archived together with the generated images. These archived were downloaded locally and extracted such that the data analysis process can be performed on a personal computer without the need for using GPUs. The data collection process resulted in a total of 27,000 generated images besides the 270 images sampled from WikiArt, as well as 27,000 captions. The process was performed in batches over several weeks. It took approximately 67 hours of runtime.

Data Analysis

For the data analysis part, the two previously defined metrics of Content and Visual Similarity are used in order to observe the influence of self-poisoning on creative novelty. Nevertheless, in previous work on autophagous loops [5] [10], the quality and diversity of the images produced by generative models throughout retraining cycles are also measured. These two properties have tremendous importance in the context of art. Is is arguable that an artwork's value is not only given by its creative novelty, but also by its qualitative standard. The Fréchet Inception Distance score (FID) [25] is a metric that computes the distance between feature vectors calculated for real and generated images, capturing the similarity of generated images to real ones in terms of visual quality and diver-

sity. The FID score is usually used for the evaluation of the performance of generative adversarial networks (GANs) at image generation, and lower scores have been shown to correlate with higher-quality images [26]. Hence, we use FID as a metric of quality and diversity within the generated datasets over iterations.

To summarize, the following metrics are used:

1. Content Similarity $CSim(\text{cap}_0, \text{cap}_k)$ is captured using BERT [18] and the cosine similarity metric. BERT is used from the Transformers library [24] of Hugging Face.
2. Visual Similarity $VSim(\text{img}_0, \text{img}_k)$ is computed using DinoV2 [19] and the cosine similarity metric. DinoV2 is used from the Transformers library [24] of Hugging Face.
3. The Fréchet Inception Distance score $FID(\mathcal{D}_k)$, for each iteration k represents the quality and diversity of the dataset generated at iteration k with respect to the original dataset of human-made artworks $\mathcal{D}_0$. It is calculated using a dedicated Python implementation [27].

Convergence

To evaluate the stability of the iterative self-poisoning process, we define convergence based on the similarity metrics over iterations relative to their respective minimum observed values. Convergence occurs when the deviation between the average similarity of a certain iteration and the minimum observed average similarity consistently falls below predefined thresholds.

The Content Similarity is considered to have converged at a point t_c if the following criteria are met:

$$|\text{CSim}(t_0, t) - \text{CSim}_{\min}| < \epsilon_c \quad \forall\, t \geq t_c \tag{5}$$

where $\text{CSim}(t_0, t)$ represents the averaged Content Similarity between each original artwork and the image generated from that artwork at iteration t, $\text{CSim}_{\min}$ is the minimum averaged Content Similarity value observed in all iterations and is defined mathematically as:

$$\text{CSim}_{\min} = \min_t \left(\text{CSim}(t_0, t)\right), \tag{6}$$

and ϵ_c is a small positive constant representing the convergence threshold.

The Visual Similarity is considered to have converged at a point t_v if the following criteria are met:

$$|\text{VSim}(t_0, t) - \text{VSim}_{\min}| < \epsilon_v \quad \forall\, t \geq t_v \tag{7}$$

where $\text{VSim}(t_0, t)$ represents the averaged Visual Similarity between each original artwork and the image generated from that artwork at iteration t, $\text{VSim}_{\min}$ is the minimum Visual Similarity value observed and is defined mathematically as:

$$\text{VSim}_{\min} = \min_t \left(\text{VSim}(t_0, t)\right), \tag{8}$$

and ϵ_v is a small positive constant representing the convergence threshold.

The convergence point t_c for Content Similarity and t_v for Visual Similarity are defined as the iterations after which all subsequent similarities remain within ϵ of the minimum observed similarity value:

$$t_c = \min\{t \mid |\text{CSim}(t_0, t') - \text{CSim}_{\min}| < \epsilon_c \quad \forall\, t' \geq t\} \tag{9}$$

$$t_v = \min\{t \mid |\text{VSim}(t_0, t') - \text{VSim}_{\min}| < \epsilon_v \quad \forall\, t' \geq t\} \tag{10}$$

In general mathematical terms, convergence refers to the property of a sequence where its elements become close to a certain value, known as the limit, as the sequence progresses. The definitions above follow this principle by checking if the similarity values $\text{CSim}(t_0, t)$ and $\text{VSim}(t_0, t)$ remain within a small threshold ϵ of the minimum similarity values $\text{CSim}_{\min}$ and $\text{VSim}_{\min}$ beyond the convergence points t_c and t_v. This ensures that the similarities are stable, presenting no significant variations.

5.2 Quantitative Results

Content and Visual Novelty

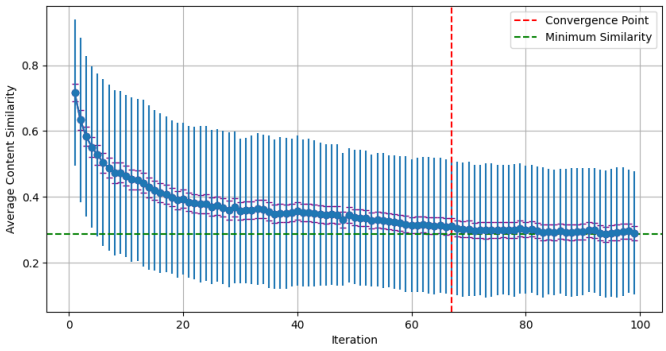

Figure 3: Content Similarity between the first and the other generated captions over iterations. The blue points represent the average Content Similarity, the blue error bars represent the standard deviation, and the purple error bars represent the 95% confidence interval.

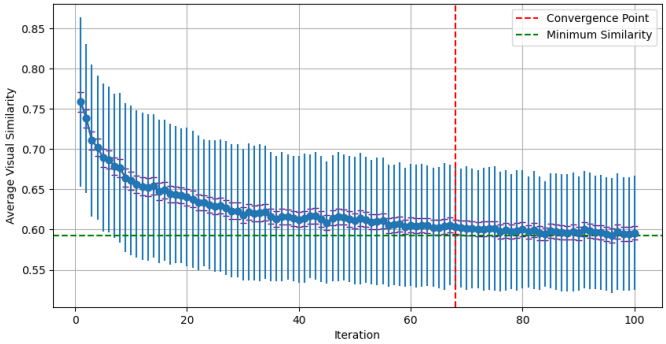

Figure 4: Visual Similarity between the initial artwork and the generated images over iterations. The blue points represent the average Content Similarity, the blue error bars represent the standard deviation, and the purple error bars represent the 95% confidence interval.

As per Figure 3 and Figure 4, both content and visual novelty show a drift from the original dataset within the first few

iterations of self-poisoning. The results indicate a significant change in the novelty metrics, particularly in the initial phases. The average Content Similarity between the generated captions and the original dataset shows a declining trend, suggesting an increase in content novelty. Over the first 20 iterations, the average Content Similarity decreases from 0.72 to 0.4, indicating a significant drift. The standard deviation of Content Similarity across iterations was observed to be between 0.2 and 0.3, suggesting significant variability around the mean. This variability indicates that the degree of content change observed through self-poisoning can differ considerably for different initial artworks. However, the narrow 95% confidence intervals (ranging from ± 0.02 to ± 0.03) around the mean Content Similarity imply high precision, meaning that we are 95% confident that the true mean Content Similarity lies within the purple error bars of Figure 3.

Visual Similarity also demonstrated a notable decline. The average Visual Similarity between the generated images and the initial dataset decreases from 0.76 to 0.64 over 20 iterations. This decline in Visual Similarity highlights the Visual Novelty introduced through self-poisoning. The standard deviation for Visual Similarity is observed between 0.07 and 0.12, underscoring a significant dispersion in Visual Similarity. This variability suggests that the visual changes introduced by self-poisoning can be quite different for various artworks. The visual features of the training data of Stable Diffusion have a direct impact on the visual features of the produced images, therefore applying self-poisoning on artworks from certain styles might have produced images that are more similar to the original than others. Nevertheless, the 95% confidence intervals for Visual Similarity are also narrow (ranging from ± 0.02 to ± 0.03), suggesting high precision of the mean Visual Similarity.

The initial average Visual Similarity of 0.76 and Content Similarity of 0.72 compare each human-made artwork to the first image generated from the respective artwork, and each initial caption generated from the human-made artwork to the caption generated from the first generated image. These initial values reflect the ability of BLIP-2 and Stable Diffusion to accurately match their input to the output. For BLIP-2, this involves accurately describing the contents of the input image, while for Stable Diffusion, it involves the accuracy and completeness with which the model includes the nuances of the input prompt in the generated image. The initial values of 0.76 and 0.72 suggest that the models do not fully maintain either content or visuals, thereby introducing novelty on both channels. This raises the question of whether maintaining content and style or creating novel content is more desirable, which can be subject to the context.

The convergence points for the two metrics are notably close to each other. The content convergence point is $t_c = 66$ and the visual convergence point is $t_v = 67$. These points represent the iterations after which all subsequent average similarity scores remain close, within ϵ to the minimum average similarity value. We can say that the Similarity metrics converge to their minimum value, which is 0.29 for Content Similarity and 0.59 for Visual Similarity. Taking into account these minimum values and their ratio, ϵ_c has been chosen to be 0.02, while ϵ_v has been chosen to be 0.01. These points represent the stagnation in average Content Novelty and Visual Novelty over time throughout the self-poisoning process. As of the current experiment, these convergence points could cannot be generalized. However, increasing the number of iterations might allow for better observations of convergence.

Quality and Diversity

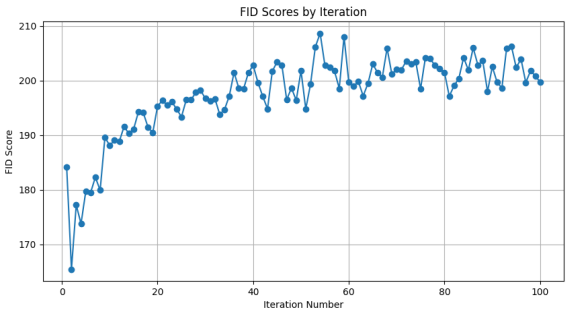

Figure 5: Fréchet Inception Distance between the original sample dataset from WikiArt and the generated datasets over iterations

The FID metric, plotted in Figure 5, shows a global upward trend in the first 20 iterations, while later oscillating above a threshold of 193. The higher the FID value, the lower the quality of the generated images and the less similar to the real artworks. Hence, self-poisoning negatively impacts the quality of the generated datasets over time. This phenomenon aligns with the concept of Model Autophagy Disorder (MAD) [5], where the continuous retraining using generated data leads to a degeneration of output quality. Despite the global upward trend, we notice the first iteration having a higher FID score than a few of the subsequent ones. This could be explained by the gaps of the models in capturing precise details about the artworks in the initial captioning phase, which are immediately used as prompts for the images of the first generation's dataset. After this first iteration, we can argue that the captioning model works with less complex images than the initial artworks and, therefore is capable of capturing more content and visuals in order to recreate it.

5.3 Qualitative Results

Analyzing self-poisoning on an individual artwork level can further support the quantitative results. Novelty can be easily observed throughout iterations, but quality and diversity are noticeably impacted. Figure 6 and Figure 7 are examples in which the self-poisoning process maintained some part of the content in the original artwork, but changed the visuals significantly. However, the connection to the initial artwork still exists. On the other hand, Figure 8 and Figure 9 completely diverged. The connection to the original artwork becomes more and more vague over time. In Figure 8, the first 43 iterations maintained the content, but in the following iteration, the models focused on a single object and further iterated on it. In Figure 9 however, the content and visuals changed drastically and continuously over time. The emotion of Figure 9 does not seem to have been captured either. From a personal

perspective, the artwork suggests that the woman is in a reflection state during a break from social responsibilities suggested by her outfit. The initial caption "The painting shows a woman sitting at a table with a cup of coffee"[6] fails to capture this reflection, therefore the next iterations also fail to reproduce it. Using a fine-tuned model for captioning in which human-produced captions for artworks are used as training data could perhaps yield results better in this aspect.

The qualitative results also indicate several notable patterns. Initially, the models tend to produce outputs that are closely related to the input images. However, as the iterations progress, a drift is observed where the generated images begin to significantly diverge from the original artistic styles and subjects. *Common themes that emerge after convergence include men reading books, women sitting on benches, red objects, forests, and snow scenes.* This drift is indicative of the models' tendency to focus on specific elements of the original images, leading to outputs that develop from these single elements. Consequently, the connection to the initial artwork becomes vague after several iterations. This suggests that the models' internal biases and the limitations in capturing the full complexity of the original artworks play a significant role in the self-poisoning process. The accuracy of the match between the models' input and what they produce is crucial. The generated images also tend to be increasingly realistic over iterations, regardless of the style of the initial artwork. This can be explained by the training data of Stable Diffusion [7], as the goal of the model was to be able to generate "photo-realistic images".

Figure 6: Self-Poisoning Example 1, starting from Konstantin Somov's artwork "Costume Sketches Of Columbine For Anna Pavlova" in style Symbolism

Figure 7: Self-Poisoning Example 2, starting from Adam Baltatu's artwork "House With Flowers" in style Post Impressionism

Figure 8: Self-Poisoning Example 3, starting from Alexey Bogolyubov's artwork "Steamship Kolkhida Fighting The Turkish Boats At The St Nicholas Fort Near Potigeor" in style Romanticism

Figure 9: Self-Poisoning Example 4, starting from Edward Hopper's artwork "Automat 1927" in style New Realism

6 Discussion

The observed drift in both Content and Visual Novelty metrics highlights the capability of self-poisoning to introduce significant changes in generated data. This suggests that generative models, when iteratively consuming their own outputs, can create novel content that diverges from the original dataset. However, the increase in FID scores suggests a trade-off between novelty and quality, as higher novelty is accompanied by reduced image quality over time.

The decrease and convergence of Content and Visual Novelty over time within the self-poisoning process underscore the limitations of generative AI models. These models are capable of maintaining content to some extent, as shown by convergence, but only if it is not too complex. Starting from an initial, visually and content-wise complex dataset of human-made art, the models struggle to capture the details that contribute to the artworks' recognition. Nevertheless, the observed results cannot be generalized, since only BLIP-2 [6] and Stable Diffusion [7] were used in self-poisoning. Investigation the process using different models could yield different results.

The performance of the models in maintaining content and visual integrity is often evaluated using Visual Question Answering frameworks like TIFA [28], which reveal common errors such as object count inaccuracies, missing or duplicated objects, and unrealistic face and body generation in complex contexts. The results of this study emphasize the importance of improving these capabilities and suggest that fine-tuning models on human-written captions for artworks or incorporating human feedback in the iterative process could

enhance the quality of AI-generated art.

The results align with the theoretical framework of blind variation and selective retention (BVSR) in creativity studies, where the generative process explores a broad range of variations without specific guidance and the most promising outcomes are selected for retention [29]. In this context, the generative models' ability to produce novel content through self-poisoning reflects the blind variation aspect, while the iterative process and convergence criteria represent the selective retention mechanism.

One notable issue encountered was the false positives generated by the safety filter in Stable Diffusion, resulting in black images when the text prompt was considered NSFW ("not safe for work"). In some situations such as certain artworks that were sampled from the Baroque artistic style, the generated captions can be traditionally considered explicit, as some of them depict naked characters, therefore justifying the generated black images. However, in many other situations, there was no explicit content and the filter was triggered purposelessly. Examples include "A black and white photograph of a towel hanging on a wooden hanger" or "a painting of a young boy looking out a window". These black images have influenced all future iterations of the self-poisoning process for the respective artworks, as the generated images in those cases were, strangely, most of the time, picturing "an old man sitting on a chair in a dark room", which also constituted the following caption. These images and captions had a direct impact on diversity since they became a recurrent theme within the generated datasets. This observation highlights a limitation in the current safety mechanisms and underscores the need for more refined filtering techniques that can better differentiate between harmful and benign content.

The findings have implications for the use of generative AI in art creation. While self-poisoning can enhance the novelty of AI-generated artworks, it also necessitates careful consideration of quality metrics to ensure the produced content remains engaging and valuable. Content generated through self-poisoning could potentially become part of the training data of generative models' future generations. As the autophagous loop has already been proven to decrease the quality and diversity of the generated data, injecting self-poisoned content into the loop might accentuate the consequences.

In summary, the study demonstrates the potential of self-poisoning to drive novelty in generative AI while highlighting the challenges associated with maintaining quality. Future research should explore alternative approaches to balance these aspects and further investigate the underlying mechanisms of creativity in AI-generated art.

7 Limitations and Future Work

This study encounters several limitations throughout the experimental scope and data analysis.

Firstly, due to limitations in hardware resources, particularly GPUs, as well as time constraints, the initial dataset was confined to a randomly selected subset of 10 images per artistic style from the WikiArt dataset [9], totaling 270 images. Consequently, the small sample size may reduce the statistical significance of the findings. Utilizing a larger dataset comprising several thousand images would enhance the reliability and generalizability of the results.

Secondly, the study conducted the self-poisoning process over 100 iterations. Increasing the number of iterations in future studies could provide a more robust confirmation of the observed trends and extend the understanding of the long-term effects of iterative generative processes.

Furthermore, the experiment only investigated self-poisoning on a particular pair of text-to-image and image-to-text models: BLIP-2 [6] and Stable Diffusion [7]. Experimenting with different models, or different fine-tuned versions of certain models might enhance the generalizability of the results in this study. Despite self-poisoning being defined as a mathematical process, the results are highly dependent on the models and their ability to match the text's contents with the image's contents and visuals.

Lastly, the definition of creative novelty employed in this paper draws solely from one prior study [17]. Future research should not only investigate this definition further but also consider incorporating and comparing other conceptual frameworks of creativity. This could enrich the analysis and offer more insights into the creative capacities of AI art generation.

8 Responsible Research

Reflecting on the reproducibility and integrity of the research, several measures have been taken to ensure that the experiments can be reliably reproduced and that the integrity of the results is maintained.

8.1 Reproducibility

A comprehensive description of the experimental setup has been provided, including the specific models used, the dataset selection, the metrics for evaluation, and the necessary libraries. This ensures that other researchers can replicate the setup accurately. Furthermore, all code and data used in the experiments are available in our GitLab repository [21]. By sharing the implementation, the reproducibility of this experiment is further facilitated. The availability of the data ensures that others can verify the results and conduct further analysis if needed. The exact platforms and hardware used for the experiments have also been mentioned. Since the platforms are free to use, the experiment can be replicated by anyone without needing access to specialized hardware.

The limitation that this study presents regarding reproducibility is due to the non-deterministic nature of the generative models. Stable Diffusion, as well as BLIP-2 may produce different results when run multiple times on the same input. For this reason, all results that were used in the analysis are provided in the GitLab repository.

8.2 Integrity

An open-source dataset (WikiArt) has been selected, and the research adhered to its usage guidelines, ensuring ethical use of data. Moreover, the tools and models used in the experiments are freely available and properly cited, respecting intellectual property rights and the contributions of the original developers. The results are presented transparently, with both

quantitative metrics and qualitative examples. A thorough analysis of content and visual similarity metrics, convergence points, and FID scores, alongside visual examples that illustrate the self-poisoning process has been provided. The negative influence that Stable Diffusion's NSFW filter had on the results has also been explained. The resulting black images were treated as valid data points and contributed to the quantitative analysis. Well-established models and evaluation techniques previously used in the literature were used throughout the experiment to strengthen the reliability of the conclusions.

However, it is essential to address the biases inherent in the training data of the models used in these experiments, as well as the dataset. Datasets often contain biases related to cultural representations, subject matter, and stylistic preferences, which can influence the outputs of generative models. These biases can lead to the reinforcement of stereotypes or the exclusion of less-represented artistic styles and cultural contexts. Stable Diffusion [7] admits such biases, specifically due to their training data primarily consisting of images with English descriptions, insufficiently accounting for other languages and reinforcing white and Western cultures as default. Furthermore, the iterative self-poisoning process may aggravate these biases, as the models repeatedly generate outputs based on their previous biased outputs. Ensuring diversity and representation in the training datasets and implementing mechanisms to identify and mitigate biases in generative models are crucial steps toward responsible AI research and development.

By incorporating these practices into the research, the principles of reproducibility and integrity are respected, ensuring that the work can be reliably built upon by the scientific community.

9 Conclusions

This study defines and investigates the phenomenon of self-poisoning, an iterative process where AI-generated outputs are repeatedly fed back into generative models. Starting with a dataset of artworks by well-known artists, the process involves captioning the artworks, generating images from the captions, and repeating these steps on the generated data. The focus is on how self-poisoning affects creative novelty, assessed through content and visual novelty. Results indicate that while self-poisoning introduces novelty in both content and visuals, it simultaneously degrades the quality of the generated artifacts. Over time, models struggle to maintain the original artworks' content and visuals, instead introducing novel elements. Qualitative analysis reveals that the outputs converge on certain content themes and realistic styles. These outcomes come from the generative models' limitations in closely aligning outputs with input prompts and generating comprehensive textual descriptions of input images. They also suggest that generative models have limitations in sustaining the complexity and creativity of original human-made artworks. Enhancing models' capabilities to capture and replicate complex artistic elements is essential for advancing AI art generation.

Future research should replicate similar experiments using different models, datasets, and parameters, and consider multiple conceptual frameworks of creativity to enrich the analysis and improve the reliability and generalizability of the results. These efforts will deepen our understanding of generative AI's potential and limitations in the creative domain and contribute to developing more robust and responsible AI art generation techniques.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[3] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proceedings of the Asian conference on computer vision*, 2020.

[4] D. Silverman, "Burying the black box: Ai image generation platforms as artists' tools in the age of google v. oracle," *Federal Communications Law Journal*, vol. 76, no. 1, pp. 115–142, 2023.

[5] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk, "Self-consuming generative models go mad," *arXiv preprint arXiv:2307.01850*, 2023.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[7] CompVis, "Stable diffusion," https://github.com/CompVis/stable-diffusion, 2022, accessed: 2024-04-23.

[8] M. A. Boden, "Creativity and artificial intelligence," *Artificial intelligence*, vol. 103, no. 1-2, pp. 347–356, 1998.

[9] "Wikiart," https://www.wikiart.org/, accessed: 2024-06-03.

[10] G. Martínez, L. Watson, P. Reviriego, J. A. Hernández, M. Juarez, and R. Sarkar, "Towards understanding the interplay of generative artificial intelligence and the internet," in *International Workshop on Epistemic Uncertainty in Artificial Intelligence*. Springer, 2023, pp. 59–73.

[11] M. Briesch, D. Sobania, and F. Rothlauf, "Large language models suffer from their own output: An analysis of the self-consuming training loop," *arXiv preprint arXiv:2311.16822*, 2023.

[12] D. Wittenberg, F. Rothlauf, and C. Gagné, "Denoising autoencoder genetic programming: strategies to control

exploration and exploitation in search," *Genetic Programming and Evolvable Machines*, vol. 24, no. 2, p. 17, 2023.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[14] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/5421e013565f7f1afa0cfe8ad87a99ab-Paper.pdf

[15] M. Simão, P. Neto, and O. Gibaru, "Improving novelty detection with generative adversarial networks on hand gesture data," *Neurocomputing*, vol. 358, pp. 437–445, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219307702

[16] S. Hiriyannaiah, A. Srinivas, G. K. Shetty, S. G.M., and K. Srinivasa, "Chapter 4 - a computationally intelligent agent for detecting fake news using generative adversarial networks," in *Hybrid Computational Intelligence*, ser. Hybrid Computational Intelligence for Pattern Analysis and Understanding, S. Bhattacharyya, V. Snášel, D. Gupta, and A. Khanna, Eds. Academic Press, 2020, pp. 69–96. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128186992000044

[17] E. Zhou and D. Lee, "Generative artificial intelligence, human creativity, and art," *PNAS Nexus*, vol. 3, no. 3, p. pgae052, 03 2024. [Online]. Available: https://doi.org/10.1093/pnasnexus/pgae052

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[20] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International conference on cyber and IT service management*. IEEE, 2016, pp. 1–6.

[21] A. Alazaroaie, "The many faces of ai art: Self-poisoning," 2024, git-Lab Repository. [Online]. Available: https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/aalazaroaie-The-Many-Faces-of-AI-Art/-/tree/master/Self-Poisoning

[22] Hugging Face's Valhalla Team, "Sd wikiart v2 model," https://huggingface.co/valhalla/sd-wikiart-v2, accessed: 2024-06-03.

[23] Hugging Face, *Hugging Face Diffusers Documentation*, 2024. [Online]. Available: https://huggingface.co/docs/diffusers

[24] ——, *Hugging Face Transformers Documentation*, 2024. [Online]. Available: https://huggingface.co/docs/transformers

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Y. Yu, W. Zhang, and Y. Deng, "Frechet inception distance (fid) for evaluating gans," *China University of Mining Technology Beijing Graduate School*, 2021.

[27] M. Seitzer, "pytorch-fid: FID Score for PyTorch," https://github.com/mseitzer/pytorch-fid, August 2020, version 0.3.0.

[28] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 406–20 417.

[29] D. K. Simonton, "The blind-variation and selective-retention theory of creativity: Recent developments and current status of bvsr," *Creativity Research Journal*, vol. 35, no. 3, pp. 304–323, 2023.